UNIVERSITY OF CALIFORNIA

Los Angeles

# On the Fluctuations of Seismicity and Uncertainties in Earthquake Catalogs: Implications and Methods for Hypothesis Testing

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Geophysics and Space Physics

by

**Maximilian Jonas Werner**

2008

The dissertation of Maximilian Jonas Werner is approved.

Peter Bird

Gilles Peltzer

Frederic Paik Schoenberg

Didier Sornette

David D. Jackson, Committee Chair

University of California, Los Angeles

2008

*To my mother and Hermann;*

*to Alice and in memory of my father*

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGMENTS

First and foremost I would like to thank my advisor and mentor Didier Sornette for his unparalleled enthusiasm and optimism, for sharing his deep insights, for his patience, for his availability at all hours of the day on all days of the year, for his interdisciplinary vision, and for training me for a career in the sciences. I am particularly thankful for his invitation and offer to spend 3 months at the Department of Management, Technology and Economics at the Swiss Federal Institute of Zurich during January to April of 2007.

I am deeply indebted to my advisor and committee chair David D. Jackson for his warm and friendly advice, for the sharp and unconventional thoughts he has shared with me, for placing his trust in my abilities and for the many great discussions. I have very fond memories of the brunches that his wife Kathy and he organized at their home for his students.

I would like to thank Peter Bird and Rick Schoenberg for stimulating discussions and for agreeing to be on my committee. I am particularly indebted to Gilles Peltzer for accepting to be on my committee. I want to thank Kayo Ide of the Department of Atmospheric and Oceanic Sciences at UCLA for our ongoing collaboration on data assimilation. The interdisciplinary work has been a great pleasure.

I am grateful to Stefan Wiemer and the Earthquake Statistics Group at the Swiss Seismological Service of ETH Zurich, in particular Annemarie Christophersen, Georgia Cua, Sjonni Jonsson, Kazu Nanjo, Thomas van Stiphout and Jochen Woessner, for making my stay there a great pleasure. I'm looking forward to a continuation of the fruitful collaboration that started during that time. Equally important have been my interactions with the people at and around the Chair of Entrepreneurial Risks in Zurich,

in particular Riley Crane, Gilles Daniel, Georges Harras, Shengsui Hu, Felix Roudier, and Uli Stark.

I thank NASA for a three-year graduate student fellowship in Earth System Sciences, which gave me the liberty and independence that made much of this thesis possible. The fellowship allowed me to attend conferences around the globe where I met many of the experts in statistical seismology. In particular, I thank the organizers of the Statistical Seismology V workshop in Erice, Sicily, and Danijel Schorlemmer and Jeremy Zechar, as organizers of the Earthquake Predictability session at the SSA meeting in Hawaii, for inviting me to give presentations.

At UCLA, I have benefited immensely from discussions with Yan Kagan, whose experience and knowledge in statistical seismology and far beyond have been a great source of inspiration during my studies. The seismology group's seismo-lunch has taught me more than most courses thanks to the lively discussions between the faculty, researchers and students. I thank Agnés Helmstetter for introducing me to the practicalities of research. I thank Jean "Bob" Grasso for his friendship, our discussion on earthquakes, and our surfing trips (and his pronunciation of the state of Oaxaca). I thank Jean Elkhoury for our many interesting and stimulating discussions about earthquakes, physics, politics and the life of a graduate student over the course of much coffee and ice cream. I'd like to thank my friends Reyna Alorro, Sabrina Benchaar, Elizabeth Cochran, Cécile Doubre, Chizoba Ezeh, Shruti Garg, Gilad Gelfond, Safeer Ghaznavi, Mark Hall, Rebecca Harrington, Allen Husker, Jan Kliemann, Robert Kozek, Gregory Owcarz, Oli Proksch, Anne Soquet, Bastien Soulé, Yuri Spritz, Jeremi Sudol, Asia Szupinska and Jelena Tomic for making my life at UCLA and in Los Angeles a wonderful time. I thank Jeremy Zechar of USC for our trip to Stromboli and for late night earthquake discussions in smoky bars in Zurich.

Thanks to Ole Peters for the many memories, from getting stuck in the sand of a

lonely beach in Baja to surfing big sets at County Line. Thanks to Thomas Decloedt for great times on Hawaii, in Sixt, in San Francisco and at C-Street, Rincon and Santa Cruz. I'm grateful to Tim Barton, Nadia Farid, Andy Hartono, Sarah Kew, Rajeev Mehta, Madhavi Nambiar, Philipp and Christian Strohm for their friendship, and Lalith Wijedoru for our culinary affair.

I'm grateful for the warmth, hospitality and friendship of Neera and Vinod Goyal, whose home has become a home away from home for me.

Thank you, Gauri, for all your support, your understanding and your sense of humor.

Finally, I'd like to thank my parents and my family for all their love and support over the many years I've been obtaining my education.

Chapter by chapter acknowledgments of academic and financial nature:

Chapter 2: We acknowledge useful discussions with A. Helmstetter, K. Felzer and J. Zhuang. This work is partially supported by NSF-EAR02-30429, by the Southern California Earthquake Center (SCEC). SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. The SCEC contribution number for this paper is 859. MJW gratefully acknowledges financial support from a NASA Earth System Science Graduate Student Fellowship.

Chapter 3: We acknowledge useful discussions with A. Helmstetter and J. Zhuang and thank the Associate Editor Frederik Simons, Ian Main, and an anonymous referee for their constructive suggestions. This work is partially supported by NSF-EAR02-30429, and by the Southern California Earthquake Center (SCEC). SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. The SCEC contribution number for this paper is 860. MJW gratefully acknowledges financial support from a NASA Earth System Science Graduate Student

Fellowship.

Chapter 5: I acknowledge insightful discussions with Ole Peters on self-organized criticality, scaling and earthquakes. I thank Gunnar Pruessner for sharing the unpublished manuscript [*Christensen et al.*, 2007] with me and for his excellent thesis on self-organized criticality.

Chapter 6: We acknowledge the Advanced National Seismic System, the Northern California Seismic Network, the US Geological Survey, the NEIC/PDE and the (Harvard) CMT project for the earthquake catalogs used in this study. We thank Margaret Hellweg of the Berkeley Seismological Laboratory and David Oppenheimer of the USGS for information about magnitude uncertainties. We gratefully acknowledge insightful inputs from Kayo Ide, David Jackson, Vladilen Pisarenko, Alexander Saichev and Jeremy Zechar. This work has been partially supported by the grant number NSF ATM-0327900. M.J.W. gratefully acknowledges financial support from a NASA Earth System Science Graduate Student Fellowship.

# VITA

| | |
|---|---|
| 1979 | Born, Hamburg, Germany |
| 1996-1998 | International Baccalaureate, International School of Beijing, PR China |
| 1998-2002 | M.Sci. in Physics, Imperial College London, UK |
| 2000-2001 | European Union Erasmus Exchange Student, Physics, École Nationale Supérieure de Physique de Grenoble, France |
| 2002-2003 | Consultant, German Agency for Technical Development (GTZ), Guatemala |
| 2003-present | Graduate Student Researcher and Teaching Assistant, Department of Earth and Space Sciences, UCLA |
| 2005 | M.S. in Geophysics and Space Physics, UCLA |
| 2006 | Ph.D. Candidate in Geophysics and Space Physics, UCLA |
| Jan-April 2007 | Visiting Scientist, Department of Management, Technology and Economics, and Institute of Geophysics, ETH Zurich, Switzerland |

# PUBLICATIONS

Sornette, D. and M. J. Werner (2005a), Constraints on the Size of the Smallest Triggering Earthquake from the Epidemic-Type Aftershock Sequence Model, Båth's Law, and Observed Aftershock Sequences, *J. of Geophys. Res., 110*, B08304, doi:10.1029/2004JB003535.

Sornette, D. and M. J. Werner (2005b), Apparent Clustering and Apparent Background Earthquakes Biased by Undetected Seismicity, *J. of Geophys. Res., 110*, B09303, doi:10.1029/2005JB00362.

Werner, M. J. and D. Sornette (2007a), Comment on "Analysis of the Spatial Distribution Between Successive Earthquakes" by Davidsen and Paczuski (2005), *Phys. Rev. Lett.*, 99, 179801, 2007.

ABSTRACT OF THE DISSERTATION

# On the Fluctuations of Seismicity and Uncertainties in Earthquake Catalogs: Implications and Methods for Hypothesis Testing

by

**Maximilian Jonas Werner**

Doctor of Philosophy in Geophysics and Space Physics

University of California, Los Angeles, 2008

Professor David D. Jackson, Chair

The randomness of the occurrences of earthquakes, together with our limited ability to detect and measure earthquakes, combine to present challenges for the testing of scientific hypothesis about earthquakes. This dissertation examines implications of these challenges and presents methods for addressing them.

In contrast to physical systems characterized by a dominating length scale, the relevant scales of earthquakes span many orders of magnitude. Our limited observations of the smallest of these scales, in the form of small, undetected earthquakes, severely impacts our ability to faithfully model observable seismicity because, as we show, small earthquakes contribute significantly to observed seismicity. Using the Epidemic-Type Aftershock Sequence model, a time-dependent model of triggered seismicity, we introduce a formalism that distinguishes between the detection threshold and a smaller size above which earthquakes may trigger others, and place constraints on its size. We derive equations that relate observed clustering parameters obtained from different thresholds. We show that parameters are biased and discuss the failure of the maximum likelihood estimator.

As an example of the power of simulation-based null hypothesis testing, we investigate a recent claim of a scaling law in the distribution of the spatial distances between successive earthquakes. Motivated by the debate on the relevance of critical phenomena to earthquakes and by the suggested contradiction of aftershock zone scaling, we analyze other regions and generate synthetic data using a realistic model that explicitly includes mainshock rupture length scales. We show that the proposed law does not hold.

Earthquake catalogs contain a wide variety of uncertainties. We quantify magnitude uncertainties and find they are more broadly distributed than a Gaussian distribution. We show their severe impact on short term forecasts by proving that the deviations of a noisy forecast from an exact forecast are power-law distributed in the tail. We further demonstrate that currently proposed consistency tests to evaluate forecasts reject noisy forecasts more often than expected at a given confidence limit. This is due to the assumed Poisson likelihood, which should be replaced by a model-specified distribution.

Finally, we propose the framework of data assimilation as a vehicle for systematically accounting for uncertainties. We review the concept of sequential Bayesian data assimilation, the purpose of which is to estimate as best as possible a desired quantity using both the noisy observations and a short-term model forecast. Sequential Monte Carlo methods are identified as a set of flexible simulation-based techniques for estimating posterior distributions. We implement a particle filter for a lognormal renewal process with noisy occurrence times and present a Bayesian solution for estimating noisy marks in a general temporal point process.

# CHAPTER 1

# Introduction

The spatial, temporal and energy patterns of earthquake occurrences display an extreme degree of randomness. The dramatic variation in the size of earthquakes, from unnoticeable events to major disasters, apparently follows no deterministic, predictable rules. Long periods of relative calm in certain regions of the world can be punctuated by extremely active periods in which large clusters of strong earthquakes reconfigure the stress field in the crust, releasing in a matter of seconds the energy slowly built up elastically in the crust during the incompatible plate movements across the globe. Their spatial distribution reflects the geometrical heterogeneity of faults and fractal plate tectonics.

But these fluctuations in seismicity can be analyzed, quantified and modeled. Since the discovery of the Omori law [*Omori*, 1894], which quantifies in a statistical distribution the random and extremely slow decay of aftershocks in time after a large (and not so large) earthquake, and the Gutenberg-Richter law [*Gutenberg and Richter*, 1944], which proposes that the energy of earthquakes is a random quantity drawn from a distribution spanning orders of magnitude but lacking a particular dominating size, seismologists have studied the large-scale patterns to quantify the interactions in space and time. Today, statistical seismicity models based on these and other empirical results are used to test their underlying scientific hypotheses about the physics of earthquakes, to calculate the risk of future earthquakes, and to probe deeper into the fluctuations and correlations.

A common theme in many of the empirical relations in seismology and employed in seismicity models is the lack of a dominating scale. Many natural phenomena can be approached by the traditional reductionist approach to isolate a process at a particular scale. For example, the waves of an ocean can be described quite accurately by a theory that entirely ignores the fact that the liquid is made out of individual molecules. Indeed, the success of most practical theories in physics depends on isolating a scale [*Wilson*, 1979], although since this recognition, much progress has been made in developing a holistic approach for processes that do not fall into this class. Given current observational evidence, earthquakes seem to belong to the set of (critical) processes characterized by a lack of characteristic length scales (i.e. scale-invariance): fluctuations on all levels are important and are in no way diminished.

The traditional reductionist approach in seismology, which, for instance, attempted to separate large (main) shocks from small (fore- or after-) shocks, is slowly giving way to the holistic approach, in which all earthquakes are created equal and seismicity is characterized by scale-invariant fluctuations of all sizes. A particularly strong model of these interactions has emerged in the concept of triggering, which places all earthquakes on the same footing: each earthquake can trigger its own events, which in turn can trigger their own events, and so on, according to the same probability distributions, and the resulting seismicity can be viewed as the cascades of triggered earthquakes that cluster in space and time.

Triggering or clustering models of seismicity are developing into community-wide agreed-upon null hypotheses that are well-understood and characterized, and are being used in a wide variety of ways: as a benchmark against which new claims may be tested; to (probabilistically) forecast earthquakes; to model and explain observations; to guide the development of improved models; to model realistically entire space-time regions of seismicity; and so on. Many of these phenomenological models belong to a

class firmly rooted in the mathematical probability theory of random point processes. The particular combination of a well-developed mathematical framework along with associated statistical tools, and the successful first-order phenomenological description of seismicity is promising in light of the obstacles still ahead.

One of these obstacles is our inability to measure earthquakes well. The obvious reason is that earthquakes initiate at a depth of several to hundreds of kilometers, which to a large degree we are unable to observe directly. Rather we rely mostly on surface instruments that measure the seismic waves excited by the earthquake in order to infer information about the event. Summary information (e.g. occurrence times, locations, magnitudes and focal mechanisms) is then usually listed in earthquake catalogs. But as a result of this complicated inverse problem, earthquake catalogs contain serious uncertainties in their representation of the data. The uncertainties are themselves often random fluctuations, but may contain strong biases. Therefore, it is critical to separate the modeling of stochastic earthquakes from the modeling of uncertain and random earthquake catalogs, which are a biased representation of earthquakes.

An example is provided by small, hard-to-detect or even undetectable earthquakes, which are not and cannot be listed in earthquake catalogs. From the point of view of seismicity as a process in which all scales are relevant and no scale can be neglected, it is apparent that even small earthquakes, undetectable by our current instruments and hence absent from earthquake catalogs, are important to the overall spatio-temporal patterns of seismicity. Small, unobserved earthquakes therefore have observable consequences. How strong is their influence, what is the nature of these consequences and what are their implications? These questions will be explored in Chapters 2, 3 and 4.

The strong stochasticity of earthquakes and our biased observations thereof combine to present unique challenges for the modeling, hypothesis-testing and forecasting of earthquakes. Some authors have argued, notably Y. Kagan [*Kagan*, 1999b], that

most studies claiming novel features are either a reformulation of already known facts about seismicity or artifacts due to earthquake catalogs errors. An efficient method for testing novel claims is to simulate data using a realistic null hypothesis, and then repeat the data analysis on the synthetic data to provide confidence limits. This simulation-based hypothesis testing can provide a benchmark for the detection of new features in seismicity. Chapter 5 presents such an example by testing a proposed scaling law in the distribution of spatial distances between subsequent earthquakes.

The fact that small earthquakes are absent from earthquake catalogs is well known. Less information is available about other uncertainties. As a result, most applications of triggering models to seismicity neglect their existence entirely and assume that the given data is exact. To develop a strong null hypothesis, in which the influence of catalog issues is minimized, we need to understand and quantify uncertainties and their effects. This is pursued in Chapter 6, with an emphasis on magnitude uncertainties and their impact on earthquake forecasts and their evaluation.

In light of the recently established earthquake forecasting and testing centers, in which the scientific hypotheses of earthquake models are tested via prospective tests, there is a growing need for robust methods that allow a systematic and rigorous treatment of data uncertainties. Chapter 7 presents the framework of data assimilation, a method developed in numerical weather prediction to estimate as best as possible the state of the atmosphere, and adapts sequential Monte Carlo methods, a set of flexible simulation-based techniques for estimating posterior distributions, to the setting of seismicity and triggering models.

## Chapter by Chapter Overview

- **Chapter 2:** Scaling laws of triggering suggest that small earthquakes are cumulatively as important for the triggered seismicity budget as rarer but larger

events. To guarantee finite seismicity, we argue for the existence of a smallest triggering earthquake, below which earthquakes do not trigger aftershocks. Introducing a formalism which distinguishes between the detection threshold and the smallest triggering earthquake, we place constraints on its size by using a simplified version of the Epidemic-Type Aftershock Sequence (ETAS) Model [*Ogata*, 1988], a powerful model of triggered seismicity, and observed aftershock sequences and Båth's law. We find that the key parameter controlling its size is given by the branching ratio of the triggering model, equal to the fraction of triggered shocks in a catalog. From upper bounds on the smallest triggering earthquake, we infer lower limits for the fraction of aftershocks.

This chapter was published in [*Sornette and Werner*, 2005a].

- **Chapter 3**: We revisit and extend the formalism introduced in Chapter 2. By considering the branching structure of one complete cascade of triggered events, we derive an apparent branching ratio and the apparent number of untriggered events, which are observed when only the structure above the detection threshold is known. We provide equations for relating (apparent) clustering parameters at different thresholds. As a result of our inability to observe the entire branching structure, inferred clustering parameters are significantly biased.

  This chapter was published in [*Sornette and Werner*, 2005b].

- **Chapter 4**: This chapter expands the discussion of the previous two chapters from the simplified ETAS model to the stochastic ETAS model. First, we argue that triggering and clustering parameters are intricately connected to the detection thresholds above which they were estimated and hence difficult to to interpret in geophysical terms. Second, separating triggered from untriggered events, commonly known as declustering, also strongly depends on the threshold, so that it cannot even in theory constitute a physically sound method. Next, we re-

view a recent article [*Saichev and Sornette*, 2006a] confirming the predictions in Chapter 3 for the stochastic number statistics of observed events. Using the fully stochastic ETAS model, we then show that the conditional intensity function, the object that defines the model, is not strictly invariant under a change of threshold due to time-dependent contributions from undetected events. Simulations and inversions confirm that parameters estimated above detection thresholds are biased. We discuss the failure of the maximum likelihood estimator, the currently preferred method for parameter estimation, to provide unbiased parameters and to provide accurate confidence limits even in an ideal setting. Finally, the recently introduced Vere-Jones model [*Vere-Jones*, 2005] is reviewed, as it eliminates the need of a detection threshold and extends the class of self-similar random measures.

This chapter extends the results of Chapters 2 and 3 to the stochastic ETAS model, proves parameter bias and summarizes developments since Chapters 2 and 3 were published. The (brief) discussion of the inaccuracy of the (Hessian-derived) confidence limits in the maximum likelihood estimator are based on [*Werner and Jackson*, 2007].

- **Chapter 5**: We review concepts and tools from the theory of critical phenomena and discuss the motivation for placing seismicity in such as framework. After briefly discussing the recent debate in the literature regarding claimed discoveries of novel and universal scaling laws in the recurrence statistics of earthquakes, we analyze a particular claim that the distribution of spatial distances between successive earthquakes obeys scale-free statistics and finite size scaling, contradicting the idea of aftershock zone scaling. We show that this "law" depends solely on one earthquake and its aftershocks, that it does not hold in other regions of the world, and that a triggering model, with aftershock zone scaling

explicitly built in, is capable of reproducing the observed power law, which supposedly contradicts the existence of aftershock zones. This chapters serves as an example of the power of simulation-based null hypothesis testing.

A section of this chapter was published in [*Werner and Sornette*, 2007a].

- **Chapter 6**: Motivated by the recent establishment of earthquake forecasting test centers, we investigate the impact of magnitude uncertainties on seismic rate estimates in seismicity models, in their forecasts and in their evaluation. First, we quantify magnitude uncertainties and find that it is more heavy-tailed than a Gaussian distribution, the commonly assumed function. Due to the fatter-than-exponential tails, we show that the impact on the forecasts of a simple clustering model, which captures the main ingredients of popular short term models, is severe. We prove that the deviations of noisy forecasts from the exact forecast are power law distributed in the tail. We study these fluctuations analytically and numerically. Finally, we show that noisy forecasts, when evaluated in currently proposed consistency tests, are rejected more frequently than expected for a given confidence limit. We discuss a first step towards making the test more adequate for short-term forecast evaluation.

  This chapter was submitted to the Journal of Geophysical Research [*Werner and Sornette*, 2007b] and is currently in review.

- **Chapter 7**: We introduce data assimilation as a framework for systematically dealing with earthquake catalog uncertainties. This chapter may be read by non-seismologists and therefore contains a brief introduction to the observational uncertainties in earthquake catalogs and to the models used for earthquake forecasting, with an emphasis on point process models. We review the few existing statistical approaches that have been applied in statistical seismology for dealing with uncertainties in point process models. So far, they remain lim-

ited to simple models and static methods. To make progress towards sequential Bayesian methods for general point processes that can be used for real-time earthquake forecasting of realistic clustering models, we review the concept of Bayesian data assimilation and focus on sequential Monte Carlo methods, a set of simulation-based techniques for estimating posterior distributions. These are described in detail to make them accessible to the seismological community. We briefly discuss the relatively thin literature of point processes in the setting of noisy observations. Sequential Monte Carlo methods typically use a state-space representation of the model, which is not entirely satisfactory for renewal processes under noisy occurrence times, but fine for more general models. We implement specific particle filters for a lognormal renewal process under noisy occurrence time observations. Finally, we present the conceptual Bayesian solution for the estimation of noisy marks in an unpredictably marked, arbitrary temporal point process.

This chapter is the basis for a manuscript under preparation [*Werner et al.*, 2007].

This thesis combines elements of seismology, statistical physics, probability theory, applied statistics and data assimilation from meteorology. The resulting interdisciplinary nature requires common terminology and concepts. However, since often relevant only to particular chapters, we chose to introduce them whenever necessary, thus making the chapters largely self-contained, but thereby increasing the length of this document. Readers familiar with certain concepts may skip the sections introducing them. The thesis contains three groups, Chapters (2, 3, 4), Chapter 5, and Chapters (6, 7). The chapters within groups should be read in succession, but the ordering of the groups is not highly important and is presented here in chronological order.

# CHAPTER 2

# Constraints on the Size of the Smallest Triggering Earthquake from the Epidemic-Type Aftershock Sequence Model, Båth's Law, and Observed Aftershock Sequences *

## 2.1 Abstract

The physics of earthquake triggering together with simple assumptions of self-similarity imply the existence of a minimum magnitude $m_0$ below which earthquakes do not trigger other earthquakes. Noting that the magnitude $m_d$ of completeness of a seismic catalog is not in general the same as the magnitude $m_0$ of the smallest triggering earthquake, we compare observed aftershock sequence parameters with the predictions made by the epidemic type aftershock sequence (ETAS) model to constrain the value of $m_0$. In particular, we use quantitative fits to observed aftershock sequences from three previous studies, as well as Båth's law, to obtain four estimates of $m_0$. We show that the branching ratio $n$ (average number of triggered earthquakes per earthquake, also equal to the fraction of aftershocks in a seismic catalog) is the key parameter con-

*An edited version of this chapter was published by AGU. Copyright (2005) American Geophysical Union. Sornette, D. and M. J. Werner (2005), Constraints on the size of the smallest triggering earthquake from the epidemic-type aftershock sequence model, Båth's law, and observed aftershock sequences, *J. of Geophys. Res., 110*, B08304, doi:10.1029/2004JB003535. Reproduced by permission of American Geophysical Union.

trolling the estimate of the minimum triggering magnitude $m_0$. Conversely, physical upper bounds for $m_0$ estimated from rate and state friction indicate that at the very least 55 percent of all earthquakes are aftershocks.

## 2.2 Introduction

Scale invariance in earthquake phenomena is widely manifested empirically, in the Gutenberg-Richter (GR) magnitude-frequency relation, in the Omori aftershock decay rate, and in many other relationships. Scale-invariance means that there are no preferred length scales in seismogenic processes and in spatio-temporal structures. However, there are many reports that purport to identify characteristic scales. As emphasized by [*Matsu'ura*, 1999; *Aki*, 2000; *Sornette*, 2002], the search for characteristic structures in specific fault zones could allow the separation of large earthquakes from small ones and thus advance earthquake prediction.

Although there is clear evidence of deviations from self-similarity at large scales [*Kagan*, 1999a; *Pisarenko and Sornette*, 2003], the issue is much murkier at small scales. For instance, *Iio* [1991] reports a lower magnitude cutoff $m_{\min} \approx -1.4$ for very small aftershocks of the 1984 Western Nagano Prefecture, Japan, earthquake ($m_{JMA} = 6.8$) in spite of the fact that the high sensitivity of the observation system (focal distances less than $1$ km and very low ground noise) would have permitted to detect much smaller magnitudes. Based on induced seismicity associated with deep gold mines, *Richardson and Jordan* [2002] find a lower magnitude cutoff $m_{\min} \approx 0$ for friction-dominated earthquakes, while fracture-dominated earthquakes have no lower cutoff but an upper cut-off of magnitude $\approx 1$. Using deep bore-hole recordings, *Abercrombie* [1995a,b] found that small earthquakes exist down to at least magnitude $0$ and that source scaling relationships hold down to at least $-1$. Based on seismic power spectra, on the evidence of a low-velocity low-Q zone reaching the top of the ductile

part of the crust and on seismic guided waves in fault zones, *Li et al.* [1994] argue for a characteristic earthquake magnitude of about 3 associated with the width of fault zones. Another characteristic magnitude in the range $4 - 5$ is proposed by *Aki* [1996], based on the simultaneous change of coda $Q^{-1}$ and the fractional rate of occurrence of earthquakes in this magnitude interval. At Parkfield, *Heimpel and Malin* [1998] found evidence of a transition from creep-dominated slip to earthquake-dominated slip taking place in the range of magnitudes close to $M = 0.9$, above their detection limit of $M = 0.3$. The authors underline that their results do not suggest the existence of a minimum earthquake size, but rather indicate a nucleation scale in their stochastic rupture model. They further consider it likely that this scale varies with the geological setting. Similarly, *Marone and Kilgore* [1993] suggested that the critical slip distance over which strength breaks down during nucleation in models of velocity-weakening friction scales with shear strain in fault zones. Therefore, if the critical slip scale fixes a minimum earthquake size, as we consider below, then the smallest earthquake may be non-universal and change with the maturity or gouge thickness of the fault.

The existence of a discrete hierarchy of scales has in addition been suggested by *Sornette and Sammis* [1995] based on the analysis of accelerated seismicity prior to large earthquakes and recently by *Pisarenko et al.* [2004] by using a non-parametric measure of deviations from power laws applied to the magnitude-frequency distributions of earthquakes in subduction zones. Evidence of a hierarchy of scales is also found in fragmentation and rupture processes [*Sadovskii*, 1999; *Geilikman and Pisarenko*, 2000; *Sahimi and Arbabi*, 1996; *Ouillon and Sornette*, 1996; *Johansen and Sornette*, 1998; *Suteanu et al.*, 2000].

From a theoretical point of view, the equation of motion for a continuum solid is scale-independent, suggesting that deformation processes in solids should produce self-similar patterns manifested in power law statistics. However, the symmetry of an

equation does not guarantee that the solutions of this equation share the same symmetry. The difference (when it exists) in the symmetry between a solution and its governing equation is known as the phenomenon of "spontaneous symmetry breaking" [*Consoli and Stevenson*, 2000] and underlies a large variety of systems (explaining for instance the non-zero masses of fundamental particles [*Englert*, 2004]. Of course, length scales associated with rheology and existing structures can produce deviations from exact self-similarity. For instance, a transition from stable creep to a dynamic instability at a nucleation size whose dimensions depend on frictional and elastic parameters defines a minimum earthquake size [*Dieterich*, 1992], estimated at magnitude $\approx -3$ by *Ben-Zion* [2003]. This minimum size corresponds only to events triggered according to the mechanism of unstable sliding controlled by slip weakening and thus concerns friction-dominated earthquakes.

A different perspective is offered by models of triggered seismicity in which earthquakes (so-called foreshocks and mainshocks) trigger other earthquakes (so-called mainshocks and aftershocks, respectively). Recent studies suggest that maybe more than 2/3 of events are triggered by previous earthquakes (see *Helmstetter and Sornette* [2003b]) and references therein). In this context, the relevant question is no longer how small is the smallest earthquake but how small is the smallest earthquake which can trigger other earthquakes (and, in particular, larger earthquakes).

The effects of seismicity below the detection threshold in models of triggered seismicity are also considered in the very closely related [*Sornette and Werner*, 2005b]. In particular, earthquakes too small to detect with the current network sensitivity are shown to bias the estimates of the branching ratio and the background event rate. The article also uses the estimates of the smallest triggering earthquake established below to link the apparent (measured) percentage of triggered quakes in a seismic catalog to the real fraction.

## 2.3 The ETAS Model and the Smallest Triggering Earthquake

To make the discussion precise, let us consider the epidemic-type aftershock sequence (ETAS) model, in which any earthquake may trigger other earthquakes, which in turn may trigger more, and so on. Introduced in slightly different forms by *Kagan and Knopoff* [1981] and *Ogata* [1988], the model describes statistically the spatio-temporal clustering of seismicity.

The ETAS model consists of three assumed laws about the nature of seismicity viewed as a marked point-process. We restrict this study to the temporal domain only, summing over the whole spatial domain of interest. First, the magnitude of any earthquake, regardless of time, space or magnitude of the mother shock, is drawn randomly from the exponential Gutenberg-Richter (GR) law. Its normalized probability density function (pdf) is expressed as

$$P(m) = \frac{b \ln(10) 10^{-bm}}{10^{-bm_0} - 10^{-bm_{max}}}, \qquad m_0 \leq m \leq m_{max}, \tag{2.1}$$

where the constant exponent $b$ is typically close to one, and the cut-offs $m_0$ (see below) and $m_{max}$ serve to normalize the pdf. The upper cut-off $m_{max}$ is introduced to avoid unphysical, infinitely large earthquakes. Its value was estimated to be in the range $8 - 9.5$ [*Kagan*, 1999]. As the impact of a finite $m_{max}$ is quite weak in the calculations below, replacing the abrupt cut-off $m_{max}$ by a smooth taper would introduce negligible corrections to our results.

Second, the model assumes that direct aftershocks are distributed in time according to the modified "direct" Omori law (see *Utsu et al.* [1995] and references therein). Denoting the usual Omori law exponent by $p = 1 + \theta$ and assuming $\theta > 0$, the normalized pdf of the Omori law can be written as

$$\Psi(t) = \frac{\theta c^{\theta}}{(t + c)^{1+\theta}}, \tag{2.2}$$

where $t$ is the time since the earthquake and $c$ is a constant.

Third, the number of direct aftershocks of an event of magnitude $m$ is assumed to follow the productivity law:

$$\rho(m) = k\ 10^{\alpha(m-m_0)}, \qquad m_0 \leq m \leq m_{max}, \tag{2.3}$$

where $k$ and $\alpha$ are constants. Note that the productivity law (2.3) is zero below the cut-off $m_0$, i.e. earthquakes smaller than $m_0$ do not trigger other earthquakes; this is typically assumed in studies using the ETAS model. The existence of the small-magnitude cut-off $m_0$ is necessary to ensure the convergence of the models of triggered seismicity (in the statistical physics of phase transitions and in particle physics, this is called an "ultra-violet" cut-off which is often necessary to make the theory convergent). Below, we show that there are observable consequences of the existence of the cut-off $m_0$ thus providing constraints on its physical value.

Since the present formulation of the ETAS model requires cut-offs to ensure its convergence, it is interesting to mention the variation recently introduced by *Vere-Jones* [2005]. This modified model is completely self-similar yet well-defined and convergent. No cut-offs break the self-similarity. To remove all scales, he first requires the Omori law constant $c$ to be a function of magnitude so that the plateau following large mainshocks lasts longer than for smaller shocks. Secondly, $\alpha$ is set equal to $b$. Thirdly, he introduces a function $S$ that penalizes a large departure of a daughter's magnitude from the mother's magnitude. This implies that the magnitudes of the daughters are distributed according to a modified GR law, with a bend around the mother magnitude. This completely self-similar model without cut-offs requires a conditioning of aftershock magnitudes on mother magnitudes, for which observational evidence remains to be established. This question is of fundamental importance in order to clarify whether the ETAS cut-off magnitude $m_0$ is of real physical relevance to earthquake triggering. We will not consider the model of *Vere-Jones* [2005] further

here.

The key parameter of the ETAS model is defined as the number $n$ of direct aftershocks per earthquake, averaged over all magnitudes. Here, we must distinguish between the two cases $\alpha = b$ and $\alpha \neq b$:

$$\begin{aligned} n &\equiv \int_{m_0}^{m_{max}} P(m)\rho(m)dm \\ &= \begin{cases} \frac{kb}{b-\alpha}\left(\frac{1-10^{-(b-\alpha)(m_{max}-m_0)}}{1-10^{-b(m_{max}-m_0)}}\right), & \alpha \neq b \\ \frac{kb\ln(10)(m_{max}-m_0)}{1-10^{-b(m_{max}-m_0)}}, & \alpha = b. \end{cases} \end{aligned} \tag{2.4}$$

Three regimes can be distinguished based on the value of $n$. The case $n < 1$ corresponds to the subcritical regime, where aftershock sequences die out with probability one. The case $n > 1$ describes unbounded, explosive seismicity that may lead to finite time singularities [*Sornette and Helmstetter*, 2002]. The critical case $n = 1$ separates the two regimes.

The fact that we use the same cut-off for the productivity cut-off and the Gutenberg-Richter (GR) cut-off is not a restriction as long as the real cut-off for the Gutenberg-Richter law is smaller than or equal to the cut-off for the productivity law. In that case, truncating the GR law at the productivity cut-off just means that all smaller earthquakes, which do not trigger any events, do not participate in the cascade of triggered events. This should not be confused with the standard but incorrect procedure in many previous studies of triggered seismicity of simply replacing the GR and productivity cut-off $m_0$ with the detection threshold $m_d$ in equations (2.1) and (2.3) (see, for example, [*Ogata*, 1988; *Kagan*, 1991; *Guo and Ogata*, 1997; *Ogata*, 1998; *Console et al.*, 2003b; *Ogata et al.*, 2003; *Ogata*, 2004; *Zhuang et al.*, 2004]). This may lead to a bias in the estimated parameters.

The realization that the detection threshold $m_d$ and the triggering threshold $m_0$ are different leads to the question of whether we can extract the size of the smallest

triggering earthquake. Here, we infer useful information on $m_0$ from the physics of earthquake triggering embodied in the simple ETAS formalism, from Båth's law, and from available catalogs.

We will assume that the detection threshold $m_d$ of a seismic catalog is (currently still) larger than the smallest triggering earthquake $m_0$. This assumption seems justified since, for instance, *Helmstetter et al.* [2005a] found that $m = 2$ earthquakes trigger their own sequences of (possibly larger) magnitudes. Their Figure 1 presents evidence that the scaling of aftershock productivity continues down to at least magnitude 2. This implies that $m_0$ has not yet been observed directly and is below the detection threshold.

There is no loss of generality in considering one (independent) branch (sequence or cascade of aftershocks) of the ETAS model. Let an independent background event of magnitude $M_1$ occur at some origin of time. We will refer to independent (non-triggered) background events as mainshocks or initial shocks and any triggered events as aftershocks, independent of magnitude. The mainshock will trigger direct aftershocks according to the productivity law (2.3). Each of the direct aftershocks will trigger their own aftershocks, which in turn produce their own, and so on. Averaged over all magnitudes, each aftershock produces $n$ direct offspring according to (2.4). Thus, over all time, we can write the average of the total number $N_{total}$ of direct and indirect aftershocks of the initial mainshock as an infinite sum over terms of (2.3) multiplied by $n$ to the power of the generation [*Helmstetter and Sornette*, 2003b], which can be expressed for $n < 1$ as:

$$\begin{aligned} N_{total} &= \rho(M_1) + \rho(M_1)n + \rho(M_1)n^2 + ... \\ &= \frac{k\ 10^{\alpha(M_1 - m_0)}}{1-n} \end{aligned} \tag{2.5}$$

However, since we can only detect events above the detection threshold $m_d$, the total number of observed aftershocks $N_{obs}$ of the sequence is simply $N_{total}$ multiplied

by the fraction of events above the detection threshold, given by $(10^{b(m_{max}-m_d)} - 1)/(10^{b(m_{max}-m_0)} - 1)$ according to the GR distribution. The observed number of events in the sequence is therefore

$$\begin{aligned} N_{obs} &= N_{total} \left( \frac{10^{b(m_{max}-m_d)} - 1}{10^{b(m_{max}-m_0)} - 1} \right) \\ &= \frac{k \ 10^{\alpha(M_1-m_0)}}{1-n} \left( \frac{10^{b(m_{max}-m_d)} - 1}{10^{b(m_{max}-m_0)} - 1} \right) . \end{aligned} \tag{2.6}$$

Equation (2.6) predicts the average observed number of direct and indirect aftershocks of a mainshock of magnitude $M_1 > m_d$. To estimate $m_0$, we need to eliminate or find estimates of the three unknowns $n$, $k$, and $N_{obs}$. We can eliminate $k$ through the expression (2.4) for $n$, leaving $n$ and $N_{obs}$. The mean number of observed aftershocks as a function of mainshock magnitude $M_1$ was estimated by *Helmstetter et al.* [2005a] and *Felzer et al.* [2003] and can also be obtained from Båth's law. In the following sections, we use these three estimates for $N_{obs}$ and thus obtain $m_0$ as a function of the only remaining unknown $n$. Acknowledging the controversy surrounding the estimation of the percentage of aftershocks in a catalog, we nevertheless use existing estimates of $n$ to finally obtain quantitative values for $m_0$.

As we rely on fits and estimates of constants to obtain $m_0$, it is useful to attempt an error estimation of $m_0$ given the variation of these constants. In particular, we can solve equation (2.6) for $m_0$ and find its variation $\Delta m_0$ with $\Delta n$, which amounts to assuming that the leading error in $m_0$ comes from the relatively poorly known $n$. This leads to

$$\Delta m_0 = \frac{1}{(b-\alpha)\ln(10)} \frac{\Delta n}{n - n^2} \tag{2.7}$$

For $\Delta n \simeq 0.2$ and $b - \alpha \simeq 0.2$, one obtains $\Delta m_0 \simeq 1.6$ for $n = 0.5$, and $\Delta m_0 \simeq 4.4$ for $n = 0.9$. Given that the other parameters may also contain errors (see also below) and that the estimates of $n$ may be biased by undetected seismicity [*Sornette and Werner*, 2005b], these error estimates may themselves contain large errors. We

therefore stress that the following sections present order of magnitude calculations.

## 2.4 Constraint on the Smallest Triggering Earthquake from the ETAS Model and Observed Estimates of Aftershock Numbers

Following the recipe outlined above, we begin by using the estimates of the observed number of aftershocks $N_{obs}$ obtained by *Helmstetter et al.* [2005a] in order to find $m_0$ as a function of $n$. *Helmstetter et al.* [2005a] sidestepped the problems associated with maximum likelihood estimates of the complete model parameters by fitting stacked observed aftershock rates within pre-defined space-time windows using the formula

$$\lambda_{fit}(t) = \frac{K_{fit}\ 10^{\alpha M_1 - b m_d}}{t^{p_{fit}}}\ , \tag{2.8}$$

based on the scaling laws (the GR law, the Omori law, and the productivity law) discussed above. The constant $K_{fit}$ includes all aftershocks, direct and indirect, and thus corresponds to a global renormalized constant different from $k$ in the ETAS productivity law (2.3). Furthermore, $p_{fit}$ is also a global exponent, which may be different from the local exponent $1+\theta$ of the ETAS model for $n$ close to 1 and at not too long times, as explained in *Sornette and Sornette* [1999b] and *Helmstetter and Sornette* [2002]. The total number of aftershocks is then obtained by integrating over an un-normalized Omori law according to [*Helmstetter et al.*, 2005a]:

$$\begin{aligned} N_{fit}(T, M_1) &= \int_c^T \lambda_{fit}(t)dt \\ &= K_{fit}\ 10^{\alpha M_1 - b m_d}\ \frac{T^{1-p_{fit}} - c^{1-p_{fit}}}{1-p_{fit}}. \end{aligned} \tag{2.9}$$

For $p_{fit} < 1$, this expression diverges as $T$ increases to infinity. But, as it has been shown that the exponent of the observed global Omori law converges to a value $1+\theta >$

1 at large times for $n < 1$ [*Sornette and Sornette*, 1999b; *Helmstetter and Sornette*, 2002] the time factor converges also to $(\theta c^{\theta})^{-1}$. Under the assumption that $p_{fit} = 1 + \theta > 1$, valid for $n$ not very close to the critical value 1 *Sornette and Sornette* [1999b]; *Helmstetter and Sornette* [2002], equation (2.9) may then be rewritten as

$$N_{fit}(M_1) = K_{fit}\, 10^{\alpha M_1 - b m_d}\, (\theta c^{\theta})^{-1}. \tag{2.10}$$

Equating the ETAS model prediction $N_{obs}(M_1)$ given by (2.6) with the empirical estimate $N_{fit}(M_1)$ given by (2.10), and eliminating the unknown $k$ through the expression for $n$ in (2.4) leads to an equation for $m_0$ as a function of $n$:

$$m_0 = \frac{1}{(b-\alpha)\ln(10)} \times$$
$$\ln(10^{(\alpha-b)m_{max}} + \frac{b-\alpha}{b}\frac{n}{1-n}\frac{\theta c^{\theta}}{K_{fit}}(1 - 10^{-b(m_{max}-m_d)}))$$

for $\alpha \neq b$ and

$$m_0 = m_{max} - (\frac{n}{1-n})\frac{\theta c^{\theta}}{K_{fit}}\frac{1 - 10^{-b(m_{max}-m_d)}}{b\ln(10)} \tag{2.11}$$

for $\alpha = b$.

Expression (2.11) shows that, provided an estimate of the branching ratio $n$ is available, we can deduce $m_0$, since the other quantities can be measured independently: $b$ is close to 1, $\alpha$ is usually between 0.5 and 1, $m_d$ depends on catalogs but is often about 3, $c$ is typically close to $0.001$ days and $K_{fit}$ in equation (2.8) is obtained from the calibration of the productivity of earthquakes as a function of their magnitude. In Table 1 of their study, *Helmstetter et al.* [2005a] report values for $K_{fit}$ in the range from $0.0009$ to $0.0193$ (days)$^{p-1}$, $0.94 \leq \alpha \leq 1.16$, $b \approx 0.95$, and $m_d = [2, 3]$. They find $c < 0.001$days and $p = 0.9$. We will thus assume $\theta = 0.1$ (see above).

We note that $m_d$ appears in the expression (2.11) for $m_0$. Clearly, a detection threshold that evolves with seismic technology should not influence the physics of triggering. We thus expect $m_0$ to be independent of $m_d$. The reason $m_d$ does appear

in the expression can be traced to the GR law (2.1), which is normalized over the magnitude interval from $m_0$ to $m_{max}$. When integrated to give the probability of $m$ lying in the range from $m_d$ to $m_{max}$, the factor involving $m_d$ does not enter as simply as in the formulation (2.8) of *Helmstetter et al.* [2005a]. Therefore the factors do not cancel out when comparing the ETAS prediction with the assumed parameterization of *Helmstetter et al.* [2005a] and $m_d$ remains in the equations. Assuming that the GR law is correctly normalized in the present ETAS model, this implies a (weak) dependence of $K_{fit}$ on $m_d$. Given the correlation between $\alpha$ and $K_{fit}$ (see below), the estimates of $\alpha$ may thus also depend on $m_d$. Finally, for practical purposes we note that for any reasonable values of the other parameters, the influence of $m_d$ is negligible.

The estimate of $m_0$ that we are trying to obtain relies on the adequacy of the model used here and on the stability and reliability of the quoted parameters. For now, we sidestep any possible difficulties in the determination of the parameters and present in Figure 2.1 the magnitude of the smallest triggering earthquake $m_0$ as a function of the average number $n$ of direct aftershocks per mainshock for a range of parameters. For $n = 0$, $m_0$ equals the largest possible earthquake $m_{max}$, representing the limit that earthquakes do not trigger any aftershocks. At the other end, for $n = 1$, the formula predicts that $m_0$ diverges to minus infinity. Recall that $n = 1$ corresponds to the system being exactly at the critical value of a branching process and the statistical average $N_{obs}(m)$ of the total number of events triggered over all generations by a mother event of magnitude $m$ becomes infinite. Of course, individual sequences have a finite lifetime and a finite progeny with probability one and the theoretical average loses its significance due to the fat-tailed nature of the corresponding distribution [*Athreya and Ney*, 1972; *Saichev et al.*, 2005; *Saichev and Sornette*, 2004]. Therefore, the prediction on $m_0$ becomes unreliable for $n$ close to 1 (how close to 1 depends on $b - \alpha$ which controls the amplitude of the fluctuations from realization to realization).

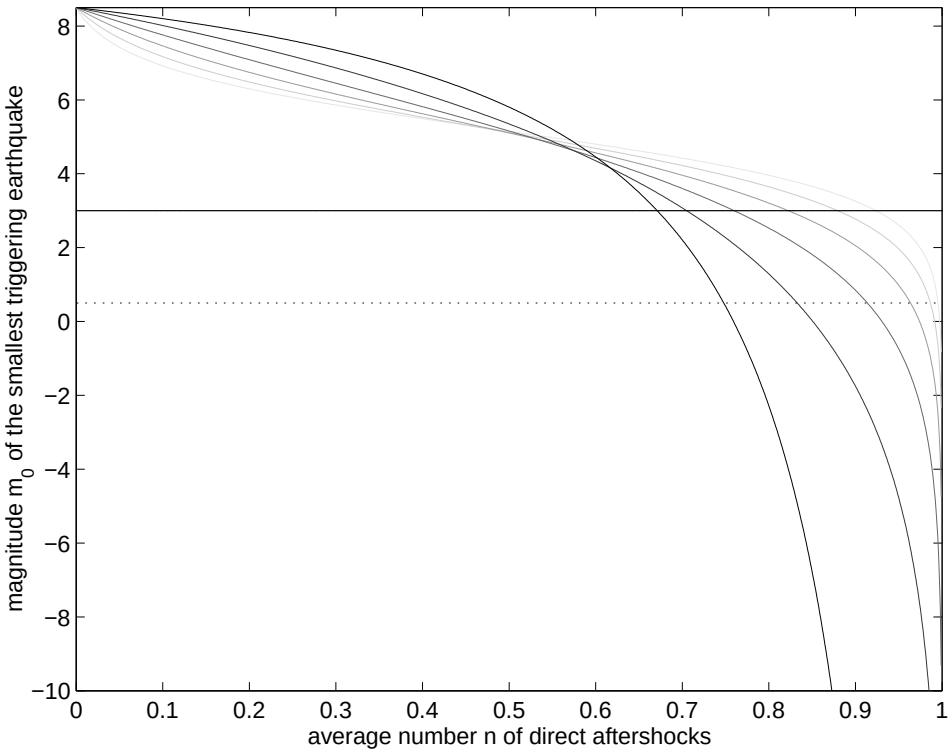


Figure 2.1: The magnitude $m_0$ of the smallest triggering earthquake as a function of the average number $n$ of direct aftershocks estimated from fits to observed aftershock sequences and the ETAS model for values of, from light to dark, $[\alpha = 0.5, K_{fit} = 9.6103]$, $[\alpha = 0.6, K_{fit} = 2.3312]$, $[\alpha = 0.7, K_{fit} = 0.5655]$, $[\alpha = 0.8, K_{fit} = 0.1372]$, $[\alpha = 0.9, K_{fit} = 0.0333]$, $[\alpha = b = 1, K_{fit} = 0.0081]$. We assume parameters $b = 1$, $m_{max} = 8.5$, $\theta = 0.1$, $c = 0.001$ days. $K_{fit}$, in days$^{1-p}$, was estimated by ***Helmstetter et al.*** [2005a] and here calculated from the values of $\alpha$ through their correlation (see Figure 2.2). The horizontal lines represent two upper limits on $m_0$: the first one is derived from a typical detection threshold at $m_d = 3$ (solid line) while the second (lower) one was obtained from estimates of the critical slip $d_c = 100\mu$m in rate and state friction resulting in $m = 0.5$ (dotted line). The upper limits on $m_0$ conversely provide a lower bound on the percentage of aftershocks in an earthquake catalog and/or $\alpha$.

For a wide range of $n$ and combinations between $\alpha$ and $K_{fit}$, the magnitude of the smallest triggering earthquake lies between $0$ and $-10$. Only for values of $n$ above $0.9$ does the size of $m_0$ become smaller than $-10$. For reference, a magnitude $-10$ event roughly corresponds to a fault of length 1mm, i.e. to grain size.

Given that we expect $m_0$ to be smaller than the detection threshold $m_d$, the horizontal line at $m_d = 3$ serves as a (very) conservative estimate of the upper limit of $m_0$ and thus provides constraints on the combination of parameters $\alpha$, $K_{fit}$ and $n$. For example, for $\alpha = 1$, at least 65 percent of all earthquakes must be aftershocks. This lower limit increases drastically to about 90 percent for $\alpha = 0.5$. Note that we extrapolate $K_{fit}$ from the observed values for $\alpha$ around one to smaller values of $\alpha$ using an exponential fit (see below).

We can obtain another external bound on $m_0$ from estimates of the minimum slip required before static friction drops to kinetic friction and unstable sliding begins, according to models of velocity-weakening friction. For example, the parameter $d_c$ in rate and state dependent friction [*Dieterich*, 1992, 1994] was estimated at $0.5$m from seismograms [*Ide and Takeo*, 1997] and similarly at $40 - 90$cm from slip-velocity records [*Mikumo et al.*, 2003], although both probably correspond to upper bounds. Estimates of $d_c$ from laboratory friction experiments give $1$ to $100\mu$m, approximately 4 to 6 orders of magnitude less than the upper bound determined by seismic studies. One could conclude that either the upper bound from seismic studies is so extreme as to render the comparison to laboratory studies meaningless, or the slip weakening process is in fact different at laboratory scales [*Kanamori and Brodsky*, 2004]. *Scholz* [1998] related the critical slip to a minimum nucleation length $L_c = G\, d_c/([B-A]\sigma)$, where $G$ is the shear modulus, $\sigma$ is the effective normal stress, and $B - A$ is a material property. Following *Lapusta and Rice* [2003], we take the values $G = 30,000$ MPa, $B - A = 0.004$, $\sigma = 50$ MPa, and $d_c = 100\mu$m to obtain $L_c = 10$ m. If we assume that

the minimum slip needed to initiate stable sliding scales with the minimum length of a friction-based earthquake, then, neglecting fracture-based earthquakes, $L_c$ corresponds to the size of the smallest earthquake. Given that the smallest triggering earthquake must be equal to or larger than the smallest earthquake, but that the estimate of $L_c$ is an upper limit, we use these values for $L_c$ as an upper limit of the smallest triggering earthquake. From the relations between fault length, moment and moment magnitude [*Kanamori and Brodsky*, 2004] with $L_c = 10$ m and a stress drop of $3$ MPa, we obtain an upper limit of magnitude $0.5$ for the smallest triggering earthquake. This upper limit is represented in Figure 2.1 as the lower, dotted horizontal line.

*Felzer et al.* [2002] have used $\alpha = b$ on the basis of an argument of self-similarity. *Helmstetter et al.* [2005a] also argue for a value of $\alpha$ essentially indistinguishable from $b$ based on fits of stacked aftershock decay rates in pre-defined space-time windows. Other studies have found $\alpha$ smaller and much smaller than $b$ (see, for example, [*Console et al.*, 2003b; *Helmstetter*, 2003; *Zhuang et al.*, 2004]. In view of the lack of consensus and to keep the discussion independent of the estimation problem, we use the correlation we found between the parameters $K_{fit}$ and $\alpha$ estimated in *Helmstetter et al.* [2005a] to extrapolate to smaller $\alpha$. The existence of such a correlation is standard in joint estimations of several parameters and can be deduced from the inverse of the Fisher matrix of the log-likelihood function *Rao* [1965]. Such correlation can also be enhanced if the model is misspecified. We performed an exponential least-square fit to the scatter plot (see Figure 2.2) to obtain a relationship between the parameters and then calculated an estimate of $K_{fit}$ for smaller values of $\alpha$ according to the best fit relationship $K_{fit} = 11441.3285 \exp(-\alpha/0.07056)$. In the absence of other estimates, this method provides one possibility to extrapolate to small $\alpha$. The resulting curves for $m_0$ are plotted in Figure 2.1.

Delaying the discussion on the estimation problem until the end of the section, we

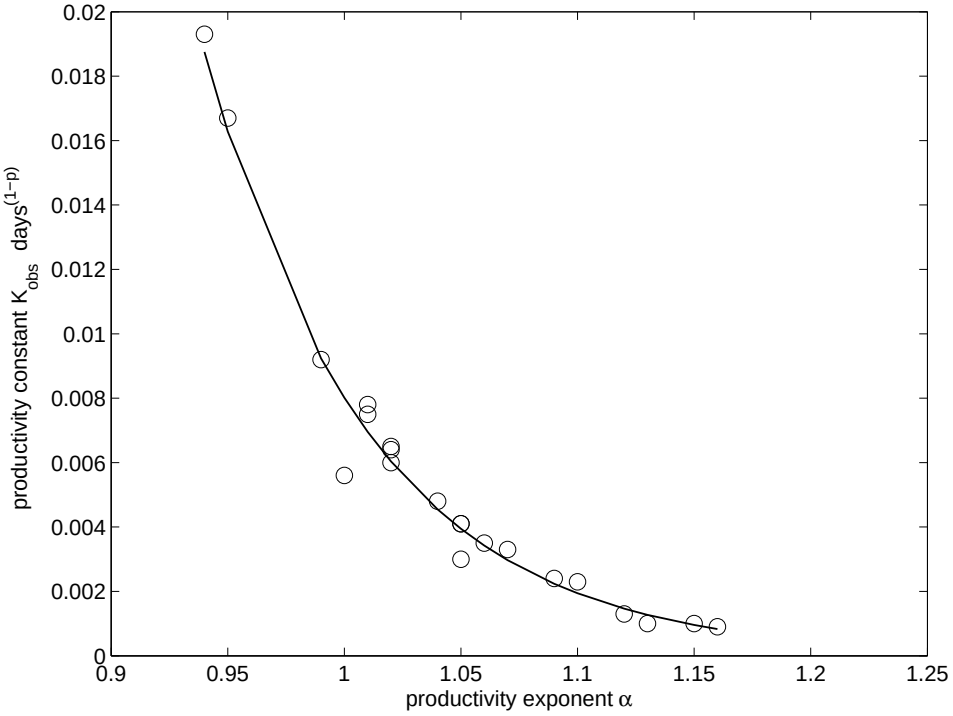


Figure 2.2: The correlation between the values of $K_{fit}$ and $\alpha$ taken from Table 1 of *Helmstetter et al.* [2005a]. The line is a least-squares exponential fit $K_{obs} = 11441.3285 \exp(-\alpha/0.07056)$. The extrapolation of this fit for smaller values of $\alpha$ was used to obtain the values for $K_{fit}$ in Figure 2.1.

use (2.11) together with existing estimates of the percentage of aftershocks in seismic catalogs (equivalent to $n$ [*Helmstetter and Sornette*, 2003b]) to constrain $m_0$. We note, however, that different declustering techniques lead to different estimates. No consensus exists on which method should be trusted most. For example, *Gardner and Knopoff* [1974] found that about 2/3 of the events in the Southern California catalog are aftershocks. With another method, *Reasenberg* [1985] found that 48% of the events belong to a seismic cluster. *Davis and Frohlich* [1991] used the ISC catalog and, out of 47500 earthquakes, found that 30% belong to a cluster, of which 76% are aftershocks and 24% are foreshocks. Recently, using different versions of the ETAS

model, *Zhuang et al.* [2004] have performed a careful inversion of the JMA catalog for Japan using a magnitude $m_d = 4.2$ for the completeness of their catalog. They provide three estimates of the branching ratio for their best model: $n = 0.42$, $0.55$, and $0.46$.

Whether any of these methods estimate $n$ correctly and without bias remains questionable. In particular, the branching ratios as calculated by *Zhuang et al.* [2004] and others may be significantly biased by the assumption that $m_d = m_0$, which can be shown to lead to an apparent branching ratio modified by the impact of hidden seismicity below the catalog completeness [*Sornette and Werner*, 2005b]. Moreover, there are problems with the maximum likelihood estimation (see for instance *Helmstetter et al.* [2005a]). However, in the absence of better estimations, we nevertheless use the above values as rough estimates of $n$. Given the range of $\alpha$ and $K_{fit}$, $m_0$ is still not very well constrained for values of $n$ near one (see Figure 2.1). For example, for 85 percent aftershocks ($n = 0.85$), $m_0$ ranges from $-10$ to an unrealistic $4$ depending on the values of $\alpha$ and $K_{fit}$. This argument could be used to rule out the combination $n = 0.85$ and $\alpha = 0.5$. In fact, for $m_0$ to be smaller than $m_d = 3$, at least 65 percent of earthquakes are aftershocks. For $m_0$ to be smaller than the upper limit estimated from $d_c$, at least 75 percent must be aftershocks. Both fractions must increase for a smaller $\alpha$.

We can also use the values obtained by *Felzer et al.* [2003] to constrain $m_0$. The authors also used finite space-time windows in which they fitted aftershock sequences with parameters for a global sequence according to

$$C_T = \frac{A_T}{1 - p_T}((t + c_T)^{1-p_T} - {c_T}^{1-p_T}), \quad p_T \neq 1, \tag{2.12}$$

where $t$ is the selected duration of the sequences, $p_T$ is the global Omori exponent, $c_T$ is the Omori constant, and $A_T$ is the productivity. Assuming that the local Omori exponent is $p = p_T = 1 + \theta_T$, expression (2.12) can be rewritten for the infinite time

limit as

$$C_T = \frac{A_T}{\theta_T {c_T}^{\theta_T}}. \tag{2.13}$$

The values obtained are listed in Table 3 of their study: $A_T = 0.116$ days$^{1-p_T}$, $p_T = 1.08$ and $c_T = 0.014$ days. These values hold for a typical California aftershock sequence of a magnitude $M_1 = 6.04$ mainshock, a detection threshold of $m_d = 4.8$, and $\alpha = b = 1$.

As before, we equate the ETAS model prediction (2.6) with the observation (2.13), eliminate $k$ through expression (2.4) for $n$ (where $\alpha = b$) and obtain an equation for $m_0$ as a function of $n$:

$$\begin{aligned} m_0 &= m_{max} - \frac{n}{1-n}\frac{(1-10^{-b(m_{max}-m_d)})}{b\ln(10)} \\ &\quad \times \frac{\theta_T c^{\theta_T}}{A_T} 10^{b(M_1-m_d)}, \qquad \alpha = b = 1. \end{aligned} \tag{2.14}$$

This expression for $m_0$ is shown in Figure 2.3 (solid curve). As in equation (2.11), $m_d$ remains artifactually in the equation due to a dependence of $A_T$ on the detection threshold. Since the parameters were obtained with $\alpha = b = 1$, we do not alter the values of $\alpha$ and obtain only one curve.

Another estimate of $m_0$ can be obtained from values estimated by ***Reasenberg and Jones*** [1989]. Their aftershock rate above $m_d$ due to a mainshock $M$ is modeled as

$$\lambda_{RJ}(t, M) = 10^{a+b(M-m_d)}\ (t+c)^{-p}\ . \tag{2.15}$$

Again, we integrate over time, assuming $p = 1 + \theta > 1$ to obtain

$$N_{RJ}(t, M) = 10^{a+b(M-m_d)}\ (\theta\ c^{\theta})^{-1}\ . \tag{2.16}$$

As before, we equate expression (2.16) with the ETAS prediction (2.6), eliminate $k$ through equation (2.4) for $n$ (where $\alpha = b$), and arrive at a third estimate of $m_0$ as a

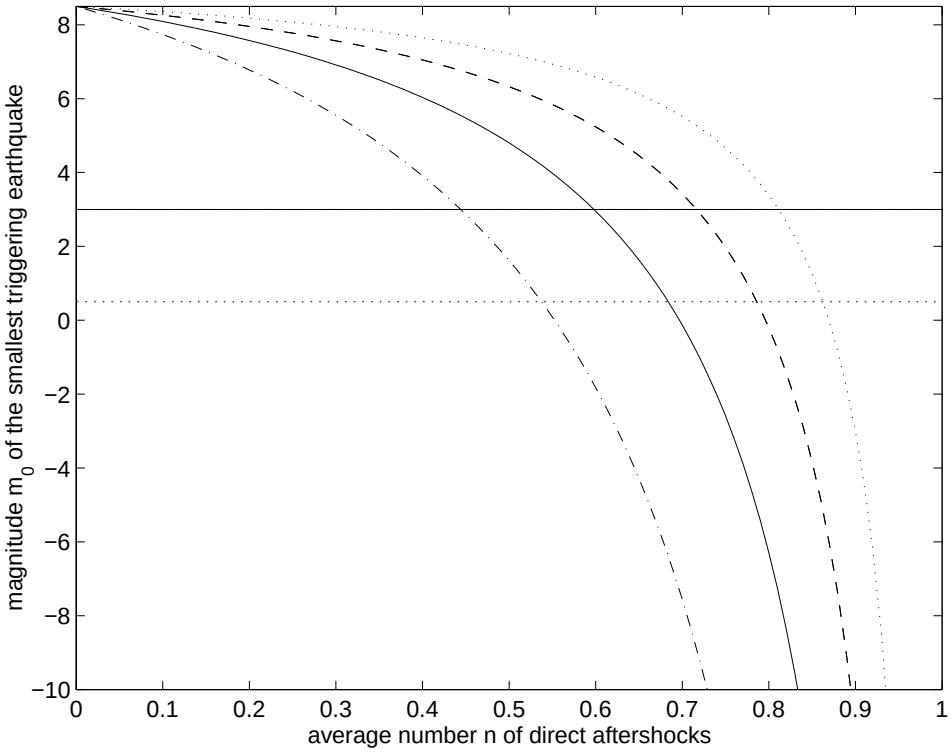


Figure 2.3: The magnitude $m_0$ of the smallest triggering earthquake as a function of the average number $n$ of direct aftershocks estimated from fits to observed aftershock sequences and the ETAS model according to *Felzer et al.* [2003] (solid) and *Reasenberg and Jones* [1989] (dotted). For the solid curve, we used the parameters of *Felzer et al.* [2003]: $\theta_T = 0.08$, $c_T = 0.014$ days, $M_1 = 6.04$ and $A_T = 0.116$ days$^{1-p_T}$. For the dotted curve, we used the parameters of *Reasenberg and Jones* [1989]: $\theta = 0.08$, $c = 0.05$ and $a = -1.67$. We also assumed $m_{max} = 8.5$, $m_d = 3$, $\alpha = b = 1$. For comparison, we include the curves corresponding to the special case $\alpha = b$ for the fit according to *Helmstetter et al.* [2005a]) (dashed) (from Figure 2.1) and the constraint due to Båth's law (dash-dotted) (from Figure 2.4). The horizontal lines represent two upper limits on $m_0$: the first one was derived from the assumption that $m_0$ is larger than a typical detection threshold $m_d = 3$ (solid line), while the second (lower) one was obtained from estimates of the critical slip $d_c = 100\mu$m in rate and state friction giving $m = 0.5$ (dotted line). The upper limits on $m_0$ conversely provide lower bounds for the percentage of aftershocks in an earthquake catalog.

function of $n$:

$$m_0 = m_{max} - \frac{n}{1-n}\frac{\theta c^{\theta} 10^{-a}}{b \ln(10)}(1 - 10^{-b(m_{max}-m_d)}) \tag{2.17}$$

assuming $\alpha = b = 1$.

We adopt here the values termed the "generic California model" according to *Reasenberg and Jones* [1989]: $a = -1.67$, $\theta = 0.08$, $c = 0.05$, and we assume $b = 1$, $m_{max} = 8.5$ and $m_d = 3$. Expression (2.17) is drawn in Figure 2.1 (dotted). For comparison, we include the curve for the case $\alpha = b$ that we obtained above in Figure 2.1, based on the fits by *Helmstetter et al.* [2005a]) (dashed curve) and the curve for the case $\alpha = b$ that results from using Båth's law (see next section) (dash-dotted). We observe the same characteristics as before in that $m_0$ approaches $m_{max}$ for small $n$ and that it diverges to minus infinity for $n$ going to one. Differences between the four curves arise only in the faster or slower decrease of $m_0$ with $n$. For example, the (conservative) upper limit $m_d = 3$ for $m_0$ constrains $n$ to be larger than 60 percent according to the values obtained by *Felzer et al.* [2003], whereas the parameters of *Helmstetter et al.* [2005a]) and *Reasenberg and Jones* [1989] for the case $\alpha = b$ impose $n$ to be at least 70 and 80 percent, respectively. For the estimate obtained from Båth's law (see next section), $n$ must be larger than about 45 percent. If we assume that the upper limit of $m_0$ can be obtained from estimates of the critical slip distance $d_c$, corresponding to $m = 0.5$, then $n$ must be at least 55 percent according to the estimate from Båth's law, 70 percent according to *Felzer et al.* [2003], 80 percent according to the fit by *Helmstetter et al.* [2005a]) and 85 percent according to the fit by *Reasenberg and Jones* [1989].

Conversely, $n = 0.7$ determines $m_0$ roughly equal to zero for the values of *Felzer et al.* [2003] and an unrealistic $6$ for the values of *Reasenberg and Jones* [1989]. $n = 0.8$ implies $m_0$ lies around $-7$ and $3$ according to *Felzer et al.* [2003] and *Reasenberg and Jones* [1989], respectively.

Since the four expressions for $m_0$ (for $\alpha = b$) show the same functional dependence on key variables and differ only in the different estimates of a few constants, this consistency provides some confidence in our results. As for the difference in the four curves, they constitute four differently formulated, empirical estimates of the rate of events of typical aftershock sequences. Given the variability of the aftershock process, the discrepancy in the estimates is to be expected.

We now point out difficulties for exploiting quantitatively the above ideas. Our conclusions for $m_0$ and $n$ are based on empirical parameter estimations that involve delicate technicalities. The constants $K_{\mathrm{fit}}$ defined in (2.8), $A_T$ defined in (2.13) and $a$ defined in (2.15) are in principle measurable. Issues that may bias the estimation of these parameters include: (i) The rate of aftershock production is estimated empirically over an apparently complete subperiod of finite space-time windows. Missed events outside the spatial delimitation may act to decrease the rate estimate. (ii) Stacking different sequences with different global Omori law decays may introduce errors. (iii) The $p$ exponent of the Omori law may intrinsically depend on the mainshock magnitude [*Sornette and Ouillon*, 2005; *Ouillon and Sornette*, 2005]. (iv) Background events may be falsely counted as aftershocks. (v) Magnitude and location uncertainties may bias the parameters. (vi) Missing events in the catalog, especially after large events, may artificially change the parameter values. (vii) Undetected seismicity may bias the estimated parameters [*Sornette and Werner*, 2005b].

Recently, *Sornette and Ouillon* [2005]; *Ouillon and Sornette* [2005] argued for a dependence of the exponent $p$ in the Omori law on the magnitude of the mainshock. According to their calculations, $p$ becomes zero, i.e. no earthquakes are triggered, at a magnitude around $-3$. From Figure 2.1, their estimate of $m_0 = -3$ constrains $n$ to be larger than roughly $80$ percent for $\alpha = 1$ and $n$ extremely close to one for $\alpha = 0.5$.

## 2.5 Constraints on the Smallest Triggering Earthquake from Båth's Law

Finally, we use the empirical Båth's law to constrain $m_0$ as a function of $n$. The law states that the average difference between a mainshock of magnitude $M_1$ and the magnitude $m_a$ of its largest aftershock is $dm = M_1 - m_a = 1.2$, regardless of the mainshock magnitude [*Båth*, 1965]. *Console et al.* [2003a], *Helmstetter and Sornette* [2003a] and *Saichev and Sornette* [2005a] showed that the law, deriving from the selection procedure used to define mainshocks and aftershocks, is consistent with the ETAS model.

Let $N_{obs}$ be the total number of aftershocks generated by the mainshock above the magnitude $m_d$ of completeness of the catalog. Assuming that the magnitudes of the aftershocks are drawn from the Gutenberg-Richter law, the largest aftershock has an average magnitude given by a simple argument of extreme value theory:

$$m_a = m_d + (1/b)(\log_{10} N_{obs}) \; . \tag{2.18}$$

Solving this expression for $N_{obs}$, equating it with the ETAS prediction (2.6) and eliminating $k$ through the expression for $n$ (2.4) provides an estimate of $m_0$ as a function of $n$:

$$m_0 = \frac{1}{(\alpha - b)\ln(10)} \times$$
$$\ln 10^{(\alpha-b)m_{max}} + \frac{(b-\alpha)n}{b(1-n)} 10^{\alpha M_1 - b m_a}(1 - 10^{-b(m_{max}-m_d)}) \tag{2.19}$$

for $\alpha \neq b$ and

$$m_0 = m_{max} - (\frac{n}{1-n})\frac{(1 - 10^{-b(m_{max}-m_d)})}{b\ln(10)} \; 10^{b(M_1-m_a)} \tag{2.20}$$

for $\alpha = b$. Note that from expression (2.19), if $\alpha$ is different from $b$, and $M_1 - m_a$ is constant with $M_1$ (Båth's law), then $m_0$ depends on $M_1$, showing the inconsistency

of the argument based solely on the average number of events, as also explained by *Saichev and Sornette* [2005a]. Only when $\alpha = b$ does the mainshock magnitude disappear in the expression of $m_0$ as shown in (2.20). The estimates of $m_0$ for $\alpha < b$ are thus dependent on $M_1$ and should thus be taken only as indications.

Figure 2.4 illustrates the behavior of $m_0$ as a function of the average number $n$ of direct aftershocks for reasonable values of the other constants ($m_{max} = 8.5$, $m_d = 3$, $b = 1$, $\alpha = [0.5, 0.6, 0.7, 0.8, 0.9, 1]$ (light to dark)), for mainshock and largest aftershock values according to $M_1 - m_a = 7 - 5.8 = 1.2$ from Båth's law. Again, as $n$ tends to one, $m_0$ tends to minus infinity, while for $n = 0$, $m_0 = m_{max}$, as expected. We also observe that $m_0$ is almost constant over a wide range of $n$ for comparatively small $\alpha$, whereas $m_0$ varies much faster for the case $\alpha = b$.

As alluded to in the last section, we obtain the same functional dependence as in both previous estimates of the last section. However, for $\alpha = b$, the decrease of $m_0$ with increasing $n$ is even faster than when using the parameters of *Felzer et al.* [2003]. Here, the upper limit $m_d = 3$ for $m_0$ (upper horizontal line) constrains $n$ to be larger than 45 percent, much smaller than the lower limit obtained previously. This discrepancy is due to the four different ways of estimating the observed number of aftershocks. However, since all four are in a similar range, they provide a test of the consistency of the results. When applying the $d_c$-derived upper limit of $0.5$ (lower horizontal line), $n$ must be larger than at least 55 percent for $\alpha = b$ and larger than 85 percent for $\alpha = 0.8$. If $n = 0.5$, $m_0$ is in the range $0$ to an unrealistic $5$, while for $n = 0.7$, $m_0$ lies between $-9$ and $5$, depending on the values of $\alpha$. Since $m_0 \geq 3$ is unrealistic, the entire region of combinations between $\alpha$ and $n$ that fall above that value can be ruled out. For example, the case $\alpha = 0.8$ leads to a reasonable $m_0$ smaller than $m_d = 3$ only for $n$ larger than about 65 percent.

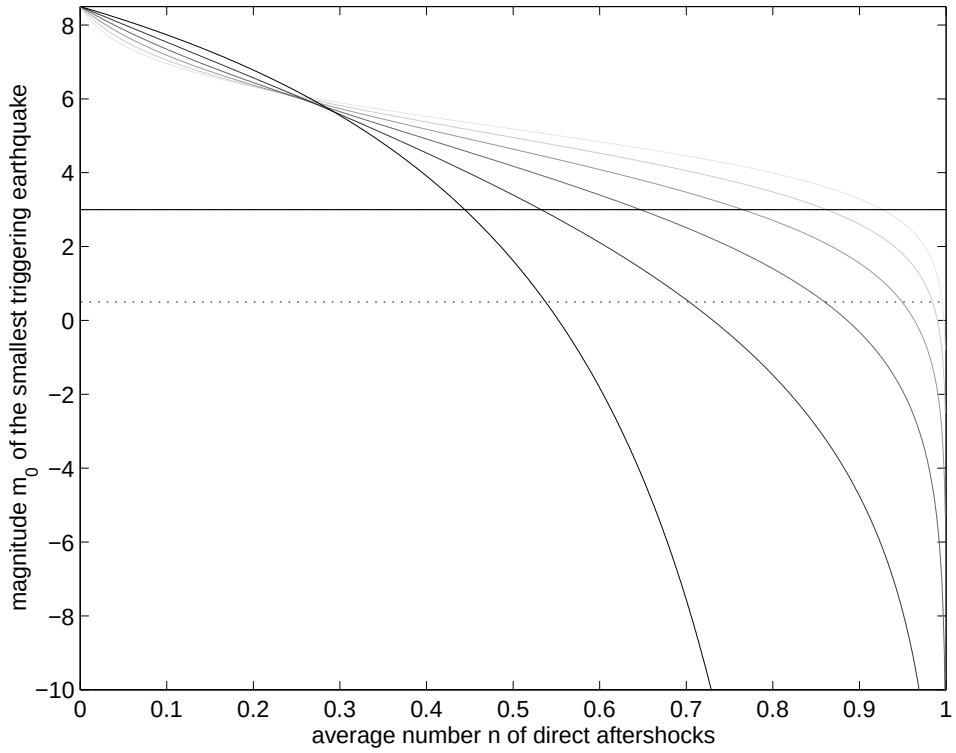


Figure 2.4: The magnitude $m_0$ of the smallest triggering earthquake as a function of the average number $n$ of direct aftershocks as estimated by the ETAS model and Båth's law for $\alpha = [0.5, 0.6, 0.7, 0.8, 0.9, 1]$, from light to dark. The horizontal lines represent two upper limits on $m_0$: the first (upper) one is derived from the assumption that $m_0$ is larger than a typical detection threshold of $m_d = 3$ (solid line), while the second (lower) one was obtained from estimates of the critical slip $d_c = 100\mu$m in rate and state friction giving $m = 0.5$ (dotted line). These two upper limits conversely provide lower bounds on the fraction $n$ of aftershocks in an earthquake catalog and/or $\alpha$. Common parameters are $b = 1$, $m_{max} = 8.5$, $m_{main} = 7$, $m_a = 5.8$.

## 2.6 Conclusions

We have shown that differentiating between the smallest triggering earthquake $m_0$ and the detection threshold $m_d$ within the ETAS model leads, together with four separate methods of estimating the observed numbers of "aftershocks" (defined as triggered events independently of their magnitude), to four estimates of $m_0$ as a function of the percentage $n$ of triggered events in a catalog (also equal to the branching ratio). We have used empirically fitted values for aftershock numbers and thereby eliminated one variable from the ETAS formalism in order to obtain an estimate of $m_0$ as a function of $n$. The three different estimates were obtained from the fits performed by *Helmstetter et al.* [2005a]), by *Felzer et al.* [2003], by *Reasenberg and Jones* [1989] and from the empirical Båth's law. All four give the same functional dependence and similar values for $m_0$. In particular, we can place bounds on $m_0$ from estimates of the percentage of aftershocks in earthquake catalogs. Conversely, we can limit the range of $n$ by observing that $m_0$ must be less than the detection threshold $m_d$, or, less conservatively, that $m_0$ must be less than the magnitude corresponding to the rate-and-state critical slip $d_c = 100\mu$m estimated in laboratories. As well as quantitative values for $m_0$, the bounds limit the possible combinations between $n$ and $\alpha$ and, in particular, indicate that at the very least 55 percent of all earthquakes are triggered events ("aftershocks").

The fact that the existence of a small magnitude cut-off $m_0$ for triggering should have observable consequences may appear surprising. But such a phenomenon of the impact of a small scale cut-off on "macroscopic" observables is not new in physics and actually permeates particle physics, field theory and condensed matter physics. In the present case, the existence of $m_0$ has an observable impact especially when $\alpha \leq b$ for which the cumulative effect of tiny earthquakes equate or dominate that of large earthquakes with respect to the physics of triggering other earthquakes [*Helmstetter*, 2003; *Helmstetter et al.*, 2005a]). We hope that the present article, together with our

companion paper [*Sornette and Werner*, 2005b], will draw the attention of the community to the important problem of the distinction between $m_d$ and $m_0$. Moreover, it will perhaps encourages re-analyses of inversion methods of models of triggered seismicity, and in particular of maximum likelihood estimations, to take into account the bias due to the unobserved seismicity below the magnitude of catalog completeness.

# CHAPTER 3

# Apparent Clustering and Apparent Background Earthquakes Biased by Undetected Seismicity*

## 3.1 Abstract

In models of triggered seismicity and in their estimation from empirical data, the detection threshold $m_d$ is commonly equated to the magnitude $m_0$ of the smallest triggering earthquake. This unjustified assumption neglects the possibility of shocks below the detection threshold triggering observable events. We introduce a formalism that distinguishes between the detection threshold $m_d$ and the minimum triggering earthquake $m_0 \leq m_d$. By considering the branching structure of one complete cascade of triggered events, we derive the apparent branching ratio $n_a$ (which is the apparent fraction of aftershocks in a given catalog) and the apparent background source $S_a$ observed when only the structure above the detection threshold $m_d$ is known due to the presence of smaller undetected events capable of triggering larger events. If, as several recent analyses have shown, earthquake triggering is controlled in large part by the smallest magnitudes, this implies that previous estimates of the clustering parameters may significantly underestimate the true values: for instance, an observed fraction of $55\%$ of aftershocks is renormalized into a true value of $75\%$ of triggered events.

*An edited version of this chapter was published by AGU. Copyright (2005) American Geophysical Union. Sornette, D. and M. J. Werner (2005), Apparent Clustering and Apparent Background Earthquakes Biased by Undetected Seismicity, *J. of Geophys. Res., 110*, B09303, doi:10.1029/2005JB00362. Reproduced by permission of American Geophysical Union.

## 3.2 Introduction

There is much evidence that a seismic event can have a significant effect on the pattern of subsequent seismicity, most obvious in aftershocks of large events. More recently an important extension of the concept of earthquake interactions has emerged: "triggered seismicity", in which the usual distinction that foreshocks are precursors of larger mainshocks, which in turn trigger smaller aftershocks, becomes blurred: an efficient description of seismicity does not seem to require the division between foreshocks, mainshocks and aftershocks, as they appear indistinguishable in many of their physical and statistical properties [*Helmstetter and Sornette*, 2003c]. An important logical consequence is that cascades of triggered seismicity ("aftershocks," "aftershocks" of "aftershocks," ...) may play an important role in the overall seismicity budget [*Helmstetter and Sornette*, 2003b; *Felzer et al.*, 2002].

There is a growing interest in phenomenological models of triggered seismicity, which use the Omori law as a coarse-grained proxy for modeling the complex and multi-faceted interactions between earthquakes, together with other robust descriptions of seismicity (clustering in space, the Gutenberg-Richter (GR) earthquake size distribution and aftershock productivity laws). The class of ETAS (Epidemic-Type Aftershock Sequences) models introduced by *Kagan and Knopoff* [1981] and *Ogata* [1988] offers a parsimonious approach that replaces the classification of foreshocks, mainshocks and aftershocks by the concept of earthquake triggering: earthquakes may trigger other earthquakes through a variety of physical mechanisms without attempting to identify the particular mechanisms.

The questions suggested by this approach include: 1) what is the fraction of triggered versus uncorrelated earthquakes? (This is linked to the problem of clustering and a partial answer is given by *Helmstetter and Sornette* [2003b]). How can one use this modeling approach to forecast future seismicity? What are the limits of predictability?

(A partial answer using only time-dependent information is given by *Helmstetter and Sornette* [2003d]). How sensitive are forecasts to catalog completeness and type of tectonic deformation? In general, to tackle any such question, one needs to estimate key parameters of the models of triggered seismicity in one way or another. Our present paper shows that there is a non-trivial and important impact of catalog incompleteness on the determination of the parameters quantifying earthquake triggering, with its expected impact on all the above questions.

The most promising approach is in general to use the maximum likelihood method to estimate the model parameters from a catalog of seismicity (time, location and magnitude) (see for instance *Ogata* [1988] and *Kagan* [1991]). The calculation of the likelihood function requires evaluating the theoretical rate of seismicity at time $t$ induced by all past events at times $t_i < t$. The maximization of the likelihood with respect to the parameters of the model, given the data, then provides an estimate of the parameters. All previous studies have considered that small earthquakes, below the detection threshold, are negligible. Thus, the rate of seismicity is calculated as if triggered only by earthquakes above the detection threshold. However, this method is not correct because it does not take into account events below the detection threshold, which may have an important role in the triggering of seismicity. Indeed, small earthquakes have a significant contribution in earthquake triggering because they are much more numerous than larger earthquakes [*Felzer et al.*, 2002; *Helmstetter*, 2003; *Helmstetter et al.*, 2005a]. This can simply be seen from the competition between the productivity law $\sim 10^{\alpha M}$ giving the number of events triggered by a mainshock of magnitude $M$ and the relative abundance $\sim 10^{-bM}$ of such mainshocks given by the Gutenberg-Richter (GR) law: the contribution of earthquakes of magnitude $M$ to the overall seismic rate is thus $\sim 10^{-(b-\alpha)M}$, which is dominated by small $M$'s for $\alpha < b$ *Helmstetter* [2003] or equally contributed by each magnitude class for $\alpha = b$ *Felzer et al.* [2002]; *Helmstetter et al.* [2005a]. Therefore, one needs to take into account small events that are

not observed in order to calibrate correctly models of seismicity and obtain reliable answers to our questions stated above. This is an essential bottleneck for the development of earthquake forecasts based on such models.

The purpose of this note is to present a general theoretical treatment of the impact of unobserved seismicity within the framework of models of triggered seismicity. We show by analyzing the branching structure of a complete cascade (cluster) triggered by an independent background event that the unobserved seismicity has the effect of decreasing the real branching ratio $n$ and of increasing the number of independent background events $S$ into apparent quantities $n_a$ and $S_a$. The bias persists in a catalog of an arbitrary number of clusters (see Appendix A.2) and may be very significant. We therefore claim that previous work should be reanalyzed from the new perspective of our approach. This leads also to important consequences for the methods presently used to forecast future seismicity based only on incomplete catalogs.

The closely related study by *Sornette and Werner* [2005a] also considered the effects of undetected seismicity in models of triggered seismicity. They found that a magnitude cut-off $m_0$ below which earthquakes do not trigger other events is necessary to make these models convergent and well-defined. If each magnitude unit of quakes collectively contributes a comparable amount of triggered events (of any magnitude) to the overall budget, then a lower cut-off $m_0$ must exist to ensure finite seismicity. *Sornette and Werner* [2005a] showed that this cut-off has observable consequences and can thus be estimated from parameters estimated from fits to the statistics of aftershock sequences and from Båth's law. They arrived at four different estimates of $m_0$ that we employ below to quantify the effects of undetected seismicity on the measured fraction of triggered events in a seismic catalog. *Sornette and Werner* [2005a] also discussed possible scenarios for this break in self-similarity. In this article, we continue to explore the effects of undetected earthquakes on the observed seismicity.

## 3.3 The ETAS Model and the Smallest Triggering Earthquake

### 3.3.1 Definition of the ETAS Model

To make this discussion precise, let us consider the epidemic-type aftershock sequence (ETAS) model, in which any earthquake may trigger other earthquakes, which in turn may trigger more, and so on. Introduced in slightly different forms by *Kagan and Knopoff* [1981] and *Ogata* [1988], the model describes statistically the spatio-temporal clustering of seismicity. We choose the ETAS model because of its increasing popularity for the statistical description of earthquake interaction [*Kagan and Knopoff*, 1981; *Ogata*, 1988; *Console et al.*, 2003b; *Zhuang et al.*, 2004], its establishment as a powerful null hypothesis for forecasting [*Helmstetter and Sornette*, 2003d; *Schorlemmer et al.*, 2007; *Helmstetter et al.*, 2006], its simplicity, and its explanatory power of features in catalogs including apparent Gutenberg-Richter $b$-value variations and Omori law exponent variations [*Helmstetter and Sornette*, 2002], foreshocks [*Helmstetter and Sornette*, 2003c], and apparent aftershock diffusion [*Helmstetter et al.*, 2005b].

The triggering process may be caused by various mechanisms that either compete or combine, such as pore-pressure changes due to pore-fluid flows coupled with stress variations, slow redistribution of stress by aseismic creep, rate-and-state dependent friction within faults, coupling between the viscoelastic lower crust and the brittle upper crust, stress-assisted micro-crack corrosion, and more. The ETAS formulation amounts to a two-scale description: these above physical processes controlling earthquake interactions enter in the determination of effective triggering laws in a first step and the overall seismicity is then seen to result from the cascade of triggering of events triggering other events triggering other events and so on [*Helmstetter and Sornette*, 2002].

The ETAS model consists of three laws about the nature of seismicity viewed as

a marked point-process. We restrict this study to the temporal domain only, summing over the whole spatial domain of interest. First, the magnitude of any earthquake, regardless of time, location, or magnitude of the mother shock, is drawn randomly from the exponential Gutenberg-Richter (GR) law. Its normalized probability density function (pdf) is expressed as

$$P(m) = \frac{b \ln(10) 10^{-bm}}{10^{-bm_0} - 10^{-bm_{max}}}, \quad m_0 \leq m \leq m_{max}, \tag{3.1}$$

where the exponent $b$ is typically close to one, and the cut-offs $m_0$ and $m_{max}$ serve to normalize the pdf. The upper cut-off $m_{max}$ is introduced to avoid unphysical, infinitely large earthquakes. Its value was estimated to be in the range $8 - 9.5$ [*Kagan*, 1999]. As the impact of a finite $m_{max}$ is quite weak in the calculations below, replacing the abrupt cut-off $m_{max}$ by a smooth taper would introduce negligible corrections to our results.

Second, the model assumes that direct aftershocks are distributed in time according to the modified "direct" Omori law (see *Utsu et al.* [1995] and references therein). Assuming $\theta > 0$, the normalized pdf of the Omori law can be written as

$$\Psi(t) = \frac{\theta c^{\theta}}{(t+c)^{1+\theta}}. \tag{3.2}$$

Third, the number of direct aftershocks of an event of magnitude $m$ is assumed to follow the productivity law:

$$\rho(m) = k \, 10^{\alpha(m-m_0)}, \quad m_0 \leq m \leq m_{max}. \tag{3.3}$$

Note that the productivity law (3.3) is zero below the cut-off $m_0$, i.e. earthquakes smaller than $m_0$ do not trigger other earthquakes, as is typically assumed in studies using the ETAS model. The existence of the small magnitude cut-off $m_0$ is necessary to ensure the convergence of these types of models of triggered seismicity (in the statistical physics of phase transitions and in particle physics, this is called an "ultra-violet"

cut-off which is often necessary to make the theory convergent). In a closely related paper, *Sornette and Werner* [2005a] showed that the existence of the cut-off $m_0$ has observable consequences which constrain its physical value. They also discuss possible scenarios for this break in self-similarity, such as a transition from fracture to friction dominated earthquakes [*Richardson and Jordan*, 2002] or a minimum earthquake size as predicted by rate-and-state friction [*Dieterich*, 1992; *Ben-Zion*, 2003].

The key parameter of the ETAS model is defined as the number $n$ of direct aftershocks per earthquake, averaged over all magnitudes. Here, we must distinguish between the two cases $\alpha = b$ and $\alpha \neq b$:

$$\begin{aligned} n &\equiv \int_{m_0}^{m_{max}} P(m)\rho(m)dm \\ &= \frac{kb}{b-\alpha}\left(\frac{1-10^{-(b-\alpha)(m_{max}-m_0)}}{1-10^{-b(m_{max}-m_0)}}\right) \quad , \end{aligned} \tag{3.4}$$

for the general case $\alpha \neq b$. The special case $\alpha = b$ gives

$$n = \frac{kb\ln(10)(m_{max}-m_0)}{1-10^{-b(m_{max}-m_0)}} \tag{3.5}$$

Three regimes can be distinguished based on the value of $n$. The case $n < 1$ corresponds to the subcritical, stationary regime, where aftershock sequences die out with probability one. The case $n > 1$ describes unbounded, exponentially growing seismicity [*Helmstetter and Sornette*, 2002]. In addition, the case $b < \alpha$ leads to explosive seismicity with finite time singularities [*Sornette and Helmstetter*, 2002]. The critical case $n = 1$ separates the two regimes $n < 1$ and $n > 1$. *Helmstetter and Sornette* [2003b] showed that the branching ratio $n$ is also equal to the fraction of triggered events in a seismic catalog. We consider the case $n < 1$ which describes stationary seismicity. The branching ratio $n$ measures the distance to the critical state of the crust ($n = 1$) which may have important implications for the self-organization of the crust.

The fact that we use the same value for the productivity cut-off and the Gutenberg-Richter (GR) cut-off is not a restriction as long as the real cut-off for the Gutenberg-Richter law is smaller than or equal to the cut-off for the productivity law. In that case, truncating the GR law at the productivity cut-off just means that all smaller earthquakes, which do not trigger any events, do not participate in the cascade of triggered events. This should not be confused with the standard incorrect procedure in many previous studies of triggered seismicity of simply replacing the GR and productivity cut-off $m_0$ with the detection threshold $m_d$ in equations (3.1) and (3.3) (see, for example, [*Ogata*, 1988; *Kagan*, 1991; *Ogata*, 1998; *Console et al.*, 2003b; *Ogata et al.*, 2003; *Zhuang et al.*, 2004]). The assumption that $m_d = m_0$ may lead to a bias in the estimated parameters. Figure 1 of *Helmstetter et al.* [2005a] shows that events of magnitude 2 trigger their own aftershock sequences. We thus expect $m_0$ to be smaller than $m_d$.

Without loss of generality, we consider one independent branch (cluster or cascade of aftershocks set off by a background event) of the ETAS model. We generalize to a seismic catalog of an arbitrary number of clusters in Appendix A.2. Let an independent background event of magnitude $M_1$ occur at some origin of time. The mainshock will trigger direct aftershocks according to the productivity law (3.3). Each of the direct aftershocks will trigger their own aftershocks, which in turn produce their own, and so on. Averaged over all magnitudes, an aftershock produces $n$ direct offspring according to (3.4). Thus, integrating over time, we can write the average of the total number $N_{total}$ of direct and indirect aftershocks of the initial mainshock as an infinite sum over terms of (3.3) multiplied by $n$ to the power of the generation [*Helmstetter and Sornette*, 2003b], which can be expressed for $n < 1$ as:

$$\begin{aligned} N_{total} &= \rho(M_1) + \rho(M_1)\, n + \rho(M_1)\, n^2 + ... \\ &= \frac{k\ 10^{\alpha(M_1 - m_0)}}{1-n} \end{aligned} \tag{3.6}$$

However, since we can only detect events above the detection threshold $m_d > m_0$, the total number of observed aftershocks $N_{obs}$ of the sequence is simply $N_{total}$ multiplied by the fraction of events above the detection threshold, given by

$$f_{obs} = \frac{10^{b(m_{max}-m_d)} - 1}{10^{b(m_{max}-m_0)} - 1} \tag{3.7}$$

according to the GR distribution. The observed number of events in the sequence is therefore

$$\begin{aligned} N_{obs} &= N_{total} \;\; f_{obs} \\ &= \frac{k \;\; 10^{\alpha(M_1-m_0)}}{1-n} \;\; \left( \frac{10^{b(m_{max}-m_d)} - 1}{10^{b(m_{max}-m_0)} - 1} \right) . \end{aligned} \tag{3.8}$$

Equation (3.8) predicts the average observed number of direct and indirect aftershocks of a mainshock of magnitude $M_1 > m_d$. *Sornette and Werner* [2005a] showed that $m_0$ may be estimated using fits of $N_{obs}$ given by (3.8) to observed aftershock sequences and Båth's law. The essential parameter needed to constrain $m_0$ is the branching ratio $n$. As we demonstrate below, typical estimates of $n$ in the literature obtained from a catalog neglect undetected seismicity and therefore cannot be used directly to constrain $m_0$.

Naturally, there is no justification for assuming that $m_d$ should equal $m_0$, as is done routinely in inversions of catalogs for the parameters of the ETAS model (see, for example, [*Ogata*, 1988; *Kagan*, 1991; *Ogata*, 1998; *Console et al.*, 2003b; *Ogata et al.*, 2003; *Zhuang et al.*, 2004]). First, detection thresholds change over time as instruments and network coverage become better, while the physical mechanisms in the Earth presumably remain the same. No significant deviation from the Gutenberg-Richter distribution or the productivity law has been recorded as the detection threshold $m_d$ decreased over time (see for example Figure 3 of *Ouillon and Sornette* [2005]). Second, studies of earthquake occurrence at small magnitude levels below the regional network cut-offs show that earthquakes follow the same Gutenberg-Richter law (for

a recent study of mining-induced seismicity, see, for example, *Sellers et al.* [2003]), while acoustic emission experiments have shown the relevance of the Omori law at small scales (see, for instance *Nechad et al.* [2004] and references therein). Within the assumption of self-similarity, i.e. a continuation of the GR and productivity laws down to a cut-off, evidence thus points towards a magnitude of the smallest triggering earthquake and a Gutenberg-Richter cut-off that lie below the detection threshold and are thus not directly observable.

The effect of undetected seismicity below the detection threshold is fundamentally different from the effect of earthquakes outside the space-time study window that may contribute to the seismicity budget inside the region. The event incompleteness below the magnitude detection threshold $m_d$ cannot be treated in analogy to the time and space detection threshold as a finite-size boundary effect problem. While events from outside the study area have a decreasing influence on the inside in time according to the Omori law and in space according to a spatial decay function (e.g. Gaussian or power-law), the influence of the many events below the detection threshold inside the study area may be very significant because each magnitude range collectively triggers a roughly equal amount of events of any size. The magnitude detection threshold is thus of a different nature than boundary effects and must be addressed.

### 3.3.2 Two Interpretations of the ETAS Model

The ETAS model may be viewed in two mathematically equivalent ways that differ in their interpretation. In this section, we develop both views to underline that our results apply in both cases and to stress the equivalence of these two views. The first describes the model as a simple branching model without loops [*Kagan*, 1991]: The independent background events, due to tectonic loading, may each independently trigger direct aftershocks, each of which may in turn trigger secondary shocks, which

in turn may trigger more. Because every triggered event, excluding of course the non-triggered background events, has exactly one mainshock (mother), but the mother may have many direct aftershocks (children), the model can be thought of as a simple branching model without loops. The background events are assumed to be a stationary Poisson process with a constant rate. The rate of the aftershocks of a background event is a non-stationary Poisson process that is updated every time another aftershock occurs until the cascade dies out. The intensity is thus conditioned on the specific history of earthquakes. The expectation of the conditional intensity is an average over an ensemble of histories. The predicted number of aftershocks of an independent background event of magnitude $M_1$ as in expression (3.8) is thus averaged over the ensemble of possible realizations of the aftershock sequence, and it is also averaged over all possible magnitudes of the aftershocks. The branching ratio $n$ is therefore an average not only over magnitudes but also over an ensemble of realizations of the non-stationary Poisson process. In summary, the model consists of statistically independent Poisson clusters of events, which are, however, dependent within one cluster.

The second view of the ETAS model does not allow a unique identification of the mother or trigger of an earthquake. Rather, each aftershock was triggered collectively by all previous earthquakes, each of which contributes a weight determined by the magnitude-dependent productivity law $\rho(m)$ that decays in time according to the Omori law $\psi(t)$ and in space according to a spatial function $R(r)$, often chosen to be an exponential or a power law centered on the event. The instantaneous conditional intensity rate at some time $t$ at location $r$ is given by

$$\lambda(t, r) = \mu + \sum_{i|t_i<t} \rho(m_i)\ \psi(t - t_i)\ R(r - r_i) \tag{3.9}$$

where the sum runs over all previous events $i$ with magnitude $m_i$ at time $t_i$ at location $r_i$. Thus the triggering contribution of a previous event to a later event at time t is given by its own weight (its specific entry in the sum) divided by the total seismic-

ity rate, including the background rate. A non-zero background rate then contributes evenly to all events and corresponds to an omnipresent loading contribution. In this way, earthquakes are seen to be the result of all previous activity including the background rate. This corresponds to a branching model in which every earthquake links to all subsequent earthquakes weighted according to the contribution to triggering. A branching ratio can then be interpreted as a contribution of a past earthquake to a future earthquake, averaged over an ensemble of realizations and all magnitudes. In contrast to the independent background events considered due solely to tectonic loading that exist in the first interpretation, all earthquakes are due to a combination of the background loading and the effect of previous events. This second view becomes the only possible one for nonlinear models whose triggering functions depend nonlinearly on previous events (see e.g. the recently introduced multi-fractal earthquake triggering model [*Ouillon and Sornette*, 2005; *Sornette and Ouillon*, 2005] and references therein).

These two views are equivalent because the linear formulation of the seismic rate of the ETAS model together with the exponential Poisson process ensures that the statistical properties of the resulting earthquake catalogs are the same. The linear sum over the individual contributions and the Poisson process formulation are the key ingredients that allow the model to be viewed as a simple branching model.

This duality of thinking about the ETAS model is reflected in the existence of two simulation codes in the community, each inspired by one of the two views. A program written by K. Felzer and Y. Gu (personal communication) calculates the background events as a stationary Poisson process and then simulates each cascade independently of the other branches as a non-stationary process. The second code by *Ogata* [1998], on the other hand, calculates the overall seismicity at each point in time by summing over all previous activity. The latter code is significantly slower because the inde-

pendence between cascades is not used, and the entire catalog is modeled as the sum of a stationary and a non-stationary process. Despite the different approach, both resulting earthquake catalogs share the same statistical properties and are thus equally acceptable.

While the simulation or forward problem is straight-forward when adopting the view of the ETAS model as a branching model with one assigned trigger for any aftershock, the inverse problem of reconstructing the branching structure from a given catalog can at best be probabilistic. Because aftershocks of one mother cannot be distinguished from those of another mother except by spatio-temporal distance, we have no way of choosing which previous earthquake triggered a particular event, or whether it is a background event. Rather, we must resort to calculating the probability of an event at time t to be triggered by any previous event according to the contribution that the previous event has at time t compared to the overall intensity at time t. This probability is of course equal to the weight or triggering contribution that a previous event has on a subsequent event when adopting the collective-triggering view. However, the interpretation remains different since the probability specifies a unique mother in a fraction of many realizations.

Having determined from catalogs a branching structure weighted according to the probability of triggering, one may of course choose to always pick as source of an event the most probable contributor, be that a previous event or the background rate. Another option is to choose randomly according to the probability distribution and thus reconstruct one possible branching structure among the ensemble of many other possible ones. The latter approach has been used by *Zhuang et al.* [2004] and labeled stochastic reconstruction.

The key point is that equating the detection threshold with the smallest triggering earthquake will most likely lead to a bias in the recovered parameters of a maximum

likelihood analysis as performed by *Zhuang et al.* [2004] and in many other studies. Therefore, the weights or probabilities of previous events triggering subsequent events were calculated from biased parameters.

In the following, we show that the branching ratio and the background source events are significantly biased when they are estimated from the apparent branching structure observed above the detection threshold $m_d$ instead of the complete tree structure down to $m_0$. We adopt the view of the simple branching model to make the derivations more illuminating but all results can be reinterpreted as contributions in the collective-triggering view.

## 3.4 The Apparent Branching Structure of the ETAS Model

### 3.4.1 The Apparent Branching Ratio $n_a$

Seismic catalogs are usually considered complete above a threshold $m_d$, which varies as a function of technology and location. For instance, $m_d \approx 2$ for modern Southern California catalogs (and for earthquakes not too close in time to a large mainshock [*Kagan*, 2003]). The analysis of the statistics of the Omori and inverse Omori laws for earthquakes of magnitude down to 3 [*Helmstetter*, 2003; *Helmstetter and Sornette*, 2003c] suggests that $m_0$ is smaller than the completeness magnitude $m_d$ and is thus not directly observable. Thus, $m_0$ is the size of the smallest triggering earthquake, which most likely differs significantly in size from the current detection threshold $m_d$. By considering the branching structure of the model, we derive the apparent branching ratio and the apparent background source that are found if only the observed (detected) part of the ETAS model is analyzed.

Since aftershock clusters are independent of each other, averages of one cluster are equal to ensemble averages, as nothing but the inherent stochasticity of the model

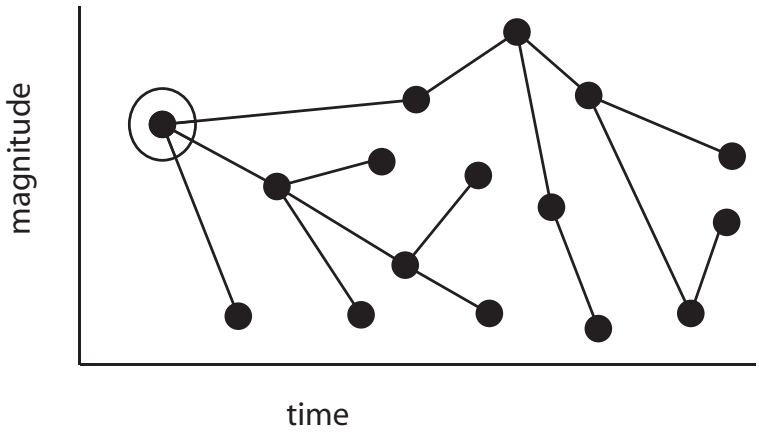


Figure 3.1: Schematic representation of the branching structure of the real ETAS model: An independent background earthquake triggers direct aftershocks, which in turn trigger second generation aftershocks, and so on. The structure is complete down to the magnitude of the smallest triggering earthquake $m_0$.

determines the properties of the clusters. One cluster consists of one independent background event (source) and its direct and indirect aftershocks (see Figure 3.1). However, if not all events of the sequence are detected, then there will appear to be less direct (and indirect) aftershocks, i.e. the branching ratio will appear different. Furthermore, some observed events will be triggered by mother-earthquakes below the detection threshold, resulting in apparently independent background events (see Figure 3.2).

This view leads to the conclusion that the average number of direct aftershocks that are observed will be less than the real branching ratio, since some of the triggered events of an observed shock will fall below $m_d$ and hence not be included in the count. Only the fraction $f_{obs}$ from equation (3.7) above $m_d$ of the total direct aftershocks $\rho(m)$ will be observed. Moreover, the pdf $P(m|m \geq m_d)$ of mother events conditioned on being larger than $m_d$ is zero for $m < m_d$ and equal to $P(m)/f_{obs}$ for $m_{max} > m \geq m_d$.

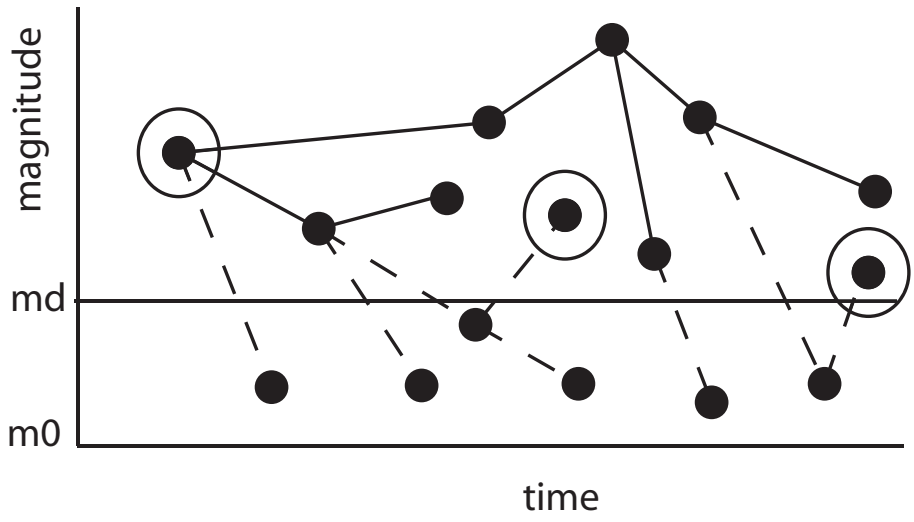


Figure 3.2: Schematic representation of the branching structure of the apparent ETAS model. The initial mainshock is circled. Only events above the detection threshold $m_d$ are observed. The apparent branching ratio does not take into account unobserved triggered events (dashed lines). An observed event triggered by a mother below $m_d$ appears as an untriggered background source event (circled).

We can thus define the apparent branching ratio as

$$\begin{aligned} n_a &\equiv \int_{m_0}^{m_{max}} P(m|m \geq m_d)\ \rho(m)\ f_{obs}\ dm \\ &= \int_{m_d}^{m_{max}} P(m)\ \rho(m)\, dm \\ &= \frac{k\, b}{b-\alpha}\left(\frac{10^{-(b-\alpha)(m_d-m_0)} - 10^{-(b-\alpha)(m_{max}-m_0)}}{1-10^{-b(m_{max}-m_0)}}\right) \end{aligned} \tag{3.10}$$

for the case $\alpha \neq b$. The special case $\alpha = b$ gives

$$n_a = \frac{kb\ln(10)(m_{max}-m_d)}{1-10^{-b(m_{max}-m_0)}} \quad . \tag{3.11}$$

Using equation (3.4) and eliminating $k$, we have $n_a$ in terms of $n$:

$$n_a = n\ \left(\frac{10^{(b-\alpha)(m_{max}-m_d)}-1}{10^{(b-\alpha)(m_{max}-m_0)}-1}\right) \quad , \tag{3.12}$$

when $\alpha \neq b$, and

$$n_a = n \left( \frac{m_{max} - m_d}{m_{max} - m_0} \right) \quad , \tag{3.13}$$

when $\alpha = b$.

According to expression (3.12), $n_a \leq n$, where the equality holds for $m_d$ equal to $m_0$. In principle, equation (3.12) also holds for $n > 1$, but we restrict this study to the regime $n < 1$ for mathematical convenience and because this gives rise to statistically stationary seismic sequences. Figure 3.3 shows $n_a$ as a function of $n$ for a range of values of $m_0$ for the case $\alpha = b$. It demonstrates that the apparent (measurable) fraction of aftershocks may significantly underestimate the true fraction of aftershocks even for $m_0$ not very small. For example, $m_0 = -5$ roughly translates a real branching ratio of $n = 0.9$ into an apparent branching ratio $n_a = 0.3$. Decreasing $\alpha$ below $b$ places more importance on the triggering from small earthquakes and therefore strongly amplifies this effect.

In Figure 3.4, we plot the ratio $n_a/n$ as a function of the unknown $m_0$. As expected, when $m_0 = m_d$, the ratio is one because there is no unobserved seismicity. As $m_0$ goes to minus infinity, $n_a$ approaches zero since almost all seismicity occurs below the threshold. We see clearly that unobserved seismicity results in a drastic underestimate of the fraction of aftershocks.

Given an estimate of the magnitude of the smallest triggering earthquake $m_0$ (see *Sornette and Werner* [2005a] and references therein), one can calculate the true branching ratio from the apparent branching ratio. In fact, *Sornette and Werner* [2005a] obtained four estimates of $m_0$ as a function of $n$ by comparing the ETAS model prediction of the number of observed aftershocks (3.8) from fits to observed aftershock sequences and from the empirical Båth's law. Their equations (10), (13), (16), and (18) are the estimates of $m_0$ as a function of $n$ and a number of known constants specific to the fits to observed aftershock sequences. We can use these relations of $m_0$ as a function

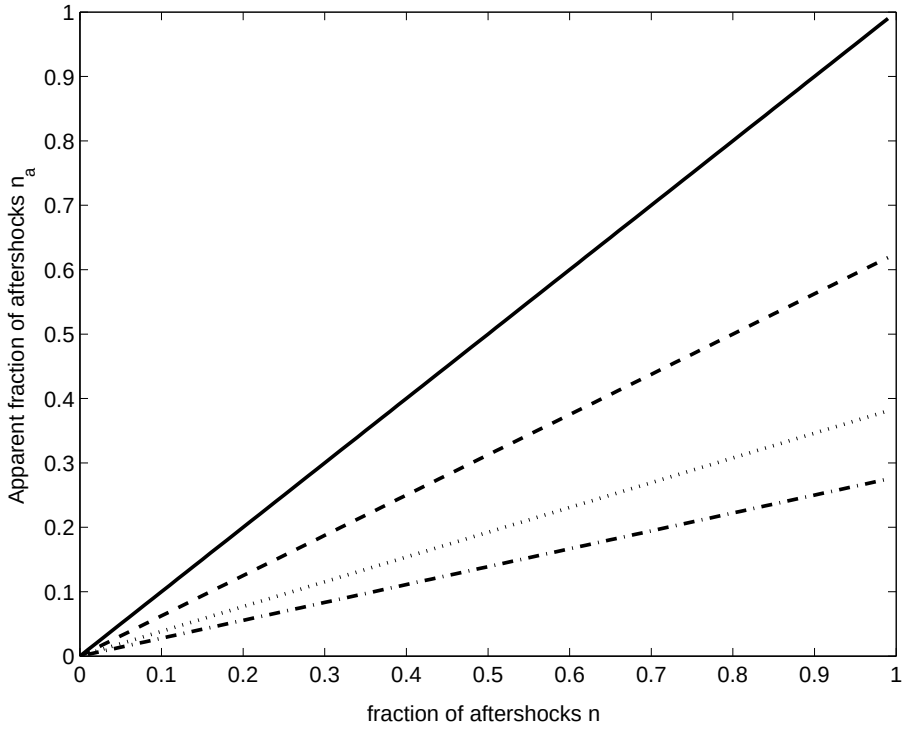


Figure 3.3: The apparent fraction of aftershocks (apparent branching ratio) $n_a$ varies linearly with the real fraction of aftershocks (real branching ratio) $n$ with a slope fixed by the smallest triggering earthquake $m_0$. As $m_0$ decreases, the apparent fraction of aftershocks significantly underestimates the real fraction. As examples, we chose $m_0 = m_d = 3$ (solid), i.e. $n_a = n$ and no events are missed; $m_0 = 0$ (dashed); $m_0 = -5$ (dotted); and $m_0 = -10$ (dash-dotted). We further assumed parameters $m_d = 3$, $m_{max} = 8$, $b = \alpha = 1$. A small value of $\alpha$ amplifies this effect (see Figure 3.4).

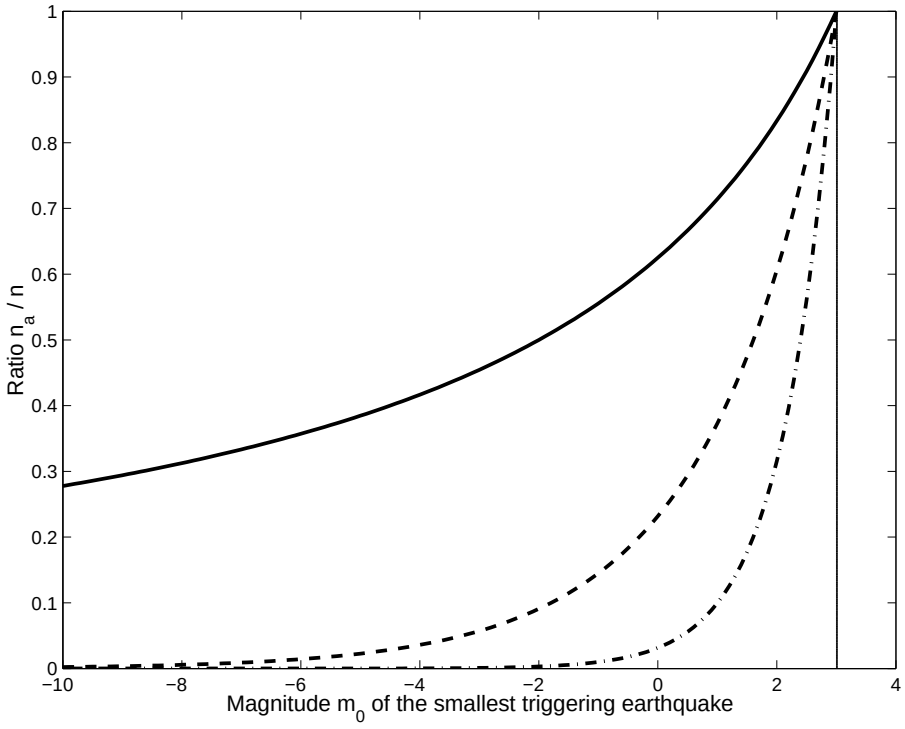


Figure 3.4: The ratio of the apparent fraction of aftershocks (apparent branching ratio) $n_a$ over the real fraction of aftershocks (real branching ratio) $n$ varies as a function of the smallest triggering earthquake $m_0$. For $m_0 = m_d$, $n_a = n$ and all events are detected above the threshold. For a small value of $m_0$, the ratio becomes small, indicating that $n_a$ significantly underestimates $n$. Decreasing $\alpha$ amplifies this effect. We used parameters $m_d = 3$ (vertical reference line), $m_{max} = 8$, $b = 1$. We varied $\alpha = 0.5$ (dash-dotted), $\alpha = 0.8$ (dashed), $\alpha = 1.0$ (solid).

of $n$ to eliminate $m_0$ from equation (3.12) to obtain direct estimates of $n$ as a function of the measurable $n_a$. For simplicity, we restrict the use of their findings to the case $\alpha = b$. The estimate resulting from the fits performed by ***Helmstetter et al.*** [2005a] yielded

$$m_0 = m_{max} - \left(\frac{n}{1-n}\right)\frac{\theta c^{\theta}}{K_{fit}}\frac{1-10^{-b(m_{max}-m_d)}}{b\ln(10)} \tag{3.14}$$

with the values $m_{max} = 8.5$, $m_d = 3$, $\theta = 0.1$, $c = 0.001$, $b = 1$ and $K_{fit} = 0.008$. The study by ***Felzer et al.*** [2002] provided another estimate

$$\begin{aligned} m_0 &= m_{max} - \left(\frac{n}{1-n}\right)\frac{(1-10^{-b(m_{max}-m_d)})}{b\ln(10)} \\ &\quad \times\frac{\theta_T c^{\theta_T}}{A_T}10^{b(M_1-m_d)} \quad , \end{aligned} \tag{3.15}$$

where $m_{max} = 8.5$, $m_d = 3$, $\theta_T = 0.08$, $A_T = 0.116$ days$^{-\theta_T}$, $b = \alpha = 1$, $c = 0.014$ and $M_1 = 6.04$. Using the declustering performed by ***Reasenberg and Jones*** [1989], ***Sornette and Werner*** [2005a] obtained

$$m_0 = m_{max} - \left(\frac{n}{1-n}\right)\frac{\theta c^{\theta}10^{-a}}{b\ln(10)}(1-10^{-b(m_{max}-m_d)}) \tag{3.16}$$

where $m_{max} = 8.5$, $m_d = 3$, $\theta = 0.08$, $a = -1.67$, $c = 0.05$ and $b = 1$. Finally, using Båth's law, ***Sornette and Werner*** [2005a] found

$$m_0 = m_{max} - \left(\frac{n}{1-n}\right)\frac{(1-10^{-b(m_{max}-m_d)})}{b\ln(10)}\ 10^{b(M_1-m_a)} \tag{3.17}$$

where $M_1 - m_a = 1.2$ according to the the law, $b = 1$, $m_{max} = 8.5$, and $m_d = 3$.

Substituting these four estimates of $m_0$ from equations (3.14), (3.15), (3.16), and (3.17) into equation (3.12) for $n_a$ provides four estimates of $n_a$ versus $n$ all in terms of known constants. These four estimates of $n$ as a function of $n_a$ can be used to find the correct fraction of aftershocks from the measurable apparent fraction of aftershocks. Figure 3.5 shows these four estimates with the above constants.

Figure 3.5 can be used to find the real fraction of aftershocks from the measured apparent fraction by assuming one of the four estimates of $m_0$ as a function of $n$.

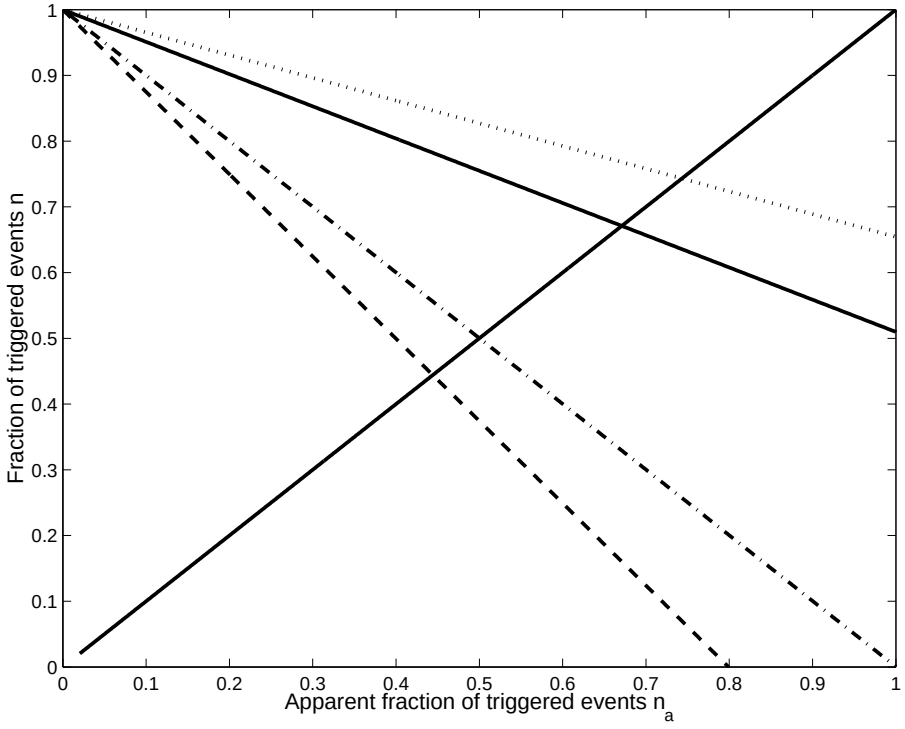


Figure 3.5: The fraction of aftershocks (branching ratio) $n$ can be estimated from the apparent fraction of aftershocks (apparent branching ratio) $n_a$ by using four estimates of the smallest triggering earthquake $m_0$ as a function of $n$ as determined in *Sornette and Werner* [2005a] (see text). The estimates of $m_0$ as a function of $n$ were obtained from comparisons of the ETAS model prediction of the number of observed aftershocks and fits to observed aftershock sequences performed by *Helmstetter et al.* [2005a] (solid), *Felzer et al.* [2002] (dash-dotted),*Reasenberg and Jones* [1989] (dotted) and from Båth's law (dashed). The additional diagonal solid line $n_a = n$ corresponds to $m_0 = m_d$ (no undetected events). Along any of the four lines, $m_0$ varies from minus infinity to $m_{max}$. Given that we can rule out $m_0 \geq m_d$, we can restrict the physical range to the left side of the diagonal $n_a = n$.

For example, *Helmstetter et al.* [2005a] find that $55$ percent of all earthquakes are aftershocks above $m_d = 3$. Using their values to estimate $m_0$ as a function of $n$, we can determine that the real fraction of aftershocks is closer to $75$ percent. Thus the size of this effect is significant. Furthermore, having determined a point on the line estimating $n$ from $n_a$ for all values of $m_0$ fixes the slope of $n(n_a)$ and therefore $m_0$. Using their values, we find $m_0 \simeq 1.2$. Similar estimates can be made using the apparent fraction of aftershock values found by *Felzer et al.* [2002] and *Reasenberg and Jones* [1989]. The uncertainty of the parameters estimated in these studies affects the estimates of $m_0$. *Sornette and Werner* [2005a] analyzed the error propagation and found that the estimates of $m_0$ are most likely order of magnitude calculations.

Assuming that current maximum likelihood estimation methods of the ETAS model parameters, which assume $m_0 = m_d$, determine a branching ratio that corresponds to the present apparent branching ratio, we can similarly correct these values to find the true fraction of aftershocks using Figure 3.5. For example, *Zhuang et al.* [2004] find a "criticality parameter" of about 45 percent, which we take as a proxy for $n_a$. Figure 3.5 shows that the true branching ratio then lies between $0.45$ and $0.80$, depending on which estimate (among the four models (3.14), (3.15), (3.16), and (3.17)) of $m_0$ as a function of $n$ is chosen. These calculations suggest that previous estimates of the fraction of aftershocks obtained by various declustering methods significantly underestimated its value.

### 3.4.2 Determination of Apparent Background Events $S_a$ of Uncorrelated Seismicity

In order to derive the number of shocks within one cascade that are not triggered by a mother above the threshold and thus appear as independent background events, we need to distinguish between the case where the initial (main) shock of magnitude $M_1$

is observable (i.e. $M_1 \geq m_d$) and the case where it is undetected (i.e. $M_1 < m_d$).

If $M_1 \geq m_d$, then the initial background event produces $\rho(M_1) f_{obs}$ observed direct aftershocks. On average, these will in turn collectively produce $\rho(M_1) \ \ f_{obs} \ \ n_a$ observed second generation aftershocks. We specifically do not consider events above $m_d$ triggered from below $m_d$, which we deal with below in the definition of the apparent background sources. By continuing this "above-water" or "above-sea-level" cascade for all generations of aftershocks, we can calculate the number of triggered events that are in direct lineage above the threshold back to the mainshock as the infinite sum of terms of $\rho(M_1) \ \ f_{obs}$ multiplied by the apparent branching ratio $n_a$ to the power of the generation. If, on the other hand, the initial background event is below $m_d$, then no such direct "above-water" cascade will be seen. Any observed shock will be triggered by an event below the water. Thus, for the two cases, the "above-water" sequence is expressed as:

$$N_{above} = \begin{cases} \frac{\rho(M_1) \ f_{obs}}{1-n_a} = N_{obs} \frac{1-n}{1-n_a} & , \ M_1 \geq m_d \\ 0 & , \ M_1 < m_d \end{cases} \tag{3.18}$$

Furthermore, since in the ETAS model, a small earthquake may trigger large earthquakes, an event below $m_d$ may produce an observed event above $m_d$. An inversion method that reconstructs the entire branching structure of the model from an earthquake catalog will identify these shocks as background events. But since in reality these events were triggered by earthquakes below the detection threshold, we will refer to them as apparent background events. These events can of course trigger their own cascades. We thus define the apparent background sources $S_a$ as the number of observed events above $m_d$ that are apparently not triggered, i.e. have "mothers" below $m_d$. Again, we distinguish between the cases where the background event magnitude is $M_1 \geq m_d$ and $M_1 < m_d$. For the first, $S_a$ is given by the total number of aftershocks below the threshold multiplied by the average number $r$ of direct aftershocks

they trigger above the threshold. For the second case, we must also include the direct aftershocks of the initial background event that are observed:

$$S_a = \begin{cases} \frac{\rho(M_1)}{1-n}\,(1-f_{obs})\,r & ,\ M_1 \geq m_d \\ \frac{\rho(M_1)}{1-n}\,(1-f_{obs})\,r + \rho(M_1)\,f_{obs} & ,\ M_1 < m_d \end{cases} \tag{3.19}$$

Now, the number $r$ of observable directly triggered shocks above $m_d$ averaged over unobserved mothers between $m_0$ and $m_d$ is given by the following conditional branching ratio:

$$r \equiv \int_{m_0}^{m_d} P(m|m<m_d)\ \rho(m)\ f_{obs}\ dm \tag{3.20}$$

$$= (n-n_a)\left(\frac{f_{obs}}{1-f_{obs}}\right) \quad , \tag{3.21}$$

where we have used that $P(m|m < m_d) = P(m)/(1-f_{obs})$ for $m < m_d$ and zero otherwise. Substituting (3.21) into the expression for the apparent source (3.19) and re-arranging using (3.8), we obtain

$$\begin{aligned} S_a &= \begin{cases} \frac{\rho(M_1)}{1-n}\ f_{obs}\ (n-n_a) & ,\ M_1 \geq m_d \\ \frac{\rho(M_1)}{1-n}\ f_{obs}\ (n-n_a) + \rho(M_1)\ f_{obs} & ,\ M_1 < m_d \end{cases} \\ &= \begin{cases} N_{obs}(n-n_a) & ,\ M_1 \geq m_d \\ N_{obs}(n-n_a) + \rho(M_1)\ f_{obs} & ,\ M_1 < m_d \end{cases} \end{aligned} \tag{3.22}$$

Equation (3.22) shows that, for each genuine background event of magnitude $M_1$, a perfect inversion method would count $S_a$ apparent background events. Figure 3.6 plots the number of apparent background events $S_a$ as a function of the branching ratio $n$ for an example aftershock cascade set off by a magnitude $m = 5$ initial shock. The figure shows that for one cascade, i.e. one independent background event, hundreds of earthquakes appear as apparent background events when $m_0 < m_d$.

In Figure 3.7, we investigate the relative importance of the apparent background events with respect to the observed number of aftershocks of one cascade. According

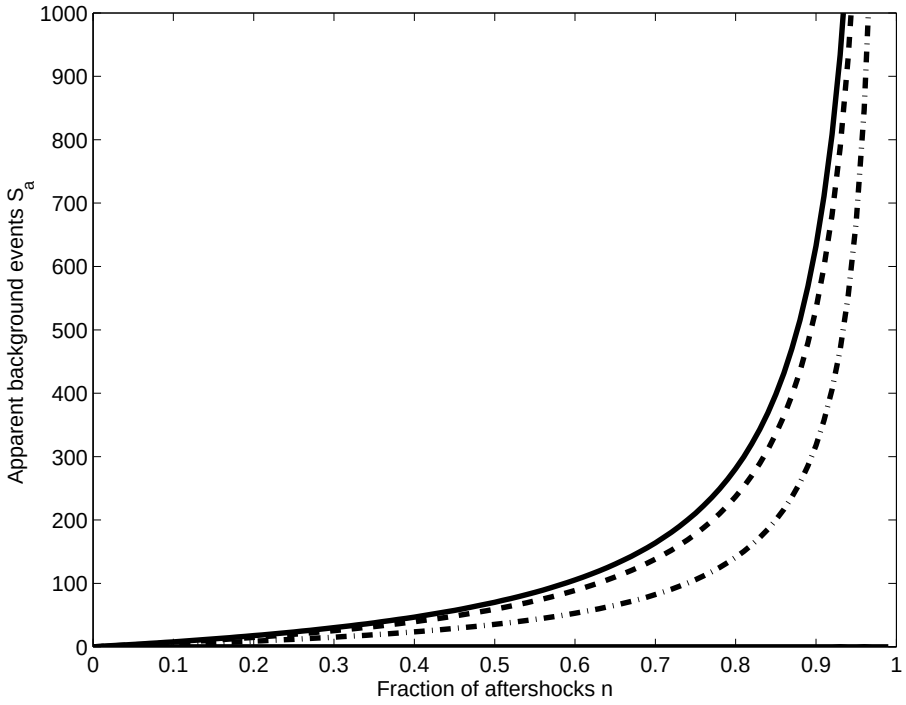


Figure 3.6: The number of apparent background events $S_a$ in an aftershock cascade due to a single background event of magnitude $M_1 = 5$ as a function of the fraction of aftershocks (branching ratio) $n$ for several values of the smallest triggering earthquake. For $m_0 = m_d$, no events are missed. Therefore the number of apparent background events is zero. As $m_0$ decreases, events below the detection threshold trigger events above the threshold and hence the number of apparent background events increases. We vary $m_0 = m_d = 3$ (solid, coinciding with x-axis), $m_0 = 0$ (dash-dotted), $m_0 = -5$ (dashed), and $m_0 = -10$ (upper solid curve). We used parameters $m_d = 3$, $m_{max} = 8$, $b = 1$, and $\alpha = 1.0$. For very small $m_0$ and $n$ close to 1, almost all events above the detection threshold are triggered from below and thus $S_a$ becomes very large (see Figure 3.7). This effect is amplified for decreasing $\alpha$ (not shown).

to equation (3.22)

$$S_a/N_{obs} = n - n_a \ , \tag{3.23}$$

i.e. a significant fraction $n - n_a$ of events of the actually observed triggered events are falsely identified as background events (since all events are really triggered from a single mainshock in our example). For $m_d = m_0$, the ratio is zero, since no events trigger below the detection threshold. However, as $m_d$ increases above $m_0$ and more and more events fall below $m_d$ and become unobserved, the fraction increases until $n_a$ goes to zero and the ratio approaches $n$. This effect increases with decreasing $\alpha$. Small values of $\alpha$ generally place more importance on the cumulative triggering of small earthquakes.

Expressions (3.11) and (3.22) show that analyzing the tree structure of triggered seismicity only above the detection threshold leads to the introduction of an apparent source $S_a$ and an apparent branching ratio $n_a$. It is important to realize that both are renormalized simultaneously by using catalogs with $m_d > m_0$. An unbiased inversion method for the parameters of this averaged, deterministic approximation of the fully stochastic ETAS model would retrieve our analytical results (3.11) and (3.22). We conjecture that our time- and space-integrated, magnitude-averaged and clustered version approximates the full ETAS model (equation (3.9)) well enough so that this bias persists for inversions of parameters of the full model. Accordingly, the value of the background source would be overestimated and the branching ratio underestimated. In fact, one single true sequence will appear as many different sequences, each apparently set off by an apparent background event. Finally, it can be shown (see Appendix A.1) that the sum of the "above-water" cascade and the cascades due to the apparent background events equal the total number of observed earthquakes, demonstrating the consistency of our decomposition.

Furthermore, we can extend the present approach to a whole catalog that consists

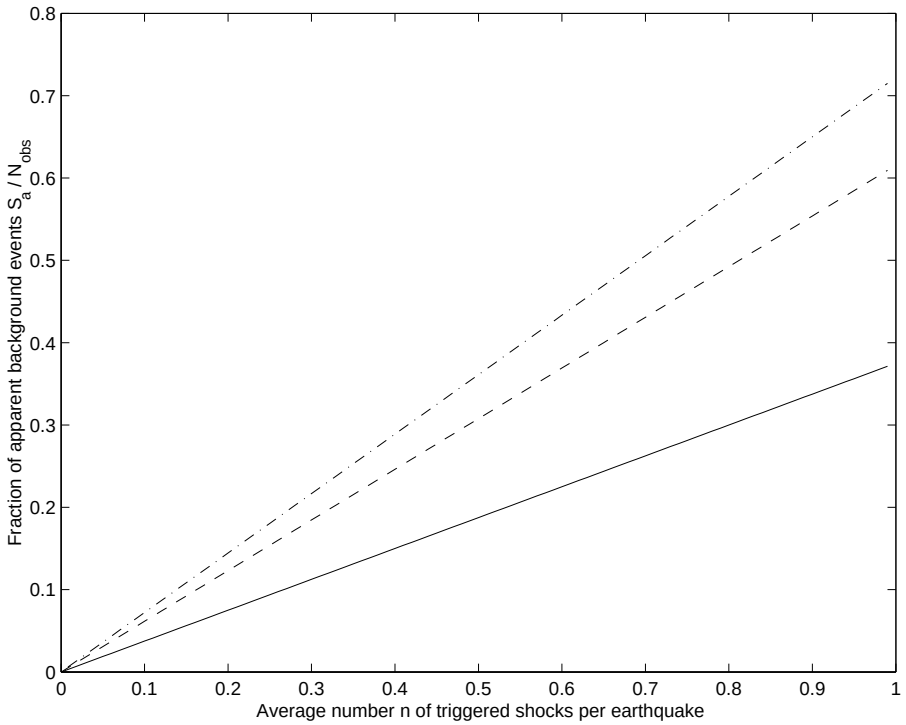


Figure 3.7: The ratio of the number of apparent background events $S_a$ over the total observed number $N_{obs}$ of aftershocks of one cascade varies as $n - n_a$. Here, we show the ratio as a function of the branching ratio $n$ by assuming a particular value of $m_0$. For $m_0 = m_d$ (solid, coinciding with x-axis), there are no apparent background sources. For $m_0$ less than $m_d$, the ratio increases as more and more of the observed events are triggered by unobserved events. As examples, we show the ratio $S_a/N_{obs}$ for $m_0 = 0$ (upper solid line), $m_0 = -5$ (dashed) and $m_0 = -10$ (dash-dotted) as a function of the branching ratio $n$ (average number of aftershocks per earthquake also equal to the fraction of aftershocks in a catalog) for parameters $m_d = 3$, $m_{max} = 8$, $b = 1$, and $\alpha = 1.0$. For very small $m_0$, $n_a$ approaches zero and the ratio $S_a/N_{obs}$ approaches its limiting value $n$, meaning that almost all observed earthquakes were triggered by events below the detection threshold $m_d$. The effect of unobserved events triggering observed quakes resulting in an apparent background source rate is further amplified by smaller values of $\alpha$ (not shown).

of many clusters, each of which is analyzed from the same point of view as above. The calculations are presented in Appendix A.2 and consist simply of summing over all clusters, each of which has been decomposed into a possible real observed source, its resulting above-water cascade, the apparent sources and their cascades. We also show the consistency of this decomposition.

We now come back to examine the assumptions made in this work. First, we assumed that events could be clustered into distinct sets that are set off by a real or an apparent source. Second, we approximated the number of aftershocks of each real or apparent source by averaging over the magnitudes of the triggered events. Third, we integrated over time and space so that we could concentrate on pure numbers of events only. In other words, we have removed all stochasticity of the model. Under these assumptions, we have shown that, in introducing a detection threshold $m_d$, one renormalizes the ETAS model onto itself but with effective parameters $S_a$ and $n_a$. The functional form of the model remains the same. However, we have not proved that the instantaneous and stochastic ETAS model as described by equation (3.9) can be renormalized exactly onto itself with effective parameters for $m_d > m_0$. For this, we would have to check that all fractional moments (that exist) and all distributions describing the stochastic seismic rates are the same (i) in the catalogs generated by ETAS with $m_d > m_0$ and (ii) in the catalogs generated by the effective ETAS with minimum magnitude of triggering taken equal to $m_d$ and with the corresponding adequate values of the effective parameters. Our present paper has just shown the already non-trivial result that the first-moment (average) of seismic rates of the catalogs of (i) and (ii) are identical for the choice of the apparent parameters (3.11,3.11) and (3.19).

## 3.5 Conclusions

We have shown that unbiased estimates of the fraction of aftershocks and the number of independent background events are simultaneously renormalized to apparent values when the smallest triggering earthquake $m_0$ is smaller than the detection threshold $m_d$. In summary, mainshocks above the threshold will appear to have fewer aftershocks, resulting in a smaller apparent branching ratio. Meanwhile, unobserved events can trigger events above the threshold giving rise to apparently independent background events that seem to increase the constant background rate to an apparent rate. Assuming that current techniques which are used to invert for the parameters of the ETAS model (for example, the maximum likelihood method) under the assumption $m_d = m_0$ are unbiased estimators of $n_a$ and $S_a$, then the obtained values for the fraction of aftershocks and the background source rate correspond to renormalized values because of the assumption that the detection threshold $m_d$ equals the smallest triggering earthquake $m_0$. We predict that $n$ will be drastically underestimated and $S$ strongly overestimated for $m_0$ much smaller than $m_d$.

# CHAPTER 4

# Effects of Undetected Seismicity: Further Implications, Extensions and Recent Developments

## 4.1 Introduction

This chapter presents further developments in the study of the effects of undetected seismicity since the previous two chapters were published in 2005. First, I discuss the implications of the two previous chapters (for the "mean" ETAS model) on (i) the geophysical interpretation of clustering parameters, and (ii) declustering methods. Second, I discuss the extension of the previously obtained results from the analysis of the first moment properties of the ETAS model to the full distribution (from the "mean" to the stochastic ETAS model). I will pay particular attention to the conditional intensity function, which completely defines a point process, under a change of detection threshold. Third, I study the implications for parameter estimation using likelihood methods. Finally, I discuss a completely self-similar class of branching models introduced by *Vere-Jones* [2005], which removes the need for the cut-off $m_0$, and its relation to the ETAS model.

## 4.2 Implications of Undetected Seismicity from Results of the Mean ETAS Model

In Chapters 2 and 3, we showed that for the "mean" ETAS model the observed clustering parameters (the fraction of triggered events) change with the detection threshold of a studied catalog. In the following sections, I consider further implications for the geophysical interpretation of clustering parameters and for declustering.

### 4.2.1 Geophysical Interpretation of Clustering Parameters

The existence of background rates in seismicity models is often justified by "tectonic loading". By this, one proposes that plate tectonic motion is a driving force in the seismicity budget. For the ETAS model, the assumed impact on earthquake probabilities is an additive term, constant in time though perhaps not in space, a term independent of the triggering (although background events trigger their own aftershocks).

Say we want to test the hypothesis that the background rate increases with plate velocity. We pick regions of the world which are tectonically similar, e.g. the subduction zones of the Japan Sea and northern South America (e.g. *Bird* [2003]). Assume for this thought experiment that other parameters are the same, including corner moment, b-value and seismic coupling (e.g. *Bird and Kagan* [2004]). Say an estimation procedure provided values of the apparent background events, $S_a^J$ above $m_d^J$ in the Japan Sea and $S_a^S$ background events above $m_d^S$ in norther South America in the same timeframe. Assume $m_d^J \neq m_d^S$ and that both areas are characterized by a similar lower cut-off $m_0$. How should one compare the two estimates for the background sources?

A direct comparison can be ruled out because the numbers depend on the detection threshold. To compare them, it would seem natural to apply the Gutenberg-Richter law to calculate the fraction $f_d$ of $S_a^S$ above $m_d^J$ and to compare $f_d \cdot S_a^S$ with $S_a^J$ on a

common basis $m_d^J$. But this is wrong according to the results of Chapter 3. There, it is shown that the apparent sources actually increase in number with increasing detection threshold, because their mothers go undetected. Instead, one would need to compare the true background sources $s^J$ and $s^S$ above $m_0$. Equivalently, one could compare the apparent background sources above a common $m_d$, say $m_d^J$. In that case, how to scale $S_a^S$? In principle, equation (3.32) tells us how to relate the different scales. But the variables in that equation are unknown, in particular the parameters $m_0$, $s$ (true background events), $k$ (productivity constant) and of course the magnitudes of the unobserved events. One can attempt to replace unknowns by averages or estimates, but this immediately introduces further complications. Therefore, it is utterly non-trivial to compare clustering parameters from different magnitude thresholds.

Rather, clustering parameters are meaningful only in relation to the detection threshold above which they were estimated. The non-uniqueness of the clustering parameters in terms of the "true" parameters and the detection threshold seems irresolvable.

Furthermore, while differences between background rates at a fixed $m_d$ indeed indicate real differences of the background rate, it remains difficult to interpret individual values in an absolute sense. An example is provided by *Hainzl and Ogata* [2005]. The region of Vogtland/NW Bohemia in central Europe is known for episodic swarm-like seismicity, which were linked with local measurements of increased isotopic content of $CO_2$, presumed due to degassing of an active magma body in the upper mantle. The hypothesis was thus formed that the swarm activity is triggered by fluid overpressure. *Hainzl and Ogata* [2005] assume that the background rate in the ETAS model adequately captures events triggered by the fluid signal, but that aftershocks are solely due to "stress-triggering" (by which they mean previous earthquakes). They concluded that "external forcing, identified with pore pressure changes due to fluid intrusion, is found to trigger only a few percent of the total activity." In light of our discussion

above, they need to qualify the statement by attaching an observer's scale, which in their case was $m_d \sim 0.2$. More importantly for their conclusion (and always following their logic), the most precise statement that can be made is that *at most* a few percent were triggered by external forcing. The formulation of the model does not allow for an absolute value. A similar problem emerges when attempting to decluster catalogs using the ETAS model.

### 4.2.2 Declustering

Guided by the idea that main shocks are driven by different mechanisms than triggered events, there have been many attempts to single out main shocks in catalogs and to study their statistics in isolation. The elimination of aftershocks from catalogs is also called declustering, as triggered events are assumed to cluster around independent events. However, there is no accepted physical mechanism separating aftershocks from main shocks that would lead to a unique method of discrimination. In fact, declustering is often motivated by attempts to constrain such a mechanism by the statistical properties of main shocks, e.g. their frequency-magnitude relationship [*Knopoff*, 2000]. Therefore, declustering methods have been devised that identify aftershocks based on belonging to a set of earthquakes defined by a space-time window around presumed mainshocks [*Utsu*, 1969; *Gardner and Knopoff*, 1974; *Felzer et al.*, 2004; *Helmstetter*, 2003] or by a link-based method *Reasenberg* [1985]; *Frohlich and Davis* [1990]; *Davis and Frohlich* [1991]; *Hainzl et al.* [2006].

Declustering and modeling should be equivalent, or at least consistent, since the declustering is an expression of some belief about seismicity relations that are also captured by the model. To use different models for predicting and declustering, for instance, only makes sense if the two are consistent with each other. This approach was followed explicitly by *Kagan* [1991] and Zhuang and co-authors [*Zhuang et al.*,

2002, 2004] because an available likelihood model could be fit to the catalog to estimate the probability of independence of a given quake. Zhuang and co-authors coined the term stochastic declustering to underline that only probabilities can be estimated, not the deterministic branching structure. A probability of being triggered by previous events and of belonging to the background is estimated. One may then roll a die to randomly choose mothers and daughters and create many replicas of declustered catalogs. Stochastic declustering in particular is often considered state-of-the-art declustering because the equivalence of the model of seismicity and the declustering go hand in hand. *Veen and Schoenberg* [2007] recently provided an expectation-maximization (EM) parameter estimation procedure which iterates between the branching structure and parameters until convergence, thereby also estimating a branching structure.

The influence of boundaries was studied to some extent by *Kagan* [1991], though he concentrates on space-time boundaries. The influence of the third boundary, in the magnitude dimension, was investigated briefly with respect to changes in the number of independent events, the constant in the productivity law, the Gutenberg-Richter exponent and the information content. His results showed that some values differed drastically, while others were more stable. He did not interpret his results in terms of the consequences for declustering or the physical interpretation of the triggering parameters. The nature of the magnitude boundary condition is fundamentally different from the space-time boundaries, because being sufficiently away from the boundary does not guarantee that results are free of its influence. This can be seen by measuring the influence of events outside the boundaries on seismicity inside the boundaries. Although the spatio-temporal trigger function decays slowly (power law in time and often power law in space), their influence relative to closer events is often small. However, if many events just below the magnitude threshold occurred close in space and time, their collective influence can be significant (I will make these statements precise in section 4.3.3, where I quantitatively measure the contribution from small earth-

quakes). Furthermore, most evidence suggests that the magnitudes of triggered events are independent of the magnitude of the triggering event (REFS and as assumed in the ETAS model), so that while the spatio-temporal location is clustered, the magnitude "location" is uncorrelated or weakly correlated. It should be stated, however, that the development of models that test for this independence of the magnitude distribution is very important and was pursued, for example, by *Schoenberg* [2003].

If the influence of small earthquakes cannot be ignored, how can it be dealt with? There are in principal three options. The first option would attempt to correct for the bias due to undetected seismicity, for instance by estimating their average effect in the likelihood formulation (this simple solution fails, see section 4.4.2).The second option is to use different models for forecasting or modeling and for declustering. Consider the suggestion to use a magnitude branching model (from large to small) (e.g. *Kagan* [1973]) for declustering. This method would be stable with respect to a change in detection threshold. However, because it is acausal in the time domain, it is awkward for prediction in that it involves many integrals. Note that the large-to-small branching allows the declustering to be robust, in contrast to the idea that small earthquakes can trigger larger earthquakes. This highlights the general property that makes declustering problematic. In any model in which undetected earthquakes contribute significantly to the seismicity budget of observable events, in particular when magnitudes are unpredictable, declustering is intrinsically tied to the detection threshold. Therefore, it would be interesting to study whether it is possible to formulate a time- and magnitude-branching model with (near-) identical properties, one for prediction and one for declustering.

The third option is to abandon the idea of "unbiased" declustering altogether and to understand the "background" rate as a quantity due to a combination of model misspecification and potentially physics. Indeed, *Brémaud and Massoulié* [2001] proved

the existence of a Hawkes process without immigrants under a condition of heavy-tailedness of the trigger function. If this model were fit above a detection threshold, the introduction of a background rate may approximate the influence of undetected events to first order (in section 4.4.2 I will test whether a change in the constant background rate suffices to fit the ETAS model to a different detection threshold and show that it fails). This model removes the suggestion that some earthquakes are different from the others, being triggered by "tectonic loading". Rather, it treats all quakes on the same footing. Clustering appears on all temporal scales, so that the background rate of the ETAS model may solely be a place holder for the model misfit and the triggering due to undetected events.

## 4.3 Analysis of the Stochastic ETAS Model

Since the publication of the previous two chapters, *Saichev and Sornette* [2006a] analyzed the statistics of the fully stochastic ETAS model, extending the study to the full distribution of the statistics, rather than looking at just the mean. After reviewing their results, I will examine the conditional intensity of the ETAS model with respect to changes of the detection threshold. It is the natural object to study because it completely determines the finite-dimensional distributions of the point process.

### 4.3.1 Summary of Results by Saichev and Sornette [2006a]

*Saichev and Sornette* [2006a] applied the formalism of probability generating functions to investigate how the statistical properties of observable earthquakes differ from the statistics of all events. They show that, to a good approximation, the statistical distribution of seismic rates of events with magnitudes above $m_d$ generated by an ETAS model with branching ratio $n$ is the same as that of events generated by another ETAS

model with effective branching ratio $n(m_d)$. For these time-independent statistics, *Saichev and Sornette* [2006a] thus recover the results of *Sornette and Werner* [2005b] in the previous chapter, based solely on the average seismic rate (first moment of the statistics). They further show that the correspondence is not exact, as there are small corrections that can be mapped onto a different branching model. Importantly, they show that the approximate correspondence holds only for the number statistics, but not the (spatio-) temporal properties. Loosely speaking, because the magnitude of an earthquakes is randomly chosen, observed and unobserved events have the same distribution in time and space. Therefore the branching parameter $n$ controls the time and spatial decay of the aftershock rate after a mainshock. For instance, the time $t^\star$ at which the global Omori law crosses over to the local Omori law [*Helmstetter and Sornette*, 2002] is determined by the true branching ratio. The effect of changing the local magnitude cut-off from $m_0$ to $m_d$ is only to rescale the the seismicity rate by the fraction of observed events, without changing the decay of aftershocks with time. The next section provides an intuitive explanation of the latter result in terms of the conditional intensity function.

### 4.3.2 Analysis of the Conditional Intensity under a Change of Detection Threshold

The conditional intensity uniquely defines all distributions of a point process [*Daley and Vere-Jones*, 2003]. It is therefore the natural object to study under a change in detection threshold. Recall that the conditional intensity of the temporal ETAS model is defined by:

$$\lambda_0(t, m|H_0) = p_m(m) \cdot \left( \mu + \sum_{i|t_i<t} \frac{k\ e^{a(m_i-m_0)}}{(t-t_i+c)^p} \right) \tag{4.1}$$

where $p_m(m) = \beta \exp(-\beta(m - m_0))$ and $H_0$ is the history of event times and marks down to $m_0$. The total rate above some minimum threshold $m_0$ is found by integrating over $m$:

$$\lambda_0(t|H_0) = \mu + \sum_{i|t_i<t} \frac{k\ e^{a(m_i-m_0)}}{(t - t_i + c)^p} \tag{4.2}$$

while the rate above $m_d$ is given by:

$$\begin{aligned}
\lambda_d(t|H_0) &= \int_{m_d}^{\infty} p_m(m)dm \cdot \left(\mu + \sum_{i|t_i<t} \frac{k\ e^{a(m_i-m_0)}}{(t - t_i + c)^p}\right) \\
&= f_d \left(\mu + \sum_{j|m_j \geq m_d} \frac{k\ e^{a(m_j-m_0)}}{(t - t_j + c)^p} + \sum_{l|m_l<m_d} \frac{k\ e^{a(m_l-m_0)}}{(t - t_l + c)^p}\right) \\
&= \mu e^{-\beta(m_d-m_0)} + \sum_{j|m_j \geq m_d} \frac{k\ e^{a(m_j-m_0)-\beta(m_d-m_0)}}{(t - t_j + c)^p} \\
&\quad + \sum_{l|m_l<m_d} \frac{k\ e^{a(m_l-m_0)-\beta(m_d-m_0)}}{(t - t_l + c)^p}
\end{aligned} \tag{4.3}$$

where the undetected events $l$ with $m_l < m_d$ were separated from the detected events $j$ with $m_j \geq m_d$ and the fraction of observed events is $f_d = \exp(-\beta(m_d - m_0))$. Now consider the conditional intensity above $m_d$ given only the history above $m_d$:

$$\lambda'_d(t|H_d) = \mu' + \sum_{j|m_j \geq m_d} \frac{k'\ e^{a'(m_j-m_d)}}{(t - t_j + c')^{p'}} \tag{4.4}$$

Is there a way of mapping $\lambda_d(t|H_0) \rightarrow \lambda'_d(t|H_d)$ by redefining the parameters only while preserving the functional dependencies? One may redefine the parameters $k' = k\ \exp[-(\beta - a)(m_d - m_0)]$, $a' = a$, $c' = c$, $p' = p$. But this forces the mapping:

$$\mu' = \mu e^{-\beta(m_d-m_0)} + \sum_{l|m_l<m_d} \frac{k\ e^{a(m_l-m_0)-\beta(m_d-m_0)}}{(t - t_l + c)^p} \tag{4.5}$$

which implies that the background rate is no longer a time-independent quantity. This failure to map the constant background onto a different, but still constant background is important because it means that one cannot reapply the ETAS model at different

thresholds by a simple redefinition of the parameters. In this sense, the model cannot be renormalized onto itself. It is not a fixed-point of the renormalization process operating via magnitude coarse-graining. The functional form of the model must change under a change in the detection threshold. In other words, if earthquakes occur according to an ETAS model above some cut-off $m_0$, then earthquakes above $m_d$ cannot be described by the ETAS model in a mathematically exact way. In practice, of course, the ETAS model may provide an excellent fit (as we will see in section 4.4.2).

Deleting unobserved events together with the fact that events have independent and identically distributed magnitudes is equivalent to randomly deleting ("thinning") events from the process. Since randomly deleting points destroys correlations, the limiting process of a highly thinned process is the Poisson process (see *Daley and Vere-Jones* [2003], for a rigorous derivation). This argument illustrates that as $m_d \to \infty$, the observed process is Poissonian, so that the conditional intensity is reduced to a constant. Fitting an ETAS model to highly thinned earthquake data would result in a non-zero background rate together with vanishing triggering. This is consistent with observations that the largest earthquakes (worldwide and regional) are less and less clustered as the threshold is raised (although *Kagan and Jackson* [1991a] found some evidence of long-term clustering of large earthquakes). One may thus guess that undetected earthquakes contribute a constant to the observable seismicity budget, i.e. as $m_0 \to m_d$, $\mu \to f_d \cdot \mu + A$. The next section will calculate the contribution of undetected events to observed seismicity by calculating (i) the time-independent average, and (ii) the exact contributions to observed events in a simulated catalog.

### 4.3.3 Contribution of Undetected Events to the Observable Seismicity Budget

The third term in (4.3) quantifies the influence of unobserved events to observed seismicity. In this section, I will first calculate the average contribution, demonstrating that

for reasonable parameter values, their influence is significant. Secondly, I will simulate a catalog and calculate their contribution to the occurrence of observed events exactly.

The influence of undetected events on the observable seismicity budget can be estimated by the average observed number of direct aftershocks from undetected mothers divided by the total observed seismicity. Denote the average number of direct aftershocks from undetected mothers by $n^<$ and the productivity law by $\rho(m)$. Assume for simplicity that the Gutenberg-Richter law $p(m)$ is truncated at an upper magnitude $m_{max}$ as in Chapters 2 and 3 (see equation 3.1). Then the branching ratio of undetected mothers is given by, assuming $\alpha < b$:

$$n^< \equiv \int_{m_0}^{m_d} p(m|m < m_d)\rho(m)dm = n\left[\frac{1 - 10^{-(b-\alpha)(m_d - m_0)}}{(1 - 10^{-(b-\alpha)(m_{max} - m_0)})(1 - f_{obs})}\right] \quad (4.6)$$

where we have used the fact that $P(m|m < m_d) = P(m)/(1 - f_{obs})$ for $m < m_d$ and zero otherwise. Also, $n$ is the branching ratio of all mothers and $f_{obs}$ is the fraction of observed events above $m_d$, given by

$$f_{obs} = \frac{10^{-b(m_d - m_0)} - 10^{-b(m_{max} - m_0)}}{1 - 10^{-b(m_{max} - m_0)}} \quad (4.7)$$

The expressions (4.6) and (4.7) reduce to simpler, more familiar forms when $m_{max} \rightarrow \infty$. When $\alpha = b$, $n^<$ reduces to:

$$n^< = \frac{n}{1 - f_{obs}}\left[\frac{m_d - m_0}{m_{max} - m_0}\right] \quad (4.8)$$

The total average seismicity per background event in a catalog is given by $1/(1 - n)$. The influence of undetected earthquakes can be measured by the ratio $\langle r \rangle$ of the observed direct and indirect aftershocks from undetected mothers $f_d \cdot n^</(1 - n)$ over all observed seismicity $f_d/(1 - n)$ per mother, given by:

$$\langle r \rangle = \frac{n^< \cdot f_{obs}}{1 - n} / \left(\frac{f_{obs}}{1 - n}\right) = n^< \quad (4.9)$$

Expression (4.9) shows that the fraction of observed events due to undetected mothers is directly equal to the average number of direct aftershocks of undetected mothers

given by (4.6). $n^<$ is smaller than $n$ because the magnitudes of the undetected mothers are smaller than all magnitudes. Figure 4.1 shows the values of $n^</n$ as a function of $\alpha$, for maximum magnitude $m_{max} = 9$, $m_d = 3$ and various choices of $m_0$. I assumed $b = 1$ and used expressions (4.6) for $\alpha < b$ and (4.8) for $\alpha = b$. The influence of small earthquakes becomes larger for small $\alpha$ because the rareness of large earthquakes is not compensated by more productive triggering. Small earthquakes also become more influential as $m_0$ decreases, simply because there are more potential mothers. For instance, for $\alpha = 0.8$ and $m_0 = 0$, undetected seismicity on average accounts for about $70\%$ of all observed triggered events. If $n$ is close to one, then the $70\%$ also describe the fraction of the total observed seismicity due to undetected events. The abundant triggering we observe ($0.5 < n < 1$, although calculated from biased parameters), together with the commonly found maximum likelihood estimates of $0.5 < \alpha < 1$, implies that small earthquakes trigger a substantial part of the observed seismicity. Even if $\alpha = 1$, as suggested by *Felzer et al.* [2003] and *Helmstetter et al.* [2005a], earthquakes in the range $0 < m < 3$ still trigger more than $30\%$ of observed events above $m_d = 3$. Therefore, their contribution to the conditional intensity in the third term of expression (4.3) is highly significant.

Thus far, we only analyzed the average contribution $n^<$ of undetected events to the observable seismicity budget. To better understand the temporal evolution of this contribution to the observable rate, I analyzed the fraction $r(t_i)$ of the observed conditional intensity due to undetected events at the observed event times $t_i$ in a simulated catalog. I generated a catalog in the interval $[0, 1000]$ days with parameters $\{k = 0.475, \alpha = 0.5, c = 0.001, p = 1.2, b = 1\}$ above $m_0 = 0$, shown in Figure 4.2. The simulated catalog contained 6402 events, while 64 events were larger than $m_d = 2$. The parameters imply that the branching ratio is $n = 0.95$ and that $n^< = 0.86$ (neglecting the upper cut-off $m_{max}$). I calculated the conditional intensity $\lambda(t_i|H_t^0)$ at each observed event $t_i$ such that $m_i \geq m_d$. Then I calculated the rate $\lambda^<(t_i|H_t^<)$ due

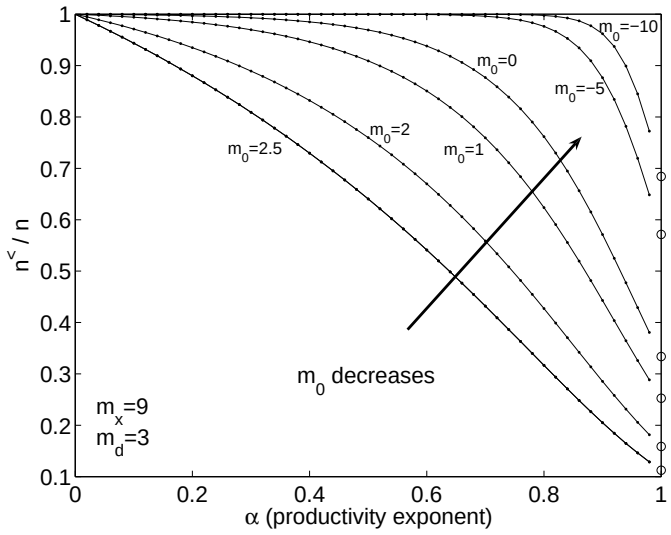


Figure 4.1: The fraction of triggered events due to undetected mothers divided by the total fraction of triggered events as a function of the parameter $\alpha$ and for several choices of $m_0$. I assumed that $b = 1$, the maximum magnitude $m_{max} = 9$ and the detection threshold $m_d = 3$. Filled markers calculated from expression (4.6) for $\alpha < b$, open markers from (4.8) for $\alpha = b$.

to unobserved events using only undetected events $m_l < m_d$. The fraction $r(t_i)$ of the total intensity due to unobserved events at each observed event measures the influence of undetected mothers in triggering observed events. Figure 4.3 shows this ratio $r(t_i)$. The average $\langle r(t_i) \rangle_{t_i} = 0.76$, not too far from the theoretical $n^< = 0.86$ given that only $64$ events were sampled. However, the ratio $r(t_i)$ clearly fluctuates strongly between $20\%$ and $100\%$ from event to event.

This section demonstrated clearly that small earthquakes contribute significantly to the observed seismicity budget. A natural next question is to investigate their impact on model fits to observed seismicity. Can their contribution be well-approximated by the correct parameters, or by effective parameters, or is the ensuing fit simply unac-

ceptable?

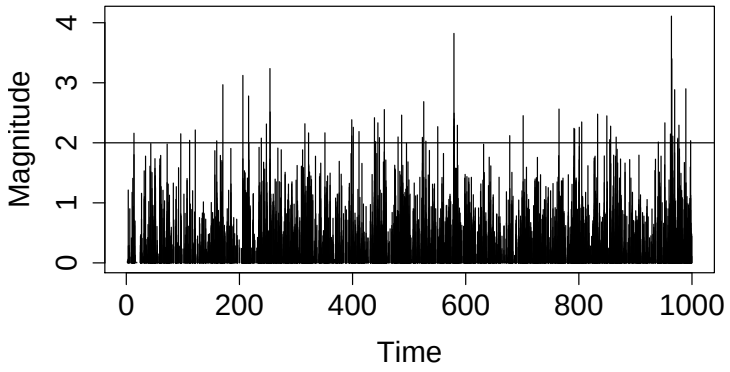


Figure 4.2: Simulated catalog with parameters $k = 0.475$, $\alpha = 0.5$, $c = 0.001$, $p = 1.2$, $b = 1$ in time interval $[0, 1000]$ days.

### 4.3.4 Residual Analysis of Observed Seismicity

Having demonstrated that the observed seismicity budget seems largely controlled by undetected earthquakes, one may wonder whether their contribution will show up as unexpected clustering, measured, for instance, by strong misfits of the model to observed events. The answer to this question is important because ETAS is increasingly being used as a null hypothesis and deviations are sometimes interpreted as significant indicators of activation or quiescence that may be used for prediction (a hypothesis defended in particular by Y. Ogata [*Ogata*, 1988, 1989, 1992, 2001, 2004, 2005, 2006a,b]. Do undetected earthquakes add a general time-independent contribution to the observed seismicity, quantified by the above average $n^<$, or are their effects strongly time-dependent?

In the following, I will begin to answer this question concerning the effect of un-

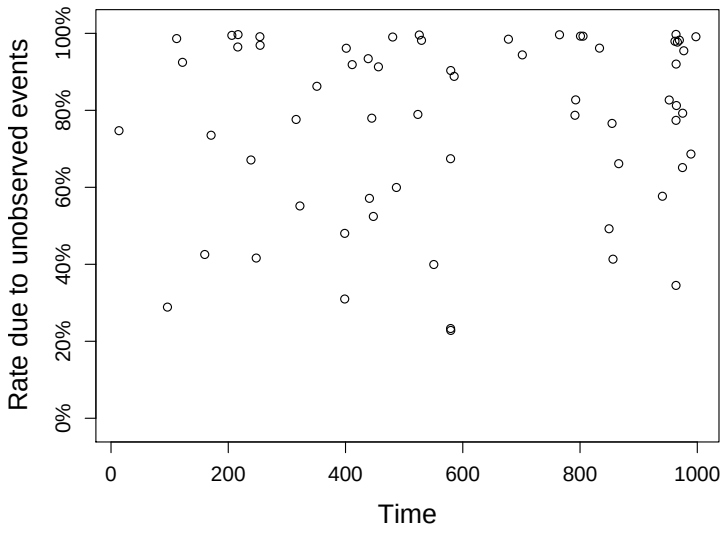


Figure 4.3: Fraction of total conditional intensity $\lambda(t_i|H_t^0)$ at observed events $m_j \geq m_d = 2$ due to unobserved mothers $0 \leq m < m_d$ from the simulated catalog in Figure 4.2.

detected seismicity on model fits. In the first part, I will fit the model to observed seismicity above $m_d$ using the original parameters used to simulate the catalog above $m_0$ and using parameters that were scaled using the Gutenberg-Richter relationship. I will show that these simple guesses fail to fit the observed catalog. This exercise proves that, if even one or two magnitude units of earthquakes exist below routinely detected thresholds, then all parameters that are estimated from the data are purely effective parameters that bear little or no relation to the "true" parameters and are intrinsically tied to the detection threshold. In the second part, I will test whether estimated parameters provide even acceptable fits. For this second part, I will use maximum likelihood to estimate the parameters. This is presented in section 4.4.1, after which I return in section 4.4.2 to address the question raised here.

The fit of the ETAS model to data can be assessed through the analysis of residuals. It can be shown [*Papangelou*, 1972; *Ogata*, 1988; *Daley and Vere-Jones*, 2003; *Schoenberg*, 2002] that, if $0 < t_1 < t_2 \ldots$ is an unbounded, increasing sequence of time points on the half-line, $N^\star$ is a simple point process with internal history $H_t$ and a monotonic, continuous $H_t$ compensator $\Lambda^\star(t)$ such that $\Lambda^\star(t) \to \infty$ almost surely (a.s.), then, with probability one, the transformed sequence

$$\tau_i = \Lambda^\star(t_i) = \int_0^{t_i} \lambda(u|H_t)du \tag{4.10}$$

is a realization of a unit-rate Poisson process if and only if the original sequence $\{t_i\}$ is a realization from the point process defined by $\Lambda^\star(t)$. Therefore $\tau_i$ should be uniformly distributed in the studied catalog time interval $(0, \Lambda^\star(T))$. If the model or parameters are misspecified, there will be detectable deviations from the Poisson process with unit rate, quantified for instance by the Kolmogorov-Smirnov test.

To illustrate quantitatively the discussion regarding biased parameter estimated because of undetected seismicity, I simulated a catalog using the ETAS model with parameters $\theta = \{\mu = 1, b = 1, k = 0.477, \alpha = 0.5, p = 1.2, c = 0.001\}$. These parameters ensured a high branching ratio of $n = 0.95$ and that small earthquakes contribute strongly to the seismicity budget, although by no means unrealistic. I generated a catalog of length $T = 100,000$ days, resulting in $1,277,028$ events. To check that the original catalog transforms correctly with its original parameters, I transformed the times of the first $100,000$ events according to (4.10). As can be seen from Figure 4.4, the resulting process is a Poisson process with unit rate (its dashed line for comparison completely coincides with the transformed times). The Kolmogorov-Smirnov test cannot reject the hypothesis that the two curves are from the same distribution (p-value of $p = 0.977$). (Since the simulation and residual analysis codes were written separately, this result helps verify the simulation code, too.)

What happens when we introduce a detection threshold $m_d > m_0$? According to

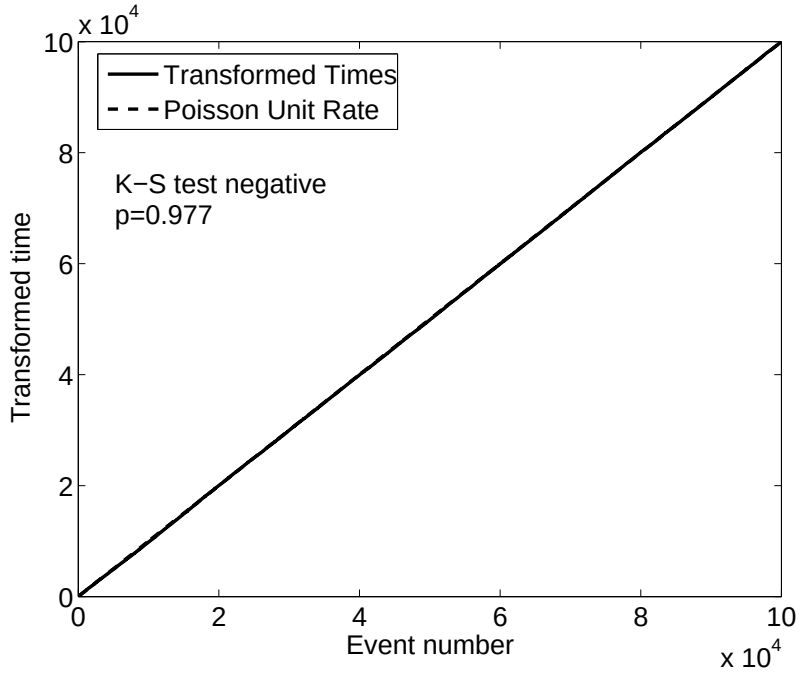


Figure 4.4: Transformed times versus event numbers and their comparison with a unit rate Poisson process: original catalog with original parameters.

equation (4.3), the only correct way for transforming the observed occurrence times $t_j$ for which $m_j \geq m_d$ is to use the true conditional intensity multiplied by the fraction of observed events $\lambda(t) \cdot f_d$ (i.e. scaling $\mu \rightarrow f_d\mu$ and $k \rightarrow f_d k$) together with the complete catalog history $H_0$. Setting $m_d - m_0 = 2$, we are left with $12,666$ events of the original catalog. Figure 4.5 shows the resulting transformed times (solid) along with a unit rate Poisson process (dashed). Again, the Kolmogorov-Smirnov (KS) test cannot reject the hypothesis that both curves are of the same distribution ($p = 0.92$).

In reality, of course, we do not have information about unobserved events. The question then becomes whether we can use the original parameters together with the observed catalog to transform the occurrence times correctly. Figure 4.6 shows clearly that this is not so (K-S test rejects the common-distribution hypothesis with $p$ essen-

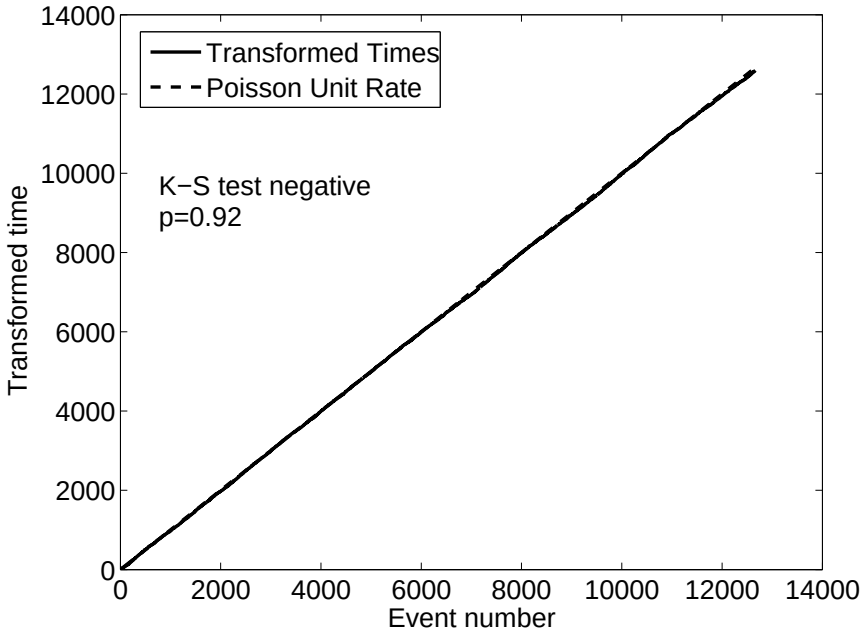


Figure 4.5: Transformed times versus event numbers and their comparison with a unit rate Poisson process: Transforming observed events $m_j \geq m_d = 2$ using "scaled" true parameters (with $\mu$ and $k$ scaled by $f_d$) and full history $H_d$ above $m_0$ in conditional intensity.

tially zero). The model expects too many events compared to the actual occurrence rates. The original parameters would have produced many more events, as expected given that they actually did produce the original, much more populated catalog.

One may be tempted to try scaling the conditional intensity function by $f_d$ (equivalently, scaling the background rate $\mu$ and $k$) when using the observed catalog to transform occurrence times. But, given that these are the correct parameters to use when the entire catalog is available to transform observed occurrence times, one expects that too few events are predicted. Indeed, the transformed times do not form a Poisson process with unit rate, as shown in Figure 4.7. The K-S test rejects this hypothesis entirely.

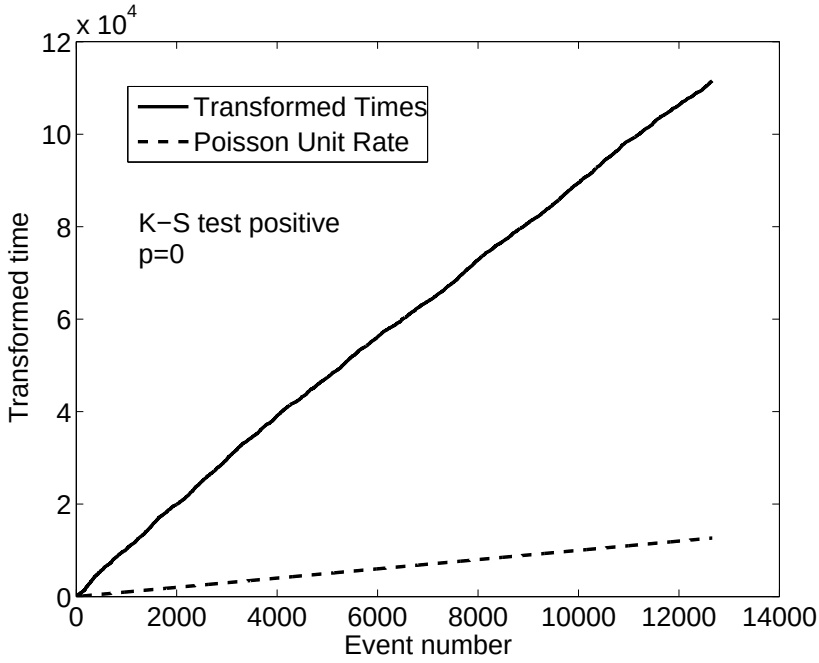


Figure 4.6: Transformed times versus event numbers and their comparison with a unit rate Poisson process: Transforming observed events $m_j \geq m_d = 2$ using the true parameters and the observed history $H_d$ above $m_d$ in the conditional intensity.

An interesting feature of Figures 4.6 and 4.7 is that the transformed times seem to follow approximately straight lines (at this resolution). This suggests that the transformed times may follow a Poisson process, but with a different rate, which in turn suggests that there exists a simple way of scaling the parameters to account for undetected seismicity. For instance there may be a constant, different from $f_d$, by which the parameters $\mu$ and $k$ can be multiplied in order to approximately fit the observed catalog. Or there may exist a constant which can be added to the background rate $\mu \rightarrow f_d \cdot \mu + A$ so as to account for undetected events, replacing the time-dependent triggering from undetected events to observable seismicity with a time-independent constant, for instance by their average contribution $n^<$. The latter is also supported by

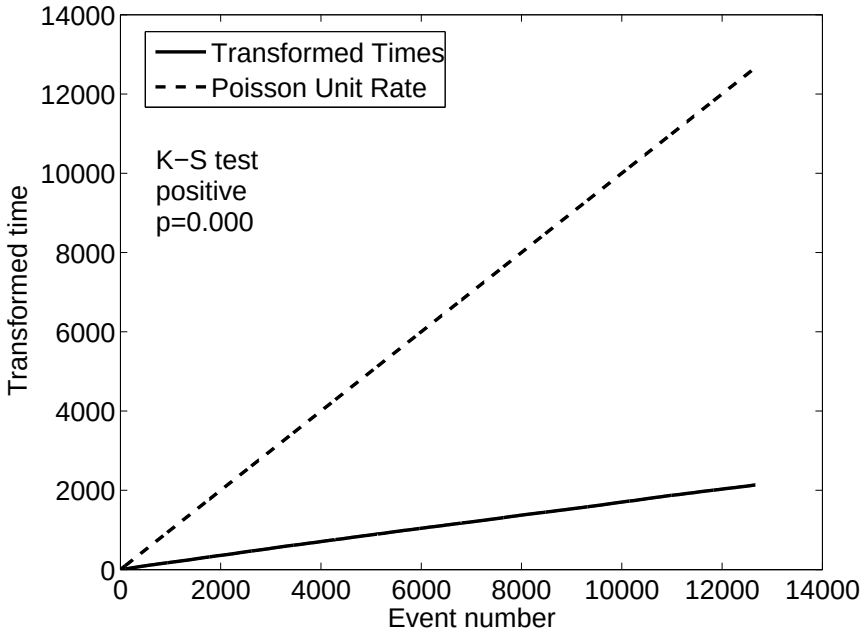


Figure 4.7: Transformed times versus event numbers and their comparison with a unit rate Poisson process: Transforming observed events $m_j \geq m_d = 2$ using the "scaled" true parameters (with $\mu$ and $k$ scaled by $f_d$) and the observed history $H_d$ above $m_d$ in the conditional intensity.

the earlier observation that thinning leads to Poisson processes. Such a simple solution seems to contradict the suggestion that the effect of undetected seismicity on observed events is time-dependent and clustered which an effective ETAS model may not be able to capture (section 4.3.2). These questions cannot be answered without estimating parameters. However, they are important because deviations from the ETAS fit may be expected solely because of the model formulation and the effect of undetected mothers triggering observable aftershock sequences. The next section presents the maximum likelihood estimation procedure before trying to address these questions.

## 4.4 Parameter Estimation

### 4.4.1 Maximum Likelihood Parameter Estimation

Parameter estimates are usually obtained by maximizing the likelihood function of the point process under consideration. *Ozaki* [1979] provided a first explicit algorithm for the Hawkes' self-exciting process [*Hawkes*, 1971a,b; *Hawkes and Oakes*, 1974], a direct ancestor of the ETAS model. *Ogata* [1988] provided a likelihood estimation procedure for the ETAS model. Recently, *Zhuang et al.* [2002, 2004] and *Veen and Schoenberg* [2007] performed iterative estimation procedures that use the branching structure of the ETAS model as well as the partial likelihood to converge to stable parameter estimates.

The likelihood of the temporal ETAS model is a special case of the more general likelihood of any marked temporal point process [*Daley and Vere-Jones*, 2003] given by:

$$L(\theta) = \left[\prod_{i=1}^{N(T)} \lambda_\theta(t_i, m_i|H_t)\right] \exp\left(-\int_{m_0}^{\infty}\int_{0}^{T} \lambda_\theta(u, m|H_u)dudm\right) . \qquad (4.11)$$

The parameters $\hat{\theta}$ which maximize the likelihood "best" explain the given data by making the observed data the most probable given the model (e.g. *Kay* [1993]).Maximizing the likelihood is equivalent to maximizing the log-likelihood $\ell$, being a monotonic function of the likelihood:

$$\ell(\theta) = \sum_{i|t_i<t} \log \lambda_\theta(t_i, m_i|H_t^0) - \int_0^T \int_{m_0}^{\infty} \lambda_\theta(t, m|H_t^0)dtdm \qquad (4.12)$$

where the conditional intensity is replaced by the specific point process under study. The dependence of the conditional intensity on the full data history $H_t^0 = \{t_i, m_i\}_{i|t_i<t}$ down to $m_0$ is made explicit. For unpredictable marks (i.e. independently and identically distributed marks independent of time), one can separate or factorize the de-

pendence of the likelihood function on the marks from the temporal "ground" process $\lambda(m, t|H) = p_m(m)\lambda(t|H)$ and estimate the parameters of the mark distribution and the ground process separately. We often assume that the Gutenberg-Richter value $b$ is known or estimated separately from the temporal model. Assuming $b$ is known, expression (4.12) simplifies to:

$$\ell(\theta) = \sum_{m_i \geq m_0} \log \lambda_\theta(t_i, m_i|H_t^0) - \int_0^T \lambda_\theta(t|H_t^0)dt \tag{4.13}$$

Before the introduction of the above-mentioned iterative estimation procedures which make use of the branching structure representation of the ETAS model, the log-likelihood was usually maximized by grid search or steepest descent methods. *Ogata et al.* [1993] proposed a method to speed up the likelihood calculation.

### 4.4.2 Parameter Bias Due to Undetected Seismicity

In practice, one replaces $m_0$ with the detection threshold $m_d$, so that the conditional intensity is replaced by $\lambda'_d(t|H_t^d)$ given the history above $m_d$ only. The likelihood is then calculated from:

$$\ell'(\theta') = \sum_{m_k \geq m_d} \log \lambda'_{\theta'}(t_k, m_k|H_t^d) - \int_0^T \lambda'_{\theta'}(t|H_t^d)dt \tag{4.14}$$

$$\lambda'_{\theta'}(t|H_t^d) = \mu' + \sum_{j|m_j \geq m_d} \frac{k'\, e^{a'(m_j - m_d)}}{(t - t_j + c')^{p'}} \tag{4.15}$$

This should be contrasted with the correct likelihood function:

$$\ell(\theta) = \sum_{m_k \geq m_d} \log(f_d \cdot \lambda_\theta(t_k, m_k|H_t^0)) - \int_0^T f_d \cdot \lambda_\theta(t|H_t^0)dt \tag{4.16}$$

$$\lambda_0(t|H_t^0) = \mu + \sum_{i|m_i \geq m_0} \frac{k\, e^{a(m_i - m_0)}}{(t - t_i + c)^p} \tag{4.17}$$

which contains information about events below $m_d$ through its history $H_t^0$, i.e. while the conditional intensity is evaluated *at* observed events, it is calculated *from* all events.

If the ETAS model description is reasonable for even one or two magnitude units below the detection threshold $m_d$, then the wrong conditional intensity is used to estimate parameters. Therefore, one can expect biased parameters.

Which of the assumptions in the derivation of the consistency of the maximum likelihood estimator fail? The (true) conditional intensity is no longer predictable. More precisely, the true conditional intensity is not available because of unavailable, unobserved events. As a consequence, the difference between the point process $N(t)$ and the compensator $\Delta(t) = \int_0^t \lambda(u)du$ based on the assumed (wrong) intensity is no longer an $\mathcal{F}$-martingale, which is a necessary assumption in the derivation of asymptotic normality and consistency [*Ogata*, 1978; *Rathbun*, 1996]. In [*Ogata*, 1978], assumption B2 (that $\lambda_\theta$ is predictable) is violated. In reality, $\lambda_\theta(t, H_t^d)$ is distributed because the unknown history below $m_d$, if left unspecified, can lead to many different values of the intensity with different probabilities.

As mentioned above, the deviation of the true intensity from the estimated intensity is not limited to areas close to the boundary (as for time and space), but pervades the entire space-time-magnitude space because marks are unpredictable and can be triggered by the numerous small earthquakes.

We return to the question posed in section 4.3.4 whether the addition of a constant to the background rate to account for the influence of undetected events is sufficient to fit an ETAS-simulated catalog above a threshold, while only scaling $k \rightarrow k \cdot f_d$. Using the catalog simulated in section 4.3.2, I performed a maximum likelihood inversion for the parameter $\mu$ of the catalog above $m_d = 3$ (1254 events) while constraining $\alpha, c$ and $p$ to their true values but rescaling $k \rightarrow k \cdot f_d = 0.00047$. 10 different random initial values all converged to very similar values, showing that the value obtained for $\hat{\mu}$ was robust. The value obtained was $\hat{\mu} = 0.0117 \pm 0.0001$ (here and below the errors are based solely on the ten different $\hat{\mu}$). The likelihood values of the resulting fits

were $LL = -6547.21 \pm 0.005$. I then performed a residual analysis of the observed occurrence times with the true parameters, rescaled $k$ and the inverted $\hat{\mu}$. Figure 4.8 shows that the fit is approximate and not terrible. However, the K-S test rejects the hypothesis that the transformed times are from the same distribution ($p = 0.002$). This suggests that to approximate the influence of undetected seismicity with a constant is not terrible to first order (visually), but by no means sufficient, as the K-S test easily rejects the model. The implication is that thinning small events does not lead solely to a higher background rate, it also impacts the triggering parameters beyond the expected Gutenberg-Richter rescaling.

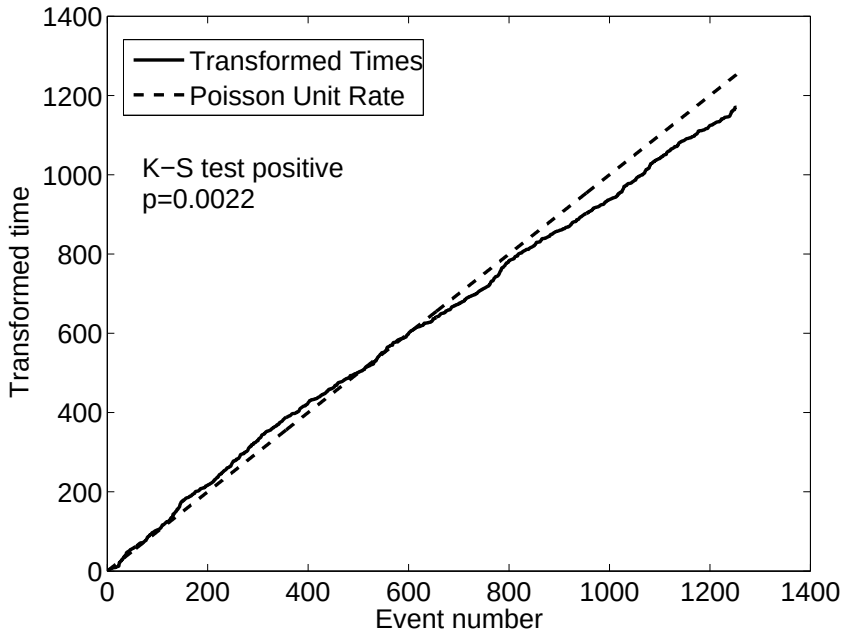


Figure 4.8: Transformed times versus event numbers and their comparison with a unit rate Poisson process: Transforming observed events $m_j \geq m_d = 3$ with estimate $\hat{\mu} = 0.0117$, scaled $f_d \cdot k$ and true $\alpha, c$ and $p$, using the observed history $H_d$ above $m_d$.

In the spirit of the previous two chapters, one may expect a trade-off between $\mu$

and $k$ (a proxy for the branching ratio $n = kb/(b - \alpha)$). I therefore repeated the previous estimation procedure but allowed $k$ to be determined by MLE as well. The first ten random initial starting values for the parameters $\mu$ and $k$ were distributed widely, and the estimation procedure estimated $\mu = 0$ in two cases and $k = 0$ in two further cases, indicating poor convergence. Only six of the ten converged to acceptable values $\hat{\mu} = 0.01135 \pm 0.00005$ and $\hat{k} = 0.0514 \pm 0.001$. A second run with ten more narrowly distributed initial values all converged to the same estimates, demonstrating a dependence on initial values even for estimating only two of the five parameters. The likelihood values of the acceptable results were $LL = -6510.729 \pm 0.003$, an improvement over the likelihood value achieved with just one parameter. Using $\hat{k} = 0.0514$ and $\hat{\mu} = 0.01134$, the K-S test cannot reject the hypothesis that the two curves belong to the same distribution (see Figure 4.9). This suggests that the observed catalog can be fit by using true values for $\alpha, p$ and $c$, while allowing $\mu$ and $k$ (the triggering parameters) to adjust. As predicted in Chapter 3, I find that the background rate and the branching ratio change simultaneously under a change of detection threshold and seem sufficient to provide an acceptable fit for this particular catalog. Both the background rate and the constant $k$ are elevated with respect to their GR-scaled values, in order to make up for the contribution from undetected mothers. But while the background rate is within $20\%$ of its GR-scaled value, $\hat{k}$ is larger by a factor of ten than its GR-scaled value. The nature of this relative increase and its dependence on other parameters and $m_d - m_0$ needs to be investigated in more detail.

In practice, the "true" values of the parameters $\alpha, p$ and $c$ are unknown and all five parameters need to be estimated. It is therefore of interest to compare the values of the estimates with the true values, i.e. their bias. I will first allow a third parameter $\alpha$ to be estimated simultaneously before estimating all five parameters.

The third parameter essential for the number statistics is $\alpha$. Again starting from

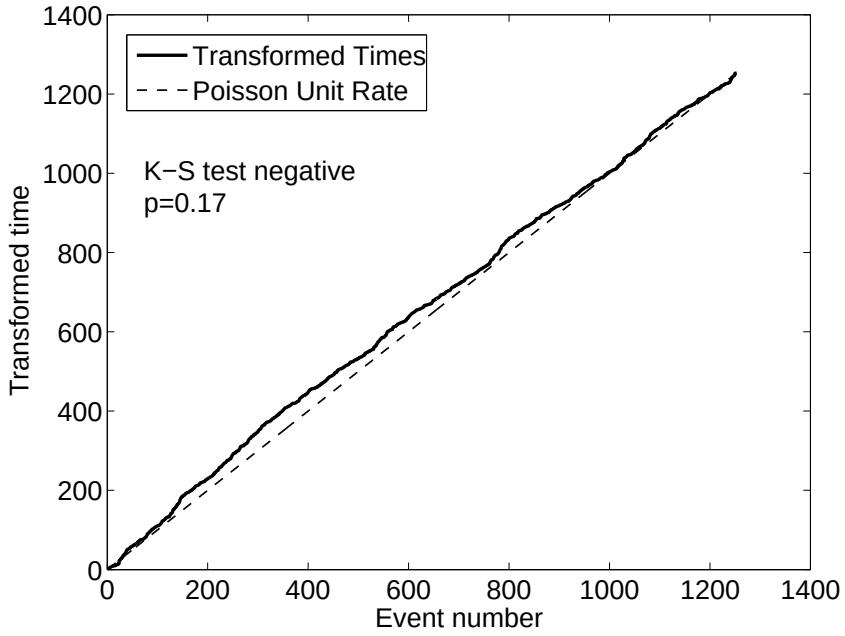


Figure 4.9: Transformed times versus event numbers and their comparison with a unit rate Poisson process: Transforming observed events $m_j \geq m_d = 3$ with estimates $\hat{\mu} = 0.01134$ and $\hat{k} = 0.0514$ and true $\alpha, c$ and $p$, using the observed history $H_d$ above $m_d$.

ten initial points, $\hat{k} = 0$ occurred in three cases, while $\alpha$ turned negative in two. Only six of the ten initial values converged to reasonable estimates $\hat{\mu} = 0.01135 \pm 0.00001$, $\hat{k} = 0.053 \pm 0.002$ and $\hat{\alpha} = 0.475 \pm 0.003$ with a log-likelihood value of $LL = -6510.675 \pm 0.004$. The increase in likelihood gained by allowing $\alpha$ to be estimated is extremely small. There is a small trade-off between $k$ and $\alpha$ which is to be expected. The transformed times are shown in Figure 4.10 and are consistent with a unit rate Poisson process.

The estimation of all five parameters using maximum likelihood above $m_d = 3$ was highly unstable. From 10 initial (random) starting points close to the true (but scaled)

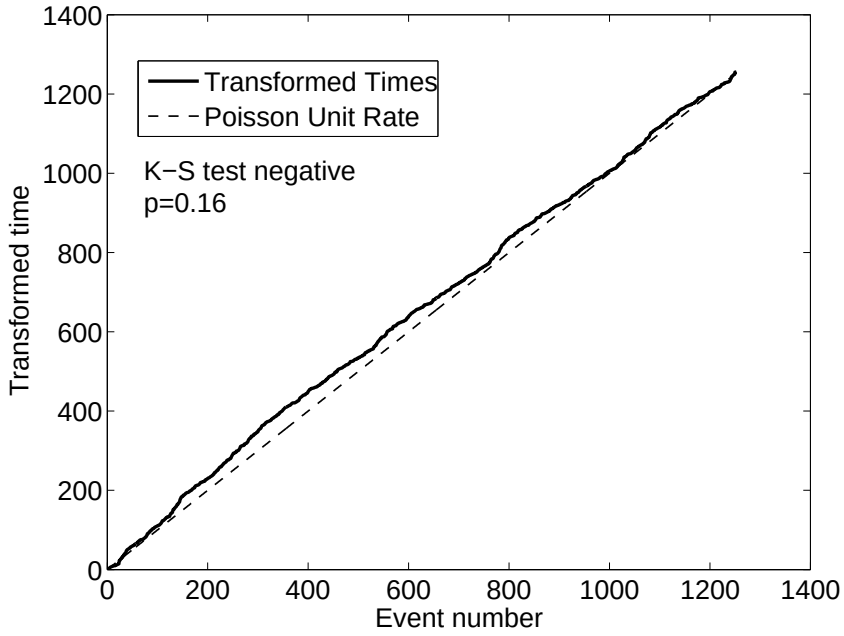


Figure 4.10: Transformed times versus event numbers and their comparison with a unit rate Poisson process: Transforming observed events $m_j \geq m_d = 3$ with estimates $\hat{\mu} = 0.01134$ and $\hat{k} = 0.0514$ and true $\alpha, c$ and $p$, using the observed history $H_d$ above $m_d$.

parameters, two failed to converge in 1000 iterations. Only four of the ten inversions estimated non-zero $\hat{k}$ (the others resulted in a zero branching ratio). Four inversions gave a negative $\hat{\alpha}$ and five gave $\hat{p} < 1$. The inversion with the highest likelihood value set $\hat{k} = 0$. Only two inversions resulted in subcritical but non-zero branching ratio. Of these, only one performed satisfactorily in a K-S test against a unit Poisson process (Figure 4.11). The other one (not shown) strongly deviated. Furthermore, the only satisfactory inversion was beaten heavily in likelihood value by two others, both of which set $\hat{k} = 0$. The relative abundance of events $(1254)$ together with such poor convergence results is a reminder of the strong non-linearity of the problem. We will

come back to the general estimation problem below.

The only acceptable inversion estimated the following parameters: $\hat{\mu} = 0.008564$, $\hat{k} = 0.05884$, $\hat{\alpha} = 0.405$, $\hat{c} = 0.000988$, $\hat{p} = 0.917$. Figure 4.11 shows the transformed times of this model. For this particular catalog, the bias $\epsilon$ in the parameters calculated by $\epsilon_\gamma = 100 * (\hat{\gamma} - \gamma)/\hat{\gamma}$ where $\gamma$ is the "true" parameter, is therefore: $\epsilon_\mu = 88\%$ too large with respect to the GR-scaled background, and $\epsilon_\mu = -116\%$ too small with respect to the original background, $\epsilon_k = 92\%$ too large with respect to the GR-scaled $k$ and $\epsilon_k = -700\%$ too small with respect to the original $k$, $\epsilon_\alpha = -23\%$ too small, $\epsilon_c = -1\%$ too small and $\epsilon_p = -31\%$ too small. The log-likelihood value of this solution was $LL = -6486.391$.

For completeness, one may consider the relative improvement of the model fit that one attains by allowing more parameters to be estimated. The Akaike Information Criterion (AIC) [*Akaike*, 1974] provides a measure of such improvement, justified theoretically as an asymptotically unbiased estimate of the Kullback-Leibler information loss [*Anderson and Burnham*, 2002; *Burnham and Anderson*, 2004]. It is defined by

$$AIC = -2\ell(\theta) + 2K \tag{4.18}$$

where $\ell(\theta)$ is the log-likelihood of the model and $K$ is the number of parameters of the model. Here, model refers to the ETAS model fit using estimated parameters and the number of parameters is the number of *estimated* parameters, treating the other parameters as fixed. Table 4.1 summarizes the above estimation procedures and gives their AIC score. Clearly, allowing all five parameters to be determined by the model fit greatly enhances the AIC score. However, it is interesting to note that the two-parameter fit scores better than the three-parameter fit. I note that the derivation of the AIC is based on asymptotic properties: for non-linear, non-Gaussian and finite-sample conditions, this estimate of the Kullback-Leibler information may be problematic.

We can draw several conclusion from the maximum likelihood inversions. Firstly,

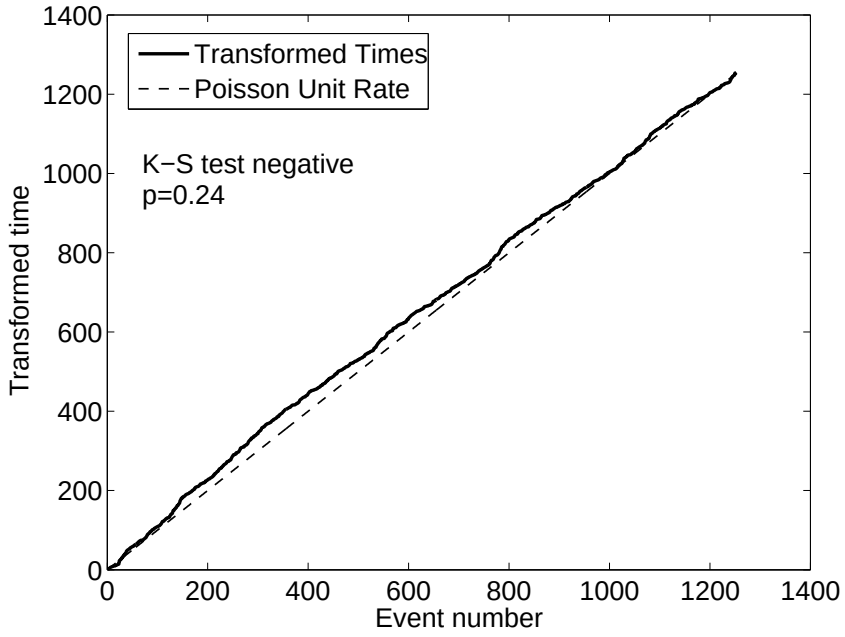


Figure 4.11: Transformed times versus event numbers and their comparison with a unit rate Poisson process: Transforming observed events $m_j \geq m_d = 3$ with estimates $\hat{\mu} = 0.008564$, $\hat{k} = 0.05884$, $\hat{\alpha} = 0.405$, $\hat{c} = 0.000988$, $\hat{p} = 0.917$, using the observed history $H_d$ above $m_d$.

| Estimated Parameters | $\ell(\theta)$ | $AIC$ |
|---|---|---|
| $\hat{\mu}$ | -6547.21 | 13095.42 |
| $\hat{\mu}, \hat{k}$ | -6510.73 | 13023.46 |
| $\hat{\mu}, \hat{k}, \hat{\alpha}$ | -6510.68 | 13024.35 |
| $\hat{\mu}, \hat{k}, \hat{\alpha}, \hat{c}, \hat{p}$ | -6486.49 | 12977.78 |

Table 4.1: Improvement of model fit measured by $AIC$.

it is demonstrated that estimated parameters above a detection threshold $m_d$ are significantly different from those above a different threshold $m_0$. One cannot really speak of bias in the parameters because no threshold is a priori the correct one that would give unbiased parameters. Secondly, to account for undetected seismicity, it seems sufficient to let $k$ and $\mu$ adapt while fixing the other parameters. This corroborates the predictions in Chapter 3 that the (inferred) branching ratio and background rate estimates trade off. Thirdly, if $p$, $c$ and $\alpha$ can be estimated separately and fixed during the MLE, then the specific nature of the trade-off between $\mu$ and $k$ should be investigated further to understand the role of undetected earthquakes triggering observable seismicity.

Another lesson from this chapter concerns the parameter estimation problem. As seen in this chapter, the five-parameter inversions using maximum likelihood are often unstable, may not converge or converge to the wrong or unreasonable estimates, in particular when the initial values are far from the true values. The stability of the estimates should therefore always be carefully tested. These issues call for better techniques than gradient based search algorithms. As already mentioned, *Zhuang et al.* [2002] and *Veen and Schoenberg* [2007] developed new techniques that take into account the branching structure of the model. After comparisons with gradient-based methods, *Veen and Schoenberg* [2007] concluded that their expectation-maximization procedure was slightly more robust. But the estimation problem involves highly non-linear and non-Gaussian data, which make estimation procedures based on asymptotic normality and linearity less reliable. In particular, the confidence bounds on parameter estimates are usually calculated from the Hessian, which assumes a linear and Gaussian problem and is accurate only asymptotically. Therefore, the development of Bayesian or simulation-based approaches to the parameter estimation and confidence limit estimation problem need to be developed. *Werner and Jackson* [2007] have taken first steps towards a simulation-based approach.

### 4.4.3 Discussion of Preliminary Results of Schoenberg, Chu and Veen (2007)

*Schoenberg et al.* [2007] are currently systematically investigating the mean bias in parameter estimates by simulating a space-time ETAS model above a cut-off $m_0$ and then re-inverting for the seven parameters of their model [*Ogata*, 1998] using the expectation-maximization algorithm [*Veen and Schoenberg*, 2007] above a higher threshold $m_d > m_0$, analogous to the results above. This procedure was repeated over a thousand times (simulated catalogs) to obtain a mean bias. An interesting feature of their results is that the parameter bias does not monotonically increase with $m_d$ for some parameters, although all parameters show bias. This may indicate the flexibility of the model to offset misfits by adjusting other parameters. Further simulations will give more definite answers.

## 4.5 Vere-Jones' Self-Similar Branching Model: An Epidemic-Type Model Without a Lower Threshold

The theory of point processes is intimately interwoven with the theory of random measures. Loosely stated, a random measure is a generalization from a point process (integer) counting measure to non-integer nonnegative counting measures [*Daley and Vere-Jones*, 2003]. Until recently, known examples of self-similar random measures (defined by the invariance of their distributions under renormalization group transformations) were limited to the class of completely random stable measures. The self-similar random measure can be constructed by summing the (self-similar) marks of a Poisson process, whence the name completely random. *Vere-Jones* [2005] extended the class of self-similar random measures by showing that the restriction to Poisson processes was not necessary. He derived conditions for the point process so as to satisfy existence, stationarity and self-similarity of the associated random measure. In

particular, he showed that if the point process is biscale-invariant (invariant under a transformation involving time and marks), then the associated random measure is self-similar. As an example, he constructed a biscale-invariant version of the ETAS model:

$$\lambda^{\dagger}(t,m) = \beta e^{-\beta m}\left(\nu + \eta \sum_{i|t_i<t} e^{-\beta m_i - \delta|m-m_i|}\left[1+\frac{t-t_i}{\sigma e^{\beta m}}\right]^{-(1+p)}\right) \tag{4.19}$$

where the quantities $p, \nu, \eta, \sigma, \beta$ and $\delta$ are positive constants that constitute the parameters of the model. Vere-Jones proved that the process exists (is stationary) and has finite first moment under a sub-criticality condition $\rho < 1$, where $\rho$ depends on the parameters of the model. If $\rho < 1$, the number of cluster members from a given ancestor is infinite in total, but most have very small marks, so that only a finite number fall into a bounded set when the space of marks is bounded away from $-\infty$ (bounded away from zero energy, which is proportional to the exponential of the magnitude).

The model (4.19) is completely self-similar in the sense that all distributions of its associated random measure are completely self-similar (invariant under the group of biscale transformations). In less technical language, the model necessarily removes the existence of a lower cut-off $m_0$, which would otherwise introduce a minimum step size in the associated random measure and therefore break self-similarity. In order to eliminate $m_0$, a number of changes had to be made to the traditional ETAS model: (i) the distribution of the daughter magnitudes had to be made dependent on the mother magnitude through $\exp(-\delta|m-m_i|)$ so that each generation does not obey the pure GR law, although their superposition may come close (see below); (ii) $\alpha$ equals $\beta$; (iii) the Omori parameter $c$ depends on the magnitude of the daughter $c \rightarrow \sigma \exp(\beta m)$.

Are these conditions consistent with observations? Condition (ii) seems to have been validated by *Felzer et al.* [2004] and *Helmstetter et al.* [2005a] amongst others. Condition (i) was studied by *Saichev and Sornette* [2005b] and is the subject of the next paragraph. Condition (iii) involving a dependence of the Omori law parameter

$c \rightarrow c(m)$ on mainshock magnitude is indeed well-observed [*Kagan*, 2004]. However, there is debate whether this variation has a physical meaning [*Shcherbakov et al.*, 2004] or is due solely to detection limitations after large events [*Kagan*, 2004; *Kagan and Houston*, 2005].

*Saichev and Sornette* [2005b] investigated the number statistics and magnitude distributions of the Vere-Jones model. They studied the conditions and nature of the sub-critical, critical and super-critical regimes in this model and a generalized version which includes the Vere-Jones model. As can be seen from the definition of the Vere-Jones model (4.19), the magnitude-frequency distributions of first generation aftershocks triggered by a mother of magnitude $m$ have two branches: for aftershock magnitude $m' < m$, the GR exponent is $\beta - \delta$, while it is $\beta + \delta$ for $m' > m$. *Saichev and Sornette* [2005b] showed that, accounting for the contributions of all generations of triggered events, this distribution is renormalized into another two-sided law: for aftershock magnitudes $m' < m$, the renormalized exponent is $\beta - h$, while it is $\beta + h$ for $m' > m$, where $h$ depends on model parameters and its criticality regime and $0 \leq h \leq \delta$. The authors suggested testing for different magnitude distributions within generations by adapting the stochastic declustering method [*Zhuang et al.*, 2002, 2004] to the Vere-Jones model.

Remarkably, *Saichev and Sornette* [2005b] also found that the distribution of magnitudes over a stationary catalog is a pure GR law and solely depends on the distribution of the background sources, which may, in general, be different be different from the exponent $\beta$ involved in the triggered events. The explanation for this result lies in the effect of the cascading generations of aftershock distributions, each conditioned on their mother magnitude. They conclude that for a significant part of the parameter space, the predicted magnitude distribution above a detection threshold may be compatible with observations.

Simulation and estimation algorithms for the Vere-Jones model are not yet available. Likelihood estimation in particular seems problematic as the conditional intensity depends on unobserved marks, thereby creating the same problems as discussed in this and the previous two chapters. *Vere-Jones* [2005] speculates that a self-similarity property with respect to the inter-event statistics (time and distances between events), instead of with respect to the underlying space as in his model, may open up routes to simulation algorithms. If and when these algorithms become available, the study of this completely self-similar branching model may be a strong alternative to the ETAS model, because it can be defined without a magnitude cut-off $m_0$.

## 4.6 Conclusions

1. Clustering parameters and background rates are difficult to interpret in terms of geophysical quantities. Rather, they are intrinsically dependent on the detection threshold above which they were estimated. Comparing estimates from different thresholds is highly non-trivial.

2. Declustering using the ETAS model is strongly dependent on the detection threshold. Results are expected to change drastically under a change of the detection threshold, making the interpretation and usefulness of declustering methods doubtful. Studying the equivalence of temporal and magnitude branching processes may provide two consistent models which can be used for prediction and declustering separately.

3. The extension of the results of Chapter 3 to full time-independent distributions by *Saichev and Sornette* [2006a] recover the results of Chapter 3 up to a small correction. However, *Saichev and Sornette* [2006a] also show that the spatio-temporal properties cannot be mapped onto a renormalized model.

4. The conditional intensity is not strictly invariant under a change of detection threshold due to time-dependent contributions from undetected earthquakes to observable seismicity. I calculated the average contribution and illustrated their role by calculating their fraction of the intensity in a simulated catalog. It is likely that small earthquakes contribute strongly to the observed seismicity budget.

5. The parameters of the ETAS model are sensitively dependent on the detection threshold. I have shown that the effect of unobserved earthquakes on observable seismicity must be more than just add a time-independent constant, as would be expected from thinning. However, based on the results presented here, it is likely that the background rate and the branching ratio together can account for the influence of unobserved mothers, trading off as predicted in Chapter 3. This relationship should be investigated in more detail. However, if all five parameters are estimated simultaneously, then the bias can be very large.

6. The current maximum likelihood estimation method is not wholly satisfactory as it may not converge, or it may be trapped in local minima with unrealistic parameter values. Furthermore, confidence intervals are often based on the Hessian, which is correct only asymptotically. Therefore, the development of robust methods that also allow accurate confidence limit estimates are important. This is especially true given the increasing popularity of the ETAS model as a null hypothesis and its potential use for detecting differences in triggering properties in different tectonic regions of the world (see, e.g. [*Bird et al.*, 2007]).

7. The Vere-Jones model [*Vere-Jones*, 2005] provides an attractive alternative as a model that does not need a detection threshold. However, the development of simulation and estimation algorithms may not be trivial. But the Vere-Jones model, which is renormalizable, provides the first wholly self-similar point pro-

cess model. This development suggests that the statistical seismology community and the part of statistical physics community working on seismicity may begin to coalesce in their conceptual approaches. Indeed, the topic of the next chapter is an approach to seismicity data analysis inspired by the theory of critical and renormalizable phenomena in statistical physics.

# CHAPTER 5

# Hypothesis Testing of a Proposed "Universal" Scaling Law in Earthquake Recurrence Statistics *

## 5.1 Introduction

Over the last five years, a group of statistical physicists, interested in earthquakes as a potential instance of Self-Organized Criticality (SOC), have claimed "novel", "universal" and "robust" scaling laws from their analysis of the spatio-temporal organization of seismicity. The authors purport to have discovered universal and hitherto unknown features of earthquakes that give new insights into the dynamics of earthquakes and add to the evidence that earthquakes are self-organized critical. This chapter focuses on one of these recent studies to add to the growing recognition that some or perhaps most of these "novel scaling laws" can be explained entirely by known statistical seismicity laws.

Much of the recent interest of the statistical physics community focused on applying scaling techniques, which are common tools in the study of critical phenomena, to the statistics of inter-event recurrence times or waiting times [*Bak et al.*, 2002; *Corral*, 2003, 2004a,b, 2005a,b; *Corral and Christensen*, 2006; *Davidsen and Goltz*, 2004; *Livina et al.*, 2006]. The lively debate over the relevance of critical phenomena to earthquakes stretches back as far as 30 years [*Vere-Jones*, 1977; *Allègre et al.*, 1982;

*An edited version of section 5.3.3 of this chapter was published in Werner, M. J. and D. Sornette, *Phys. Rev. Lett.*, 99, 179801, 2007. Copyright (2007) by the American Physical Society.

*Smalley et al.*, 1985; *Kagan*, 1989; *Sornette and Sornette*, 1989; *Bak and Tang*, 1989; *Sornette and Sornette*, 1999a; *Kagan*, 1992; *Olam et al.*, 1992; *Kagan*, 1994; *Sornette and Sammis*, 1995; *Saleur et al.*, 1996a; *Bak*, 1996; *Jensen*, 1998; *Nature Debates*, 1999; *Hergarten*, 2002; *Sornette*, 2002, 2004; *Kagan*, 2006]. The current debate on recurrence statistics is the latest tack in the evolving string of arguments. As discussed below, many of the claims made in the recent articles on recurrence statistics have either been challenged, refuted or explained by previously known facts about earthquake statistics [*Lindman et al.*, 2005, 2006; *Molchan*, 2005; *Saichev and Sornette*, 2006b, 2007]. This chapter focuses on the work by *Davidsen and Paczuski* [2005], who extended the analysis from recurrence times to the spatial dimension and to the statistics of the spatial distances between successive earthquakes. As will be discussed below, this debate in the literature is important because of the potential consequences for understanding earthquakes, but it needs to be pursued with rigorous scientific arguments accessible to both the seismological and the statistical physics communities.

*Davidsen and Paczuski* [2005] (hereafter DP) claimed that (i) the probability density function of the spatial distances between successive earthquakes in southern California obeys finite size scaling with a power law scaling function, (ii) the associated critical exponent $\delta \simeq 0.6$ is a novel dynamical scaling exponent that characterizes the critical behavior of seismicity, (iii) the scale-free statistics of the distribution contradict the theory of aftershock zone scaling, and (iv) their results are consistent with SOC. Their results are a priori interesting for a variety of reasons. First, they present a new turn in the debate on the relevance of non-equilibrium statistical mechanics to earthquake seismology. Second, they contradict the idea that main shock rupture lengths strongly influence the spatial distribution of aftershocks. These claims will be analyzed in detail in this chapter.

We will show that (i) if there truly were a dynamical exponent, it would have to

equal 1, based on the recent results by *Christensen et al.* [2007]; but (ii) the power law scaling function breaks down in other regions of the world; (iii) the results obtained by DP for southern California depend crucially on a single earthquake (the June 28, 1992, M7.3 Landers earthquake): Without Landers and its aftershocks, the power law disappears; (iv) a model of clustered seismicity, with aftershock zone scaling explicitly built in, is able reproduce the apparent power law, indicating that an apparent lack of scales in the data does not necessarily contradict aftershock zone scaling and the existence of scales associated with mainshock rupture length scales.

The last point, in particular, underlines the strength of testing hypotheses with simulations to establish null hypotheses and benchmarks: seismicity patterns are sufficiently stochastic and earthquake catalogs contain a sufficient amount of observational uncertainties so as to make inference difficult. It is often not straightforward to predict the signal of well-known statistical features such as clustering in new data analysis techniques. Therefore, testing the purported claims by realistic simulations of earthquake catalogs can provide a strong benchmark against which the claims can be evaluated. This view and its criticism of many studies has been put forward and defended for a long time by Y. Kagan [*Kagan*, 1999b].

Such a model-dependent approach may be at odds with the philosophy of a so-called ”model-free” analysis, which the community of statistical physicists claim to take in their analysis. For instance, network theory-based approaches, space-time window-based finite size scaling, box-covering methods and other techniques used in the study of critical and fractal phenomena are said to be ”model-free” because no assumptions about seismicity are supposedly made at the outset. By using model-free analysis techniques, the often uncertain and sometimes clearly wrong assumptions of flawed models and resulting biased results are meant to be circumvented.

However, as is almost always the case in statistical hypothesis testing, the less

assumptions are made about the test, the less powerful the test statistic. More importantly, seismicity is sufficiently stochastic so that well-known features may appear as novel in new analysis methods. Furthermore, to convince the seismological community of new data analysis techniques, the methods need to be tested on established knowledge and show the improvement over traditional methods. These types of initial tests are rarely performed by the statistical physics community. This chapter provides an example of how useful such tests can be.

The chapter is organized as follows. In section 5.2, I attempt to convey the strong appeal that motivates the application of methods from statistical mechanics and critical phenomena to earthquake seismology. To understand this appeal and its potential, I will briefly review elements of the theory of critical phenomena, such as phase transitions, finite size scaling and SOC, which will become important in section 5.3. After this short introduction, I will summarize the recent work on recurrence time statistics and SOC, which sets the stage for the main topic of this chapter in section 5.3: testing the hypotheses and claims by *Davidsen and Paczuski* [2005] as already stated above.

## 5.2 Critical Phenomena, Self-Organized Criticality, and Earthquakes

A comprehensive review of any of these topics is outside the scope of this thesis. I shall only introduce the minimum necessary for understanding the motivation for the research into the proposed link between earthquakes and SOC. The technical language of finite size scaling, in particular, will be useful for section 5.3. More comprehensive, excellent introductions on these topics can be found in many texts, for instance in [*Cardy*, 1996; *Jensen*, 1998; *Landau and Lifshitz*, 1980; *Sornette*, 2004; *Stanley*, 1999; *Yeomans*, 1992]. *Hergarten* [2002] and *Sornette* [2004] focus on applying these concepts to earth sciences.

- **Phase transitions**: In (equilibrium) statistical mechanics, a phase transition occurs when there is a singularity in the free energy or one of its derivatives. Examples include the freezing of water, the transition from ferromagnetic to paramagnetic behavior in magnets, and the transition from a normal conductor to a superconductor [*Landau and Lifshitz*, 1980; *Yeomans*, 1992].

- **Continuous phase transitions**: If there is a finite discontinuity in the first derivative of the thermodynamic potentia, then the phase transition is termed first-order. During such a transition, a system either absorbs or releases a fixed amount of latent heat (e.g. the freezing/melting of water/ice). If the first derivative is continuous but higher derivatives are discontinuous or infinite, then the phase transition is called continuous, of the second kind, or critical. Examples include the critical point of the liquid-gas transition, the Curie point of the ferromagnetic transition, or the superfluid transition [*Landau and Lifshitz*, 1980; *Yeomans*, 1992].

- **Critical phenomena**: Phenomena observed in systems that undergo a continuous phase transition. They are characterized by scale invariance: the statistical properties of a system on one scale are related to those at another scale only through the ratio of the two scales, not on the scales themselves. The scale invariance is a result of fluctuations and correlations at all scales, all of which are important and in no way diminished [*Cardy*, 1996; *Sornette*, 2004; *Yeomans*, 1992].

- **Critical exponents**: Near the critical point, various thermodynamic quantities diverge as power laws with associated critical exponents. In equilibrium systems, there are scaling relations that connect some of the critical exponents of different thermodynamic quantities [*Cardy*, 1996; *Landau and Lifshitz*, 1980; *Sornette*, 2004; *Stanley*, 1999; *Yeomans*, 1992].

- **Universality**: In systems with little or no frozen disorder, equilibrium continuous phase transitions fall into a small set of universality classes that are characterized by the same critical exponents and certain scaling functions become identical near the critical point. The class depends only on the dimension of the space and the dimension of the order parameter. For instance, the critical point of the liquid-gas transition falls into the same universality class as the 3D Ising model. Even phase transitions occurring in high-energy physics are expected to belong to the Ising class. Universality justifies the development and study of extremely simplified models (caricatures) of Nature, since the behavior of the system at the critical point can nevertheless be captured (in some cases exactly). Non-universal features remain even at the critical point but are less important, e.g. amplitudes of fluctuations or system-specific corrections to the scaling that appear to sub-leading order [*Cardy*, 1996; *Stanley*, 1999; *Yeomans*, 1992; *Zee*, 2003].

- **Renormalization group theory**: A mathematical theory built on the idea that the critical point can be mapped onto a fixed point of a suitably chosen transformation on the system's Hamiltonian. It provides a foundation for understanding scaling and universality and provides tools for calculating exponents and scaling functions. Renormalization group theory provides the basis for our understanding of critical phenomena [*Cardy*, 1996; *Stanley*, 1999; *Yeomans*, 1992].

- **The thermodynamic limit and criticality:** The thermodynamic limit corresponds to an infinite system size or an infinite number of particles, such that the correlation length is truly infinite at the critical point and not cut short by the finite system size. Any finite physical system therefore cannot be truly critical. However, one may pretend that the system is critical and apply, in this thermodynamic limit, renormalization group techniques to calculate exponents and

obtain scaling functions, and then apply finite size corrections that are system specific [*Cardy*, 1996; *Pruessner*, 2004; *Yeomans*, 1992]. The concept of the thermodynamic limit will be used in section 5.3.

- **Finite size scaling**: If a thermodynamic or other quantity is investigated at the critical point under a change of the system size, the scaling behavior of the quantity with respect to the system size is known as finite size scaling [*Cardy*, 1996]. The quantity may refer to a thermodynamic quantity such as the free energy or it may refer to an entire probability distribution function. At criticality, the sole length scale in a finite system is the upper cut-off $s_c$, which diverges in the thermodynamic limit $L \to \infty$. Assuming a lower cut-off $s_0 \ll s_c, s$, a finite size scaling ansatz for the distribution $P(s; s_c)$ of the observable variable $s$, which depends on the upper cut-off $s_c$ is then given by:

$$P(s; s_c) = as^{-\tau} G(s/s_c) \qquad \text{for} s, s_c \gg s_0 \tag{5.1}$$

  where the parameter $a$ is a non-universal metric factor, $\tau$ is a universal (critical) exponent, and $G$ is a universal scaling function that decays sufficiently fast for $s \gg s_c$ [*Cardy*, 1996; *Christensen et al.*, 2007]. *Pruessner* [2004] provides a simple yet instructive and concise introduction to scaling theory and finding associated exponents. As already stated above, system-specific corrections appear to sub-leading order. Finite size scaling will play an important role in section 5.3.

- **Non-equilibrium phase transitions**: In contrast to systems at equilibrium, non-equilibrium phase transitions involve dynamics, energy input and dissipation. Detailed balance is violated and no known equivalent of the partition function exists, from which all thermodynamic quantities of interest derive in equilibrium. Examples of non-equilibrium phase transitions include absorbing state

phase transitions, reaction-diffusion models, morphological transitions of growing surfaces, and percolation in porous media [*Hinrichsen*, 2000; *Luebeck*, 2004].

- **Dynamical scaling and exponents**: Non-equilibrium critical phase transitions are also characterized by scale invariance, scaling functions and critical exponents. Furthermore, some evidence supports the claim that universality classes also exist for non-equilibrium phase transitions (e.g. the directed percolation and the Manna universality class), although a complete classification of classes is lacking and may in fact not exist at all. Much interest has recently focused on directed percolation, which, as the most common universality class of absorbing state phase transitions, is expected to occur in many physical, chemical and biological systems [*Hinrichsen*, 2000; *Luebeck*, 2004; *Sornette*, 2004].

- **Self-Organized Criticality (SOC)**: Despite almost two decades of research since its inception by *Bak et al.* [1987] and the ambitious claim by *Bak* [1996] that, as a mechanism for the ubiquitous power laws in Nature, SOC was "How Nature Works", a commonly accepted definition along with necessary and sufficient conditions for SOC is still lacking [*Jensen*, 1998; *Pruessner*, 2004; *Sornette*, 2004]. A less rigorous definition may be the following: Self-organized criticality refers to a non-equilibrium, critical and marginally stable steady-state, which is attained spontaneously and without (explicit) tuning of parameters. It is characterized by power law event distributions and fractal geometry and may be expected in slowly driven, interaction-dominated threshold systems [*Jensen*, 1998]. Some authors additionally require that temporal and/or spatial correlations decay algebraically (e.g. [*Hergarten*, 2002], but see *Pruessner* [2004]). Definitions in the literature range from broad (simply the absence of characteristic length scales in non-equilibrium systems) to narrow (the criticality is due to an underlying continuous phase transition with all of its expected properties)

(see, e.g., *Peters and Neelin* [2006] for evidence that precipitation is an instance of the latter definition of SOC in which a non-linear feedback of the order parameter on the control parameter turns a critical phase transition into a self-organized one attracting the dynamics [*Sornette*, 1992]).

- **Other mechanisms for power laws**: Power laws may be the hallmark of critical phenomena, but there are a host of other mechanisms that can lead to power laws (see Chapter 14 of *Sornette* [2004] for a list of power law mechanisms). Observations of scale invariant statistics therefore do not necessarily imply SOC, of course.

The abundance of power laws in earthquake seismology ignited a keen interest amongst statistical physicists to explain the spatio-temporal organization of earthquakes by the theory of critical phenomena and, in particular, by SOC. *Sornette and Sornette* [1989] and *Bak and Tang* [1989] first suggested that SOC may be relevant for earthquakes. The appeal of placing the study of earthquakes in the framework of critical phenomena may perhaps be summarized as follows. Power law distributions can be understood as a result of an underlying continuous phase transition into which the crust has organized itself. Applying the methods of renormalization group theory may help calculate exponents and scaling functions and rationalize the spatio-temporal organization of seismicity along with its highly correlated structures. Perhaps earthquakes fall into a universality class which can be solved exactly and/or investigated in toy models. Moreover, studying the detailed and highly complicated microphysics involved in earthquakes may not lead to insights about the spatio-temporal organization, because, as a critical phenomenon, the traditional approach of separating length scales to describe systems is inadequate. As another reason for the importance of the topic, interesting consequences for the predictability of earthquakes might be derived, for instance by mapping earthquakes to a genuine critical point (the accelerating moment

release hypothesis, e.g. [*Sornette and Sammis*, 1995; *Sornette*, 2002]) or by mapping earthquakes to SOC (e.g. [*Geller et al.*, 1997a; *Nature Debates*, 1999]).

It should be noted at this point that the statistical physics approach to earthquake science is not limited to SOC. Various mechanisms drawn conceptually from statistical mechanics but not necessary even limited to critical (phase transition) phenomena have been proposed and are being pursued. Such approaches include the concept of the critical point earthquake related to accelerated moment release, network theory, percolation and fiber models as models for fracture, and many more, some of which can be found in [*Hergarten*, 2002; *Sornette*, 2004; *Turcotte*, 1997; *Turcotte et al.*, 2000].

A slightly different approach has been favored by Yan Kagan [*Kagan*, 1989, 1992, 1994, 2006], who described seismicity as the turbulence of solids - attesting to the potentially far more complex problems that need to be solved than SOC promises to deliver, although renormalization group methods and scaling theory have contributed immensely to the study of turbulence [*Frisch*, 1995].

We are not going to settle the issue here, of course. However, in our opinion, this discussion is highly relevant to earthquakes and therefore needs to be pursued with rigorous scientific arguments that are accessible to both the seismological and the statistical physics community. The present chapter aims to contribute to this discussion by testing some of the claims made by *Davidsen and Paczuski* [2005] that the probability density function of spatial distances between successive earthquakes obeys finite size scaling with a novel dynamical scaling exponent, suggesting that earthquakes are self-organized critical.

### 5.2.1 "Unified" Scaling Laws of Recurrence Times of Earthquakes

*Davidsen and Paczuski* [2005] followed a recent interest in inter-event statistics, started by *Bak et al.* [2002], who analyzed the scaling of the probability density function of

waiting times between successive earthquakes in southern California as a function of "box size" or small regions in which subsequent earthquakes are considered. They found an approximate collapse of the pdfs for different magnitude thresholds $S$ and box sizes $L$ which suggested the following scaling ansatz for the waiting times $T$:

$$T^{\alpha} P_{S,L}(T) = f(TS^{-b}L^{d_f}) \tag{5.2}$$

where $b = 1$ is the Gutenberg-Richter exponent, $d_f \simeq 1.2$ was claimed to be a spatial fractal dimension of seismicity (see *Molchan and Kronrod* [2005] and *Kagan* [2007] for more in-depth studies), $\alpha = 1$ was identified as the exponent in the Omori law and $f(\cdot)$ is a scaling function which was proposed to be roughly constant up to a constant ("kink") beyond which it quickly decays. The scaling (5.2) was claimed to be a unified law for earthquakes that revealed a novel feature in the spatio-temporal organization of seismicity in that the Gutenberg-Richter, the Omori law and the spatial distribution of earthquakes were unified into a single picture that made no distinction between fore-, main- and aftershocks. The scaling relations and critical exponents were claimed to be contained in the scaling ansatz. *Corral* [2003, 2004a,b, 2005a] and others broadened the analysis to other regions of the world. *Corral* [2004a] proposed a slightly different scaling ansatz for a modified data analysis.

Early criticism came from *Lindman et al.* [2005], who noted that generating synthetic data using a non-homogeneous Poisson process derived from Omori's law was able to reproduce some of the results of *Bak et al.* [2002], indicating a rather trivial origin of the unified scaling law. *Molchan* [2005] showed that, if at least two regions in the data set are independent, then, if a scaling relation were to hold exactly, this scaling function could only be exponential. All other functions could only result in approximate data collapses. Proponents of the unified scaling law, e.g. *Corral* [2005b], argued that indeed all regions were correlated, as expected by systems near a critical point so that the assumption of independence between different regions should not hold. But

Molchan also showed that a simple Poisson cluster model (Poissonian mainshocks that trigger Omori-type aftershock sequences) could reproduce the short and long time limits of the observed statistics, indicating that the Omori law, the Gutenberg-Richter relationship and simple clustering were the sole ingredients necessary for the observed short and long time limit, and no spatial correlation was needed.

*Saichev and Sornette* [2006b, 2007] extended Molchan's arguments to show that the approximate data collapse of the waiting times could be explained completely by the Epidemic-Type Aftershock Sequence (ETAS) model of *Ogata* [1988]. This provided further evidence that the apparent data collapse was only approximate. Remarkably, the theoretical predictions of the ETAS model seem to fit the observed data better than the scaling function postulated by *Corral* [2004a]. *Saichev and Sornette* [2006b, 2007] thus showed that a benchmark model of seismicity was able to reproduce the apparent unified scaling law and that therefore the distribution of interevent times did not reveal new information beyond what was already known via statistical laws: The combination of the Gutenberg-Richter law, the Omori law, and the concept of clustering suffice to explain the apparent "universal" scaling of the waiting times.

In the next section, we will extend this debate to the spatial dimension where similar claims of universal scaling functions and novel features of the organization of seismicity have been made [*Davidsen and Paczuski*, 2005; *Davidsen et al.*, 2006; *Corral*, 2006].

## 5.3 On The Spatial Distances Between Successive Earthquakes

The search for robust features in seismicity that are independent of the model assumptions and space-time window boundaries (not unlike the problem of the detection threshold and its influence on the parameters in chapters 2, 3 and 4) is motivated by

the concepts introduced in the last section. As an analogy, the length of the coast of Britain depends on the size of the ruler, but the exponent (fractal dimension) which relates the coastal length to the size of the ruler is independent of the resolution. The approach of *Davidsen and Paczuski* [2005] followed in this line of thought by asking whether the distance to the next earthquake depends solely on the observer's resolution (i.e. the spatial window).

This section is organized as follows. In section 5.3.1, we review their results and conclusions. Next, we comment on their finite size scaling ansatz and the proposed scaling exponent in section 5.3.2. In section 5.3.3, we criticize some of their conclusions regarding the "universal" scaling law and the purported evidence contradicting the theory of aftershock zone scaling. Finally, we summarize DP's reply [*Davidsen and Paczuski*, 2007] to our comment [*Werner and Sornette*, 2007a] before concluding the chapter.

### 5.3.1 Results by Davidsen and Paczuski (2005) (DP): Scale-Free Distribution Contradicting Aftershock Zone Scaling Consistent with SOC

*Davidsen and Paczuski* [2005] (hereafter DP) analyzed the spatial distances between subsequent earthquake epicenters in southern California. Their data set included $23,374$ earthquakes above $m_d = 2.4$ from January 1984 to December 2000 in the region $(120.5^oW, 115.0^oW) \times (32.5^oN, 36.0^oN)$. Their major results can be summarized as follows:

- DP found that spatial distances between subsequent earthquakes in southern California exhibit scale-free (power-law) statistics with a critical exponent $\delta \simeq 0.6$ in the range 2 km to $\sim$ 500 km. Their results are shown in Figure 5.1.
- The probability density function (pdf) $p_{m_d,L}(\Delta r)$ of the spatial distances $\Delta r =$

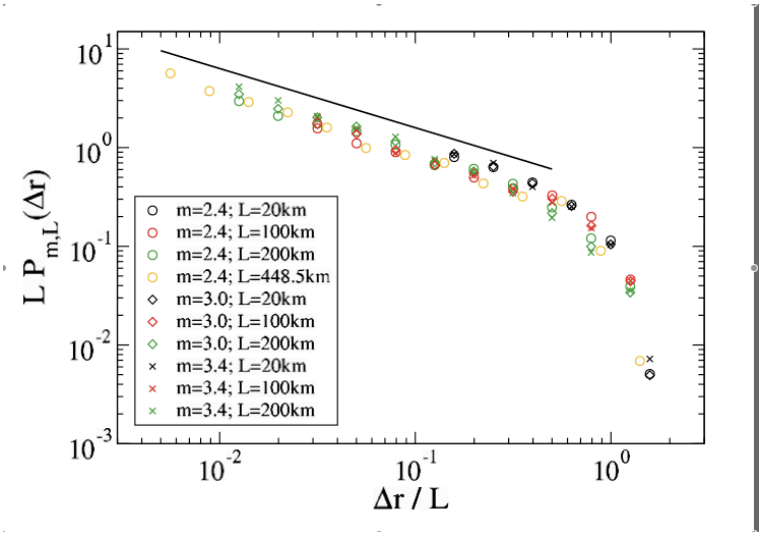


Figure 5.1: The probability density function $P_{m,L}(\Delta r)$ of the spatial distances between successive earthquakes in southern California as a function of the scaled variable $\Delta r/L$ for various magnitude thresholds and linear box sizes $L$. This figure was copied from *Davidsen and Paczuski* [2005] (their Figure 1).

$|\mathbf{r}_{i+1} - \mathbf{r}_i|$ between successive earthquakes obeys finite size scaling (FSS) under a change of the linear extent $L$ of the box in which earthquakes are studied. The proposed scaling ansatz is:

$$p_{m_d,L}(\Delta r) = \frac{f(\Delta r/L)}{L} \tag{5.3}$$

where the scaling function $f(x)$ decays as $x^{-\delta}$ with $\delta \simeq 0.6$ for $x < 0.5$. For $x > 0.5$, it decays quickly since the finite cell size requires that $f(x) = 0$ for $x > \sqrt{2}$. The subscripts $m_d = (2.4, 3.0, 3.4)$ and $L = (20\text{km}, 100\text{km}, 200\text{km}, 448.5\text{km})$ indicate the detection threshold above which the pdf was estimated and the linear size of the boxes in which subsequent earthquakes were defined. For a given $L$, the distances between successive earthquakes in each box were concatenated, i.e. the successive-event distances from different boxes were mixed together

to compute the overall pdf. They claimed that the pdf did not depend on the different thresholds $m_d$.

Based on these observations, the authors made the following conclusions:

1. DP claim that "the appearance of FSS precludes the existence of any other length scale over the range where FSS holds. Thus no physical length scale exists in the range from 20 km to $\simeq 500$ km, in contrast to the theory of aftershock zones [*Kagan*, 2002a]. According to this theory, main shocks generate aftershocks within finite aftershock zones, whose extent is comparable to the rupture length $l_r = 0.02 \times 10^{0.5m}$ km of the main event. This implies that the distance between subsequent aftershocks would be limited to the size of the largest aftershock zone, which is less than 90 km for the catalog analyzed here." Using their formula for the rupture length $l_r$, they find that the largest rupture lengths of this period are given by the M7.3 Landers and M7.1 Hector Mine earthquakes, both of which are smaller than 90 km. "However, we find no break or change in scaling behavior for larger distances extending all the way up to the size of the region considered, of the order of 500 km." They thus claim that "these observations contradict the theory of aftershock zone scaling with main shock magnitude."

2. The exponent $\delta \simeq 0.6$ is claimed to be a dynamical critical exponent characterizing the dynamical features of (self-organized) critical seismicity, rather than the static, geometrical features. The latter was recovered in the form of the correlation dimension by temporally randomizing the catalog, thereby destroying any temporal correlations. The critical exponent $\delta$ is thus purported to be a novel, dynamical exponent describing the finite size scaling of the underlying self-organized critical phenomenon.

In section 5.3.3, it will be shown that their model-free method of analysis is not

sensitive enough to sustain the claim in DP's conclusion 1 that aftershock zone scaling does not exist. But first, DP's conclusion 2 is examined in section 5.3.2.

### 5.3.2 On DP's Proposed Finite Size Scaling and the Dynamical Exponent

Assume for a moment that the pdf of distances between successive earthquakes really did obey finite size scaling with a power law scaling function. Although section 5.3.3 will cast doubt on this hypothesis, this exercise provides insight into the pitfalls of finite size scaling. The authors note that "as $\delta$ is unambiguously less than one, the distribution $p_{m_d,L}(\Delta r)$ becomes non-normalizable for large L. Extrapolating our results, the finite size of the earth may play an important role in the definition of distances between subsequent earthquakes."

As discussed in section 5.2, the thermodynamic limit is a central pillar in the study of critical phenomena. If indeed functions obey FSS and the physical system is critical, then the thermodynamic limit must be well defined. Else, there is a problem, either with the criticality hypothesis or the FSS and most likely both. In the present case, the thermodynamic limit corresponds to $L \to \infty$, i.e. the linear size of the spatial area tends to infinity. In other words, if the limit $L \to \infty$ cannot be taken, something is wrong.

It is important to distinguish between a well-defined thermodynamic limit (in a mathematical sense) and the finiteness of any physical system. Since only an infinite system can be truly critical without any finite size corrections, no physical system can ever attain that status. However, while there is a physical size of any real critical system which restricts its properties strictly to off-but-near criticality, the thermodynamic limit can (mathematically) be taken, so that the toolbox of critical phenomena can be applied in the (abstract) thermodynamic limit. In principle, renormalization group theory can be used to derive the exponent of the moments of the pdf in this abstract limit. One

can then introduce system-specific corrections to the scaling, including the finite size effects of the observed system.

A geophysical example is provided by the Gutenberg-Richter distribution, whose probability density function is a power law with exponent $1 + \mu \simeq 1 + 2/3$ after changing variables from magnitudes to moments (we use the non-standard terminology of the full exponent $1 + \mu$ here to enable a direct comparison with DP's $\delta$). The law cannot extend to infinite magnitudes because of the finite size of the Earth. But its exponent is larger than one so that, in principle, the same pdf can be normalized in the thermodynamic limit as the size of the Earth tends to infinity. However, the infinite mathematical expectation for $\mu < 1$ would require an infinite amount of energy to be available in the Earth [*Knopoff and Kagan*, 1977], which has led to the debate over whether the observed roll-off in the distribution may be due to something other than finite size effects [*Knopoff and Kagan*, 1977; *Sornette et al.*, 1996; *Kagan*, 1999a; *Pisarenko and Sornette*, 2003; *Pisarenko et al.*, 2004, 2008].

With this discussion in mind, I return to the statement by DP that the pdf becomes "non-normalizable" because the supposed scaling exponent $\delta$ is less than one. As just mentioned, if the thermodynamic limit cannot be taken or if the pdf cannot be normalized, this constitutes a very serious problem for the proposed criticality and FSS. Recognizing a potential issue, the authors state that "the finite size of the earth may play an important role in the definition of distances between subsequent earthquakes." The thermodynamic limit corresponds to an infinite-size critical "Super-Earth" on which earthquakes occur. In the following it is shown that (i) the pdf can always be normalized so that the thermodynamic limit is well-defined and there is no need to (wrongly) invoke the finite size of the Earth, and (ii) the true scaling exponent $\delta^{\dagger}$ equals one under a more rigorous scaling ansatz. The latter is an application of the results on the FSS of pdfs by *Christensen et al.* [2007].

Assume that the power law scaling function holds above a fixed minimum threshold $r_0$ (I will henceforth denote $\Delta r$ by $r$ for notational simplicity). In fact we know that the pdf must tend to zero for $r \to 0$ because the probability of a subsequent earthquake occurring at the same exact point vanishes (assuming a point source nature of earthquakes). For any $L > r_0$, the normalization of the unnormalized pdf is given by the zeroth moment:

$$\langle r^0 \rangle_L = \int_{-\infty}^{+\infty} p_L(r) dr \tag{5.4}$$

where $p_L(r) = f(r/L)/L$ with $f(x) \sim x^{-\delta}$ for 2 km $< r < 0.5L$ km, as reported by DP and given by expression (5.3). I will show that $\lim_{L\to\infty} \langle r^0 \rangle_L$ exists and is finite so that the pdf never becomes "non-normalizable."

The normalization can be performed in three parts:

$$\langle r^0 \rangle_L = \int_0^{r_0} p_L(r) dr + \int_{r_0}^{L/2} p_L(r) dr + \int_{L/2}^{\infty} p_L(r) dr \tag{5.5}$$

We have already argued that $p_L(r)$ decays to zero for $r \to 0$ so that the first integral is always a finite constant, say $c_1$. Furthermore, the third integral is taken over the range in which the scaling function $f(x)$ is quickly decaying until $x = \sqrt{2}L$, after which the pdf is zero. Therefore, the third integral is also a finite constant, say $c_3$, unless $L = \infty$, in which case the third integral is zero because the entire support of the pdf is taken up by the first and second integral. The second integral can be evaluated as follows (where $A$ is a constant of proportionality that depends on $L$ and $\delta$ but not $r$):

$$\begin{aligned} \int_{r_0}^{L/2} p_L(r) dr \quad &= \int_{r_0}^{L/2} AL^{-1} f(\frac{r}{L}) dr = \int_{r_0}^{L/2} AL^{-1} \left(\frac{r}{L}\right)^{-\delta} dr \\ &= AL^{\delta-1} \left[\frac{r^{1-\delta}}{1-\delta}\right]_{r_0}^{L/2} \qquad , \delta \neq 1 \\ &= \frac{A}{1-\delta} \left(1/2^{1-\delta} - \left(\frac{r_0}{L}\right)^{1-\delta}\right) \end{aligned} \tag{5.6}$$

which is positive since $r_0 < L/2$ and $\delta < 1$.

Taking the limit $L \to \infty$ of the three integrals gives the following result:

$$\begin{aligned}
\lim_{L\to\infty} \langle r^0 \rangle_L &= \lim_{L\to\infty} c_1 + \lim_{L\to\infty} \left[ \frac{A}{1-\delta} \left( 1/2^{1-\delta} - \left( \frac{r_0}{L} \right)^{1-\delta} \right) \right] + \lim_{L\to\infty} c_3 \\
&= c_1 + \frac{1/2^{1-\delta}}{1-\delta} \cdot \lim_{L\to\infty} A + \frac{r_0^{1-\delta}}{1-\delta} \cdot \lim_{L\to\infty} \left( \frac{A}{L^{1-\delta}} \right) + 0 \\
&= c_1 + c' \lim_{L\to\infty} A + c'' \lim_{L\to\infty} \left( \frac{A}{L^{1-\delta}} \right) \qquad (5.7)
\end{aligned}$$

where $c'$ and $c''$ are finite constants. Expression (5.7) shows that as long as $\lim_{L\to\infty}(A) < \infty$, the normalization constant can be calculated. Therefore, the pdf in expression (5.3) can always be normalized, even in the thermodynamic limit. Hence, the statement by DP that the pdf becomes "non-normalizable" for large $L$ is wrong and there is no need to wrongly invoke the finite size of the Earth.

Let us briefly consider a more rigorous scaling ansatz, which further clarifies the the perceived problem of the non-normalizability and demonstrates that the apparent scaling exponent $\delta < 1$ should be absorbed into the scaling function, while the true scaling exponent $\delta^\dagger = 1$. This presentation follows *Christensen et al.* [2007].

We begin by repeating the finite size scaling ansatz for probability density functions already introduced in section 5.2:

$$P(s; s_c) = a s^{-\tau} G(s/s_c) \qquad \text{for} \quad s, s_c \gg s_0 \qquad (5.8)$$

where the parameter $a$ is a non-universal metric factor, $\tau$ is a universal (critical) exponent, and $G$ is a universal scaling function that decays sufficiently fast for $s \gg s_c$. In the regime $s_0 \ll s \ll s_c$, the pdf shows power law behavior:

$$P(s; s_c) \propto s^{-\tilde{\tau}} \qquad (5.9)$$

where the apparent exponent $\tilde{\tau}$ is the slope of the fit to the data when plotting $\log P(s; s_c)$ versus $\log s$, as is the case in DP's work (see Figure 5.1). In the informal expression (5.9), the scaling function is missing, so that to derive (5.9), one needs to assume that

the scaling function is roughly constant in this regime. If that is indeed the case, then the apparent exponent equals the true exponent $\tilde{\tau} = \tau$. But if the scaling function is itself a power law in the scaling regime, then $\tilde{\tau} \neq \tau$. To illustrate, one can define a "cut-off" function $\tilde{G}$, which can be a Heaviside step function $H(1 - s/s_c)$ which is constant up to $s_c$ and zero for $s > s_c$, and assume that the scaling function is of the form $G(x) = x^{\alpha}\tilde{G}(x)$. Hence, the pdf behaves as $as^{-\tau}(s/s_c)^{\alpha}$ in the scaling regime, so that the actual scaling exponent is given by $\tau = \tilde{\tau} + \alpha$.

*Christensen et al.* [2007] demonstrate that, if the apparent exponent $\tilde{\tau}$ is less than one, then the scaling exponent $\tau = 1$ exactly. This can be seen from the following. Assume the pdf has the form:

$$P(s; s_c) = \tilde{a}s^{-\tilde{\tau}}\tilde{G}(s/s_c) \qquad \text{for} s, s_c \gg s_0 \tag{5.10}$$

which corresponds exactly to the ansatz by DP. Naively, one would equate $\tilde{\tau} = \tau$, but for $\tilde{\tau} < 1$, this is incorrect. The unknown prefactor $\tilde{a}$ can be derived from the normalization condition:

$$\int_{s_0}^{s_c} \tilde{a}s^{-\tilde{\tau}}ds = \frac{\tilde{a}}{1-\tilde{\tau}}(s_c^{1-\tilde{\tau}} - s_0^{1-\tilde{\tau}}) \equiv 1 \qquad \text{for} \quad \tilde{\tau} \neq 1 \tag{5.11}$$

Note that if $\tilde{\tau} < 1$, then $\tilde{a} \to 0$ as $s_c \to \infty$. By comparison with the normalization constant $A$ above in expressions (5.6) and (5.7), we can identify $\tilde{a} = AL^{-1}$, which indeed decays to zero as $L \to \infty$. This provides another way to show that the pdf form proposed by DP in (5.3) can always be normalized.

Now substitute (5.11) into (5.10), and simplify for $\tilde{\tau} < 1$ to obtain:

$$P(s; s_c) = (1 - \tilde{\tau})s^{-1}(s/s_c)^{-\tilde{\tau}}\tilde{G}(s/s_c) \tag{5.12}$$

By comparison with the original scaling ansatz (5.8), we can identify $a = (1 - \tilde{\tau})$, $\tau = 1$ and $G = (s/s_c)^{-\tilde{\tau}}\tilde{G}$. We can then make the identifications $\delta = \tilde{\tau}$ and $G(\cdot) =$

$f(\cdot)$, proving that DP's $\delta \simeq 0.6$ is an apparent exponent that is in reality related to the scaling function, while the scaling exponent is $\delta^{\dagger} = 1$.

In summary, we have clarified the finite size scaling ansatz of DP by showing (i) the pdf can always be normalized so that there is no need to wrongly invoke the finite size of the Earth, and (ii) if finite size scaling were to hold, the scaling exponent would equal one exactly, while the apparent exponent $\delta$ is part of the power law scaling function. Issues such as the two discussed here appear rather frequently in the literature (see the discussion by *Christensen et al.* [2007] and Chapter 2 of *Pruessner* [2004]), which justified the lengthy treatment. In particular, this section serves as an example of potential pitfalls and is not intended solely as a critique of DP because, as we will see in the next section, we will present evidence that challenges the existence of DP's proposed power law finite size scaling and the universal scaling law.

### 5.3.3 Comment on "Analysis of the Spatial Distribution Between Successive Earthquakes" by Davidsen and Paczuski (2005) *

*Davidsen and Paczuski* [2005] claim evidence contradicting the theory of aftershock zone scaling (AZS) in favor of scale-free statistics. DP cite *Kagan* [2002a] on AZS but their definition differs: DP claim the"theory" states "main-shocks generate aftershocks within finite [...] zones [...] comparable to the [main-shock's] rupture length $l_r = 0.02 \times 10^{0.5m}$". *Kagan* [2002a] fit aftershocks with a Gaussian and found scaling of the standard deviation (a proxy for rupture length) with moment. He did not dispute the well-known fact that aftershocks occur far beyond $l_r$ [*Hill et al.*, 1993]. We further present three elements showing that DP's analysis may not detect rupture length scales.

First, their power law depends on one single earthquake. We use the same data

*An edited version of this section was published in Werner, M. J. and D. Sornette, *Phys. Rev. Lett.*, 99, 179801, 2007. Copyright (2007) by the American Physical Society.

as *Davidsen and Paczuski* [2005] but contrast in Figure 5.2 the active 6-month period from June 1 until December 31, 1992 including the June 28 M7.3 Landers earthquake (crosses) with the remaining catalog (circles). First, removing only 6 months from a 17 year period causes the power law to disappear. Second, the Landers aftershocks show clear signs of scales, such as the bump marked by an arrow, which may be due to the simultaneous aftershocks of the June 28 M$6.4$ Big Bear and the July 11 M$5.7$ Mojave earthquake and the rupture length of Landers. Rather than a period of "heterogeneous" rates, DP's results are dominated by a single event.

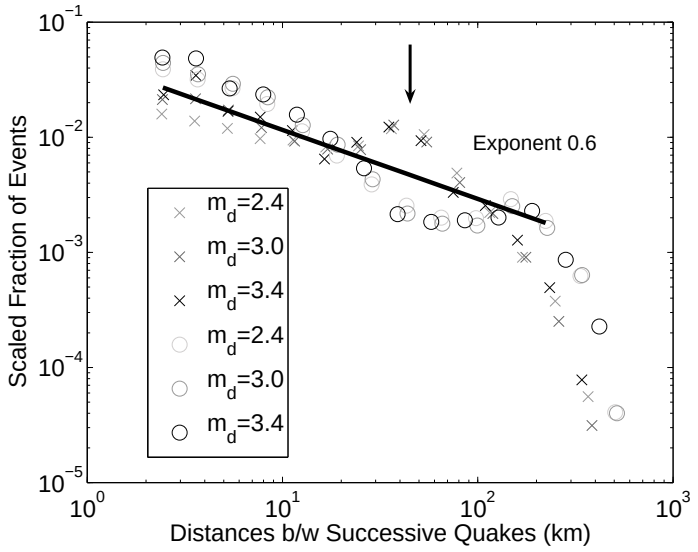


Figure 5.2: Distribution of the epicentral distances between successive quakes in southern California: (crosses) Landers and its aftershocks; (circles) the remainder of the 17 year catalog.

Second, we find no power laws in other regions. For Japan (Figure 5.3, circles), we use the JMA catalog from Jan 1984 to Dec 2001 within $(120^o\text{E}, 150^o\text{E})$ by $(25^o\text{N}, 45^o\text{N})$. Despite being more seismically active (and hence "heterogeneous", as DP require), no power law emerges. For northern California (Figure 5.3, crosses), we

use NCSN data from Jan 1984 until Dec 2004. We did not repeat DP's analysis for smaller boxes because without a power law at the largest scale, DP's finite size scaling (of a power law scaling function $\sim x^{-\delta}$) cannot hold.

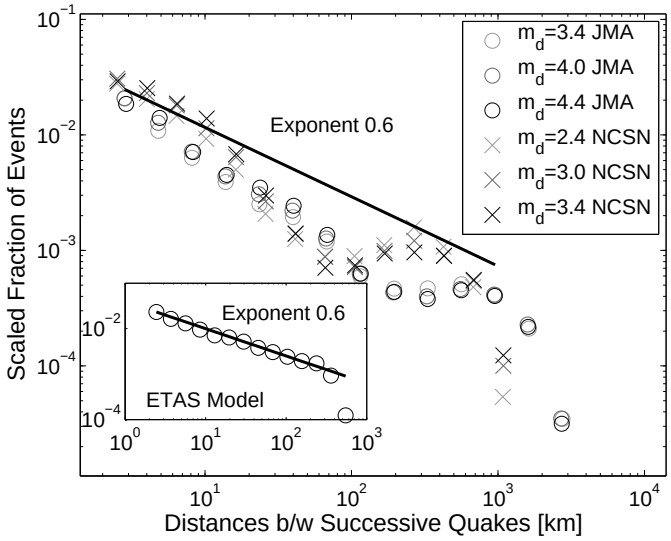


Figure 5.3: Same as Figure 5.2 for Japan (circles) and northern California (crosses). Inset: ETAS model simulation.

Third, we show in Figure 5.3 (inset) that a model [*Ogata*, 1988] with explicit AZS consistent with Kagan's definition can reproduce the power law. Therefore, an apparent power law in interevent distances does not necessarily contradict AZS. We simulated a catalog of comparable duration and number of events as in DP's data set with an ETAS model [*Ogata*, 1988], explicitly including rupture length scales $l_r(m) = 0.02 \times 10^{0.5m}$ in the spatial aftershock decay with distance $d$ according to $P(d) \sim (l_r(m) + d)^{-(1+\mu)}$ (e.g. ***Helmstetter et al.*** [2005a]) and constrained the catalog to include a Landers-like $M7.3$ main-shock. We used the parameters ($b = \alpha = 1, p = 1.1, c = 0.0001, k = 0.0025, \mu = 2$, background=$1.0$ per day, $m_d = 2.4$ in a $500$ km by $500$ km window). We observe an apparent power law irrespective of the largest

aftershock zone scale $l_r \sim 90$ km.

### 5.3.4 DP's Reply and Discussion

*Davidsen and Paczuski* [2007] (hereafter DP2) composed a reply to our comment [*Werner and Sornette*, 2007a] (hereafter WS). Their reply consists of four arguments, which we summarize briefly and discuss point by point.

- DP2 reply that "almost all of [WS'] arguments are based on an analysis that is completely insensitive to finite size scaling [...]. While we studied the probability density function (PDF) $P_L(\Delta r)$ of spatial distances over all boxes of a given linear size $L$ and its variation with $L$, WS do not take into account any variation with box size. Instead, they consider all events in the given catalog being in a single "box" whose size and shape is determined by the catalog. Since the scaling function $f$ is not constrained to a particular form a priori, the presence or absence of finite size scaling cannot be established by considering a single scale $L$ as claimed by WS. Thus, they cannot make any statements about the variation with linear size $L$ and the existence of finite size scaling."

  DP2 are wrong in saying that we cannot make any statements about FSS: we can say that the scaling function cannot be a power law, as already stated in our comment. It is true that, in principle, there could indeed be FSS, but without a power law at our chosen scale, there cannot be a data collapse onto a power law and hence no FSS with a power law scaling function, as claimed by DP. Thus, at the very least, the "universal" scaling function was reduced to a "regional" scaling function that changes shape from a power law in southern California to something different in Japan and northern California. This strongly challenges the idea that universal features are being observed that may be related to the criticality in the earth's crust. To support their claim, DP must now find and

explain FSS with a different scaling function in different regions of the world.

- DP claim that our results for Japan and northern California are "strikingly similar for periods of quasistationary seismic activity", as defined by *Corral* [2006]. "This suggests that the observation periods in Japan and northern California [...] show rates of seismic activity which are rather homogeneous" compared to the "heterogeneous" southern California data set. DP repeat that their proposed power law scaling ansatz "only holds for very long observation periods where the rate of earthquake activity is highly heterogeneous in space and time".

  As already stated in the comment, Japan is more seismically active than southern California. Therefore, Japan is expected to be more heterogeneous in terms of seismic rates. As a consequence, the power law scaling should be easily observed in Japan. The fact that we do not observe a power law is therefore a direct contradiction of their qualification (that heterogeneous rates must be present for the power law to appear). This observation, in turn, suggests that the power law in southern California is a coincidence rather than a universal feature that seems difficult to reproduce elsewhere. This point also highlights the subjective definition of homogeneous and heterogeneous rates.

- Their third point is as follows: "Another point raised by WS is that aftershocks occur at distances larger than the main shock's rupture length. This is well known, yet the vast majority of what are typically considered aftershocks occur within distances which are comparable to and no more than a few times the rupture length (see, for example, [*Shcherbakov et al.*, 2005] and references therein). In particular, the largest rupture length in the catalog from southern California we studied is $\simeq 90$ km. This is significantly less than the spatial extent of the area studied ($\simeq 500$ km), which allows us to test the hypothesis of aftershock zone scaling with magnitude. As our results [...] show, no physical length scale

exists in the range from 20 to $\simeq 500$ km".

We are not entirely sure about the argument of DP here. Perhaps they are calling attention to the fact that many other authors do assume that aftershocks occur within a few rupture lengths of the main shock, so that their definition of aftershock zones was indeed justified. In that case, the misunderstanding of aftershock zone scaling is that of the authors of such studies, not ours, and presents a weak and uninformative null hypothesis to test.

- Finally, DP state that "While [WS] claim that this extended [ETAS] model is an accurate description of aftershock zone scaling with main shock magnitude, a comparison of the Landers sequence shown in [Figure 5.2] and the ETAS model in [Figure 5.3] suggests otherwise. More importantly, the particular form of $P(d)$ is speculative - it does not follow from the work by Kagan – and it is by no means generally accepted. Despite some indication for a power law with $\mu \simeq 1.35$ for short times after the main shock [*Felzer and Brodsky*, 2006], other results even directly contradict the form of a power-law decay for distances larger than the main shock rupture length [*Davidsen et al.*, 2006, 2007] if the activity is considered over the long time scales relevant for our study in [*Davidsen and Paczuski*, 2005]. To summarize, the behavior of a model, which is not an accurate description of aftershock zone scaling, cannot prove that our earlier results are insensitive to the existence of physical length scales associated with aftershock zones."

  First, we do not claim that the extended ETAS model is an accurate description of aftershock zone scaling. Rather, we show simply that a particular (reasonable) form of aftershock zone scaling is not inconsistent with the apparent power law. This suggests that aftershock zones may be present in the power law, but that DP's power of resolution is not good enough to resolve these scales. The idea

is to challenge DP into thoroughly testing their results based on simulations. Second, the simulations are supposed to reproduce the entire data set, not simply the Landers sequence, as DP seem to believe. As stated in the comment, we use a similar number of earthquakes in a similar period of time as for the original DP data set for southern California and constrain the catalog to include a Landers-like event, thereby modeling the entire data set, not just subsections. Third, as already stated, it is true that the spatial kernel of the ETAS model which we use does not follow from Kagan's work: in fact no particular form follows from his work, except that the kernel must decay somehow with distance and contain a scale associated with the main shock rupture length that scales with magnitude. Neither do we claim that the spatial kernel is a highly accurate or accepted form. In fact, the spatial kernel in the ETAS model is perhaps the least agreed upon building block of the model. It is a testament of the lack of power of DP's data analysis that we can nevertheless reproduce the observations. But our point is different: we have shown that aftershock zones can be hidden in the apparent power law, so that an apparent power law does not imply that aftershock zone scaling is wrong. It is now in DP's hands to resolve this non-uniqueness to continue their claim that their results contradict aftershock zone scaling.

We did not attempt to settle the question of how aftershocks distribute spatio-temporally around main shocks. This is an important topic that needs to be addressed. In our comment [*Werner and Sornette*, 2007a], we solely intended to challenge DP to test their results elsewhere and by simulations. As long as finite size scaling has not been shown to hold in other regions of the world, and as long as a reasonable model with explicit aftershock zone scaling can reproduce their observations, their claims do not stand.

## 5.4 Conclusion

This chapter contributed to the ongoing debate in the literature on the hypothesis that earthquakes are self-organized critical and on the existence, validity and interpretation of finite size scaling relations and universal scaling laws of probability density functions of inter-earthquake statistics. In agreement with previous criticism of the purported scaling laws, we find little evidence that supports such a claim for the spatial distances between successive earthquakes. Rather, previously known statistical laws about seismicity are sufficient to explain the apparently novel features.

In particular, we closely analyzed the work of *Davidsen and Paczuski* [2005], who extended the analysis from temporal recurrence times to the spatial dimension by suggesting that the probability density function of spatial distances between successive earthquakes obeyed finite size scaling with a power law scaling function. This result was interpreted to contradict aftershock zone scaling. First, we analyzed the proposed scaling ansatz and were able to clarify a perceived problem regarding the normalization of the pdf. We used the results of *Christensen et al.* [2007] to show that the actual scaling exponent must equal 1 and that the proposed exponent $\delta \simeq 0.6$ is an apparent exponent. However, a re-analysis of the data challenges the claim of finite size scaling itself and suggests that the scales associated with main shock rupture lengths may be hidden in the data and that the analysis is not sensitive to their existence.

The conclusion is therefore that the results of *Davidsen and Paczuski* [2005] do not add to our knowledge, similar to previous criticism of universal scaling laws in seismicity. The study highlights the need for thorough testing of new data analysis methods before claims are made contradicting established knowledge. By simulating data that was analyzed in the same way as the real data, we showed that the null hypothesis could reproduce the data, which were supposed to be evidence contradicting the null hypothesis. This underlines the need for rigorous testing of hypotheses on

simulated benchmark data. Models with realistic clustering, such as the ETAS model, can provide powerful null hypotheses against which observed patterns can be tested. To provide better and better benchmarks, the null hypotheses must themselves be well-understood, tested and calibrated. This is the topic of the next chapter.

# CHAPTER 6

# Magnitude Uncertainties Impact Seismic Rate Estimates, Forecasts and Predictability Experiments *

## 6.1 Abstract

The Collaboratory for the Study of Earthquake Predictability (CSEP) aims to prospectively test time-dependent earthquake probability forecasts on their consistency with observations. To compete, time-dependent seismicity models are calibrated on earthquake catalog data. But catalogs contain much observational uncertainty. We study the impact of magnitude uncertainties on rate estimates in clustering models, on their forecasts and on their evaluation by CSEP's consistency tests. First, we quantify magnitude uncertainties. We find that magnitude uncertainty is more heavy-tailed than a Gaussian, such as a double-sided exponential distribution, with scale parameter $\nu_c = 0.1 - 0.3$. Second, we study the impact of such noise on the forecasts of a simple clustering model which captures the main ingredients of popular short term models. We prove that the deviations of noisy forecasts from an exact forecast are power law distributed in the tail with exponent $\alpha = (a\nu_c)^{-1}$, where $a$ is the exponent of the productivity law of aftershocks. We further prove that the typical scale of the fluctuations remains sensitively dependent on the specific catalog. Third, we study how noisy forecasts are evaluated in CSEP consistency tests. Noisy forecasts are

*A differently formatted version of this chapter was submitted to *J. Geophys. Res.* on October 9, 2007 and is currently under review. A preprint is available from http://arxiv.org/abs/0710.4196

rejected more frequently than expected for a given confidence limit. The Poisson assumption of the consistency tests is inadequate for short-term forecast evaluations. To capture the idiosyncrasies of each model together with any propagating uncertainties, the forecasts need to specify the entire likelihood distribution of seismic rates.

## 6.2 Introduction

Earthquake prediction experiments such as the recently formed Collaboratory for the Study of Earthquake Predictability (CSEP) [*Jordan*, 2006] and the Working Group on Regional Earthquake Likelihood Models (RELM) *Schorlemmer et al.* [2007] aim to investigate scientific hypotheses about seismicity in a systematic, rigorous and truly prospective manner by evaluating the forecasts of models against observed earthquake parameters (time, location, magnitude, focal mechanism, etc) that are taken from earthquake catalogs. After the optimism of the 1970s followed by pessimism on earthquake prediction [e.g. *Geller* [1997], the lesson for the next generation of earthquake forecasters is clear: model formulation, calibration and hypothesis testing must be robust, especially with respect to data quality issues.

Paleoseismology provides a good example of the importance of uncertainties in data analysis and hypothesis testing. *Rhoades et al.* [1994] showed that the hazard rate on the Pallett Creek segment of the San Andreas fault can vary by a factor of three depending on parameter estimates of any chosen model for the earthquake cycle, all consistent with the data. *Davis et al.* [1989] showed that parameter uncertainties and their continuous updating with time lead to vastly different probability forecasts of the Parkfield segment. In the context of renewal processes, *Sornette and Knopoff* [1997] showed that different distributions of inter-event times, which may all be compatible with the data, have drastically diverging implications for the conditional waiting time until the next earthquake. *Ogata* [1999a, 2002] concluded that uncertainties in the

occurrence times of historical quakes make differentiating between different renewal distributions for the hazard on one fault inconclusive.

CSEP and RELM and future experiments have at their disposal earthquake catalogs of much higher quality than paleoseismic studies, of course. Nevertheless, these modern earthquake catalogs are ridden with their own observational uncertainties, biases, arbitrary conventions and spatio-temporally varying quality characteristics. While RELM acknowledges data problems and plans to simulate so-called "modified observations" from the actual observations and their error estimates in order to test the forecasts against these alternative, potentially equally likely scenarios, their proposed statistics and tests were not shown to be sufficient to solve data quality issues at the various stages of the hypothesis testing process, in particular with respect to the generation of model forecasts.

But models for short-term forecasts are typically quite sensitive to recent earthquake catalog data that the models are calibrated on. For instance, the seismic rate forecasts in popular clustering and aftershock models are exponentially dependent on main shock magnitudes. These models include the Epidemic Type Aftershock Sequence (ETAS) Model [*Ogata* [1988]; see *Helmstetter et al.* [2006] for an implementation], Short Term Earthquake Probabilities (STEP) [*Gerstenberger et al.*, 2005], the model by *Reasenberg and Jones* [1989, 1994], the Epidemic Rate and State (ERS) model by *Console et al.* [2007] and the model by *Kagan and Knopoff* [1987]. Small errors in reported magnitudes can therefore have a large, detrimental effect on forecasts. Rather than entering at the evaluation phase, the magnitude uncertainties impact during the input or calibration stage - not accounted for by the current RELM/CSEP evaluation tests. To the best of our knowledge, no study has addressed the impact of data quality on seismic rate estimates, model forecasts and their evaluation.

Here, we analyze how observational errors in magnitude estimates propagate to

seismic rate estimates and forecasts in common aftershock models. While other data quality issues are certainly important, we focus here on magnitude uncertainties, which seem to be the most important source of noise for short term clustering models because of the models' exponential sensitivity on past magnitudes. However, because magnitude uncertainties are not routinely reported in earthquake catalogs, we first study the accuracy of magnitude estimates in section 6.3. In section 6.4, we then use a simple aftershock model containing the basic ingredients of most operational short term clustering models and investigate the impact of the magnitude errors on seismic rate estimates and forecasts. Finally, we conduct a numerical experiment in which we mimic the RELM/CSEP testing center to investigate the impact of noise on the evaluation of forecasts generated from noisy magnitude data.

## 6.3 Magnitude Uncertainties

### 6.3.1 Different Types of Magnitudes and their Errors

Magnitude is a measurement unit originally introduced by Charles Richter and Beno Gutenberg during the 1930s at the California Institute of Technology to measure the "size" of earthquakes. Different magnitudes measure different characteristics of seismic waves (or of earthquake-generated tsunamis or surface deformation), although all aim to quantify the "size" of an earthquake. For example, the original Richter magnitude ($M_L$) uses the maximum amplitude of seismic waves and the time difference between P and S waves recorded on a Wood-Anderson seismograph along with correction factors for geometric spreading and attenuation. Body wave magnitudes ($m_b$) and surface wave magnitudes ($M_S$) are based on amplitudes of P and surface waves often near periods of 1s and 20s, respectively. Both may include corrections for focal depth, the period used, the attenuation and potentially a station correction for the

inhomogeneity of the soil.

The moment magnitude [*Kanamori*, 1977; *Hanks and Kanamori*, 1979] is based on the scalar seismic moment, which is the magnitude of the seismic moment tensor. The moment tensor, in turn, is a representation of the equivalent body forces that would produce the same radiated seismic wave pattern as the observed one. Under a point source approximation, the seismic moment is equal to the product of the rigidity of the material, the average displacement on the fault and the average fault area that slipped. The moment is a long period (zero frequency) measure of the total energy of the earthquake and in theory does not saturate. The moment magnitude has a clear physical interpretation, while the relationship of other magnitudes to earthquake source parameters is less (if at all) established.

Many other definitions and conventions exist for magnitudes. For instance, the Advanced National Seismic System (ANSS) [*Benz et al.*, 2005] reports various magnitudes, including local (Richter-type) magnitudes, body wave magnitudes, moment magnitudes and coda duration magnitudes. Few earthquake catalogs homogeneously use the same magnitude type to measure earthquakes. The (Harvard) Centroid Moment Tensor (CMT) project catalog [e.g. *Ekstrom et al.* [2005]] is a rare exception of a relatively uniform global catalog.

In principle, each magnitude type needs to be addressed separately to establish uncertainties. For non-physics-based magnitudes (i.e. all but the moment magnitude), this can be particularly difficult as they are conventions by definition and cannot be verified independently. However, even in these cases it is possible to establish estimates of the errors in the convention-based magnitude estimate. For instance, it is possible to analyze the effects of discretization of media and equations, of the measurement precision of seismometers, of the assumed velocity and attenuation models of the Earth (spherically symmetric, 3D, etc) and of the resolution of the inversion

algorithm depending, e.g., on station coverage. We will use the term intra-magnitude uncertainties to refer to such individual magnitude error estimates. These uncertainties measure how close a magnitude estimate may be to its convention-based "true" value.

More fundamentally, the definition of an earthquake "event" (and hence an associated magnitude) can be questioned. The identification of one earthquake seems inherently tied to the particular time, space and frequency resolution of the observer. *Kagan and Knopoff* [1981] constructed a branching model which mimics a continuous deformation flow. When a Green's function is applied to the deformation and a scale imposed, the resulting seismograms seem to show separate events. *Peng et al.* [2007] analyzed the properties of catalogs after large earthquakes and found by hand-picking events from the waveforms that many more events are present than detected by catalog routines, providing further observational evidence to the idea that the scale of the observer may determine the definition of an event. In this article, we will assume that listed catalog magnitudes are nevertheless useful for extrapolating and forecasting seismic rates. This is the working assumption of all seismicity-based earthquake forecasting experiments.

Earthquake prediction experiments such as RELM and CSEP use a so-called "authorized data stream" for their "natural laboratories" [*Schorlemmer et al.*, 2007]. For California, this data stream is the ANSS catalog. Models accept the listed magnitudes to generate forecasts of future events, irrespective of the type of magnitude listed. The forecast validation is also performed against the listed magnitudes. Apart from the intra-magnitude uncertainties, one should therefore also consider the uncertainties of one particular magnitude estimate in relation to the magnitude that best forecasts future events (if there exists such a "forecast magnitude"). For instance, the moment magnitude may be more relevant in predicting aftershocks than a body wave magnitude: A. Helmstetter (personal communication, 2007) observed unbroken scaling of

the number of aftershocks with moment magnitude $\propto 10^{\alpha M_W}$ up to $M_W = 9.3$ (the great December 26 2004 Sumatra-Andaman earthquake). Physical mechanisms for earthquake triggering might also be constrained by the knowledge of the "forecast magnitude."

But lacking this magnitude, we need to consider the uncertainties between the different types of magnitudes: the inter-magnitude uncertainties. We will study both inter- and intra-magnitude uncertainties to get a sense of the scale of the uncertainties. We then use these error estimates to simulate noisy magnitudes and to study their impact on seismic rate estimates and forecasts in section 6.4.

### 6.3.2 Intra-Magnitude Uncertainties

Ideally, intra-magnitude uncertainties are reported by earthquake catalogs based on knowledge about the seismic instruments and the inversion algorithm. Unfortunately, such information is often lacking in catalogs. A rare exception is provided by the Northern California Seismic Network (NCSN) catalog operated by the U.S. Geological Survey (USGS) and the Berkeley Seismological Laboratory at UC Berkeley. We study the reported uncertainties in section 6.3.2.2.

A simple alternative to studying these errors independently is to compare the magnitude estimates for the same event from different networks and inversion algorithms, e.g. from different catalogs. While one cannot assume that one measure is correct and indicative of the error in the other, one can make some inferences, especially if the catalogs seem uniform and trustworthy (established, e.g., by verifying completeness, stability and other known statistical properties). We therefore study the differences in moment magnitudes as reported by two relatively well-studied and trusted catalogs: the (Harvard) Centroid Moment Tensor (CMT) catalog [*Dziewonski et al.*, 1981; *Dziewonski and Woodhouse*, 1983; *Ekstrom et al.*, 2005] and the USGS moment tensor

(MT) catalog [*Sipkin*, 1986, 1994].

#### 6.3.2.1 Moment Magnitude Uncertainty From the (Harvard) CMT and USGS MT Catalogs

*Sipkin* [1986] (his Figure 7) compared the CMT scalar moment tensor with the USGS-MT scalar moment tensor. He generally found comparable values, but CMT moments were smaller by a factor of two for small events and slightly larger for the largest events. The scatter shows deviations of 0.25 in logarithmic moment units. *Helffrich* [1997] compared the moment tensor solutions provided by three organizations: the Harvard group, the USGS and the Earthquake Research Institute of the University of Tokyo. He found a standard deviation of $0.21$ in $\log_{10}$ moment units between the three data sets. He also showed that the moment estimates systematically improved (converged) for deeper events. *Kagan* [2003] found that the differences in moment magnitude estimates of matched events reported by the Harvard CMT and the USGS MT catalog have a standard deviation of $0.08$ and $0.12$ for deep and shallow events, respectively. If both catalogs contain an equal amount of noise, then the standard deviation of each estimate is equal to 0.05 and 0.08. *Kagan* [2002b] concluded that the standard deviations for moment magnitude estimates in California was 0.08, while conventional (first motion) catalogs provided less accurate estimates with a standard deviation equal to 0.23. Since we are interested in simulating noisy magnitudes to study their impact on seismic rate forecasts and prediction experiments, we update and expand these analyses of moment magnitude to determine the entire distribution.

We used the Harvard CMT catalog from 1 January 1977 until 31 July 2006, which contains 25066 events above $M_W \geq 3$ and wrote an algorithm to match its events with the USGS MT catalog from January 1980 until 31 July 2006, which contains 4952 events above $M_W \geq 3$. Both catalogs are available from `http://neic.usgs.`

gov/neis/sopar/. Neither catalog contains events between 1 December 2005 and 31 March 2006. But since we are not interested in temporal properties of the catalogs, this gap should not bias our results.

We consider two listings from the two catalogs to refer to the same event if they are separated in time by less than 1 minute and in space by less than 150 km. *Kagan* [2003] used the same definitions. In agreement with his findings, the matches are quite robust with respect to space but less robust with respect to the condition on time.

Using these conditions, we match $4923$ pairs of events. Only 29 events listed in the USGS MT catalog cannot be matched with events in the Harvard CMT catalog. Of these, 5 events pass the time requirement but fail the spatial condition. Thus one might suspect extreme errors in locating the events. However, increasing the spatial limit up to 1000 km does not change the results, i.e. the differences do not seem to be due simply to large location errors. They seem to be events listed in the USGS MT catalog that are entirely absent from the Harvard CMT catalog. Most of the other 24 events listed in the USGS MT that could not be matched with events Harvard CMT seem to be events in complex aftershock sequences where the identification of single events may sensitively depend on certain different network characteristics and on choices in the two computer algorithms. Vice versa, many events listed in the Harvard CMT are absent from the USGS MT, often due to a lower detection threshold in the Harvard CMT catalog.

For matched events, we calculate the moment magnitude $M_W$ from the scalar moment $M_0$ (in Newton-meter) using the relation $M_W = 2/3 \ \log_{10}(M_0) - 6$ [*Kagan*, 2003] and analyze the differences in $M_W$ between Harvard CMT and USGS MT estimates. Figure 6.1 shows the distribution of the differences in the moment magnitude estimates from the two different catalogs. Figure 6.1a) shows a fixed kernel density estimate [e.g. *Izenman* [1991]] of the probability density function (pdf) of the mag-

nitude uncertainties (solid). Figure 6.1b) shows the same data in a semi-logarithmic plot. Figures 6.1c) and d) show semi-logarithmic plots of the survivor function and cumulative distribution function, respectively, in order to emphasize both tails.

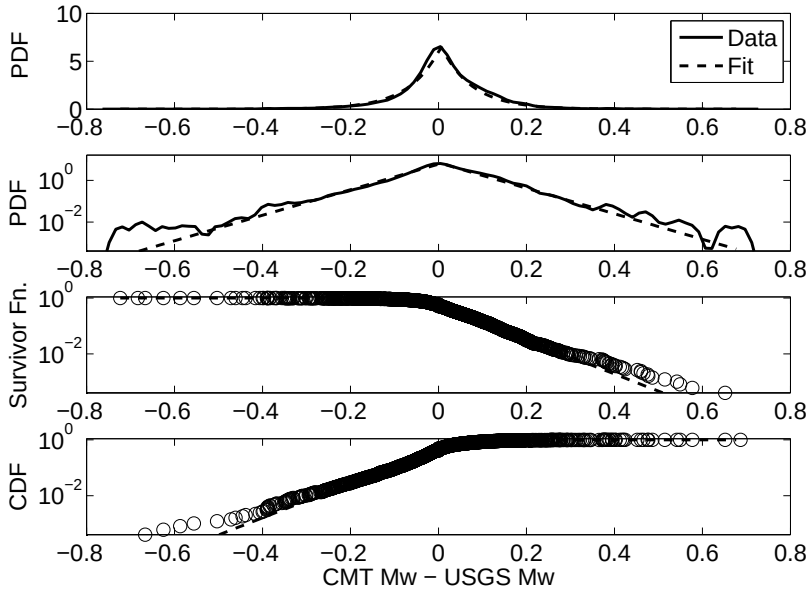


Figure 6.1: Estimating intra-magnitude uncertainty by comparing moment magnitude estimates for the same event from the Harvard CMT and the USGS MT catalogs. **a)** Fixed kernel density estimate (solid) of the probability density function of the differences in moment magnitudes and maximum likelihood fit (dashed) of a Laplace (double-sided exponential) distribution given by equation (6.1) with scale parameter $\nu_c = 0.07$. **b)** Same as a) but in semi-logarithmic scale. **c)** Semi-logarithmic plot of the survivor function (complementary cumulative distribution function). **d)** Semi-logarithmic plot of the cumulative distribution function.

We performed a maximum likelihood fit of the data to a Laplace distribution (a

double-sided exponential distribution), defined by:

$$p_\epsilon(\epsilon) = \frac{1}{2\nu_c} e^{\left(-\frac{|\epsilon - \langle\epsilon\rangle|}{\nu_c}\right)} \tag{6.1}$$

where $\nu_c$ is the scale (e-folding) parameter indicating the strength of the noise and $\langle\epsilon\rangle$ is a shift parameter equal to the median and mean of the distribution. The maximum likelihood estimate of $\langle\epsilon\rangle = 0.006$ is given by the median and essentially indistinguishable from zero. To estimate the scale parameter $\nu_c$ (e-folding scale), we then took absolute values of the deviations from the median and fit the resulting positive data set to an exponential using maximum likelihood. We find that $\nu_c = 0.07$ for the entire data set. The dashed lines in all plots of Figure 6.1 correspond to this maximum likelihood fit (using $\langle\epsilon\rangle = 0.006$ and $\nu_c = 0.07$). The fit approximates the data well in the body, but underestimates the tails as can be seen in Figure 6.1b), c) and d).

To estimate the effect of the tails which are fatter than exponential, we determined the scale parameter $\nu_c$ as a function of the threshold above which we fit the distribution. We determined the median of the data (corresponding to the threshold "0"), took absolute values of the data and performed a maximum likelihood fit to an exponential distribution. We then increased the threshold (with respect to the median) and re-calculated the scale parameter for each threshold value. Figure 6.2 shows the resulting maximum likelihood estimates of $\nu_c$ as a function of the threshold. The error bars show estimated 95 percent confidence bounds. Confirming the presence of the fat tails, the estimated e-folding scale $\nu_c$ increases with the threshold from about $0.07$ to about $0.1$.

The distribution of the differences between magnitude estimates is not the same as the distribution of the individual magnitude uncertainties in one estimate (see the next section 6.3.2.2 for such a direct estimate). To obtain individual uncertainties, one can assume, for instance, that both uncertainties are identically and independently distributed (i.i.d.). In this case, the distribution of the differences can be assumed to be the

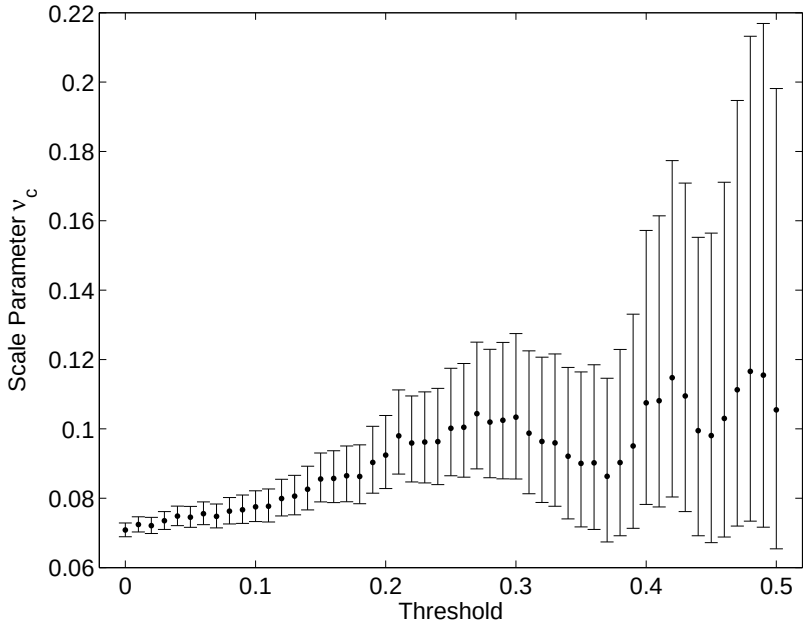


Figure 6.2: Estimates of the e-folding scale parameter $\nu_c$ of the Laplace (double-sided exponential) distribution of equation (6.1) as a function of the threshold above which the data is fit. $\nu_c$ increases from $0.07$ to about $0.1$ due to fatter-than-exponential tails before starting to fluctuate more strongly due to finite sample effects. Error bars show $95\%$ confidence intervals.

convolution of the two individual distributions. In the case of Gaussian distributions, the convolution is also a Gaussian with variance equal to the sum of the individual variances. Unfortunately, the Laplace distribution is not endowed with the same property. The difference between two Laplace distributed random variables is not exactly another Laplace distributed random variable.

While we cannot determine the distributions of the individual uncertainties exactly, we can, however, constrain them. For instance, they cannot be Gaussian for the above-mentioned property that their differences would be Gaussian. The presence of

exponentially or even more slowly decaying tails indicates that the individual uncertainties have at least exponentially decaying tails. For instance, to obtain a distribution for the difference of two random variables which looks approximately Laplace with scale parameter $0.1$, we found that using Laplace distributions for the individual variables requires them to have an e-folding parameter equal to $0.07$. We even found that the convolution of power law distributions was able to produce the observed tails more accurately. In summary, the individual uncertainty distributions must have tails that decay as slowly as or more slowly than exponential.

#### 6.3.2.2 Intra-Magnitude Uncertainties Reported by the NCSN

In the context of the CSEP earthquake prediction experiment, the important intra-magnitude uncertainties should be evaluated in California and in the regions where natural laboratories are being established around the world (e.g. Europe, New Zealand, West Pacific). For California, the earthquake catalog that provides the data for the RELM and CSEP experiments is the ANSS composite catalog.

Many regional networks feed their data into the ANSS composite catalog. It is essentially a computer program with rules for merging the data [e.g. *Oppenheimer* [2007]]. The ANSS assigns so-called "authoritative" regions to each seismic network, meaning that in those regions only data from its designated network is accepted into the composite catalog. The Northern California Seismic Network (NCSN) fulfills this role for northern California. The earthquake data that is passed on to the ANSS by the NCSN in turn comes from two sources: the Berkeley Seismological Laboratory at the University of California, Berkeley (UCB) and the USGS at Menlo Park.

The UCB reports mainly moment magnitudes and local magnitudes, while the USGS reports coda duration magnitudes. Their policy for merging into ANSS is: moment magnitude supercedes local magnitude supercedes coda duration magnitude

(David Oppenheimer, 2007, personal communication). UCB does not report uncertainties, but the moment magnitude is believed to be the most stable with uncertainties around $0.1$, while the scatter in local magnitudes is strongly affected by radiation pattern (approximately 0.15 magnitude units) and can be as large as $\pm 0.5$ (Margaret Hellweg, 2007, personal communication).

The USGS, on the other hand, does provide uncertainties to the ANSS catalog, based on inversions by the Hypoinverse program, written by Fred W. Klein of the USGS [*Klein*, 2002]. The Hypoinverse code "processes files of seismic station data for an earthquake (like P-wave arrival times and seismogram amplitudes and durations) into earthquake locations and magnitudes." The summary magnitude for an event is the weighted median of the station magnitudes. Each station can report a magnitude for an event if its user-specified weight is non-zero. The final reported magnitude is the value for which half of the total weights are higher and half lower.

The measure of uncertainty reported by the Hypoinverse program, available from the NCSN in its hypoinverse format output and from the ANSS in its "raw" format, is the Median Absolute Difference (MAD) between the summary magnitude and the magnitudes from the other reporting stations.

The MAD value measures the variability of the magnitude estimates across several stations. It therefore probes the assumed velocity, attenuation and geometrical spreading models and the differences in station properties (e.g. frequency response, gain etc). Systematic biases due to other reasons in the magnitude inversion, however, cannot be captured by this measure. For instance, phase picking or instrument calibration may be systematically biased. Furthermore, only one measure (the median) of the entire distribution of the magnitudes from the different stations calculated for one event is reported. If the distributions are heavy-tailed, then the median may give a false sense of good measurement in the presence of large variability.

Nevertheless, some inference can be made about the accuracy of determined magnitudes. We collected all earthquakes in the NCSN's authoritative region of the ANSS, defined by a polygon with the following latitude and longitude coordinates: { 34.5, -121.25, 37.2167, -118.0167, 37.75, -118.25, 37.75, -119.5, 39.5, -120.75, 42.0, -121.4167, 42.0, -122.7, 43.0, -125.0, 40.0, -125.5, 34.5, -121.25 }. We selected data from 1/1/1984 to 31/12/2006 (inclusive) above a magnitude threshold $m_{th} = 3$. The data with MAD values can be retrieved from the website `http://www.ncedc.org/ncedc/catalog-search.html` by selecting output in the "raw" format. This yielded 6125 earthquakes. But, as already mentioned, only the USGS reports MAD values for its magnitudes, leaving 3073 events.

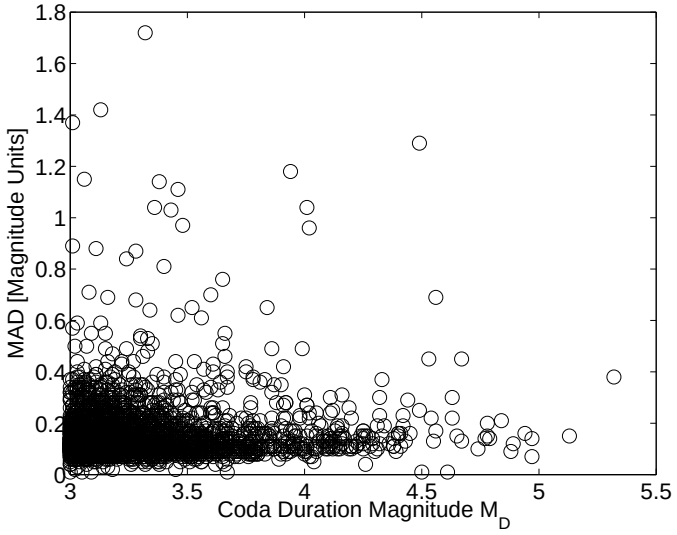


Figure 6.3: Median Absolute Differences (MAD) versus their associated coda duration magnitudes as reported by the NCSN in its authoritative region of the ANSS composite catalog. MAD values measure the variability of the magnitude estimates for the same event from different stations as computed from the HYPOINVERSE program of the USGS.

In Figure 6.3, we show a scatter plot of the MAD versus their associated duration magnitudes. To test for a decrease of MAD with increasing magnitude, we divided the magnitudes into bins of size $0.5$ and calculated the mean MAD for each bin. In the range $3.5 < m < 4.0$ (2420 events), the mean MAD was 0.15; for $4.0 < m < 4.5$ (372 events), the mean was 0.16; for $4.5 < m < 5.0$ (22 events), the mean was 0.20. The bin $5.0 < m < 5.5$ had mean 0.27 but counted only 2 events. Rather than a decrease of MAD with increasing magnitude, we see some evidence for an increase.

Figure 6.4 shows the kernel density estimate of the probability density function (pdf) of the MAD values, the cumulative distribution function (cdf) and the survivor function. For reference, we also included the $99\%$ confidence limit (the MAD value for which $99\%$ of all reported MAD values are smaller). While the mean of the values was $0.15$ and the standard deviation $0.1$, the $99\%$ confidence limit was only reached at $0.59$. That the distribution is strongly non-Gaussian can also be seen from the bottom panel. The tails decay more slowly than an exponential, indicating that outliers occur relatively frequently. Indeed, the maximum MAD value was $1.72$.

Figure 6.5 shows a scatter plot of the MAD values versus the number of stations involved in the computation of each coda duration magnitude and its MAD value. When the number of stations involved is very small, we see large scatter - very small MAD values of less than 0.1 and very large values above 0.5. On the other hand, as more stations are involved, the smallest MAD values increase to about 0.1. This may indicate that MAD values less than $0.1$ are due to too few stations involved in the computation and probably unreliable. (At the same time, we note that a $M_D = 5.32$ event with MAD $0.38$ was recorded by 328 stations, suggesting that large MAD values are real.)

When the number of stations registering an event is small, this presumably implies that the event is small and/or remote. Given that many events are located by less

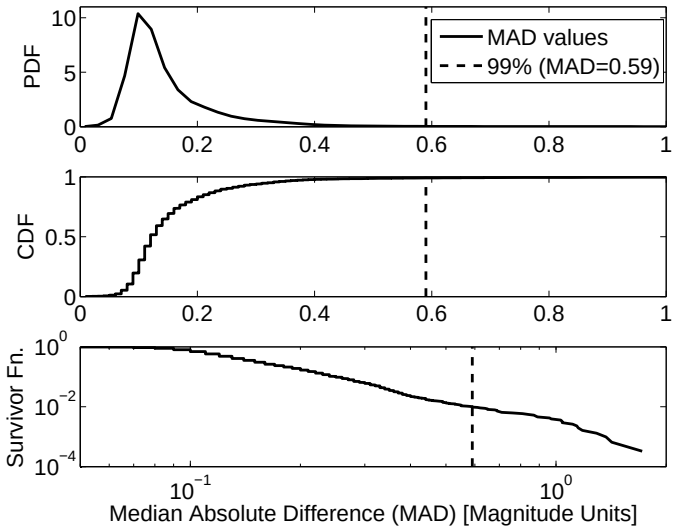


Figure 6.4: Median Absolute Differences (MAD) reported by the NCSN in its authoritative region of the ANSS composite catalog. Top: kernel density estimate of the probability density function (pdf). Middle: cumulative distribution function (CDF). Bottom: survivor function plotted on logarithmic axes. The dashed line at MAD=0.59 corresponds to the 99th percentile of the distribution.

than 10 stations, we may have detected evidence that the ANSS is not complete in the authoritative region of the NCSN down to $m = 3$, because even some $M_D \sim 5$ events are constrained only by few stations.

Finally, it is difficult to interpret the group of large MAD values reported when few stations are involved. Perhaps the events belong to a particular group defined by a region or period with special properties that are not well modeled by the Hypoinverse program.

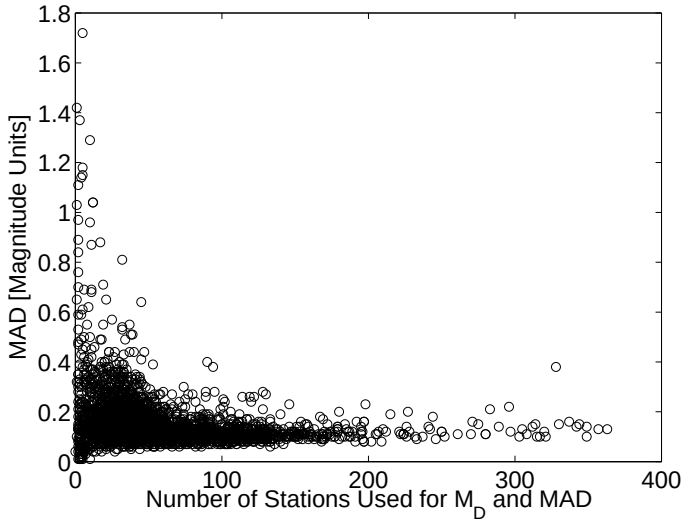


Figure 6.5: Median Absolute Differences (MAD) reported by the NCSN in its authoritative region of the ANSS composite catalog as a function of the number of stations involved in the computation of the magnitude and the MAD value.

### 6.3.3 Inter-Magnitude Uncertainties

As was mentioned before in this article, models in earthquake prediction experiments indiscriminately use whatever type of magnitude is listed in the authorized data stream. Given the random and systematic differences between the magnitudes, the relevant magnitude uncertainties in the context of forecasting are in fact inter-magnitude uncertainties. Many studies have investigated the relation of one magnitude scale to another and their random scatter. We review some of those here. We then analyze two data sets: (i) the differences between the CMT moment magnitudes and their corresponding body or surface wave magnitudes from the catalog provided by the Preliminary Determination of Epicenters [*Preliminary Determination of Epicenters (PDE)*, 2001] (section 6.3.3.1); and (ii) the differences between duration and local magnitudes in the NCSN (section 6.3.3.2).

*Sipkin* [1986] (his Figures 1 and 2) compared the body wave magnitudes $m_b$ and surface wave magnitudes $M_S$ from the PDE catalog with early USGS scalar moments. Body wave magnitudes were scattered by up to one magnitude unit for the same value of the seismic moment, while surface wave magnitudes were less scattered but also substantial. *Dziewonski and Woodhouse* [1983] (their Figures 14a and 14b) compared the CMT moment with the PDE body wave magnitudes and found a similar amount of scatter.

*Harte and Vere-Jones* [1999] compared the properties of the local New Zealand catalog with the data from the PDE catalog, made available by the National Earthquake Information Center (NEIC). They concluded that the differences between the PDE's $m_b$ and the local catalog's $M_L$ were random for shallow events, but up to 1 unit of magnitude in size. For deeper events, $m_b$ was systematically smaller than the local $M_L$.

*Kuge* [1992] found a systematic bias in the body wave magnitude $m_b$ reported by the International Seismic Center (ISC) and the converted seismic moment taken from the Harvard CMT catalog for intermediate and deep earthquakes in Japan. He computed a "theoretical" $m_b$ based on the CMT seismic moment and regression relationships between the moment and NEIS (National Earthquake Information Service) and ISC body wave magnitudes [*Giardini*, 1988]. The systematic differences were on the order of 0.2 to 0.3 units.

*Patton* [2001] investigated the transportability of the Nuttli magnitude scale based on 1-Hz Lg waves to different regions of the world. He routinely found differences of different types of body wave magnitudes on the order of 0.3 magnitude units.

*Kagan* [2003] found that $m_b$ to $M_W$ conversions could result in scatter with a standard deviation of $0.41$. He also concluded that converting conventional magnitudes into moment magnitude leads to uncertainties which are three to four times the errors

in moment magnitude estimates ($0.05$ to $0.08$).

#### 6.3.3.1 Moment Magnitude vs Body and Surface Wave Magnitude in the CMT and PDE Catalogs

The Harvard CMT catalog calculates moment magnitudes when it receives notification of a large event either from the NEIC via the PDE system or from the ISC. We compared the seismic moments of the Harvard CMT with the original PDE body ($m_b$) and/or surface wave ($M_S$) magnitudes. That large systematic differences exist between these magnitudes is well-known. Here, we look at the differences between the Harvard $M_W$ and the PDE $m_b$ and $M_S$ estimates to evaluate their scale.

We used the global Harvard CMT catalog from 1 January 1976 until 31 December 2005 (available from http://www.globalcmt.org/CMTfiles.html in gzipped "ndk" format). We found $24583$ events listed. Of these, we selected all events that were assigned the source "PDE" ($21450$ events) and converted their scalar seismic moments to moment magnitudes using the equation $M_W = 2/3\ \log_{10}(M_0) - 6$ [*Kagan*, 2003]. We found $21435$ $m_b$ and $13363$ $M_S$ values which we subtracted from their corresponding $M_W$ magnitudes. Figure 6.6 shows the resulting differences as a function of the Harvard CMT $M_W$. There are systematic trends and random scatter. The body-wave magnitude $m_b$ systematically underpredicts $M_W$ for about $M_W > 5.2$. Since $m_b$ is based on relatively short periods, the energy in these frequency bands does not increase beyond this value and the scale saturates. The S-wave magnitude $M_S$, on the other hand, underpredicts $M_W$ systematically but especially for $M_W < 7$.

Figure 6.7 shows fixed kernel density estimates of the probability density functions (pdfs) of the two sets of differences. The systematic shifts of both pdfs indicate systematic under-estimation of $M_W$. The means of the data are 0.26 for $m_b$ and 0.42 for $M_S$. The widths of the pdfs indicate the random scatter. The standard deviations are

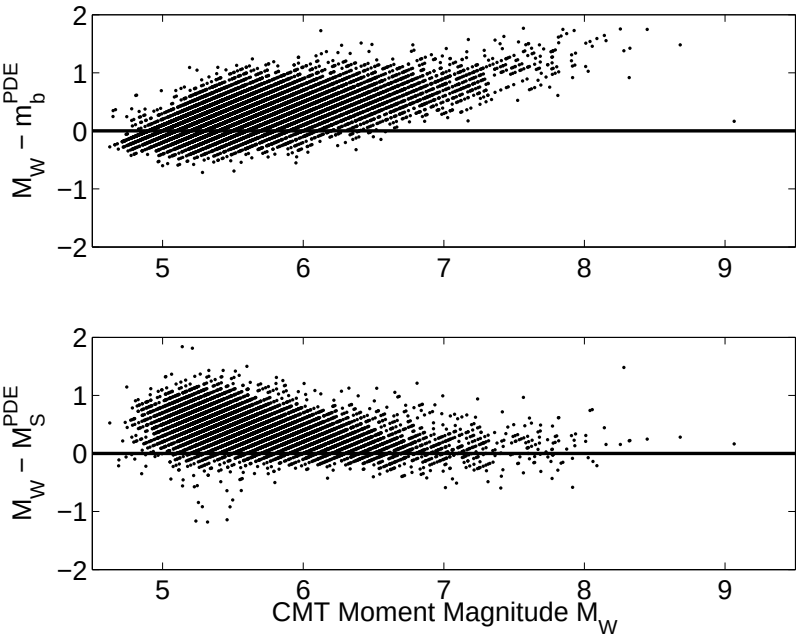


Figure 6.6: Estimating *inter*-magnitude uncertainties by comparing the CMT moment magnitude $M_W$ with its corresponding body wave magnitude $m_b$ (a) and surface wave magnitude $M_S$ (b) from the PDE catalog.

approximately 0.29 for $m_b$ and 0.26 for $M_S$.

For context, an ETAS model would predict 10 times as many aftershocks if the magnitude unit were inflated spuriously by one magnitude unit! These differences can have a profound impact on global testing experiments.

#### 6.3.3.2 Duration Magnitude vs. Local Magnitude in the NCSN Catalog

The NCSN catalog reports both coda duration magnitude $M_D$ and maximum amplitude (local) magnitude $M_L$ in its Hypoinverse output format, available from `http://www.ncedc.org/ncedc/catalog-search.html`. We used data from 1 January 1984 until 31 December 2006 in the region $33^o$ to $43^o$ latitude and -$120^o$ to -

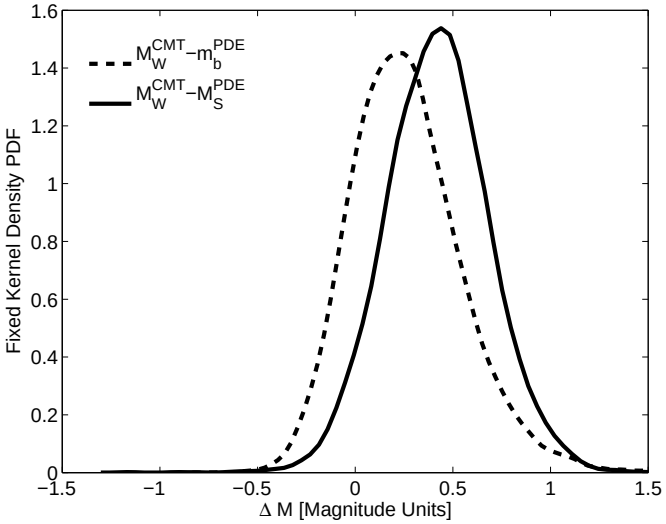


Figure 6.7: Kernel density estimates of the probability density functions of the differences between the Harvard CMT moment magnitudes $M_W$ and their corresponding body wave $m_b$ (dashed) and surface wave $M_S$ (solid) magnitude estimates from the PDE. The means of the data are 0.26 for $m_b$ and 0.42 for $M_S$. The standard deviations are about 0.29 for $m_b$ and 0.26 for $M_S$.

$115^o$ longitude. We selected earthquakes larger than the magnitude threshold $m_{th} = 3$, leaving a total of 4679 events. We found 4595 reported MAD values for duration magnitudes $M_d$ and 2711 MAD values reported for local magnitudes $M_L$.

Whenever both magnitudes were available for the same event, we compared their values. However, we additionally required that at least one of the two magnitudes be equal to or larger than $M_{(\cdot)} = 3$. Despite selecting a magnitude cut-off $m_{th} = 3$ in the search algorithm on the website `http://www.ncedc.org/ncedc/catalog-search.html`, we found 74 events out of the total 4679 where both $M_D$ and $M_L$ were smaller than the prescribed cut-off. The extreme case was an event with $M_D = 0.35$ and $M_L = -0.43$ (we presume that the cut-off magnitude corresponds to the magnitude

reported to the ANSS, which may be different, see beginning of section 6.3.2.2). Removing these 74 events, we are left with 4605 events which obey the condition that at least one of the two magnitudes be equal to or larger than 3. Out of these events, we found 2733 events for which both $M_D$ and $M_L$ were reported. We then calculated the difference $\Delta = M_D - M_L$ for these 2733 events.

Figure 6.8 shows the differences as a function of $M_D$. Recall that at least one of the two magnitudes must be larger than 3 (but not necessarily both). Few events are reported with $M_L > 3$ and $M_D < 3$. On the other hand, many events are reported with $M_L < 3$ and $M_D > 3$. It is hard to discern a trend visually, but $M_L$ seems to under-predict $M_D$ up to about $M_D = 3.5$, then over-predict until about $M_D = 5.5$.

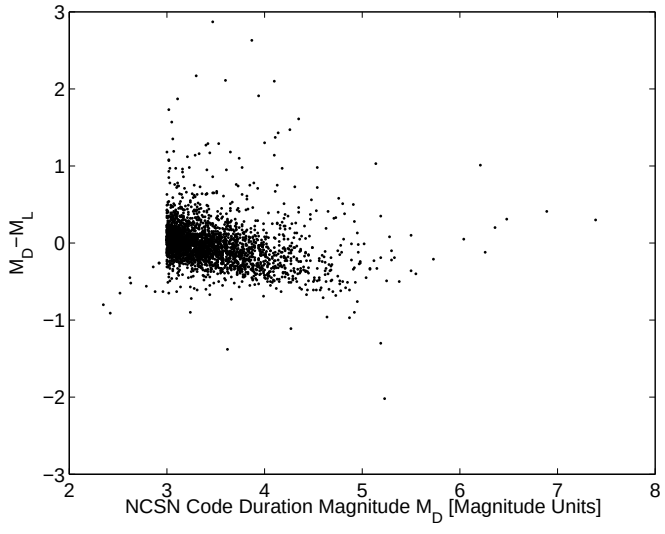


Figure 6.8: Differences between the coda duration magnitude $M_D$ and the local amplitude magnitude $M_L$ versus $M_D$ in the NCSN catalog whenever at least one is larger than 3.

Figure 6.9 shows a fixed kernel density estimate (solid) of the probability density function of the differences $\Delta$. The largest difference between the two was 2.87 mag-

nitude units (note that the x-axis was cut at $\pm 1$). The mean of their differences was $-0.015$, essentially showing no systematic bias. The standard deviation was $0.3$, while the e-folding scale parameter is 0.2. Also shown are fits of the data to a Gaussian distribution (dashed; mean equal to $-0.015$ and standard deviation equal to $0.3$) and to a Laplace distribution (dash-dotted; median equal to $-0.04$ and scale parameter equal to 0.2). While neither fit approximates well the central part of the pdf, the Laplace distribution provides much better fits to the bulk and tails of the data. The Gaussian distribution significantly underestimates the probability of outliers.

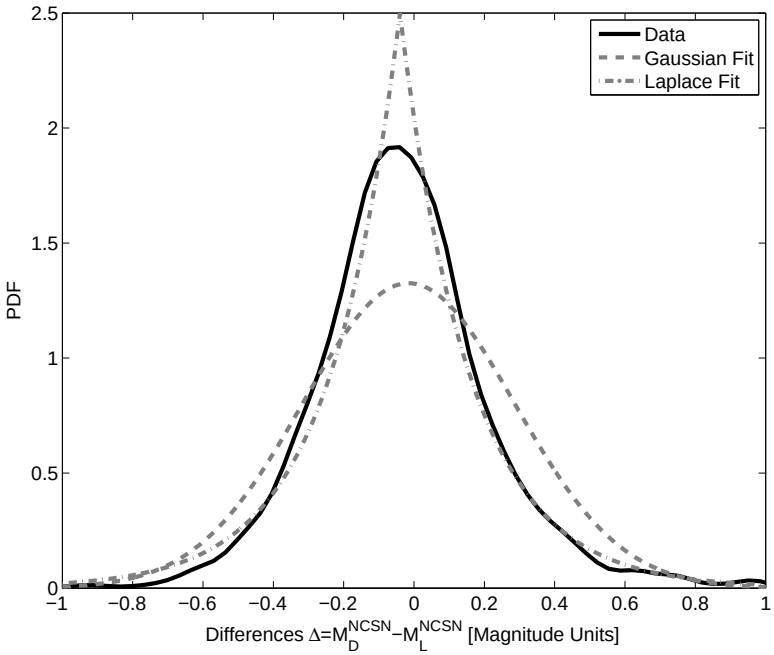


Figure 6.9: Kernel density estimate (solid) of the probability density function of the differences $\Delta$ between the duration magnitude $M_D$ and the local magnitude $M_L$ reported by the NCSN. Also shown are a Gaussian fit with mean $-0.0153$ and standard deviation $0.3$ (dashed); and a fit to a Laplace pdf with median $-0.04$ and e-folding scale parameter $0.2$ (dotted).

### 6.3.4 Summary of Magnitude Uncertainty

Firstly, we compared estimates of the moment magnitude from the CMT and USGS MT catalogs. We found that the Laplace distribution is a good approximation to the bulk of the magnitude differences but it underestimates the tails. Our characterization of the entire distribution of the magnitude differences implies that individual uncertainties are distributed with tails that decay exponentially or even more slowly.

Secondly, we analyzed a data set directly relevant to CSEP predictability experiments. We analyzed MAD values, a magnitude uncertainty measure, reported in the NCSN's authoritative region in the ANSS. We found (i) MAD values fluctuate up to $1.71$ with an average of $0.15$, (ii) there is no evidence that magnitude uncertainty decreases with increased magnitude, (iii) the region may not be complete down to $m_d = 3$, (iv) MAD values less than $0.1$ seem unreliable, and (v) the 99th percentile of MAD values is only reached at $0.59$.

We also considered inter-magnitude uncertainties. These can be extremely large and include systematic differences. We found that PDE body and surface wave magnitudes are systematically smaller (with mean 0.26 and 0.42, respectively) and randomly scattered (with standard deviations 0.29 and 0.26, respectively).

Finally, we studied the differences between NCSN's duration and local magnitudes. We found that the Laplace distribution again provided an adequate fit to the differences with scale factor $0.2$ so that individual uncertainties have exponential or fatter-than-exponential tails.

## 6.4 Impact of Magnitude Noise on Seismic Rate Estimates

In the previous section, we studied magnitude uncertainties. How do these magnitude uncertainties propagate to seismic rate estimates and to forecasts? How can they in-

fluence forecast evaluation tests and the rate of type I and II errors? In particular, how could they influence the CSEP predictability experiments? In this section, we address the first question of the impact of magnitude noise on the estimates of seismic rates in short term clustering models. We use the knowledge of magnitude uncertainties we have gained in the previous section to model to simulate magnitude noise. In section 6.5, we conduct a numerical experiment to begin answering the second and third question.

### 6.4.1 A Simple Aftershock Clustering Model

Most short term seismicity models use three statistical laws to extrapolate rates into the future. These are the Gutenberg-Richter law for magnitude-frequency statistics, the Omori-Utsu law for the temporal distribution of aftershocks and the productivity law for the expected number of offspring of a mother-shock. Models based on these laws include the Epidemic Type Aftershock Sequence (ETAS) Model [*Ogata*, 1988], Short Term Earthquake Probabilities (STEP) [*Gerstenberger et al.*, 2005], the model by *Reasenberg and Jones* [1989, 1994], the Epidemic Rate and State (ERS) model by *Console et al.* [2007] and the model by *Kagan and Knopoff* [1987]. Although there are important differences between these models, all of them employ the above-mentioned three laws to forecast events. In particular, all of them use the so-called productivity law, defined below in equation (6.3), in which the number of expected aftershocks is an exponential function of the mother magnitude.

We choose the Poisson cluster model as the basis of our analysis [see *Daley and Vere-Jones* [2003]]. It is simpler than the above models, yet it captures the essence of the clustering phenomenon through the above three laws. In particular, it preserves the exponential sensitivity of the number of expected aftershocks on the magnitude of the mother-shock. We expect that the conclusions obtained below concerning the impact

the magnitude errors on the uncertainty of predicted seismic rates carry over to the above mentioned models. For models that include secondary triggering, the impact of magnitude noise may even be strongly amplified.

The model can be described as follows. It consists of two processes: a cluster center process and an aftershock process. The cluster center process creates the mother events (also called main shocks or background) which have not been triggered. This process is Poissonian and governed by a homogeneous rate $\lambda_c$. The magnitudes (or marks) of the cluster centers are drawn independently from the Gutenberg-Richter distribution [*Gutenberg and Richter*, 1944]:

$$p_m(m) = \beta e^{-\beta(m-m_d)}, \quad m \geq m_d \tag{6.2}$$

where $m_d$ is an arbitrary cut-off determined by the detection threshold and $\beta = b\log(10)$ is a constant. We denote the marked cluster center process by $\{t_{i_c}, m_{i_c}\}_{1\leq i_c \leq N_c}$. Each of these mothers can trigger direct aftershocks. In contrast to cascading models such as ETAS, there is no secondary triggering, allowing for a simplified analytical treatment and faster simulations. The number of aftershocks is a random number drawn from a Poisson distribution with expectation given by the productivity law:

$$\rho(m_{i_c}) = k\ e^{a(m_{i_c}-m_d)} \tag{6.3}$$

where $k$ and $a$ are constants and $m_{i_c}$ are the magnitudes of the mothers. The productivity law captures the exponential dependence of the number of aftershocks on the magnitude of the mother. This exponential dependence suggests that fluctuations in the magnitudes due to noise will strongly affect any forecast, as we are going to demonstrate analytically and by numerical simulations.

The threshold $m_d$, which measures the size of the smallest triggering earthquake below which earthquakes do not trigger, is arbitrarily set to the detection threshold.

This unjustified assumption is known to bias the clustering parameters [*Sornette and Werner*, 2005a,b].

The triggered events are distributed in time according to the Omori-Utsu law [*Utsu et al.*, 1995]:

$$\Phi(t - t_{i_c}) = \frac{1}{(t - t_{i_c} + c)^p} \tag{6.4}$$

where $c$ and $p$ are constants and the $t_{i_c}$ are the occurrence times of the cluster centers. The marks of the aftershocks are distributed in the same manner as their mothers according to the Gutenberg-Richter law (6.2).

In summary, the rate of events (including aftershocks) of the marked Poisson cluster process is completely defined by its conditional intensity (or rate) [see *Daley and Vere-Jones* [2003]]:

$$\lambda(t, m|H_t^c, \theta) = p_m(m) \left( \lambda_c + \sum_{i_c|t_{i_c}<t} \frac{k\ e^{a(m_{i_c} - m_d)}}{(t - t_{i_c} + c)^p} \right) \tag{6.5}$$

where $H_t^c$ is the history up to time $t$ which need only include information about the cluster centers $\{t_{i_c}, m_{i_c}\}_{1 \le i_c \le N_c}$, as aftershocks do not trigger their own aftershocks and do not influence the future. The set of parameters $\theta = \{\beta, \lambda_c, k, a, m_d, c, p\}$ are assumed to be known.

The (unmarked) intensity above the detection threshold $m_d$ is simply given by integrating over $m$:

$$\lambda(t|H_t^c, \theta) = \lambda_c + \sum_{i_c|t_{i_c}<t} \frac{k\ e^{a(m_{i_c} - m_d)}}{(t - t_{i_c} + c)^p} \tag{6.6}$$

In section 6.5, we will use a spatio-temporal Poisson cluster process to mimic the CSEP experiment and its evaluation as closely as possible. The model is defined by the conditional intensity at time $t$ and location $\mathbf{r}$:

$$\lambda(t,\ \mathbf{r}, m|H_t^c) = p_m(m) \times$$

$$\left( \lambda_c + \sum_{i_c | t_{i_c} < t} \frac{k\, e^{a(m_{i_c} - m_d)}}{(t - t_{i_c} + c)^p} \frac{\mu\, (0.02 \cdot 10^{0.5 m_{i_c}})^{\mu}}{(0.02 \cdot 10^{0.5 m_{i_c}} + |\, \mathbf{r} - r_{i_c}|)^{1+\mu}} \right) \tag{6.7}$$

where we added a commonly used spatial decay function [see e.g. *Helmstetter et al.* [2005a]].

### 6.4.2 Magnitude Noise

To study the impact of magnitude noise on seismic rate estimates and forecasts, we assume that each (true) magnitude is perturbed identically and independently by an additive noise term $\epsilon$

$$m_{i_c}^o = m_{i_c}^t + \epsilon_{i_c} \, . \tag{6.8}$$

We do not need to perturb the magnitudes of the aftershocks because they have no influence on the future seismicity rate.

We use our results on magnitude uncertainties from section 6.3 by modeling noise according to a Laplace distribution characterized by zero mean (unbiased) and a scale (e-folding) parameter $\nu_c$:

$$p_\epsilon(\epsilon) = \frac{1}{2\nu_c} e^{\left(-\frac{|\epsilon|}{\nu_c}\right)} \tag{6.9}$$

We believe that this is a conservative estimate of the distribution of the magnitude noise because the Laplace distribution under-estimates the occurrence of large error outliers. In general, therefore, the fluctuations of the seismic rate are likely to be even larger than calculated below.

We assume the parameters are known. In particular, we assume below that the seismic rate estimates from noisy magnitudes use the same parameters as the "true" rate. We do so to isolate the effect of magnitude uncertainties. A comprehensive analysis of uncertainties including trade-offs between parameter and data uncertainty is beyond the scope of this article and is most likely extremely model-dependent.

### 6.4.3 Fluctuations in the Seismic Rates Due to Noisy Magnitudes

To investigate the fluctuations of the seismic rate estimates due to noisy magnitudes, we compare the seismic rates estimated from perturbed magnitudes with the true seismic rate. The perturbed rates from the observable, noisy magnitudes $m_i^o$ are given by:

$$\lambda^p(t|H_t^c) = \lambda_c + \sum_{i_c|t_{i_c}<t} \frac{k\ e^{a(m_{i_c}^o - m_d)}}{(t - t_{i_c} + c)^p} \tag{6.10}$$

We consider the deviation $\Delta\lambda(t)$ of the perturbed rates from the true rate, given by:

$$\Delta\lambda(t|H_t^c) = \lambda^p(t|H_t^c) - \lambda(t|H_t^c) \tag{6.11}$$

In the remainder of this section, we characterize the fluctuations $p(\Delta\lambda)$ of the deviations from the true rate $\Delta\lambda(t|H_t^c)$. In particular, we would like to know how strong the deviations can be for a given catalog and whether one can make general statements for any catalog.

To obtain some properties of the distribution of $\Delta\lambda(t|H_t)$, we rewrite equation (6.11) as a finite sum over the product $w_i \cdot z_i$ (see Appendix B.1):

$$\Delta\lambda(t|H_t^c) = \sum_{i_c|t_{i_c}<t} w_i z_i \tag{6.12}$$

where

$$w_i = \frac{k e^{a(m_i^t - m_d)}}{(t - t_i + c)^p} \tag{6.13}$$

are quenched weights that depend on the true magnitudes and occurrence times of the cluster centers, while

$$z_i = e^{a\epsilon_i} - 1 \tag{6.14}$$

are random variables due to the random noise $\epsilon$. From equation (B.7) in Appendix B.2, we see that the $z_i$ are power-law distributed with an exponent $\alpha = 1/\nu_c a$ that depends inversely on the size of the noise $\nu_c$ and the exponent of the productivity law $a$. In

Figure 6.10, we show the theoretical and simulated pdf of the random variable $z$ for several values of the noise level $\nu_c$.

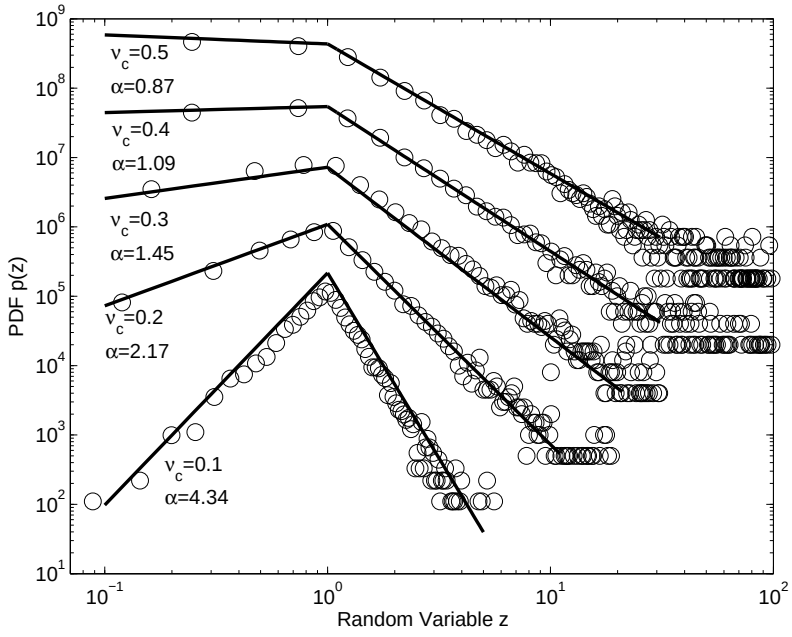


Figure 6.10: Theoretical and simulated probability density function (pdf) of the random variable $z = (\exp(a \cdot \epsilon) - 1)$ for various choices of the scale parameter of the noise $\nu_c$ and assuming $a = 2.3$. The curves are shifted for clarity.

Equation (6.12) together with the knowledge that the random variables $z_i$'s are power law distributed implies that the following proposition is true:

*Proposition 1:* The deviations $\Delta\lambda(t|H_t^c)$ of the perturbed seismic rates from the true seismic rate due to magnitude noise converge to bounded random variables with a distribution having a power law tail for $\Delta\lambda \to \infty$:

$$p(\Delta\lambda) \sim \frac{C_{\Delta\lambda}}{(\Delta\lambda)^{1+\alpha}} \qquad with \quad \alpha = \frac{1}{\nu_c a} \tag{6.15}$$

and with a scale factor given by

$$C_{\Delta\lambda} = \sum_{i=1}^{N_c} w_i^{\alpha} \,, \tag{6.16}$$

where the sum is over the $N_c$ earthquakes in the catalog that occurred before the present time $t$ and the $w_i$'s are given by (6.13).

*Proof:* See Appendix B.3.

*Remarks:*

1. The deviations of a rate estimate based on noisy magnitudes from the exact rate are power-law distributed in the tail. Estimates of a seismic rate may therefore be wildly different from the "true" rate, simply because of magnitude uncertainties.

2. The exponent $\alpha$ determines how broadly the seismic rate estimates are distributed around the "true" rate. The exponent sensitively depends on the noise level $\nu_c$ and the productivity exponent $a$. For the often quoted value $a = \ln(10) = 2.3$ [e.g. *Felzer et al.* [2004]] and for $\nu_c = (0.1, 0.2, 0.3, 0.5)$, one obtains $\alpha = (4.34, 2.17, 1.45, 0.87)$, respectively. Even for relatively small levels of noise $\nu_c \geq 0.22$, the variance of $\Delta\lambda$ does not exist ($\alpha \leq 2$), while for $\nu_c \geq 0.44$, the average of $\Delta\lambda$ does not exist ($\alpha \leq 1$)!

3. The power law tail does not depend on the specific history of the magnitudes and occurrence times. The same power law tail holds for any catalog. On the other hand, the scale factor $C_{\Delta\lambda}$ depends sensitively on the specific realization via the magnitudes and occurrence times (see Proposition 2).

4. Short term earthquake forecasts are directly (sometimes indirectly) taken from such seismic rate estimates. The accuracy of the forecasts cannot be better than the seismic rate estimates.

As a demonstration of the strong fluctuations, Figure 6.11 shows simulations of the seismic rate deviations for various levels of noise $\nu_c = (0.1, 0.2, 0.3, 0.5)$. In each case, we simulate $N_c = 100$ cluster centers according to a Poisson process and randomly draw their magnitudes from the GR law. We then choose a time lag $dt$ (days) between the last event in the cluster center process and the time at which we want to evaluate (forecast) the seismic rate. We calculate the "true" seismic rate from equation (6.6), using the parameter set $\theta = \{\beta = 2.3, \lambda_c = 1, k = 0.01, a = 2.3, m_d = 3, c = 0.001, p = 1.2\}$ and using the "true" simulated magnitudes. Next, we generate "perturbed" catalogs by adding noise to the simulated magnitudes $100,000$ times according to the double exponential noise defined in (6.1). For each perturbed catalog, we recalculate the "perturbed" seismic rate as before using equation (6.6) but replacing the "true" magnitudes by noisy magnitudes.

Figure 6.11 shows fixed kernel density estimates of the normalized differences between the perturbed and the true rates. In each panel, the deviations are shown for four different noise levels $\nu_c = (0.1, 0.2, 0.3, 0.5)$. The different panels (top to bottom) are different choices of the time lag $dt = \{0.0001, 0.1, 1, 10\}$ days since the last event in the process at which the rates are calculated. Figures 6.12 and 6.13 show double logarithmic plots of the survivor functions for the two extreme cases $dt = 0.0001$ and $dt = 10$, respectively.

The seismic rate estimates are very broadly distributed, but with different dependence on the specific catalog, depending on the time horizon $dt$. For small $dt$ (smaller than the average inter-event time $1/\lambda_c$ between successive main shocks), the only relevant event that determines the distribution of the perturbed rates is the last one. Therefore, the (normalized) distributions for small $dt$ are almost identical to the distribution of the random variable $z$ (up to the weight pre-factor associated with the last event). However, as $dt$ increases to become comparable to and

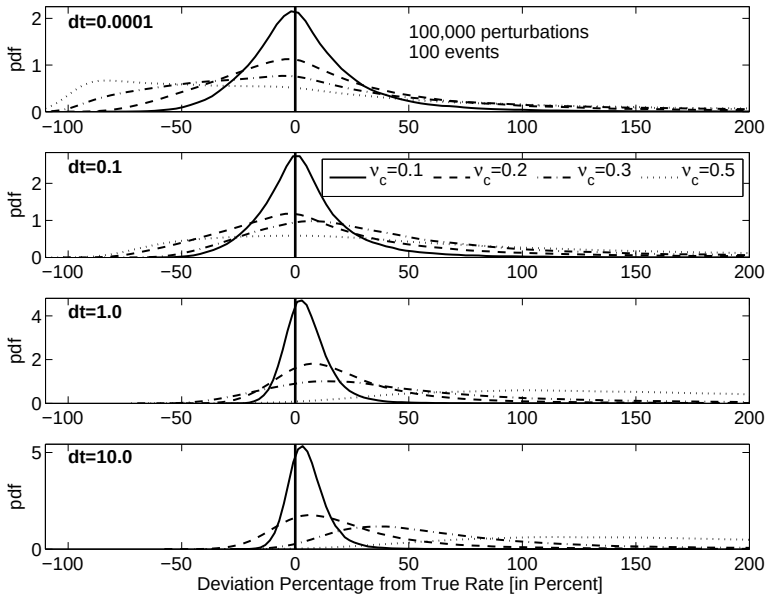


Figure 6.11: Simulated probability density functions of the "perturbed" seismic rates due to noisy magnitudes shown as a percent deviation from the true rate for noise levels $\nu_c = (0.1, 0.2, 0.3, 0.5)$. We calculate seismic rates at a time lag $dt$ after the 100th cluster center event. From top to bottom, $dt = (0.0001, 0.1, 1, 10)$. Seismic rate estimates (and hence forecasts) are wildly distributed around the true value (solid vertical lines).

larger than $1/\lambda_c$, the rate is no longer dominated solely by the last event. Previous events can become increasingly significant in determining the sum compared to the last event. Consider two earthquakes of similar magnitudes occurring at $t_1 = 0$ and $t_2 = 10$ respectively. At time $t = 11$, $t - t_2 = 1$ and $t - t_1 = 11$ so that earthquake 2 will dominate at early times by the effect of the Omori law on the weights, since $w_2/w_1(t = 11) = (11 + c)^p/(1 + c)^p \sim 11$. But at $t = 20$, say, the weight $w_1$ of the first earthquake has decreased only by a factor $2^p$ while the

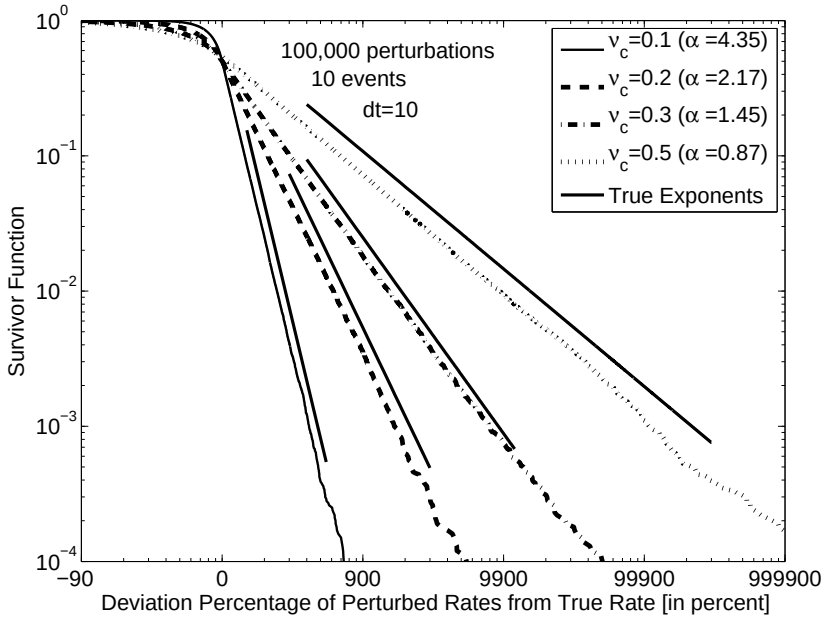


Figure 6.12: Survivor functions of the simulated "perturbed" seismic rate estimates (and forecasts) shown as a deviation in percent from the simulated "true" rate for noise levels $\nu_c = (0.1, 0.2, 0.3, 0.5)$ and time lag since the last event in the process $dt = 0.0001$. The survivor functions are approximately parallel to the straight lines which are guides to the eye with theoretically predicted exponents given by $\alpha = (4.34, 2.17, 1.45, 0.87)$, respectively. Even for small noise level $\nu_c = 0.2$, 10 percent of the rate estimates over-predict the rate by 100 percent!

weight $w_2$ of the second earthquake has decreased by a factor $10^p$, so that the ratio is now $w_2/w_1(t = 20) = (20 + c)^p/(10 + c)^p \sim 2$. Since the weights are so strongly stochastic, each particular catalog realization will have a different number of events with vastly different weights that are "felt" by the rate as $dt$ increases.

We give some concrete numbers for illustration for two cases, $dt = 0.0001$ (Figure 6.12) and $dt = 10$ (Figure 6.13). With our parameter choices $c = 0.001$ and $\lambda_c = 1$,

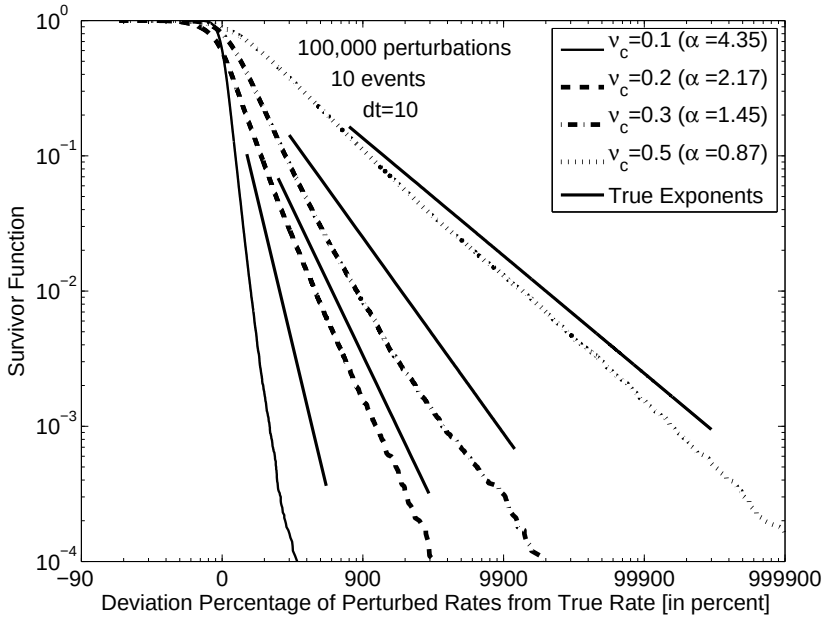


Figure 6.13: Same as Figure 6.12 except that perturbed and true rates are evaluated at $dt = 10$: log-log plots of the survivor functions of the simulated "perturbed" seismic rate estimates (and forecasts) shown as a deviation in percent from the simulated "true" rate for noise levels $\nu_c = (0.1, 0.2, 0.3, 0.5)$. The slopes of the straight lines are given by the asymptotic predicted exponents $\alpha = (4.34, 2.17, 1.45, 0.87)$, respectively.

this ensures that $dt = 0.0001$ corresponds to the regime dominated by the last event in the catalog while the case $dt = 10$ will show the impact of many past events. For $dt = 0.0001$ and realistic $\nu_c = 0.2$, $80\%$ of the seismic rate estimates (and hence forecasts) deviate by more than $10\%$; almost two thirds are off by more than $20\%$; and almost one third are off by more than $50\%$. Even fluctuations of $100\%$ occur $11\%$ of the time. For larger levels of noise, much stronger fluctuations can occur. For instance, $28\%$ of rate estimates are off by more than $100\%$ for $\nu_c = 0.5$. For the case where seismic rate estimates (forecasts) are made for $dt = 10$ after the last event, the

percentages depend strongly on the particular cluster center realization.

Apart from the dependence on a particular realization, there is another consequence of the increasing importance of previous events with time since the last event. As previous events are summed over, the distribution of the sum tends towards its stable law (either Gaussian or Lévy). For $\alpha > 2$ ($\nu_c < 0.22$), the central limit theorem applies to the body of the distribution (though not the tails, see e.g. *Sornette* [2004]) and tends to organize the distribution of the deviations towards a Gaussian. At the same time, only a finite number of terms effectively control the rate so that there is no asymptotic convergence. Nevertheless, there may be a tendency for the body of the distribution to become Gaussian. For $0 < \alpha < 2$ ($\nu_c > 0.22$), in contrast, the distribution will tend towards a Lévy law with a power law tail exponent equal to $\alpha$, keeping intact the original power law tail.

This discussion can be summarized by the following proposition, which emphasizes that the scale factor $C_{\Delta\lambda}$ is "non-self-averaging" [*Mézard et al.*, 1987].

*Proposition 2:*

1. For $p\alpha = \frac{p}{\nu_c a} > 1$, the scale factor $C_{\Delta\lambda} = \sum_i^{N_c} w_i^\alpha$ given in *Proposition 1* (or the typical scale $C_{\Delta\lambda}^{1/\alpha}$) of the distribution of the deviations of the perturbed rates from the true rate converges to a finite value as $N_c \to \infty$.

2. For $p\alpha = \frac{p}{\nu_c a} \leq 1$, the scale factor $C_{\Delta\lambda} = \sum_i^{N_c} w_i^\alpha$ diverges $\to \infty$ as $N_c \to \infty$. This means that the deviations $\Delta\lambda(t|H_t)$ of the perturbed rates from the true rate diverge as the duration of the catalog used to calculate them increases without bound.

3. In the regime $p\alpha > 1$ for which $C_{\Delta\lambda}$ converges almost surely to a finite value as $N_c \to \infty$, $C_{\Delta\lambda}$ remains a random variable sensitively dependent on the specific

catalog, $C_{\Delta\lambda}$ being distributed according to a non-degenerate distribution. In symbols,

$$C_{\Delta\lambda} \;\rightarrow\; \langle C_{\Delta\lambda} \rangle \qquad as \qquad N_c \rightarrow \infty \;, \tag{6.17}$$

where $\langle C_{\Delta\lambda} \rangle$ denotes the average of $C_{\Delta\lambda}$ over many different catalogs. This means that the value $C_{\Delta\lambda}$ is catalog specific and changes from catalog to catalog and from realization to realization. This property is known as "lack of self-averaging" [*Mézard et al.*, 1987].

*Proof:* See Appendix B.4.

The divergence of the scale factor $C_{\Delta\lambda}$ and of the deviations $\Delta\lambda(t|H_t)$ described by item 2 of Proposition 2 occurs only for rather large magnitude errors. For instance, for $p = 1.2$ and $a = 2.3$, the noise level needs to be larger than $\nu_c = 0.52$ for $p\alpha = \frac{p}{\nu_c a} \leq 1$ to hold. According to our previous analysis in Section 2, we can expect $\nu_c$ to lie in the range $0.1 - 0.3$ typically, so that the regime $p\alpha = \frac{p}{\nu_c a} > 1$ in Proposition 2 is likely to be the most relevant one.

In summary, for large $dt$, there is a trade-off between some averaging for small noise levels (i.e. a tendency towards a Gaussian shape), the dependence on the specific catalog realization (in particular, the last few events) and the power law tail proven in Proposition 1. These effects can be seen by comparing the survivor functions in Figure 6.12, which for $dt = 0.001$ are pure power laws for $\Delta\lambda > 0$, with Figure 6.13 for $dt = 10$. For $\nu_c = 0.3$ and $\nu_c = 0.5$ (Lévy regime), the power law remains largely intact even for the body of the distribution. For $\nu_c = 0.1$ and $\nu_c = 0.2$, the body is not well approximated by the power law, while the tails still show significantly larger outliers than expected from a Gaussian.

We now return to the explicit values and percentages of the deviations for the case

$dt = 10$. For realistic $\nu_c = 0.2$, about $70\%$ are off by more than $10\%$; $45\% - 55\%$ deviate by more than $20\%$; roughly $20\%$ of seismic rate estimates over-predict the true value by more than $50\%$.

In contrast to the noise levels for which $\alpha > 2$, the fluctuations actually increase for $\nu_c = 0.3$ and $\nu_c = 0.5$ compared to the case $dt = 0.0001$. For $\nu_c = 0.5$, $90\% - 95\%$ are off by more than $10\%$; $80\% - 90\%$ are off by more than $20\%$; $70\% - 75\%$ are off by more than $50\%$; and roughly one half are off by more than $100\%$!

To conclude this section, we restate the main results: (i) Seismic rates estimated from noisy data deviate strongly from their "true" value; (ii) If magnitude uncertainties are exponentially distributed (or more broadly), as we have shown in section 2, then these deviations are power law distributed in the tail with an exponent $\alpha = 1/a\nu_c$; (iii) We believe there is no law of large numbers that can adequately describe the entire distribution of the deviations. However, the scale factor (or the typical size of the deviations) can be shown to diverge almost surely if $p\alpha < 1$ and to be distributed extremely broadly if $p\alpha > 1$. These results demonstrated rigorously have been illustrated with numerical simulations. The next section investigates the impact of the propagating uncertainties on the test scores that forecasts receive in earthquake predictability experiments such as CSEP.

## 6.5 Impact of Magnitude Uncertainties on Model Forecasts and Their Evaluation in Consistency Tests

Seismic rates estimated via some probabilistic model at time $t$ in the future as a function of past seismicity are routinely used as direct estimates of a forecast (the expected number of events at time $t$) [e.g. *Helmstetter et al.* [2005a]; *Gerstenberger et al.* [2005]]. In rare cases, forecasts are generated by averaging over Monte Carlo simula-

tions which nevertheless remain parametrized by the estimated seismic rate. Forecasts are therefore either equal to, proportional to or strongly correlated with seismic rate estimates. As a consequence, they suffer from the same fluctuations as seismic rate estimates.

How important are these spurious fluctuations in the evaluation of forecasts in prediction experiments? Can good forecasts perform poorly solely due to magnitude noise? Can accurate models be rejected by test scores because noisy magnitudes influenced the rate estimation? How sensitive are evaluation procedures to this source of noise?

We address these questions by designing the following numerical experiment: We test forecasts based on noisy magnitudes against a hypothetical "reality", chosen as the rate based on exact magnitudes, to see whether the noisy forecasts are rejected. We pretend that the exact seismic rate as calculated from the model equation (6.7) in each space-time bin is "reality" according to which earthquakes actually occur. Since the exact rate is not an integer, we assume that observations are drawn from a Poisson distribution with mean equal to the seismic rate in each bin. We further call the noisy forecasts "models" ($j = 1 \ldots M$). We then constructed a miniature CSEP testing center which tests these "models" against "reality" according to their consistency with the "observations" in a manner entirely equivalent to the proposed scenario [*Schorlemmer et al.*, 2007; *Schorlemmer and Gerstenberger*, 2007].

The testing center uses the likelihood (L) and number (N) tests as consistency criteria for evaluating forecasts. These are currently used for the five year time-independent forecasts which have already been submitted [*Schorlemmer et al.*, 2007]. The test were used in previous studies of forecasting [*Kagan and Jackson*, 1994, 1995] and were further explained by *Jackson* [1996].

We calibrated the numerical experiment to mimic California as much as possible to

create realistic conditions under which actual forecast evaluation will take place. This meant choosing a realistic model, realistic model parameters and realistic space-time bins.

If we find that "models" (i.e. forecasts based on noisy data) are rejected more frequently by the consistency tests than according to the chosen confidence limit, we can draw the following conclusions: (i) the "true" model with "true" parameters may be rejected in a realistic one year test period because noisy magnitude observations affected its forecasts, (ii) the outcomes of the likelihood and number consistency tests are therefore sensitive to observational uncertainties that affected the generation of the forecasts.

First we describe the simulations and their evaluation by likelihood methods before presenting our results. Our numerical experiment can be separated into two steps: (i) the simulation of the exact seismic rates in each space-time bin and the corresponding noisy forecasts, and (ii) the evaluation of the noisy forecasts with respect to the exact forecast using likelihood tests. Figure 6.14 graphically explains our numerical experiment and should be used for reference.

### 6.5.1 Simulation of Exact Seismic Rates and Noisy Forecasts

The simulation of the exact seismic rates and the noisy forecasts proceed according to the following steps.

1. **Choose a model, its parameters and the test area**: The impact of magnitude uncertainties on forecasts will depend on the specific model and potentially its parameters. We chose a spatio-temporal Poisson cluster model defined in equation (6.7) to capture the main ingredients of the popular short-term forecasting models. We used parameters and a spatial test area consistent with Californian

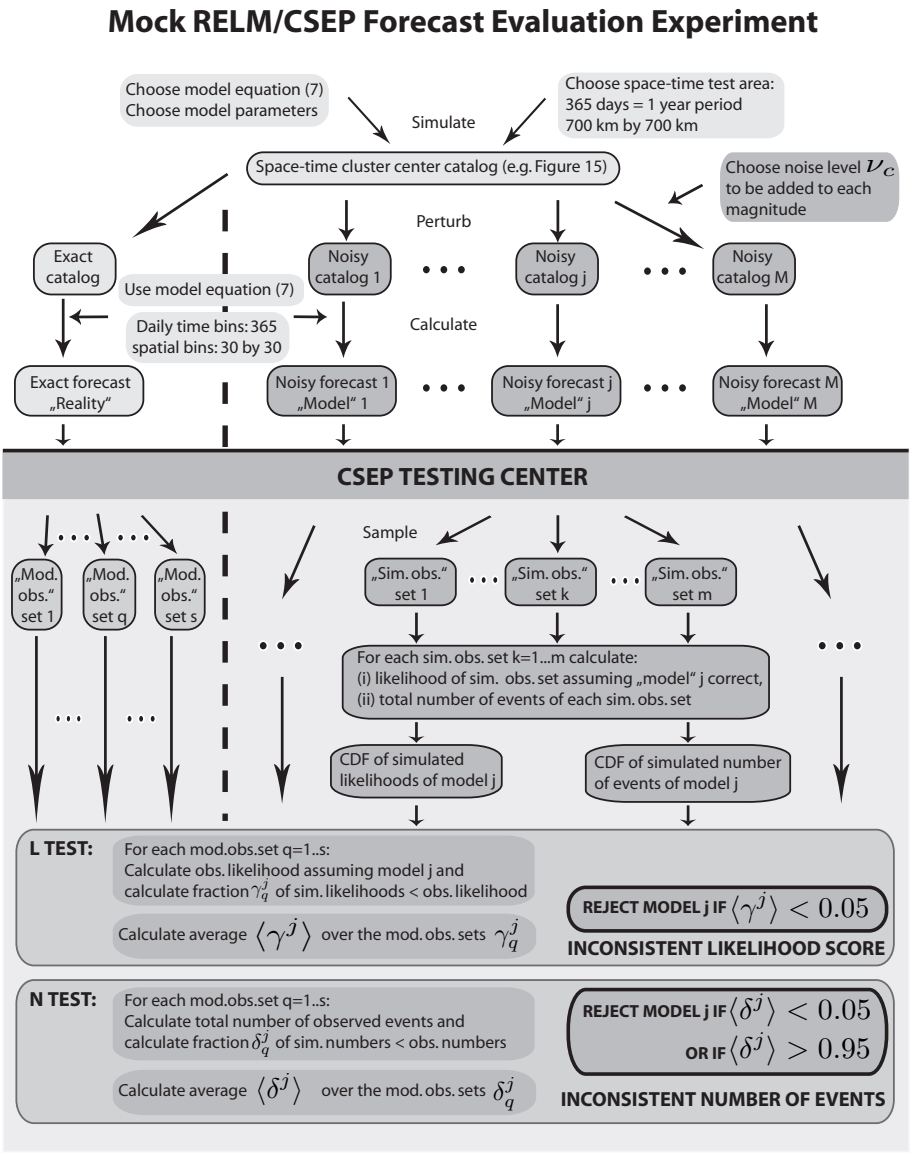


Figure 6.14: Diagram explaining the numerical experiment designed to study the influence of magnitude noise on forecasts and their evaluation in daily earthquake forecast experiments such as RELM or CSEP. We mimic California both in spatial test area and in model parameters and perform the likelihood (L) and number (N) test for daily forecasts over the period of one year using a spatio-temporal Poisson cluster center model that captures essential ingredients of most short-term seismicity models. The abbreviations mod., sim. and obs. stand for modified, simulated and observations, respectively.

earthquake data to create realistic conditions (see section 6.5.3 below for parameter values).

2. **Simulate a "learning" catalog:** Using a set of parameters and the model (6.7), we simulate a "learning" catalog from which forecasts are to be generated. Because aftershocks of the cluster centers do not trigger their own aftershocks, they do not affect future seismicity and do not need to be simulated for the "learning" catalog. Only the Poisson cluster center process needs to be simulated. It is a homogeneous space-time Poisson process of constant rate $\lambda_c$ (per unit time and area). Independent magnitudes are drawn from the Gutenberg-Richter law (6.2). See Figure 6.15 for an example of a simulated cluster center process.

3. **Calculate the exact daily seismic rate over one year in each space-time bin $i$:** Divide the test area into spatial cells (bins). Use equation (6.7) and the original parameters to calculate the intensity $\lambda(t, \mathbf{r}|H_t^c)$ for every day over the course of one year at the centers of the spatial cells. Past seismicity (times, locations and magnitudes) enter into equation (6.7). The rate in each space-time bin is calculated by multiplying the rate at the center by the area of the spatial cell for simplicity. Pretend these rates are "reality": the number of actually observed events is drawn from a Poisson distribution with mean equal to this rate (see below).

4. **Create $M$ noisy replicas of the "learning" catalog by perturbing the magnitudes:** Generate $j = 1 \dots M$ different noisy replicas of the original catalog by adding random noise to each magnitude. The random noise is simulated from the Laplace distribution defined in (6.1) characterized by the scale parameter $\nu_c$.

5. **Generate $M$ "noisy" forecasts based on the $M$ noisy catalogs:** Using equation (6.7) and the original parameters, calculate the daily seismic rate for the whole

year in the same manner as the exact rate but using the noisy catalogs. Denote these noisy forecasts as "models" $j = 1 \ldots M$ which we want to compare to the exact seismic rate in a mock CSEP/RELM testing center.

### 6.5.2 Evaluation of the Noisy Forecasts in a Hypothetical RELM/CSEP Testing Center

In the evaluation of the "models" (or noisy forecasts), we follow the recipe of the RELM/CSEP testing procedures, as described below. But first we address an important issue regarding our choice not to use magnitude bins.

When we simulate the learning catalog above, we must choose a magnitude threshold. We set $m_d = 4$ as for the proposed daily forecast competition. Magnitudes for the simulated cluster center events were drawn from the Gutenberg-Richter distribution with $m_d = 4$. We then perturbed this catalog according to a symmetric Laplacian distribution. As a result, some of the simulated cluster centers now have magnitudes that are smaller than the magnitude threshold $m_d = 4$. We note that their influence is very small in the calculation of the noisy seismic rates (forecasts) with respect to the events above the threshold because of the exponential dependence of the productivity on the difference $\exp(a(m_{i_c} - m_d))$. Rather than re-applying the cut-off $m_d = 4$ after the perturbation, we kept the number of events in the learning catalog fixed. We therefore also did not take into account the possibility of magnitudes originally below the threshold becoming "visible" above the threshold due to the addition of noise. These more realistic conditions should be pursued as a next step. In our scenario, the mean magnitude before and after the perturbation remains constant. Any results in the performance of the forecasts are therefore not due to an overall shift in the magnitude distribution.

As a consequence, we did not introduce magnitude bins at the testing stage. Rather,

we assumed that the seismic rates and the forecasts were for one and the same large magnitude bin. This also meant that we did not need to generate magnitudes of the aftershocks of the cluster centers, neither of the exact catalog nor of the noisy catalogs. This helped the numerical efficiency of the computer program. Furthermore, by not testing the magnitude values, we effectively assumed that all simulated magnitudes were completely consistent with the "observed" ones, as the sum over the likelihood values of the magnitude bins equals 1. In other words, we are less stringent than the CSEP/RELM center by not rejecting models because of under- or over-predicted magnitudes.

The CSEP/RELM testing procedure acknowledges the existence of observational uncertainties in the observed earthquake parameters such as magnitude, location or focal parameters. The test center therefore creates so-called "modified" observations from the actually observed earthquake parameters by sampling parameters that are consistent with the observed ones and their uncertainties. In this way, many alternative "realities" or observations are created that are consistent with the actual observations and their uncertainties. The forecasts are tested against each alternative reality and the final rejection is based on the average performance of the forecast against all alternative "realities".

In our hypothetical center, we should therefore create many "modified" observations consistent with the actual observation in the same manner. However, as just stated above, we did not actually generate magnitudes of the events and do not test a forecast's ability to predict magnitudes. The magnitude dimension is effectively eliminated from the evaluation phase by assuming that all simulated magnitudes are consistent with the "observed" ones.

At the same time, we did generate "modified" observations in each space-time bin according to the following logic. The CSEP/RELM test center generates so-called

"simulated" observations from a model's forecast that are consistent with the forecast by drawing numbers from a Poisson distribution with mean equal to the forecast in each space-time bin. This assumes that earthquakes are independent and occur according to a Poisson process in each space-time bin. This is not true for the Earth. However, we decided to assume that our "reality" did indeed occur according to a Poisson process in each bin so that we create favorable conditions for the test to work and to be sure that "models" were not rejected because they violate this assumption of the test.

We therefore created many alternative realizations of observed earthquakes that are consistent with the exact seismic rate in each bin according to a Poisson distribution with mean equal to the seismic rate. In analogy with the RELM/CSEP terminology, we called these alternative "realities" modified observations. For the Earth, these favorable conditions do not hold and may also influence whether a model is rejected.

With reference to Figure 6.14, the various steps of the RELM/CSEP testing procedure are described below. We follow the notation of *Schorlemmer et al.* [2007].

1. **For each of the $\{j = 1 \dots M\}$ "models", generate $\{k = 1 \dots m\}$ "simulated" observations**: The RELM/CSEP test center assumes that earthquakes are independent in each space-time bin $i$ and generates Poisson distributed random variables $\hat{w}^j_{k,i}$ with mean equal to the model's forecast $\lambda^j_i$ in that bin. This is repeated $m$-times to generate $\{k = 1 \dots m\}$ sets of simulated observations for each model.

2. **For each of the $M$ "models", calculate the likelihoods of the $m$ simulated observations**: Assume that the "model" is true and calculate the likelihood of each of the $\{k = 1 \dots m\}$ sets of observations generated from its forecast $\lambda^j_i$ and sum over all space-time bins. These are the simulated likelihood values $\{\hat{L^j_k}, k = 1 \dots m\}$ consistent with a particular model $j$ which will be tested

against the likelihood of the observations below. The simulated likelihood of the jth model of the kth observation set is given by:

$$\hat{L}_k^j = \sum_i -\lambda_i^j + \hat{w}_{k,i} \log \lambda_i^j - \log(\hat{w}_{k,i}!) \tag{6.18}$$

3. **Generate $\{q = 1 \ldots s\}$ "modified" observations ("alternative realities") from the exact seismic rate**: In each space-time bin $i$, generate a "modified" observation $\tilde{w}_i^q$ by drawing a random variable from a Poisson distribution with mean equal to the exact seismic rate in that bin. Repeat $s$-times to generate $s$ sets of "modified observations" against each of which each "model" is tested.

4. **For each of the $M$ "models", calculate the likelihoods of the $s$ "modified" observations**: Assume that the jth "model" is true and calculate the likelihood $\tilde{L}_q^j$ of observing each of the $\{q = 1 \ldots s\}$ "modified" observations $\tilde{w}_i^q$:

$$\tilde{L}_q^j = \sum_i -\lambda_i^j + \tilde{w}_i^q \log \lambda_i^j - \log(\tilde{w}_i^q!) \tag{6.19}$$

5. **For each of the $M$ "models", calculate the fraction $\gamma_q^j$ of simulated likelihoods $\hat{L}_k^j$ less than each of the $s$ observed likelihoods $\tilde{L}_q^j$**: For each model $j$ and each set $q$ of modified observations, compute the fraction $\gamma_q^j$ of simulated likelihoods $\{\hat{L}_k^j, k = 1 \ldots m\}$ less than the observed likelihood $\tilde{L}_q^j$. Repeat for each of the $q = 1 \ldots s$ observed likelihoods for each of the $M$ "models". The fraction $\gamma_q^j$ measures whether an observed likelihood value is consistent with the values expected from a particular "model".

6. **For each of the $M$ "models", calculate the mean fraction $\langle \gamma^j \rangle$ from the $s$ values of $\gamma_q^j$**: Because many "modified" observations are consistent with our chosen "reality", we average the fraction $\gamma^j$ over all sets of "modified" observations for each model (noisy forecast). The mean fraction $\langle \gamma^j \rangle$ measures whether

the observed likelihood values fall within the center of the simulated likelihoods given a particular model.

7. **For each of the $M$ "models", calculate the simulated total number of events $\hat{N}_k^j$ from each of the $\{k = 1 \ldots m\}$ simulated observations**: Generate the distribution of the simulated total number of events in the test region consistent with each model.

8. **Calculate the "modified" total number of events $\tilde{N}_q$ from each of the $\{q = 1 \ldots s\}$ modified observations**: For each of the possible "realities" or modified observations, sum the modified observations over all bins to obtain the modified total number of events $\tilde{N}_q$.

9. **For each of the $M$ "models", calculate the fraction $\delta_q^j$ of simulated total numbers of events $\hat{N}_k^j$ less than the modified number of events $\tilde{N}_q$ for each of the $\{q = 1 \ldots s\}$ modified observations**: Compute the fraction $\delta_q^j$ of simulated numbers of events $\{\hat{N}_k^j, k = 1 \ldots m\}$ less than the qth modified total number of events $\tilde{N}_q$ for each modified observation and each model. The fraction $\delta_q^j$ measures whether the simulated numbers of events are consistent with the observed number.

10. **For each of the $M$ "models", calculate the mean fraction $\langle \delta^j \rangle$ from the $s$ values of $\delta_q^j$**: Again, because many modified observations are consistent with the actual observation, average the fraction $\delta_q^j$ over all of its $s$ values for each model (noisy forecast). The mean fraction $\langle \delta^j \rangle$ measures whether the modified observations are on average consistent with the simulations.

11. **Perform L test:** Reject model $j$ if $\langle \gamma^j \rangle < 0.05$. This indicates that the observed likelihood values are inconsistent with the model. According to Schorlemmer et al. (2007), the L test is one-sided.

12. **Perform N test:** Reject model $j$ if $\langle \delta^j \rangle < 0.05$ or if $\langle \delta^j \rangle > 0.95$. This indicates that the observed number of events are inconsistent with the model.

### 6.5.3 Test Area and Model Parameter Choices

To simulate catalogs and generate exact forecasts, we needed to choose the spatial test area, the number of spatial bins, the parameters of the model and the temporal bins (for which forecasts are issued and evaluated). To mimic a daily forecast competition in California in an earthquake prediction experiment such as RELM, we chose temporal bins of one day and issued forecasts over the course of an entire year always using all information available up to the day of the forecast. We assumed a square spatial area of $700$ km by $700$ km to approximate the size of California.

To generate synthetic catalogs and forecasts, we needed to decide on a set of parameters $\theta = \{\beta, m_d, \lambda_c, k, a, c, p, \mu\}$ for the model (6.7). We decided on $m_d = 4$ to mimic the proposed RELM/CSEP daily forecast competition and on $\beta = 2.3$. Inverting the parameters of the Poisson cluster process requires knowledge of the entire branching structure (being able to identify which events are cluster centers and which are their dependent aftershocks). Lacking such knowledge, the maximization of the likelihood must be performed over all possible branching structures. Given the complexity, we decided instead to use parameter values based on an ETAS model inversion for southern California by *Helmstetter et al.* [2006] [their model 2 in Table 1] and adjusted them appropriately for our model, magnitude threshold and spatial test area. Their background rate was multiplied by $10^{-2}$ to adjust to the higher magnitude threshold, and multiplied by $3$ for the increased spatial area, resulting in $\lambda_c = 0.063$. We also multiplied $k$ by $10^{-2}$ to obtain $k = 0.013$. We used their other parameters without change $\{a = 0.43 \times \log(10), c = 0.0035, p = 1.19, \mu = 2\}$. We checked that the

simulated total number of expected events per year over the entire region given these parameter choices was comparable to yearly rates above M4 in California over the last twenty years (from about 50 up to 250).

Because of computational limitations, we had to restrict the number of spatial bins to 30 by 30 and the number of perturbations of the original catalog to M=10. However, we were still able to generate $m = 1000$ simulated observations for each noisy forecast and $s = 1000$ modified observations for the exact forecast.

In this article, we aim to establish only whether examples exist in which good models are rejected based solely on realistic magnitude uncertainties. A more in-depth study which explores the model, parameter and bin space is certainly required to establish robust confidence limits for testing the consistency of model forecasts with observations.

### 6.5.4 Simulations and Results

Figure 6.15 shows an example of a simulated cluster center catalog: the top panel shows the spatial distribution of the cluster centers in the spatial test area $700$ km by $700$ km. The middle panel shows the magnitudes of the cluster centers against time in days over the course of one year. The bottom panel shows the rate for both cluster centers and their aftershocks calculated from the model (6.7), summed over all spatial bins.

We checked that setting the noise level to zero $\nu_c = 0.0$ does not result in any "models" being rejected. Table 6.1 shows the results of the mock RELM/CSEP evaluation of the daily forecasts of 10 unperturbed forecasts over the period of one year. None of the "models" are rejected by the L and N tests. Apart from the mean fractions $\langle \delta^j \rangle$ and $\langle \gamma^j \rangle$, we also calculated their standard deviations $\sigma_\gamma$ and $\sigma_\delta$. Table 6.1 shows both fractions to be right in the middle of the simulated distributions, indicating strong

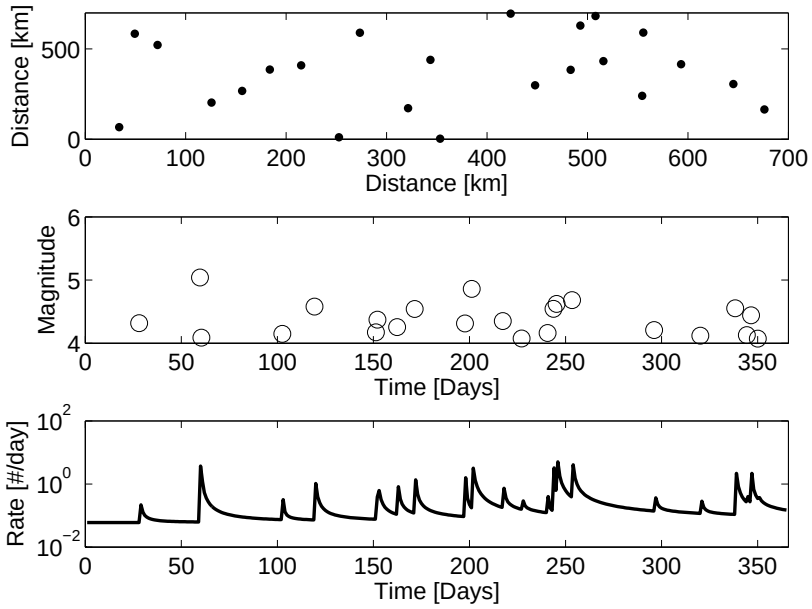


Figure 6.15: Example of a simulated cluster center catalog and resulting conditional intensity including aftershocks. Top: spatial distribution. Middle: Magnitudes of cluster centers against time over the course of one year. Bottom: Conditional intensity rate per day calculated using the model equation (6.7) but summed over all bins. The rate is used as the "exact forecast" from which modified observations are generated.

consistency between the models and the observations, as should be expected.

Introducing just a little bit of noise changes the situation. For $\nu_c = 0.1$, we found that the N test rejects 2 models at the 90 $\%$ confidence limit (see Table 6.2). We found that the difference between the simulated total number of expected events of a model and the actual expected number based on the exact rate was a good indicator for the model's performance in the N test, as should be expected. In keeping with the statement by *Schorlemmer et al.* [2007] that the L test is one-sided, we do not reject models for which $\langle \gamma^j \rangle > 0.95$. In this case, the (two-sided) L test would have rejected

| "Model" | E(N) | $\langle\gamma\rangle$ | $\sigma_\gamma$ | $\langle\delta\rangle$ | $\sigma_\delta$ |
|---|---|---|---|---|---|
| 1 | 143.61 | 0.4996 | 0.2835 | 0.5361 | 0.2849 |
| 2 | 143.61 | 0.5106 | 0.2795 | 0.5162 | 0.2866 |
| 3 | 143.61 | 0.4895 | 0.2887 | 0.5279 | 0.2866 |
| 4 | 143.61 | 0.5051 | 0.2791 | 0.5274 | 0.2841 |
| 5 | 143.61 | 0.5028 | 0.2837 | 0.5185 | 0.2910 |
| 6 | 143.61 | 0.5138 | 0.2864 | 0.5177 | 0.2870 |
| 7 | 143.61 | 0.4913 | 0.2897 | 0.5350 | 0.2975 |
| 8 | 143.61 | 0.5131 | 0.2882 | 0.5157 | 0.2921 |
| 9 | 143.61 | 0.4948 | 0.2866 | 0.5136 | 0.2870 |
| 10 | 143.61 | 0.5110 | 0.2848 | 0.5154 | 0.2846 |

Table 6.1: Results of the mock RELM/CSEP experiment of daily forecasts over the period of one year for $\nu_c = 0.0$: We checked that "models" are not rejected by the tests when the data is exact and no noise is present. "Models" are forecasts generated from equation (6.7) using a noisy cluster center process which was perturbed from the original one by adding random noise of scale $\nu_c$ to the magnitudes. The first column contains different perturbations of the original catalog corresponding to different forecasts or "models" (which, for $\nu_c = 0.0$, are all equal). The second column is the total expected number of events obtained by summing all daily forecasts over all spatial bins over the one year period. The expected number of events of the original catalog was $143.61$. The third column shows the fraction $\langle\gamma\rangle$ of the $m$ simulated likelihoods less than the observed likelihood, averaged over all $s$ modified observations. The fourth column shows the standard deviation of the $\gamma$ values. The fifth column shows the fraction $\langle\delta\rangle$ of the $m$ simulated numbers of events less than the observed number of events, averaged over all $s$ modified observations. The sixth column shows the standard deviation of the $\delta$ values.

| "Model" | E(N) | $\langle\gamma\rangle$ | $\sigma_\gamma$ | $\langle\delta\rangle$ | $\sigma_\delta$ |
|---|---|---|---|---|---|
| 1 | 117.31 | 0.3361 | 0.2647 | 0.9011 | 0.1552 |
| 2 | 187.36 | 0.9546 | 0.0880 | 0.0018 | 0.0109 |
| 3 | 150.23 | 0.5262 | 0.2921 | 0.2038 | 0.2106 |
| 4 | 145.38 | 0.5981 | 0.2729 | 0.2858 | 0.2461 |
| 5 | 665.09 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 6 | 143.25 | 0.4935 | 0.2870 | 0.3321 | 0.2574 |
| 7 | 139.40 | 0.4826 | 0.2812 | 0.4335 | 0.2749 |
| 8 | 162.28 | 0.6158 | 0.2705 | 0.0627 | 0.1054 |
| 9 | 124.38 | 0.2920 | 0.2573 | 0.7817 | 0.2321 |
| 10 | 130.70 | 0.3203 | 0.2639 | 0.6454 | 0.2759 |

Table 6.2: Same as Table 6.1 but now perturbing the original catalog from which forecasts ("models") are generated by introducing noise $\nu_c = 0.1$. The total expected number of events based on the exact forecast was $E(N) = 136.31$. The L test does not reject any models while the N test rejects 2 models.

the same models as the N test. Because rejecting 2 models out of 10 at 90% confidence may simply be the expected false negative errors, we additionally performed simulations for 50 models. Of these, 8 models were rejected by the N test, none by the L test. Thus in total, 10 models out of 60 models were rejected, indicating slightly higher than expected rejections at 90% confidence.

Table 6.3 shows the results for simulations with the noise level set to $\nu_c = 0.2$. The cluster center process of this simulation is shown in Figure 6.15 along with the calculated seismic rate according to model (6.6). The N test rejects 9 out of 10 models because they over-predict the number of observed events. A two-sided L test would have rejected 3 models, but the one-sided L test does not reject any models. In contrast

| "Model" | E(N) | $\langle\gamma\rangle$ | $\sigma_\gamma$ | $\langle\delta\rangle$ | $\sigma_\delta$ |
|---|---|---|---|---|---|
| 1 | 124.11 | 0.8077 | 0.2072 | 0.0292 | 0.0577 |
| 2 | 120.29 | 0.5917 | 0.2845 | 0.0495 | 0.0892 |
| 3 | 135.62 | 0.8260 | 0.1984 | 0.0042 | 0.0133 |
| 4 | 98.97 | 0.3816 | 0.2730 | 0.4288 | 0.2813 |
| 5 | 156.55 | 0.9810 | 0.0547 | 0.0000 | 0.0004 |
| 6 | 170.75 | 0.9949 | 0.0222 | 0.0000 | 0.0000 |
| 7 | 246.27 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 8 | 128.59 | 0.8307 | 0.2033 | 0.0150 | 0.0369 |
| 9 | 121.60 | 0.7972 | 0.2208 | 0.0367 | 0.0713 |
| 10 | 139.99 | 0.8710 | 0.1719 | 0.0020 | 0.0074 |

Table 6.3: Same as Table 6.2 but now perturbing with stronger noise $\nu_c = 0.2$. The total expected number of events based on the exact forecast was $E(N) = 95.81$. The L test does not reject any models while the N test rejects 9 models.

to the case for $\nu_c = 0.1$, the fluctuations strongly impact the forecasts.

Results for simulations with noise level $\nu_c = 0.3$ are summarized in Table 6.4. Note that the values of $\langle\gamma^j\rangle$ are now fluctuating very strongly. The one-sided L test now rejects one model while the N test rejects 7 models. Clearly, the noisy forecasts are no longer consistent with the observed likelihood values and numbers of events. Note that model 5 predicts more than 10 times the actually expected number of events, reflecting the extreme fluctuations we proved in the previous sections. It is also interesting that the L test rejects model 2, which passes the N test. This shows that the daily expected numbers can fluctuate but in a sense "average out" over the course of one year, but the likelihood scores of each day keep a "memory" of the bad predictions. This exemplifies the complementary properties of the two tests.

| "Model" | E(N) | $\langle\gamma\rangle$ | $\sigma_\gamma$ | $\langle\delta\rangle$ | $\sigma_\delta$ |
|---|---|---|---|---|---|
| 1 | 329.28 | 1.0000 | 0.0007 | 0.0000 | 0.0000 |
| 2 | 179.76 | 0.0406 | 0.0714 | 0.6250 | 0.2806 |
| 3 | 398.07 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 4 | 162.82 | 0.0932 | 0.1311 | 0.8936 | 0.1533 |
| 5 | 3258.11 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 6 | 249.14 | 0.8388 | 0.1896 | 0.0015 | 0.0058 |
| 7 | 323.39 | 0.9863 | 0.0418 | 0.0000 | 0.0000 |
| 8 | 204.90 | 0.1300 | 0.1600 | 0.1853 | 0.2159 |
| 9 | 257.31 | 0.8528 | 0.1830 | 0.0007 | 0.0021 |
| 10 | 288.11 | 0.6957 | 0.2588 | 0.0000 | 0.0000 |

Table 6.4: Same as Table 6.2 but now perturbing with stronger noise $\nu_c = 0.3$. The total expected number of events based on the exact forecast was $E(N) = 186.65$. The L test rejects 1 model while the N test rejects 7 models.

The case $\nu_c = 0.5$ is shown in Table 6.5. All models are rejected by the N test, indicating systematically larger forecasts.

These simulations show that, as $\nu_c$ increases, the models tend to forecast numbers of events that are larger than the true value: the larger $\nu_c$, the larger the effect, and therefore the more probable the rejection by the N test. This results from the fact that, while the distribution (6.1) of magnitude errors is found (and assumed in our simulations) to be approximately symmetric, the impact of a magnitude error is strongly asymmetric when comparing negative and positive deviations from the true magnitude, due to the exponential dependence of the productivity law (6.3).

Because the tests rejected more models for $\nu_c = 0.2$ than for $\nu_c = 0.3$, we believe that we are not sampling the actual fluctuations with 10 models. More models (per-

| "Model" | E(N) | $\langle\gamma\rangle$ | $\sigma_\gamma$ | $\langle\delta\rangle$ | $\sigma_\delta$ |
|---|---|---|---|---|---|
| 1 | 349.73 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 2 | 11715.37 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 3 | 691.43 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 4 | 9150.17 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 5 | 1248.60 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 6 | 341.29 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 7 | 457.51 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 8 | 645.55 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 9 | 5593.47 | 1.0000 | 0.0000 | 0.0000 | 0.0000 |
| 10 | 217.19 | 0.4032 | 0.2537 | 0.0000 | 0.0000 |

Table 6.5: Same as Table 6.2 but now perturbing with stronger noise $\nu_c = 0.5$. The total expected number of events based on the exact forecast was $E(N) = 92.15$. The L test does not reject any models while the N test rejects all 10 models.

turbed catalogs) are needed to characterize precisely how the confidence limits of the tests are affected. In this article, we showed simply that the stated confidence limits as we would like to interpret them (that the model is inconsistent with 90 $\%$ confidence) are clearly not adequate.

### 6.5.5 Discussion of Mock RELM/CSEP Predictability Experiment

The results from the mock RELM/CSEP evaluation of forecasts generated from noisy data against the exact seismic rate based on exact data indicate that noisy forecasts fluctuate wildly and may therefore be rejected by the RELM/CSEP consistency tests. Even conservative levels of magnitude noise of $\nu_c = 0.2$ impact forecasts so strongly

that they are easily rejected by the likelihood and number tests. In other words, the L and N tests are sensitive to observational uncertainties that entered in the creation of the forecasts. As a consequence, when considering actual results from L and N tests based on comparing a real model with real data, one should keep in mind the possibility that the forecast contains noise which may influence severely the performance of a short-term model. The supposed confidence limits may be misleading as they do not take into account uncertainties in the forecast. Some models may be rejected purely because of forecasts generated from noisy earthquake catalogs, while others may appear to be consistent with the data (more often than expected given the RELM/CSEP confidence limits).

We do not expect these "wrongful" rejections to stop if the tests are performed over a longer period of time. Each day is separately scored according to forecast and observations and the daily forecast will always be strongly fluctuating. Extending the evaluation period to two years, for instance, would not solve the problem.

We emphasize that we have completely isolated the effect of magnitude uncertainties and assumed everything else to be known. We have shown that, in this scenario, magnitude uncertainties lead to strongly fluctuating forecasts. While a comprehensive study of uncertainty in data, parameters, forecasts, observations and their trade-offs should be encouraged, we expect that there will be no simple formula to "correct" the forecasts, tests or interpretations.

Rather, the propagation of data and parameter uncertainties needs to be carefully examined in each specific model and accounted for in the forecasts. The resulting distribution of forecasts can most likely not be captured by one value such as the expectation. In fact, we have shown above that, depending on the noise, we should expect extremely large variations that cannot be represented by one number per bin.

The L and N tests both assume that earthquakes are independent in each space-time

bin and that observations consistent with the forecast are Poisson distributed random variables. In this article, we did not study the implications of the first assumption. It would seem, however, that the assumption of independence will be strongly violated during active aftershock sequences - the days when clustering models can actually be tested on observed clustering.

But from the standpoint of our results, the second assumption (that observations consistent with a model are Poisson distributed with mean equal to the forecast) may need to be relaxed because of uncertainties in the forecasts, whether due to noisy magnitudes used to generate forecasts, parameter uncertainties or other sources. Instead, models will need to provide the entire likelihood distribution in each space-time bin. Apart from likelihood and number consistency tests, methods for alarm-based earthquake predictions are also equipped to deal with full forecast distributions [see, e.g., *Zechar and Jordan* [2007]].

There is a second reason for allowing forecasts to be specified as full distributions. The idiosyncrasies of a model may cause consistent observations to be distributed completely differently than a Poisson distribution. While a certain actual observation may not be consistent with a Poisson distribution given the mean rate of a model forecast, the observation may still be consistent with the model. Specifying the entire distribution of a forecast is computationally much more demanding, but it is the only way to guarantee that forecasts accurately reflect the scientific hypotheses of the model along with all sources of uncertainties involved in the generation of the forecast.

The power of the "non-Poisson" L and N tests may appear weaker because less models are rejected, but this is simply a reflection of the potential stochasticity of the model and its real uncertainties. If the tests are adjusted to the distribution of each forecast, then the confidence limits can be interpreted appropriately.

A trivial way for a model to pass the "non-Poisson" consistency tests would be to

specify extremely broad distributions so that whatever is observed falls into the center of the distribution. Most clustering models are in fact already very broadly distributed so that this may reflect some scientific truth about seismicity. But in case these distributions are too broad, there exists also the likelihood ratio test, which compares models against each other and would be able to reject overly "dilute" forecasts against peaked ones that are accurate.

Before concluding, we briefly mention two techniques that seem suitable for generating the entire distribution of forecasts from a model. The first is a simple simulation-based method and is essentially an extension of the method by *Rhoades et al.* [1994] from renewal processes to clustering models. The idea is to acknowledge that data and parameters are uncertain and hence distributed and then to repeatedly sample parameters and data randomly from these distributions to generate many forecasts.

The simulation-based method is simple but computationally expensive. Furthermore, past forecasts of the model are "thrown away" whenever new observations become available. A second method, data assimilation, provides an optimal and more efficient solution by making use of all available information, including previous model forecasts [see e.g. *Kalnay* [2003]]. The goal of data assimilation is to estimate the state of the physical system (and/or parameters) through a statistical combination of the noisy observation and the distributed model forecast according to Bayes' theorem. The state (e.g. past earthquake data and parameters of the model) are sequentially updated through time by correcting the model forecast (the prior) with the observations (the likelihood).

## 6.6 Conclusions

In this article, we analyzed magnitude uncertainties and their impact on seismic rate estimates in short-term clustering models, on their forecasts and on their evaluation in predictability experiments such as RELM or CSEP. In the first part, we quantified magnitude uncertainties. We estimated moment magnitude uncertainties by comparing the estimates for the same events from the CMT and USGS MT catalogs. We found that a double-sided exponential (Laplace) distribution with a scale parameter $0.1$ fit the distribution of the estimate differences significantly better than a Gaussian, reflecting the higher probability of outliers. If the distributions of independent, individual magnitude uncertainties decay much more slowly than a Gaussian, they have at least exponential or fatter-than-exponential tails. We also analyzed MAD values, a measure of magnitude uncertainty, reported by the NCSN in its authoritative region of the ANSS. We found that MAD estimates below $0.1$ may be unreliable. Typical values were between $0.1$ and $0.3$ but outliers occur often.

Because short-term seismicity models indiscriminately use any listed magnitude in earthquake catalogs for seismic rate projections, inter-magnitude uncertainties reflect the true errors better. We compared the CMT moment magnitudes with the PDE body and/or surface wave magnitudes and found scatter with standard deviations of $0.29$ and $0.26$, respectively. We further found that the NCSN local and coda duration magnitude estimates for the same events fit a Laplace distribution with scale parameter $0.2$ better than a Gaussian (with standard deviation $0.3$).

The relative lack of available quantitative magnitude uncertainty estimates coupled with their importance for hypothesis testing underscore the need for increased (funding for) data quality assessment and control by network operators.

In the second part, we studied the impact of magnitude noise on seismic rate projec-

tions in a simple clustering model that captures the main ingredients of popular short-term models. We proved that seismic rate estimates based on noisy catalog data deviate from their exact rate by power law fluctuations in the tail with exponent $\alpha = (a\nu_c)^{-1}$, where $a$ is the exponent of the productivity law of aftershocks and $\nu_c$ is the scale parameter of the Laplace distribution of the magnitude noise. Thus seismic rate projections and forecasts can fluctuate extremely due to magnitude noise. We further proved that the scale factor $C_{\Delta\lambda}$, which characterizes the typical scale of the fluctuations, remains a random variable and does not converge to a unique, fixed constant. Rather, there are subtle trade-offs between the power law tail, a tendency for the sum of random variables to converge to its stable law (Gaussian or Levy) and the strong quenched disorder due to particular catalog realizations and the stochasticity of the model.

In the last part, we studied how forecasts based on noisy data would fare in RELM/CSEP predictability experiments. We conducted a numerical experiment in which we constructed a hypothetical testing center and performed a one year test of daily forecasts. We assumed that earthquakes happen according to the seismic rate of a simple clustering model calibrated on an exact catalog data set. We then perturbed the catalog but used the same model to generate forecasts from the noisy data. These were submitted to the mock testing center as "models" that were tested for the consistency with the hypothetical observations. We found that noisy forecasts were rejected much more frequently than would be expected for a given confidence limit. We concluded that the current RELM number and likelihood consistency tests were sensitive to noisy forecasts and could wrongly reject the "true" model due to magnitude noise.

To robustly reject models at specified confidence limit, tests cannot assume that observations consistent with a model are Poisson distributed about its mean rate forecast. To properly capture the idiosyncrasies of each model together with all propagating uncertainties, the forecasts need to specify the entire distribution for each space-time-

magnitude bin. Based on our results that forecasts are power law distributed, we expect the deviations from a Poisson distribution to be severe. We noted that data assimilation techniques were particularly useful for propagating entire probability distributions while taking into account all uncertainties.

# CHAPTER 7

# Earthquake Forecasting Based on Data Assimilation

## 7.1 Introduction

This chapter focuses on the development of earthquake forecasting methods that systematically account for observational uncertainties, such as the magnitude uncertainties presented in Chapter 6. We use the framework of Bayesian data assimilation, a general method for estimating evolving quantities in the presence of noise and prior (model) information. In particular, we implement sequential Monte Carlo methods, a set of simulation-based techniques for estimating posterior distributions, to estimate occurrence times and magnitudes from noisy (synthetic) data in point process models relevant to seismicity.

The chapter is structured as follows. Section 7.2 explains the motivation for pursuing earthquake forecasting based on data assimilation. Sections 7.3 and 7.4 give a broad but brief overview over the different types of observational uncertainties in earthquake catalogs and over the different point process models that are used for forecasting. Due to the interdisciplinary nature of this chapter, these sections are primarily intended for non-seismologists.

Section 7.5 reviews some of the existing methods that have been applied to treat uncertainties in earthquake "hindcasting" (retrospective forecasting) using point process models. The techniques are currently limited to renewal processes, the simplest class of point processes, and were applied in a static setting. Section 7.5.4 thus

states the general problem of developing sequential (near real-time) methods for multi-dimensional marked point processes of clustered seismicity.

Section 7.6 introduces the concept of data assimilation. Due to the non-Gaussian, nonlinear nature of seismicity models, we immediately use the language of probabilistic methods to set the stage. In particular, we discuss sequential Monte Carlo (SMC) methods, a set of simulation-based techniques for estimating posterior distributions that have recently been quite successful at making sequential Bayesian estimation problems feasible. Section 7.7 describes several such methods, known as particle filters, along with their algorithms. The two sections 7.6 and 7.7 are intended as an introduction to data assimilation and SMC methods for seismologists and others. SMC methods are relatively new even in applied statistics.

Section 7.8 presents a brief review of the literature in which point process models are used in data assimilation applications. Although very recent work involves (simple) point processes in SMC methods, research does not seem to have involved more complex processes or applications in seismology.

After this review, the remainder of the chapter is devoted to extending the SMC methods to point process models for earthquake forecasting. In section 7.9, we contrast the mathematical framework of state-space models, which underlie most data assimilation methods, with point processes. We note that noisy occurrence times together with renewal processes present some theoretical challenges, which do not present themselves for more general processes and/or other noisy observations such as location or magnitude. Section 7.10 presents an implementation of several particle filters for a renewal process under noisy occurrence times. We use a lognormal model, which is relevant to earthquakes, and compare the improvement of the particle filter over the benchmark forecast, which ignores uncertainties. Section 7.11 discusses the generalization to multi-dimensional marked point processes. It presents the Bayesian solution

for estimating noisy magnitudes in a marked cluster process with exact occurrence times. Section 7.12 concludes.

## 7.2 Motivation

This section summarizes some of the motivational grounds for pursuing earthquake forecasting based on data assimilation.

1. Earthquake forecasting under observational uncertainties: As already emphasized in the last chapter, the current surge in earthquake predictability experiments [*Schorlemmer et al.*, 2007; *Schorlemmer and Gerstenberger*, 2007; *Jordan*, 2006] provides strong motivational grounds for developing earthquake forecasting methods that are robust with respect to observational uncertainties in earthquake catalogs. As was partially established in the last chapter, the uncertainties in earthquake catalogs are sufficiently serious to call for the development of a general framework that quantifies our degree of confidence in both forecasts and observations.

2. Seismic hazard calculations under observational uncertainties: On a practical level, the inclusion of uncertainties for seismic hazard calculations may provide better scientific support for the decision-making in the insurance industry, governmental and non-governmental disaster agencies, and risk mitigation strategies.

3. Scientific hypothesis testing under observational uncertainties: Observational uncertainties may bias conclusions regarding scientific hypotheses. The only way to accurately provide confidence limits to test our ideas about earthquakes is to take uncertainties into account – data assimilation provides the ideal vehicle for such a systematic treatment.

4. The development of a powerful null hypothesis: Seismicity forecasts are often compared to uninformative null hypotheses. The class of epidemic-type triggering models is fast gaining momentum as a powerful null hypothesis against which new claims can be tested (as seen in Chapter 5). However, there are outstanding issues in the formulation and understanding of this null hypothesis that have led to a variety of different implementations. Some of the differences may be insignificant in the face of data uncertainty.

5. Model development: Data assimilation can be used as a framework for likelihood-based model testing and development, fully accounting for uncertainties.

6. Estimating physical quantities of physics-based models from seismicity: In its general formulation as a state and parameter estimation problem, data assimilation may also be viewed as a method for estimating physical quantities ("states") and model parameters, directly related to physics-based models.

7. Optimally integrating different types of data for earthquake forecasts: The models submitted to the five-year forecast competition of the Working Group on Regional Earthquake Likelihood Models (RELM) contain a variety of data from which forecasts are generated. In the future, the coupled integration of several types of different data is highly desirable. Numerical weather prediction has a long history of integrating such different types of data – earthquake seismology may be able to adapt these methods.

8. Developing statistical theory and methodology for noisy point processes: The theory of point processes has so far largely focused on exact data (with some notable exceptions). The development of the statistical theory and practical methodology for taking into account noisy observations is therefore interesting in its own right.

9. Applications beyond earthquake forecasting: Point process models are being increasingly used in a large variety of fields. Much of the methodology is likely to be directly applicable in those fields.

## 7.3 Observational Uncertainties in Earthquake Catalogs

This section presents a brief overview of the observational uncertainties in earthquake catalogs. The discussion is kept broad but short since the implemented techniques in sections 7.7 and 7.11 are restricted to random occurrence times and magnitudes and the emphasis in this chapter is on the development of the methodology. In fact, each point below may be (and is) the subject of many articles.

Uncertainties often include both systematic and random errors; uncertainties due to the point source representation of earthquakes; uncertainties due to the very definition of earthquakes; and the list goes on. Uncertainty estimates were compiled from the following articles and references therein: [*Harte and Vere-Jones*, 1999; *Kagan*, 2002b, 2003; *Werner and Sornette*, 2007b]. Note that uncertainties vary strongly between different catalogs, which may measure different quantities (e.g. moment magnitude versus body wave magnitude, or earthquake centroids versus hypocenters).

- Occurrence times: The standard deviation of reported occurrence times depends on the type of catalog (centroid or first motion). Estimates vary between $1-3$ seconds depending on the catalog and the geographical location (i.e. the seismic network coverage).

- Magnitudes: Magnitude uncertainty was discussed in detail in Chapter 6. Random errors seem to be distributed more broadly than a Gaussian distribution. A Laplacian distribution with scale parameter $0.1-0.3$ seems to capture the errors.

- Locations: Location uncertainties strongly depend on the location method. From cross-correlated catalogs to global catalogs, the range spans from several meters to dozens of kilometers. Depth uncertainty is even harder to constrain: it ranges from hundreds of meters to essentially impossible to constrain, typically on the order of several to tens of kilometers.

- Focal mechanisms: The orientation of the moment tensor of earthquakes may be wrong by as much as $15 - 20^o$ in terms of a 3D rotation angle.

- Missing events (missing observations): Earthquake catalogs, even during optimal conditions, can only report earthquakes down to the size that can be detected by their networks. This optimum threshold varies strongly depending on the network. Regional catalogs may detect earthquakes as small as magnitude $0$ while global catalogs usually do not report below $M5$. However, this detection threshold can vary strongly spatially, due to uneven network coverage, and temporally, due to strong shocks that "swamp" the instruments so that even large events may go unnoticed.

- Historical seismicity (initial conditions): Large earthquakes influence seismicity for years and decades (and, in some cases like the 1891 Nobi earthquake in Japan, for centuries). Reliable homogeneous catalogs are usually only available for several decades at best (depending on the required threshold). The lack of precise information about large earthquakes prior to a catalog start date may thus influence model calibration and forecasts.

- Uncertainties due to the catalog representation: Catalogs list earthquakes as points, while of course they are spatio-temporally extended objects. The very definition of an earthquake event may be questioned as the identification depends on the temporal and spatial resolution of seismic instruments. Further-

more, other moment-releasing events (slow earthquakes, tremors, etc) are not yet included in catalogs.

- Non-catalog-based data: Earthquakes leave clear imprints on other observables, such as deformation measurable by GPS and InSAR, permeability changes in the crust, and are expected to be strongly influenced by pore-fluids, stress changes etc. The inclusion of such (equally uncertain) data may help improve forecasts.

## 7.4 Earthquake Forecasting and Point Process Models of Seismicity

In this section we can only give a brief flavor of some of the models used for the forecasting of earthquakes. Obviously there are an enormous range of models that can be and are being used to forecast earthquakes, ranging from deterministic to probabilistic, from alarm-based to likelihood-based, from physics-based to empirical, models that predict from past seismicity or from other data. Reviews, debates and views on earthquake prediction can for instance be found in [*Geller*, 1997; *Geller et al.*, 1997a,b; *Helmstetter and Sornette*, 2003d; *Jordan*, 2006; *Kagan*, 1997b, 1999b, 2006; *Keilis-Borok and Soloviev*, 2003; *Nature Debates*, 1999; *Sammis and Sornette*, 2002; *Sornette*, 2002]. The famous failed Parkfield prediction is discussed in [*Harris and Arrowsmith*, 2006; *Kagan*, 1997a; *Jackson and Kagan*, 2006; *Davis et al.*, 1989]. We restrict this brief overview to seismicity-based forecasting with a bias towards point process models. Thorough treatments of the theory of point processes can be found in [*Daley and Vere-Jones*, 2003; *Karr*, 1986].

*Vere-Jones* [1970] and *Kagan* [1973] were among the first to study seismicity using point processes. Other early stochastic models were developed in [*Vere-Jones*, 1978; *Kagan and Knopoff*, 1981; *Kagan*, 1982]. Since then, many different models of point

processes have been proposed and used in studies of seismicity. Reviews are provided by [*Kagan*, 2006; *Ogata*, 1988, 1999a; *Utsu et al.*, 1995] and especially in [*Vere-Jones*, 1995]. *Vere-Jones* [2000, 2006] provided personal accounts of the development of statistical seismology.

Simple models include renewal processes, often used for the seismic gap hypothesis [*McCann et al.*, 1979; *Nishenko*, 1991; *Kagan and Jackson*, 1991b, 1995; *Rong et al.*, 2003], models of characteristic earthquakes [*Wesnousky*, 1994; *Bakun et al.*, 2005; *Scholz*, 2002; *Kagan*, 1993], seismic hazard [*Field*, 2007] and earthquakes on a particular fault segment [*Bakun et al.*, 2005; *Davis et al.*, 1989; *Rhoades et al.*, 1994; *Ogata*, 1999a, 2002].

Long term probabilistic models were provided for example by [*Vere-Jones*, 1978; *Kagan and Jackson*, 1994; *Jackson and Kagan*, 1999; *Kagan and Jackson*, 2000] and [*Evison and Rhoades*, 2004; *Rhoades and Evison*, 2004].

Recently, five-year probabilistic forecasts were submitted to the Collaboratory for the Study of Earthquake Predictability (CSEP) and the Working Group on Regional Earthquake Likelihood Models (RELM) for community-agreed prospective testing [*Schorlemmer et al.*, 2007; *Jordan*, 2006]. The 19 submitted models are time-independent (Poissonian) and based on a variety of data and models, including past seismicity and constitute some of the current state of the art in time-independent forecasts.

Short-term models include non-homogeneous poisson processes (i.e. Poisson processes with a time-varying rate), which are often used for modeling individual aftershock sequences to mimic Omori's law of the power law decay of aftershocks [*Utsu et al.*, 1995; *Ogata and Katsura*, 2006].

Cluster or triggering models in which a renewal process generates its own offspring, are often used to model entire earthquake catalogs [*Vere-Jones*, 1970; *Kagan*, 1973; *Kagan and Knopoff*, 1987; *Ogata*, 1988, 1998, 1999b; *Felzer et al.*, 2002; *Helm-*

*stetter and Sornette*, 2002; *Console et al.*, 2003b; *Felzer et al.*, 2003; *Zhuang et al.*, 2004; *Gerstenberger et al.*, 2005; *Hainzl and Ogata*, 2005; *Sornette and Ouillon*, 2005; *Ouillon and Sornette*, 2005; *Helmstetter et al.*, 2006; *Console et al.*, 2007].

Many studies have focused on accelerating seismicity before major earthquakes or volcanic eruptions [*Voight*, 1988; *Sykes and Jaumé*, 1990; *Bufe and Varnes*, 1993], the concept of large earthquakes as genuine critical points [*Sornette and Sammis*, 1995; *Saleur et al.*, 1996a,b; *Johansen et al.*, 1996; *Huang et al.*, 1998; *Bowman et al.*, 1998] and accelerating moment release [*Vere-Jones et al.*, 2001; *Mignan et al.*, 2006].

Other physics-based models based on rate-and-state friction and/or Coulomb stress transfer were for instance studied by [*Dieterich*, 1994; *Stein*, 1999; *Stein et al.*, 1997].

Methods for the difficult task of assessing probabilistic earthquake forecasts include likelihood-based methods [*Bebbington*, 2005; *Harte and Vere-Jones*, 2005; *Kagan and Jackson*, 1994, 1995; *Jackson*, 1996; *Schorlemmer et al.*, 2007; *Werner and Sornette*, 2007b], and alarm-based methods [*Molchan*, 1990; *Molchan and Kagan*, 1992; *Zechar and Jordan*, 2007].

### 7.4.1 Some Basic Notions and Classes of Point Processes

[*Daley and Vere-Jones*, 2003] provide a rigorous development of the theory of point processes. Here, we solely introduce the conditional intensity function and some standard classes of point processes that will be relevant in this chapter.

- **Conditional intensity function:** A point process is completely defined by its conditional intensity function (an instantaneous probability of an event occurring), defined here for a one-dimensional, temporal point process

$$\lambda(t|H_t) = \lim_{\Delta \to 0} P(\text{ event occurs in } [t, t+\Delta]|H_t)/\Delta \tag{7.1}$$

where $H_t$ is the entire history up to time $t$, which includes all previous occurrence times.

Let $f(t|H_t)$ denote the probability density function (pdf) time until the next event (potentially dependent on more than just the last event, i.e. non-Markovian) and let $F(t|H_t)$ be the corresponding survivor function. The relationship given between the conditional intensity and these two quantities is given by

$$\lambda(t|H_t) = f(t|H_t)/F(t|H_t) \tag{7.2}$$

and the probability of an event in the time interval $(t_c, t_c + s)$ is given by

$$P(t_c; s|H_t) = 1 - \exp\left(-\int_{t_c}^{t_c+s} \lambda(u|H_{t_c})du\right) \tag{7.3}$$

When an event occurs, the history $H_t$ changes and therefore $\lambda(t|H_t)$ may change in form - it is defined piecewise continuously between events. Finally, another useful relationship is that the pdf $f(t_i|H_t)$ can be directly related to the conditional intensity by differentiating the equation (7.3):

$$f(t_i|H_t) = \lambda(t_i|H_t)\exp\left(-\int_{t_{i-1}}^{t_i} \lambda(u|H_t)du\right) H(t_i - t_{i-1}) \tag{7.4}$$

where $H(\cdot)$ is the Heaviside step function. The conditional intensity can be generalized to include locations and marks.

- **Renewal Processes:** A renewal process is a particular class of temporal point process in which the probability of occurrence of the next event only depends on the time since the last event. Their pdf for the occurrence time of the $i^{th}$ event given $H_t$ is simply given by:

$$f(t_i|H_t) = f(t_i|t_{i-1}) = f(t_i - t_{i-1})H(t_i - t_{i-1}) \tag{7.5}$$

where, again, $H(\cdot)$ is the Heaviside step function. Here, the dependence on $H_t$ is reduced to $t_{i-1}$.

Renewal processes are the simplest class of point processes. One can equivalently define them by the class of conditional intensities that depend only on the time of the last event:

$$\lambda(t|H_t) = \lambda(t - t_l) \tag{7.6}$$

where $t_l$ is the time of the last event and the rest of the history $H_t$ is irrelevant for the probability of the next event. One can also define the interval $\tau = t - t_l$ as the relevant variable. Examples of such processes in terms of their pdf are:

$$f_{Poisson}(\tau;\lambda) = \lambda \exp(-\lambda\tau) \tag{7.7}$$

$$f_{lognormal}(\tau;\mu,\sigma) = \frac{1}{\tau\sqrt{2\pi}\sigma} \exp(-(\log\tau - \mu)^2/2\sigma^2) \tag{7.8}$$

$$f_{Weibull}(\tau;k,s) = \frac{k}{s}(\frac{\tau}{s})^{k-1} \exp -(\tau/s)^k \tag{7.9}$$

where the quantities after the colon are parameters of the pdf, and $\tau$ is strictly positive. For the Poisson process, the parameter $\lambda$ is equal to its (constant) hazard rate or conditional intensity. Note that the Poisson process, because of its lack of memory, is a special case of the renewal processes: if parameters are known, then more accurate estimates of past occurrences do not lead to a different forecast. The Poisson process is hence trivial in data assimilation methods.

- **Clustering Models:** Clustering models are usually constructed from two processes: a cluster center process, which is often a renewal or Poisson process, and a cluster member process. The cluster member process consists of events that are triggered by the cluster center via a triggering or response function $h(t - t_{i_c}, \xi)$ which usually depends only on the time since the cluster center and a stochastic amplitude $\xi$ drawn from a usually time-invariant distribution. An example is given by the simple aftershock model considered in Chapter 6. A generic form is given by

$$\lambda(t|H_t^c, \theta) = \lambda_c + \sum_{i_c|t_{i_c}<t} h(t - t_{i_c}, \xi_{i_c}) \tag{7.10}$$

where $H_t^c$ is the history up to time $t$ which need only include information about the cluster centers $\{t_{i_c}, \xi_{i_c}\}_{1 \leq i_c \leq N_c}$, as cluster members do not trigger their own events and do not influence the future. Here, the cluster center process is a Poisson process with rate $\lambda_c$ and the model is characterized by a set of parameters $\theta$. In contrast to shot noise models, $\lambda$ is not the observable, but drives the observed point process characterized by $\lambda$ as a conditional intensity function.

- **Self-Exciting Models:** Self-exciting models were first proposed by [*Hawkes*, 1971a,b; *Hawkes and Oakes*, 1974]. They are a generalization of the cluster models by allowing each event, including cluster members, to trigger their own events according to some response function or kernel $h(t - t_i)$:

$$\lambda(t|H_t, \theta) = \lambda_c + \sum_{i|t_i<t} h(t - t_i) \tag{7.11}$$

where the history $H_t = \{t_i\}_{1 \leq i \leq N}$ now includes all events and the sum in equation (7.11) runs over all events. The major introduction in the model is a (linear) dependence on all previous activity. The model can equivalently be viewed as a branching model [*Hawkes and Oakes*, 1974]. Nonlinear generalizations are much more difficult to handle analytically (see, for instance, [*Ouillon and Sornette*, 2005; *Sornette and Ouillon*, 2005].

- **Marked Self-Exciting Point Processes:** This class is a multidimensional extension of the self-exciting process obtained by associating with each point $t_i$ a random mark $m$ drawn from a usually invariant distribution $p(m)$:

$$\lambda(t, m|H_t, \theta) = p(m) \left( \lambda_c + \sum_{i|t_i<t} h(t - t_i, m_i) \right) \tag{7.12}$$

where the stochastic amplitude in the response function $h(t - t_i, m_i)$ is usually the mark of the $i^{th}$ event and the history is given by $H_t = \{t_i, m_i\}_{1 \leq i \leq N}$. This

class can be further extended to include locations in space. Under a particular choice of the response function and the mark distribution, one recovers the Epidemic-Type-Aftershock-Sequence (ETAS) model [*Ogata*, 1988, 1998]:

$$\lambda(t, m|H_t, \theta) = p(m) \left( \lambda_c + \sum_{i|t_i<t} \frac{k\ e^{a(m_i - m_d)}}{(t - t_i + c)^p} \right) \tag{7.13}$$

where $p(m)$ is given by the Gutenberg-Richter law and the model is characterized by the set of parameters $\theta = \{\beta, \lambda_c, k, a, m_d, c, p\}$.

These basic ideas suffice for the remainder of the chapter. A more comprehensive treatment can be found in [*Daley and Vere-Jones*, 2003].

## 7.5 Existing Methods for Uncertainties in Point Process Seismicity Models

Little research seems to have been devoted to dealing explicitly and systematically with observational uncertainties in point process models of seismicity. Notable exceptions are [*Rhoades et al.*, 1994; *Ogata*, 1999a, 2002]. These authors focused exclusively on paleoseismological data and their impact on renewal process forecasts for individual faults and fault segments. As opposed to more ad hoc approaches to uncertainties (e.g. California or national seismic hazard estimates), these authors introduced systematic and rigorous methods, which are reviewed briefly below.

### 7.5.1 Ignoring Observational Uncertainties: The "Benchmark"

The obvious first approach to uncertainties is, of course, to ignore them. We mention this "benchmark" method here solely to introduce the "straw man", against which improvements are to be considered. Most if not all current operational earthquake

forecasts based on marked point processes follow this "strategy" (e.g. [*Gerstenberger et al.*, 2005; *Helmstetter et al.*, 2006]). It should be noted, however, that the parameters, which are estimated from earthquake catalog data, may be able to adjust to some of the uncertainties. Such effective parameters may appear to reduce the effects of errors. It is therefore important to consider the joint parameter estimation and forecasting problem. Such a comprehensive approach should be developed, but is outside the scope of this dissertation: In our numerical experiments, we consider parameters as known.

### 7.5.2 Sampling from the Uncertainty Distribution: The "Bootstrap"

In an application to the hazard calculations of earthquakes on individual faults from paleoseismic data, *Rhoades et al.* [1994] average the probability $P(t_c; s|y^o)$ of an event in the interval $[t_c, s]$, given by equation (7.3), which is conditional on the data set $y^o$ (the occurrence times of the earthquakes in the data set), over the data uncertainty distribution $f_y(y^o)$:

$$P(t_c; s|H_t) = \int P(t_c; s|y^o) f_y(y^o) dy^o \tag{7.14}$$

which is equivalently to averaging the hazard (conditional intensity) rate:

$$\lambda(t|H_t) = \int \lambda(t|y^o) f_y(y^o) dy^o \tag{7.15}$$

where the (multidimensional) data uncertainty distribution may contain correlations between the uncertainties in the data. *Rhoades et al.* [1994], while outlining the generality of the approach, apply their method to un-correlated data uncertainties, i.e. each occurrence time uncertainty is independent of the others. It is not a recursive approach.

*Ogata* [2002] also considered uncertain paleoseismic data and used renewal processes to attempt to identify which process fit the data significantly better. This may provide insight into the nature of earthquake recurrences on individual faults. *Ogata*

[2002] averaged the pdf's instead of the probability distribution or the conditional intensity:

$$f(t|H_t) = \int f(t|y^o) f_y(y^o) dy^o \tag{7.16}$$

*Rhoades et al.* [1994] commented on the difference between averaging over the hazard functions as opposed to the probability density functions.

### 7.5.3 Informing the Data Uncertainty via the Model Forecast: A Static Bayesian Method

*Ogata* [1999a] presented a Bayesian approach including parameter estimation for an uncertain paleoseismic data set assumed to be governed by a renewal process. He considered the entire data set at once:

$$\phi(t_1, ..., t_n | t_1^o, ..., t_n^o) = L(t_1, ...., t_n) \Pi_{i=1}^n \psi_i(t_i)/N \tag{7.17}$$

where the product of the independent data uncertainty densities $\psi_i$ is the prior, $N$ is a normalization constant and the "likelihood" $L$ is defined by the product of the individual interval densities given by equation (7.5). This approach does not integrate or take averages but multiplies the densities directly for the entire data set. This allows an estimation of a Bayesian likelihood, which was used to evaluate the goodness of fit between different models.

### 7.5.4 Towards Sequential Bayesian Methods for General Point Processes: Problem Statement

The methods by *Rhoades et al.* [1994], *Ogata* [1999a] and *Ogata* [2002] were applied to simple renewal processes and to an entire data set at once. Only *Ogata* [2002] made optimal use of his proposed models by "correcting" the observation by the model forecast in a Bayesian framework. First, we would like to generalize the approach to

general marked point processes that are relevant to time-dependent earthquake forecasting today (e.g. in the framework of RELM/CSEP). Second, we would like to use a sequential method that recursively assimilates observations into the model forecast so as to optimally estimate the actual occurrence times, magnitudes, locations, focal mechanisms, and potentially other data. We can thus state the general statistical problem that will be addressed in this chapter:

*Estimate as best as possible and sequentially in time, earthquake parameters such as occurrence time, magnitude, location, focal mechanism etc, using all available information, including noisy observations of these parameters and statistical laws combined into spatio-temporal marked point processes that govern their behavior.*

Data assimilation provides the ideal vehicle for such an approach, as we shall see in the next section.

## 7.6 A Brief Introduction to Data Assimilation

Data assimilation is a statistical technique that integrates observations with model-based forecasts in an optimal way [*Kalnay*, 2003; *Daley*, 1991]. Data assimilation came about in dynamical meteorology and physical oceanography, but has quickly spread into an active and growing research field in engineering, mathematics, finance and geophysics (see, for example, the references in *Ide et al.* [1997]). Numerical weather prediction provides the prime example for the method: an uncertain estimate of the state of the atmosphere, characterized by e.g. temperature, pressure, and velocities, is evolved according to physical model, e.g. the Navier-Stokes equation, to produce an uncertain forecast of the state for the next discrete time step. Observations, which are also characterized by uncertainty, are made at this new time step. Data assimilation then statistically combines both observation and forecast in an optimal way

so as to estimate the current state of the atmosphere as accurately as possible using all available information. Here, the available information consists of both the actual observations and the physical laws that govern the evolution of the atmospheric flow.

A variety of specific data assimilation methods exist, which differ in their practical implementation of the optimal combination and in their assumptions about the model, the observations and their uncertainties. The most general approach in state evolution problems of the kind found in numerical weather prediction combines the Liouville equation for the model forecast together with Bayes' theorem for the assimilation (or analysis) of the observations [*Ehrendorfer*, 1994]. A review of methods from early techniques to current weather forecast methods of national agencies to 4D variational approaches, Kalman filtering, extended Kalman filtering and ensemble Kalman filtering can be found in Chapter 5 of *Kalnay* [2003]. The Kalman filter presents the optimal solution to the data assimilation problem under the assumptions of linear models with Gaussian observation and model noise. These assumptions are strongly violated in stochastic point process models for earthquake forecasting. We therefore need a more general approach based on propagating the entire probability distribution, rather than solely mean and covariance. Such a scheme is presented in detail in the next section.

### 7.6.1 Bayesian Data Assimilation

In this section, we state the general problem of Bayesian data assimilation that will be solved for specific model and observation assumptions in section 7.10. The presentation borrows from [*Doucet et al.*, 2000, 2001] and [*Arulampalam et al.*, 2002].

For ease of presentation, we restrict ourselves to signals modeled as Markovian, nonlinear, non-Gaussian state-space models. The unobserved signal (hidden states) $\{x_t\}_{t\geq 1}$ is modeled as a Markov process ($x_t$ may be a vector). The initial state $x_0$ has initial distribution $p(x_0)$. The transition from $x_t$ to $x_{t+1}$ is governed by a transition

probability distribution $p(x_{t+1}|x_t)$. The observations $\{y_t\}_{t\geq 1}$ are assumed to be conditionally independent given the process $\{x_t\}_{t\geq 1}$ and of marginal distribution $p(y_t|x_t)$ (the observations may also be vectors, in general of different dimension than the state). The model can be summarized by

$$\begin{aligned} &\text{Initial condition:} && p(x_0) && \\ &\text{Model forecast:} && p(x_{t+1}|x_t) && t \geq 1 \\ &\text{Data likelihood:} && p(y_t|x_t) && t \geq 1 \end{aligned}$$

We denote $x_{0:t} = \{x_0, \ldots, x_t\}$ and $y_{1:t} = \{y_1, \ldots, y_t\}$. The problem statement is then as follows: the aim is to estimate sequentially in time the posterior distribution $p(x_{0:t}|y_{1:t})$, its associated features (e.g. the marginal distribution $p(x_t|y_{1:t})$, also known as the filtering distribution) and potentially the expectation (mathematical average)

$$I(f_t) = E_{p(x_{0:t}|y_{1:t})}[f_t(x_{0:t})] \equiv \int f_t(x_{0:t}) p(x_{0:t}|y_{1:t}) dx_{0:t} \tag{7.18}$$

where the function $f_t$ may be chosen to obtain the conditional mean of $x_{0:t}$ or other quantities of interest such as the covariance. For broadly distributed posteriors, the mean may not be particularly meaningful and is therefore largely ignored here.

At any time $t$, the posterior distribution is given by Bayes' theorem

$$p(x_{0:t}|y_{1:t}) = \frac{p(y_{1:t}|x_{0:t})\; p(x_{0:t})}{\int p(y_{1:t}|x_{0:t})\; p(x_{0:t}) dx_{0:t}} \tag{7.19}$$

A recursive or sequential formula is given by

$$p(x_{0:t+1}|y_{1:t+1}) = p(x_{0:t}|y_{1:t})\; \frac{p(y_{t+1}|x_{t+1})\; p(x_{t+1}|x_t)}{p(y_{t+1}|y_t)} \tag{7.20}$$

The marginal distribution $p(x_t|y_{1:t-1})$ also satisfies the following recursion:

$$\text{Prediction:} \quad p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1}) p(x_{t-1}|y_{1:t-1}) dx_{t-1} \tag{7.21}$$

$$\text{Updating:} \quad p(x_t|y_{1:t}) = \frac{p(y_t|x_t)\; p(x_t|y_{1:t-1})}{\int p(y_t|x_t)\; p(x_t|y_{1:t-1}) dx_t} \tag{7.22}$$

Expressions (7.21) and (7.22) are the essential steps in sequential data assimilation. Using the last update (the posterior, also often called analysis) as initial condition, the Chapman-Kolmogorov (prediction) equation is used to forecast the state at the next time step. When observations $y_t$ become available, they are assimilated into the model forecast by the update equation. This cycle constitutes data assimilation.

Only in very special cases are the prediction and update equations (7.21) and (7.22) amenable to analytical solutions. When the model is linear with Gaussian system noise and the observations are Gaussian, the analytical and optimal solution is given by the Kalman filter. The extended Kalman filter, the ensemble Kalman filter and the Kalman-Levy filter [*Sornette and Ide*, 2001] generalize the setting of models and distributions somewhat, but remain limited to near-Gaussian models or at least locally Gaussian models (extended and ensemble Kalman filter), and a Gaussian model with Levy-like errors (Kalman-Levy filter).

In general, however, such a simplification is impossible. Furthermore, one cannot usually calculate the normalizing constant $p(y_{1:t})$, the marginal $p(x_t|y_t)$ of the posterior or $I(f_t)$, since one needs to evaluate complex high-dimensional integrals. Thus, the recurrence relations (7.21) and (7.22) only form a conceptual solution, since in practice they cannot be determined analytically and one needs to resort to sub-optimal approximation methods such as extended Kalman filtering, ensemble Kalman filtering, approximate grid-based methods or sequential Monte Carlo methods.

### 7.6.2 Sequential Monte Carlo (SMC) Methods

Earthquake statistics clearly violate Gaussian approximations in terms of their temporal, spatial and magnitude occurrences, so much so that approximate algorithms based on local Gaussian approximations (e.g. the extended Kalman filter) are highly unlikely to produce good results. Furthermore, the continuous state space of seismicity rules

out methods in which that space is assumed to be discrete (such as grid-based methods). This leaves us with numerical integration techniques and Monte Carlo methods. The former are numerically accurate but computationally expensive in problems with high dimensionality.

Sequential Monte Carlo (SMC) methods bridge the gap between these cost-intensive methods and the methods based on Gaussian approximations. They are a set of simulation-based methods that provide a flexible alternative to computing posterior distributions. They are applicable in very general settings, parallelisable and often relatively easy to implement. Early methods were developed in the 70s, but only with the advent of cheap computational power in the mid 90s did they become a widespread tool. Since then, however, SMC methods have been applied in target tracking, financial analysis, diagnostic measures of fit, missing data problems, communications and audio engineering, population biology, neuroscience, and many more. SMC methods are also known under the names of particle filters, bootstrap filters, condensation, Monte Carlo filters, interacting particle approximations and survival of the fittest. Good introductions can be found in [*Arulampalam et al.*, 2002; *Doucet et al.*, 2000, 2001; *Liu*, 2001; *Liu and Chen*, 1998] and, in particular, in Chapter 6 of *de Freitas* [1999].

In the following, we describe some of the central building blocks of SMC methods, including Monte Carlo sampling, importance sampling and sequential importance sampling. For a more comprehensive introduction, we refer the interested reader to the above references, from which this presentation borrows.

#### 7.6.2.1 Monte Carlo Sampling

In Monte Carlo (MC) simulation, a set of $N$ weighted "particles" (samples) $x_{0:t}^{(i)}$ are drawn identically and independently from a distribution (say, the posterior $p(x_{0:t|y_{1:t}})$.

Then, an empirical estimate of the distribution is given by

$$\hat{p}_N(x_{0:t}|y_{1:t}) = \frac{1}{N}\sum_{i}^{N} \delta_{x_{0:t}^{(i)}}(x_{0:t}) \tag{7.23}$$

where $\delta_{x_{0:t}^{(i)}}(x_{0:t})$ denotes the Dirac mass located at $x_{0:t}^{i}$. The essential idea of Monte Carlo sampling is to convert an integral into a discrete sum. One is often interested in some function of the posterior distribution, say its expectation, covariance, marginal or another distribution. Estimates of such functions $I(f_t)$ can be obtained from

$$I_N(f_t) = \int f_t(x_{0:t})\hat{p}_N(x_{0:t}|y_{1:t})dx_{0:t} = \frac{1}{N}\sum_{i}^{N} f_t(x_{0:t}^{(i)}) \tag{7.24}$$

This estimate is unbiased. If the posterior variance of $f_t(x_{0:t})$ is finite, say $\sigma_{f_t}^2$, then the variance of $I_N(f_t)$ is equal to $\sigma_{f_t}^2/N$. From the law of large numbers,

$$I_N(f_t) \xrightarrow[N\to\infty]{a.s.} I(f_t) \tag{7.25}$$

where a.s. denotes almost sure convergence. That is, the probability that the estimate $I_N(f_t)$ converges to the "true" value $I(f_t)$ equals one in the limit of infinite number of particles. Furthermore, if the posterior variance $\sigma_{f_t}^2 < \infty$, then a central limit theorem holds:

$$\sqrt{N}(I_N(f_t) - I(f_t) \xrightarrow[N\to\infty]{\Delta} \mathcal{N}(0, \sigma_{f_t}^2) \tag{7.26}$$

where $\xrightarrow[N\to\infty]{\Delta}$ denotes convergence in distribution and $\mathcal{N}(0, \sigma_{f_t}^2)$ is the normal (Gaussian) distribution with mean zero and variance $\sigma_{f_t}^2$). The advantage of this perfect Monte Carlo method is therefore that the rate of convergence of the MC estimate is independent of the dimension of the integrand. This stands in contrast to any deterministic numerical integration method, whose rate of convergence increases with the dimensionality of the integrand.

Unfortunately, because the posterior distribution is usually highly complex, multidimensional and only known up to a normalizing constant, it is often impossible to

sample directly from the posterior. One very successful solution for generating samples from such distribution is Markov Chain Monte Carlo (MCMC). Its key idea is to generate samples from a proposal distribution, different from the posterior, and then to cause the proposal samples to migrate, so that their final distribution is the target distribution. The migration of the samples is caused by the transition probabilities of a Markov chain (see Appendix D of *de Freitas* [1999] for an introduction that is easy to follow). However, MCMC are iterative algorithms unsuited to sequential estimation problems and will not be pursued here. Rather, SMC methods primarily rely on a sequential version of importance sampling.

#### 7.6.2.2 Importance Sampling (IS)

Importance Sampling (IS) introduced the idea of generating samples from a known, easy-to-sample probability density function (pdf) $q(x)$, called the importance density or proposal density, and then "correcting" the weights of each sample so that the weighted samples approximate the desired density. As long as the support of the proposal density includes the support of the target density, one can make use of the substitution

$$p(x_{0:t}|y_{0:t}) = \frac{p(x_{0:t}|y_{0:t})}{q(x_{0:t}|y_{0:t})} q(x_{0:t}|y_{0:t}) \tag{7.27}$$

to obtain the identity

$$I(f_t) = \frac{\int f_t(x_{0:t}) w(x_{0:t}) q(x_{0:t}|y_{0:t}) dx_{0:t}}{\int w(x_{0:t}) q(x_{0:t}|y_{0:t}) dx_{0:t}} \tag{7.28}$$

where $w(x_{0:t})$ is known as the importance weight

$$w(x_{0:t}) = \frac{p(x_{0:t}|y_{0:t})}{q(x_{0:t}|y_{0:t})} \tag{7.29}$$

Therefore, if one can generate $N$ independently and identically distributed samples $x_{0:t}^{(i)}$ from the importance density $q(x_{0:t}|y_{0:t})$, a Monte Carlo estimate of $I(f_t)$ is given

by

$$\hat{I}_N(f_t) = \frac{\frac{1}{N}\sum_i^N f_t(x_{0:t}^{(i)})w(x_{0:t}^{(i)})}{\frac{1}{N}\sum_i^N w(x_{0:t}^{(i)})} = \sum_i^N f_t(x_{0:t}^{(i)})\tilde{w}_t^{(i)} \tag{7.30}$$

where the normalized importance weights $\tilde{w}_t^{(i)}$ are given by

$$\tilde{w}_t^{(i)} = \frac{w(x_{0:t}^{(i)})}{\sum_{j=1}^N w(x_{0:t}^{(j)})} \tag{7.31}$$

For finite $N$, the estimate $\hat{I}_N(f_t)$ is biased, as it is the ratio of two estimates. However, it is possible to obtain asymptotic almost sure convergence $\hat{I}_N(f_t) \xrightarrow[N\to\infty]{a.s.} I(f_t)$ and a central limit theorem provided (i) the importance density support contains the posterior density support, and (ii) the expectations of the weights $w_t$ and $w_t f_t^2(x_{0:t})$ exist and are finite.

Thus, the posterior density function can be approximated arbitrarily well by the point-mass estimate

$$\hat{p}(x_{0:t}|y_{1:t}) = \sum_i^N \tilde{w}_t^{(i)} \delta_{x_{0:t}^{(i)}}(x_{0:t}) \tag{7.32}$$

#### 7.6.2.3 Sequential Importance Sampling (SIS)

In its simplest form, IS is not adequate for sequential estimation. Whenever new data $z_t$ become available, one needs to recompute the importance weights over the entire state sequence. Sequential Importance Sampling (SIS) modifies IS so that it becomes possible to compute an estimate of the posterior without modifying the past simulated trajectories. It requires that the importance density $q(x_{0:t}|y_{1:t})$ at time $t$ admits as marginal distribution at time $t-1$ the importance function $q(x_{0:t-1}|y_{1:t-1})$:

$$q(x_{0:t}|y_{1:t}) = q(x_{0:t-1}|y_{1:t-1})q(x_t|x_{0:t-1}, y_{1:t}) \tag{7.33}$$

after iterating, one obtains:

$$q(x_{0:t}|y_{1:t}) = q(x_0)\prod_{k=1}^t q(x_k|x_{0:k-1}, y_{1:k}) \tag{7.34}$$

Assuming that the state evolves according to a Markov process and that the observations are conditionally independent given the states, one can obtain

$$p(x_{0:t}) = p(x_0) \prod_{k=1}^{t} p(x_k|x_{k-1}) \quad \text{and} \quad p(y_{1:t}|x_{0:t}) = \prod_{k=1}^{t} p(y_k|x_k) \tag{7.35}$$

Substituting (7.34) and (7.35) into (7.31) and using Bayes' theorem, we arrive at a recursive estimate of the importance weights

$$\tilde{w}_t^{(i)} \propto \tilde{w}_{t-1}^{(i)} \frac{p(y_t|x_t^{(i)}) p(x_t^{(i)}|x_{t-1}^{(i)})}{q(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{1:t})} \tag{7.36}$$

where the normalization is provided by $\sum_{j=1}^{N} \tilde{w}_t^{(j)}$. Equation (7.36) provides a mechanism for sequentially updating the importance weights. In summary, SIS provides a method to approximate the posterior density function (7.32) (or some function thereof) sequentially in time without having to draw samples directly from the posterior. All that is required is (i) sampling from the importance density and evaluating it up to some constant, (ii) evaluating the likelihood $p(y_t|x_t^{(i)})$ up to some proportionality constant, (iii) evaluating the forecast $p(x_t^{(i)}|x_{t-1}^{(i)})$ up to some constant, and (iv) normalizing the importance weights via $\sum_{j=1}^{N} \tilde{w}_t^{(j)}$. SIS thus makes sequential Bayesian estimation feasible.

#### 7.6.2.4 Choice of the Importance Density and Resampling

The problem encountered by the SIS method is that, as $t$ increases, the distribution of the importance weights becomes more and more skewed. For instance, if the support of the importance density is broader than the posterior density, then some particles will have their weights set to zero in the update stage. But even if the supports coincide exactly, many particles will over time decrease in weight so that after a few time steps, only a few lucky survivors have significant weights, while a large computational effort is spent on propagating unimportant particles. It has been shown that the variance of

the weights can only increase over time, thus it is impossible to overcome the degeneracy problem [*Kong et al.*, 1994]. Two solutions exist to this minimize this problem: (i) a good choice of the importance density, and (ii) resampling.

- **Importance Density**: The optimal importance density is given by the posterior itself:

$$q_{opt}(x_t|x_{0:t-1}, y_{1:t}) = p(x_t|x_{0:t-1}, y_{1:t}) = \frac{p(y_t|x_t, x_{t-1}^{(i)})p(x_t|x_{t-1}^{(i)})}{p(y_t|x_{t-1}^{(i)})} \tag{7.37}$$

  because it can be proven to minimize the variance of the importance weights (see *Kong et al.* [1994] and in particular Chapter 6 of *de Freitas* [1999] for an insightful discussion). However, using the optimal importance density requires the ability to sample from $p(x_t|x_{t-1}^{(i)}, y_t)$ and to evaluate the integral over the new state $p(y_t|x_{t-1}^{(i)})$ [*Arulampalam et al.*, 2002; *Doucet et al.*, 2001; *de Freitas*, 1999]. In many situations, this is impossible or very difficult, prompting the use of other importance densities. Perhaps the simplest and most common one is given by the prior:

$$q(x_t|x_{0:t-1}, y_{1:t}) = p(x_t|x_{t-1}) \tag{7.38}$$

  which, although resulting in a higher variance of the Monte Carlo estimator, is usually easy to implement. Many other choices are possible, including additional MCMC steps to sample from the importance density and bridging densities and progressive corrections to herd the particles to the important part of the state space [*Arulampalam et al.*, 2002; *Doucet et al.*, 2001; *Liu*, 2001].

- **Resampling**: Even the optimal importance density will lead to this "degeneracy" of the particles (few important ones and many unimportant ones). One therefore introduces an additional selection or resampling step, in which particles with little weight are eliminated and new particles are sampled in the impor-

tant regions of the posterior. *de Freitas* [1999] and *Arulampalam et al.* [2002] provide an overview of different resampling methods.

Resampling introduces its own problems. Since particles are sampled from discrete approximations to density functions, the particles with high weights are statistically selected many times. This leads to a loss of diversity among the particles as the resultant sample will contain many repeated points. This is known as "sample impoverishment" [*Arulampalam et al.*, 2002] and is severe when the model forecast is very narrow. There are various methods to deal with this problem, including sophisticated methods that recalculate past states and weights via a recursion and MCMC methods. Systematic techniques include the Resample-Move algorithm and the Regularized Particle Filter (RPF). The popular RPF uses a kernel estimation of the discrete posterior so that samples can be generated from a smoothed, continuous posterior.

Because of the additional problems introduced by resampling, it makes sense to resample only when the variance of the weights has decreased appreciably. A suitable measure of degeneracy of an algorithm is the effective sample size $N_{eff}$ introduced by *Liu and Chen* [1998] and defined by

$$N_{eff} = \frac{N}{1 + var(w_t^{\star i})} \tag{7.39}$$

where $w_t^{\star i} = p(x_t^{(i)}|y_{1:t})/q(x_t^{(i)}|x_{k-1}^{(i)}, y_k)$ is referred to as the true weight. This may not be available, but an estimate $\hat{N}_{eff}$ can be obtained by:

$$\hat{N}_{eff} = \frac{1}{\sum_{i=1}^{N}(w_t^{(i)})^2} \tag{7.40}$$

Thus, resampling can be applied when $\hat{N}_{eff}$ falls below a certain threshold $\hat{N}_{eff} < N_{thres}$.

## 7.7 Particle Filters and Algorithms

This section outlines a few choices for the importance density and resampling strategy of particle filters. More on particular particle filters can be found in [*Arulampalam et al.*, 2002; *de Freitas*, 1999; *Doucet et al.*, 2000, 2001; *Liu*, 2001]. We present four particle filters.

1. The Simple Sequential Importance Sampling (SSIS) particle filter: The simplest particle filter, it uses the prior (7.38) as the (sub-optimal) importance density and does not include a resampling step.

2. The Optimal Sequential Importance Sampling (OSIS) particle filter: This filter improves on the SSIS by using the optimal importance sampling density (7.37), but does not include resampling.

3. The Sequential Importance Resampling (SIR) particle filter: This filter improves on the SIS filters by including a resampling step to counteract the degeneracy of particles. The importance density may either be the prior or the optimal importance density.

4. The Regularized Particle Filter (RPF): This filter improves on the SIR filters by resampling from a smoothed, continuous posterior based on a kernel density estimate of the discrete posterior. It thereby counteracts the problem of sample impoverishment. The importance density may either be the prior or the optimal importance density.

### 7.7.1 Simple Sequential Importance Sampling (SSIS) Filter

The Simple SIS (SSIS) particle filter is characterized by the lack of resampling and by choosing the prior $p(x_t|x_{t-1}^{(i)})$ as the importance density:

$$q(x_t|x_{0:t-1}, y_{1:t}) = p(x_t|x_{t-1}^{(i)}) \tag{7.41}$$

It can be shown [*Arulampalam et al.*, 2002] that the SSIS can be reduced to the pseudo-code given by Algorithm 7.1, where the weights are given by:

$$w_t^{(i)} \propto w_{t-1}^{(i)} p(y_t|x_t^{(i)}) \tag{7.42}$$

where $p(y_t|x_t^{(i)})$ is simply the likelihood and the weights are normalized by

$$\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}} \tag{7.43}$$

**Algorithm 7.1** *Simple SIS Particle Filter*

$[\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N]$ =SSIS$[\{x_{t-1}^{(i)}, w_{t-1}^{(i)}\}_{i=1}^N, y_t]$

**for** $i$=1 to $N$ **do**

Draw $x_t^i \sim p(x_t|x_{t-1}^i)$

Assign the particle a weight, $w_t^i$, according to (7.42)

**end for**

This filter is simple and easy to implement. However, if the likelihood has a much narrower support than the importance density, then weights of many particles will be set to zero so that only few active particles are left to approximate the posterior. Depending on the overlap of the support of the two density functions, this particle filter may quickly degenerate in quality.

### 7.7.2 Optimal Sequential Importance Sampling (OSIS) Filter

The Optimal Simple SIS (OSIS) improves on the SSIS by using the optimal sampling density:

$$q_{opt}(x_t|x_{0:t-1}, y_{1:t}) = p(x_t|x_{0:t-1}, y_{1:t}) = \frac{p(y_t|x_t, x_{t-1}^{(i)})p(x_t|x_{t-1}^{(i)})}{p(y_t|x_{t-1}^{(i)})} \tag{7.44}$$

Then the algorithm of the OSIS filter is given by Algorithm 7.2, where the weights are given by substituting the optimal importance density (7.44) into the recursive weight equation (7.36) to obtain:

$$w_t^{(i)} \propto w_{t-1}^{(i)} p(y_t|x_{t-1}^{(i)}) = w_{t-1}^{(i)} \int p(y_t|x_t')p(x_t'|x_{k-1}^{(i)})dx_t' \tag{7.45}$$

Weights are normalized as in equation (7.43). As was already mentioned, the optimal density suffers from two difficulties: (i) generating samples from the posterior (7.44), and (ii) calculating the integral in (7.45).

---

**Algorithm 7.2** *Optimal SIS Particle Filter*

---

$[\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N]$ =OSIS$[\{x_{t-1}^{(i)}, w_{t-1}^{(i)}\}_{i=1}^N, y_t]$

**for** $i$=1 to $N$ **do**

Draw $x_t^{(i)} \sim q_{opt}(x_t|x_{t-1}^{(i)}, y_t) \propto p(y_t|x_t, x_{t-1}^{(i)})p(x_t|x_{t-1}^{(i)})$

Assign the particle a weight, $w_t^{(i)}$, according to (7.45)

**end for**

---

### 7.7.3 Optimal Sampling Importance Resampling (OSIR) Filter

To counter the inevitable problem of particle degeneracy, we can use resampling to generate a new set of particles from the (discrete) posterior. Setting the importance

density equal to the optimal importance density as in the OSIS particle filter described above, we recover the Optimal Sampling Importance Resampling (OSIR) algorithm given by Algorithm 7.3.

In the literature, the SIR filter is usually an implementation with the prior as the (suboptimal) importance density. Such a filter is called the "bootstrap" filter by *Doucet et al.* [2001].

---

**Algorithm 7.3** *Optimal SIR Particle Filter*

---

$[\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N] =$SIR$[\{x_{t-1}^{(i)}, w_{t-1}^{(i)}\}_{i=1}^N, y_t]$

**for** $i$=1 to $N$ **do**

Draw $x_t^{(i)} \sim q_{opt}(x_t|x_{t-1}^{(i)}, y_t) \propto p(y_t|x_t, x_{t-1}^{(i)})p(x_t|x_{t-1}^{(i)})$

Assign the particle a weight, $w_t^{(i)}$, according to (7.45)

**end for**

Calculate total weight: $W =$SUM$[\{w_t^{(i)}\}_{i=1}^N]$

**for** $i$=1 to $N$ **do**

Normalize: $w_t^{(i)} = W^{-1}w_t^{(i)}$

**end for**

Calculate $\hat{N}_{eff} = \frac{1}{\sum_{i=1}^N (w_t^{(i)})^2}$

**if** $\hat{N}_{eff} < N_{thres}$ **then**

Resample using Algorithm 7.4: $[\{x_k^{(i)}, w_t^{(i)}, -\}_{i=1}^N] =$RESAMPLE$[\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N]$

**end if**

---

There are many methods to resample from the posterior (see *Doucet et al.* [2001] or Chapter 6 of *de Freitas* [1999] for a discussion of methods, and *Arulampalam et al.* [2002] for a brief overview). The basic idea is to eliminate particles that have small weights and to concentrate on particles with large weights. It involves generating a new

set of particles and associated weights by resampling (with replacement) $N$ times from an approximate discrete representation of the posterior. The resulting sample is an independently and identically distributed sample so that the weights are reset to $1/N$. The method of choice in *Arulampalam et al.* [2002] is systematic resampling since "it is easy to implement, takes $O(N)$ time and minimizes the Monte Carlo variation." Its operation is described in Algorithm 7.4, where $U[a, b]$ is the uniform distribution on the interval $[a, b]$. For each resampled particle $x_t^{j\star}$, this resampling algorithm also stores the index of its parents, which is denoted $i^j$. This is unnecessary and can easily be suppressed, but may be useful in some situations.

### 7.7.4 Regularized Particle Filter (RPF)

We introduced resampling to counter particle degeneracy. However, resampling introduces the problem of a loss of diversity among the particles. This arises due to the fact that in the resampling stage, samples are drawn from a discrete distribution rather than a continuous one. If this problem is not addressed properly, it may lead to "particle collapse," which is a severe case of sample impoverishment where all $N$ particles occupy the same point in the state space, resulting in a poor representation of the posterior. The regularized particle filter was introduced as a potential solution to this problem.

The RPF filter is identical to the (O)SIR filter, except for the resampling. The RPF samples from a continuous approximation of the posterior density, whereas the (O)SIR samples from the discrete one. Specifically, samples in the RPF are drawn from the approximation:

$$p(x_t|y_{1:t}) \approx \sum_{i}^{N} w_k^{(i)} K_h(x_t - x_t^{(i)}) \tag{7.46}$$

**Algorithm 7.4** *Systematic Resampling*

$[\{x_t^{(j\star)}, w_t^{(j)}, i^j\}_{j=1}^N]$ =RESAMPLE$[\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N]$

Initialize the CDF: $c_1 = 0$

**for** $i$=2 to $N$ **do**

Construct CDF: $c_i = c_{i-1} + w_t^{(i)}$

**end for**

Start at the bottom of the CDF: $i = 1$

Draw a starting point: $u_1 \sim U[0, N^{-1}]$

**for** $j$=1 to $N$ **do**

Move along the CDF: $u_j = u_1 + N^{-1}(j-1)$

**while** $u_j > c_i$ **do**

$i = i + 1$

**end while**

Assign sample: $x_t^{(j\star)} = x_t^{(i)}$

Assign weight: $w_t^{(j)} = N^{-1}$

Assign parent: $i^j = i$

**end for**

where

$$K_h(x) = \frac{1}{h} K(\frac{x}{h}) \tag{7.47}$$

is the rescaled Kernel density $K(\cdot), h > 0$ is the Kernel bandwidth (a scalar parameter) and $w_t^{(i)}$ are normalized weights. The Kernel density is symmetric such that its average (over $x$) is zero and its variance is finite. The Kernel $K(\cdot)$ and bandwidth $h$ are chosen to minimize the mean integrated square error (MISE) between the true posterior density and the corresponding regularized empirical representation in equation (7.46),

which is defined as

$$\text{MISE}(\hat{p}) = E\left[\int [\hat{p}(x_t|y_{1:t}) - p(x_t|y_{1:t})]^2 dx_t\right] \tag{7.48}$$

where $\hat{p}(\cdot|\cdot)$ denotes the approximation to $p(x_t|y_{1:t})$ given by the right hand side of equation (7.46). In the special case of all the samples having the same weight, the optimal choice of the kernel is the Epanechnikov kernel. Taking its expression from equation (76) of *Arulampalam et al.* [2002] and setting the dimensionality of the state vector $x$ to $n_x = 1$ and substituting the volume of the unit hypersphere $c_{n_x} = 1$ for a 1D sphere gives

$$K_{opt} = \begin{cases} \frac{3}{4}(1 - ||x||^2) & \text{if } ||x|| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{7.49}$$

Furthermore, when the underlying density is Gaussian with a unit covariance matrix, the optimal choice for the bandwidth is

$$h_{opt} = AN^{1/5} \tag{7.50}$$

$$A = [40\sqrt{\pi}]^{1/5} \tag{7.51}$$

Although the results of equations (7.49), (7.50) and (7.51) are optimal only in the special case of equally weighted particles and underlying Gaussian density, these results can still be used in the general case to obtain a suboptimal filter. Note that the choice of kernels is less important than the choice of bandwidth. One may have to experiment with several bandwidths (and kernels) to gain experience. One iteration of the RPF is described by Algorithm 7.5. Here, we did not specify the importance density, so that the weights are calculated by the general equation (7.36), reproduced here for convenience:

$$\tilde{w}_t^{(i)} \propto \tilde{w}_{t-1}^{(i)} \frac{p(y_t|x_t^{(i)})p(x_t^{(i)}|x_{t-1}^{(i)})}{q(x_t^{(i)}|x_{0:t-1}^{(i)}, y_{1:t})} \tag{7.52}$$

**Algorithm 7.5** *Regularized Particle Filter*

$[\{x_t^{(i\star)}, w_t^{(i)}\}_{i=1}^N]$ =RPF$[\{x_{t-1}^{(i)}, w_{t-1}^{(i)}\}_{i=1}^N, y_t]$

**for** $i$=1 to $N$ **do**

Draw $x_t^{(i)} \sim q(x_t|x_{t-1}^{(i)}, y_t)$

Assign the particle a weight, $w_t^{(i)}$, according to (7.52)

**end for**

Calculate total weight: $W$ =SUM$[\{w_t^{(i)}\}_{i=1}^N]$

**for** $i$=1 to $N$ **do**

Normalize: $w_t^{(i)} = W^{-1} w_t^{(i)}$

**end for**

Calculate $\hat{N}_{eff} = \frac{1}{\sum_{i=1}^N (w_t^{(i)})^2}$

**if** $\hat{N}_{eff} < N_{thres}$ **then**

Calculate the empirical covariance matrix $S_t$ of $\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N$

Compute $D_t$ such that $D_t D_t^T = S_t$

Resample using Algorithm 7.4:

$[\{x_t^{(i)}, w_t^{(i)}, -\}_{i=1}^N]$ =RESAMPLE$[\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^N]$

**for** $i$=1 to $N$ **do**

Draw $\epsilon^{(i)} \sim K$ from the Epanechnikov Kernel

$x_t^{(i)^\star} = x_t^{(i)} + h_{opt} D_t \epsilon^{(i)}$

**end for**

**end if**

In terms of complexity, the RPF is comparable to the SIR since it only requires $N$ additional generations from the kernel $K(\cdot)$ at each time step. The RPF has the theoretic disadvantage that the samples are no longer guaranteed to asymptotically approximate those from the posterior. In practical scenarios, the performance is better than the SIR in cases where sample impoverishment is severe, for example, when the process noise is small.

While there are of course many more particle filters, each suited to particular applications, we have here presented the standard algorithms. For more advanced particle filters, see for instance [*Arulampalam et al.*, 2002; *de Freitas*, 1999; *Doucet et al.*, 2000] and references therein. The particle filters described above will be used below for seismicity models based on point processes.

## 7.8 A Brief Literature Review of Point Process Models in Data Assimilation Applications

Section 7.6 provided an introduction to data assimilation and sequential Monte Carlo, while section 7.7 illustrated some examples of standard particle filters, which will be used later in section 7.10. This section briefly reviews the relatively scarce literature in which point process models appeared in data assimilation problems.

There has been some interest in estimation problems for filtered point processes that are observed in noisy conditions [*Andrieu et al.*, 2001; *Hero*, 1991; *Kwakernaak*, 1980]. Filtered point processes are often of the form

$$x(t) = \sum_{j=1}^{k} h(t - t_j; \chi_j) \tag{7.53}$$

where the occurrence times $t_j$ belong to an underlying point process (e.g. Poisson or renewal process) and the amplitude of the resulting signal is given by the response

function $h(\cdot)$ which may depend on a stochastic variable $\chi$. Noisy filtered point processes are then given by:

$$G(t) = \sum_{j=1}^{k} h(t - t_j; \chi_j) + n_k(t) \tag{7.54}$$

where $n_k(t)$ is a random noise term, often assumed to be Gaussian white noise. This class is closely related to shot noise processes [*Daley and Vere-Jones*, 2003]. The filtering problem is then to estimate parameters, including the number of events $k$ and their arrival times $t_j$ given the observations (either discrete or continuous) of $G(t)$. Many estimation methods, including SMC, have been applied to these types of problems. However, in contrast to seismicity models, the signal $y(t)$ is the observable, unlike the (unobservable) conditional intensity function that drives another (seismologically observable) point process. (Note that fractal shot noise [*Lowen and Teich*, 1990], when driving another point process, is indeed highly relevant for seismicity).

Neural spike trains in rats show modified behavior depending on the animals spatial position and its experience with its spatial environment. *Barbieri et al.* [2001] and *Brown et al.* [1998, 2001] developed point process models and sequential filter algorithms that can be summarized in two steps: (i) the modeling of the firing patterns based on a biological signal, and (ii) the estimation of the biological signal based on neural spike observations. Thus the conditional intensity of the point process is a function of an underlying state (the biological signal), which is usually assumed to evolve as a Markov state-space model. The authors developed adaptive estimators that were based on Gaussian approximations. *Wang et al.* [2006] implemented a particle filter based on sequential Monte Carlo methods that outperformed their method. These studies did not explicitly include uncertainties in their estimation problems. Furthermore, their point processes were usually non-homogeneous Poisson processes (Poisson processes with a time-dependent rate) or Markov renewal processes. We mention these studies because they highlight that point processes are usually assumed to be driven

by an underlying state. In the Earth's crust, the underlying state may be related to the stress field, rate-and-state friction, chemical and physical properties of rock, etc. The further development of probabilistic point process models that relate these states to seismicity (e.g. estimating the frictional rate-and-state properties of faults from seismicity observations) remain an interesting future avenue. Some work in this vein was reported in [*Vere-Jones*, 1978; *Zheng and Vere-Jones*, 1991; *Dieterich*, 1994; *Bebbington and Harte*, 2003; *Ouillon and Sornette*, 2005].

Estimation procedures have also been applied to partially observed queues. *Doucet et al.* [2006] developed a trans-dimensional SMC method that was designed to estimate the number of clients in a queue based on the observations of customer departure times. Arrivals to the queue were based on a Poisson process, as were the service times, and the maximum waiting times for customers before they decided to leave. This example seems to be the only article (published in conference proceedings in 2006) known to us in which the occurrence (arrival) times are estimated based on some perturbed observation thereof. Their method is thus the closest in spirit to our approach to estimate true occurrence times, magnitudes and locations of earthquakes.

Finally, it is worth mentioning that in the spatial point process community, several papers have been published on estimating the type of noisy perturbations in the setting where both perturbed and original data are given (see [*Lund and Rudemo*, 2000] and references therein). It should be possible to generalize their approach to the case where the perturbation and the perturbed data is known to estimate the original data in a marked spatio-temporal context.

The relative scarcity of literature in which point process models are applied in a data assimilation setting may be related to two factors. First, the mathematical and computational techniques for Bayesian data assimilation have only become available to a broad community over the last ten years or so. Secondly, the theory of point

processes is also a relatively new field, especially outside the mathematical statistics community.

The remainder of this chapter is our contribution to the implementation of sequential Monte Carlo methods for point process models. Sections 7.9 and 7.10 concern the simplest class of point processes: renewal processes under noisy occurrence times observations. While section 7.9 discusses their mathematical formulation as a state-space model (which is the preferred representation in SMC methods), in section 7.10 we will implement particle filters on synthetic data and test their improvement over techniques that neglect uncertainties in numerical experiments. Finally, section 7.11 will discuss the generalization to general marked point processes before concluding the chapter.

## 7.9 Renewal Point Processes with Noisy Occurrence Times as State-Space Models: On the Mathematical Framework

In section 7.10, we will present an implementation of a particle filter for a 1D temporal renewal process under noisy observation of occurrence times. In this numerical experiment, we will simply assume that the SMC methods presented in section 7.7 are applicable to point processes. One particular assumption is that occurrence times can be described in the framework of state-space models. This section discusses this assumption, as it leads to some theoretical challenges in the particle filter. The challenges arise from the particular combination of renewal processes and noisy occurrence times and do not present a problem when (i) occurrence times are assumed to be exact and other variables (magnitude, locations, focal mechanisms) are noisy, or (ii) occurrence times are noisy but the point process model does not care about the particular ordering of events, including which event is the last one. The latter condition precludes renewal processes. This section thus discusses theoretical problems from the perspective of

temporal point processes with noisy occurrence times, which may not apply to other point process models or when the state estimation problem does not involve occurrence times. The issues can be traced back to the issue that point processes are not naturally state-space models.

State-space models are a mathematical representation of physical systems as coupled first order differential equations. The state-space refers to the space whose axes are the state variables (known as phase-space in dynamical systems theory). An example of such a model is

$$x_t = f_t(x_{t-1}, \eta_t) \tag{7.55}$$

$$y_t = g_t(x_t, \nu_t) \tag{7.56}$$

where $x_t$ (potentially a vector) is called the state and $y_t$ (potentially a vector of different dimensionality) is called the observation, and where the (deterministic) functions $f_t$ and $g_t$ may in general be nonlinear and depend on time. The measurement noise $\nu_t$ and the process noise $\eta_t$ are random variables that are often modeled as Gaussian. This particular example model is Markovian as the state $x_t$ only depends on $x_{t-1}$. The state space model usually contains parameters that may need to be estimated (including the parameters of the distributions of the measurement and process noise).

Although a more general approach exists (using Liouville's equation together with Bayesian updating [*Ehrendorfer*, 1994]), most sequential data assimilation methods rely on the representation given by (7.55) and (7.56), including nonlinear/non-Gaussian "stochastic" methods [*Liu and Chen*, 1998; *Miller et al.*, 1999; *Pham*, 2001; *Alexander et al.*, 2005; *Doucet et al.*, 2001; *Sornette and Ide*, 2001]. Given the field's origins in numerical weather prediction, physical oceanography, and object tracking, it is not surprising that model representations based on differential equations underlie much of data assimilation [*Ide et al.*, 1997; *Kalnay*, 2003].

When the occurrence times, locations and marks represent the "state", as in our

problem statement, then point process models do not seem to be naturally formulated as in equation (7.55). Unlike continuous stochastic processes such as Wiener or Ornstein-Uhlenbeck processes, occurrence times are discrete and random. Assume the last event occurred at time $t_k$, where $k$ marks the number of the event, rather than the discrete observation interval $t_t = t_{t-1} + dt$, where $dt$ is a small fixed interval, which is often the case in methods that periodically measure and update the state via equations (7.55) and (7.56). As stated in section 7.4.1, the probability density function of the time until the next event is was given by equation (7.5):

$$f(t_k|H_{t_k}) = \lambda(t_k|H_{t_k}) \exp\left(-\int_{t_{k-1}}^{t_k} \lambda(u|H_u)du\right) \quad t_k > t_{k-1} \tag{7.57}$$

where $H_{t_k} = \{t_j\}_{j=1}^{k}$ is the history of the process and we have for this discussion assumed that the process is temporal. The distribution (7.57) is determined by the particular conditional intensity function (i.e. the point process) and is of a wide variety. The evolution of the process in terms of the state (the occurrence times), may perhaps be written as

$$t_k = t_{k-1} + \tau_k(H_{t_k}) \tag{7.58}$$

where $\tau_k(H_{t_{k-1}})$ is the random interval between the $(k-1)$st and the $k$th event that depends on the history $H_{t_k}$ up to time $t_k$.

Data assimilation methods often assume that the true dynamics of the system can be separated into a deterministic (large scale) evolution and a stochastic (small scale) evolution. For example, the large scale climate (deterministic) may be locally changed by turbulent structures. Hence, the state evolution is often separated into a deterministic part and a small perturbative stochastic part (process noise). For the point process models under considerations for earthquakes, this separation of scales is no longer possible. The dynamics is entirely stochastic and discrete. For illustrative purposes, we

may write

$$\begin{aligned} x_t = f_t(x_{t-1}, \eta_t) \quad &\longleftrightarrow \quad t_k = t_{k-1} + \tau_k(H_{t_k}) \\ \text{state-space model} \quad &\longleftrightarrow \quad \text{point process} \end{aligned} \tag{7.59}$$

Remarks:

1. **The state is given by occurrence times:** The state, usually assumed to be a physical quantity to evolve as a function of time, is now a temporal quantity itself. The index $k$ refers to the $k^{\text{th}}$ event and not the standard $k^{\text{th}}$ discretized time interval. Thus the state evolves discretely, but in unequal intervals, which are given by the pdf of the point process.

2. **The state strictly increases:** Since $\tau_k(H_{t_k}) > 0$, the state is monotonically increasing so that $t_k > t_{k-1} \; \forall \; k > 1$. The support of the posterior distribution of the $k^{\text{th}}$ occurrence time is always on the same space as the posterior of previous occurrence times.

3. **Markov point processes:** Many data assimilation methods, including sequential Monte Carlo methods, assume that the state-space model is a Markov process, i.e. that $x_t$ only depends on $x_{t-1}$ and not previous states. Only a small class of point processes, renewal processes, are Markovian in the sense that the next occurrence time depends solely on the last occurrence time.

4. **Overlapping posteriors (important for noisy occurrence times in Markov point processes):** Suppose that the occurrence times of the temporal point process are noisy. A comparison with the state space model yields:

$$\begin{aligned} y_t = g_t(x_t, \nu_t) \quad &\longleftrightarrow \quad t^o = t^t + \epsilon \\ \text{state-space model} \quad &\longleftrightarrow \quad \text{noisy temporal point process} \end{aligned} \tag{7.60}$$

where $\epsilon$ is an additive noise term distributed according to some distribution $p_\epsilon(\epsilon)$. If occurrence times are narrowly-spaced and the noise term is large, we are faced with the possibility of two posteriors of estimated occurrence times overlapping. If, additionally, the chosen model is Markovian (renewal process), then identifying the last event is no longer trivial. There is a finite probability for each of the two posteriors to represent the last event. One can in principle compute the probability of each event being the last one, but this can result in a major combinatorial challenge when several posteriors overlap. However, this does not present a problem for point processes with long memory, such as those used for short-term earthquake forecasting that include clustering, since all occurrence times determine the future, not just the last one.

5. **True versus observed ordering of events (important for noisy occurrence times in Markov point processes):** Suppose again that the occurrence times are narrowly-spaced and the observational noise is large. In principle, it is possible that the observed ordering of events is not equal to the true ordering of events. As above, epidemic-type models are not affected by the re-ordering, but other models, especially renewal processes, are. The theoretical challenge is then to solve the general mathematical problem of devising a forecasting technique that takes into account the possibility of re-ordering. Depending on the application, this effect may be negligible.

6. **The completeness of the observational history and conditioning forecasts (important for uncertain occurrence times in all point processes):** Depending on the particular application, the history of observations may not be complete. By this, we mean that information about an occurrence time $t_k < t$ may only be available to us at a later time $t$ (say, the present). This is the standard situation for earthquake catalogs: there is always a time lag between the end of

the observational period during which information is available in catalogs, and the present. Worse still, some but not all information might be available about the recent past. Finally, one may wish to truly forecast from the present for the next 24 hours, but if information is only available up to some time in the past, then one needs to condition the forecast on the forecast window. Any of these issues may not be trivial to solve exactly or approximately.

As can be seen from this discussion, it is not wholly satisfactory to use the state-space model representation for temporal renewal point processes when the occurrence times are noisy. However, we underline that most of the issues discussed are only relevant in the context of this particular class of process and uncertain occurrence times. In the case of self-exciting processes or in situations where the occurrence times can be treated as exact, all but the last issue no longer present obstacles. Finally, it may be noted that the particular impact on the performance of the filters and forecasts may be negligible depending on the application. We thus side-step these issues for now and present in section 7.10 an implementation of a particle filter for a lognormal process.

## 7.10 A Particle Filter for Estimating Noisy Occurrence Times in Renewal Processes

This section presents numerical implementations of particle filters for renewal processes under noisy occurrence times. Although we want to formulate, in general, an algorithm for sequentially estimating the occurrence times, locations and magnitudes of earthquakes from noisy observations of the same quantities using relevant earthquake models, our strategy is to begin with the simplest models relevant for earthquakes and to later add complexity. This dissertation will only present various particle filters for a renewal process and comment in section 7.11 on generalizing the point

process models.

The simplest models relevant to earthquakes are one-dimensional (temporal) renewal processes. The models are designed to capture large earthquakes, effectively eliminating all the small (and not so small) events following big quakes. They are routinely used for purposes of seismic hazard analysis, including, for example, the US national seismic hazard map. Furthermore, renewal processes and their underlying scientific hypotheses are often compared on paleoseismological data.

This section is organized as follows. We discuss the particular renewal process in section 7.10.1 and the noisy observations in section 7.10.2. In section 7.10.3, we state the Bayesian solution of the sequential estimation problem as a special case of the general results of 7.6.1. Section 7.10.4 presents the results of the numerical experiments.

### 7.10.1 The Model: Lognormal Renewal Process

Motivated by its relevance to seismicity on fault segments and in seismic hazards (see section 7.4), we use a lognormal renewal process as our model of intervals between subsequent earthquakes:

$$f_{lognormal}(\tau; \mu, \sigma) = \frac{1}{\tau\sqrt{2\pi}\sigma} \exp(-(\log \tau - \mu)^2/2\sigma^2) \tag{7.61}$$

where we shall assume for simplicity that the parameters $\mu$ and $\sigma$ are known.

### 7.10.2 The Observations: Noisy Occurrence Times

We suppose that the observed occurrence times are noisy:

$$t^o = t^t + \epsilon \tag{7.62}$$

where $\epsilon$ is an additive noise term distributed according to some distribution $p_\epsilon(\epsilon)$.

Typically, earthquake catalogs report root-mean-square uncertainties calculated

from P-wave arrivals at different stations under a crustal velocity model. These are on the order of milliseconds to a few seconds (see section 7.3).

What is the appropriate density of the occurrence time uncertainty? An unbounded distribution is clearly wrong: When the earthquake is registered by seismometers, it must have happened before that time with probability one. One can probably make more accurate absolute limits, given that the velocity structure of the Earth is known to within a few percent. The distribution is necessarily bounded on both ends, say $-a < \epsilon < b$ where the constants $a, b > 0$. Simple densities are the uniform (uni) density, a triangle (tri) density, and a truncated Gaussian (tg) density:

$$p^o_{uni}(\epsilon) = \begin{cases} \frac{1}{b-a}, & \text{if } a < \epsilon < b \\ 0, & \text{otherwise} \end{cases} \tag{7.63}$$

$$p^o_{tri}(\epsilon) = \begin{cases} \frac{2}{b+a}(\epsilon/a + 1), & \text{for } -a < \epsilon < 0 \\ \frac{2}{b+a}(-\epsilon/b + 1), & \text{for } 0 < \epsilon < b \\ 0, & \text{otherwise} \end{cases} \tag{7.64}$$

$$p^o_{tg}(\epsilon) = \begin{cases} C \exp(-\epsilon^2/2\sigma^2_\epsilon) & \text{for } -a < \epsilon < b \\ 0, & \text{otherwise} \end{cases} \tag{7.65}$$

where $C$ is a normalization constant so that $\int_a^b p^o_{tg}(\epsilon) d\epsilon = 1$. The limits $a$ and $b$ may be chosen as multiples of the root-mean-square error reported in catalogs, which also determines $\sigma_\epsilon$. For simplicity, we choose as a first step the uniform distribution given by

$$p_\epsilon(\epsilon) = \frac{1}{\Delta} H(\epsilon + \frac{\Delta}{2}) H(\frac{\Delta}{2} - \epsilon) \tag{7.66}$$

where $H(\cdot)$ is the Heaviside step function. Substituting $\epsilon = t^o - t^t$ gives the densities (likelihoods) of the data given the true occurrence time:

$$p^o_{(\epsilon)}(\epsilon) \leftrightarrow p^L(t^o - t^t) \tag{7.67}$$

### 7.10.3 The Bayesian Solution

The Bayesian solution presented here is a straightforward application of the general solution presented in section 7.6.1. The prior (forecast) can be written using the Chapman-Kolmogorov equation:

$$p(t_k^t|t_{k-1}^o,\ldots,t_1^o) = \int p(t_k^t|t_{k-1}^t,t_{k-1}^o,\ldots,t_1^o)\cdot p(t_{k-1}^t|t_{k-1}^o,\ldots,t_1^o)dt_{k-1}^t \quad (7.68)$$

$$= \int p(t_k^t|t_{k-1}^t)\cdot p(t_{k-1}^t|t_{k-1}^o,\ldots,t_1^o)dt_{k-1}^t \quad (7.69)$$

Note that in (7.69), use has been made of the fact that $p(t_k^t|t_{k-1}^t,t_{k-1}^o,\ldots,t_1^o) = p(t_k^t|t_{k-1}^t)$ for a renewal (Markov) process. Non-Markov models (e.g. ETAS) will need to integrate over the joint pdf.

The update equation is given by

$$p(t_k^t|t_k^o,\ldots,t_1^o) = \frac{p(t_k^o|t_k^t)\cdot p(t_k^t|t_{k-1}^o,\ldots,t_1^o)}{p(t_k^o|t_{k-1}^o,\ldots,t_1^o)} \quad (7.70)$$

where the prior is given by the forecast equation (7.69), the likelihood is given by the noise process (7.67) and the normalization constant is given by

$$p(t_k^o|t_{k-1}^o,\ldots,t_1^o) = \int p(t_k^o|t_k^t)\cdot p(t_k^t|t_{k-1}^o,\ldots,t_1^o)dt_k^t \quad (7.71)$$

The recurrence relations (7.69) and (7.70) form the basis for the optimal Bayesian solution.

### 7.10.4 Numerical Experiments

#### 7.10.4.1 SSIS Filter for the Lognormal Renewal Process

We first present results for the simple sequential importance sampling (SSIS) algorithm, given by Algorithm 7.1 defined in section 7.7.1. We will find that the SSIS filter deteriorates quickly, as discussed below. We therefore only include it as a pedagogical example.

We initialized the algorithm with 1000 particles with a perfectly observed event at time $t = 0$. We used the parameters $\mu = 2$ and $\sigma = 1/8$ for the lognormal model (7.61). The uniform noise distribution (7.66) is parametrized by the parameter $\Delta = 1$. We introduce the "naive" forecast, which assumes that observed occurrence times are exact and uses the analytical form (7.61), as the benchmark against which the SSIS filter should be evaluated. Because the first event at $t_1 = 0$ is perfectly observed, the "naive forecast" for the second event $t_2$ is actually entirely correct. This quickly changes as noise is introduced for the second event.

Figure 7.1 shows the results for step 1. The left panel shows the forecast for the second event obtained by propagating the particles from the perfectly observed first event $t_1 = 0$. Black crosses underneath the histogram refer to active particles with non-zero weight. The true event (dashed) and the observed event (dotted) almost coincide. The "naive" forecast (solid curve) should coincide with the discrete Monte Carlo estimate, since we assumed no noise for the initial event. It is apparent that 1000 particles are not sufficient to obtain a good approximation. Although increasing the number of particles to 100,000 does not present a computational problem and can indeed approximate the analytical forecast very well (not shown), there is little reason to attempt it, because the number of active particles decays very quickly, as discussed below.

In the right panel of Figure 7.1 we show the posterior (analysis) obtained by combining the forecast with the likelihood. Since the SSIS filter uses the prior (forecast) as the importance density to sample from the posterior, most of the particles "die" (grey crosses) at this point and only about 400 of the original 1000 particles remain active (black crosses). In situations where the forecast is broad and the likelihood narrow, most of the particles will "die" in this sampling scheme.

Figure 7.2 shows the forecast and the posterior for the third event. Since there is

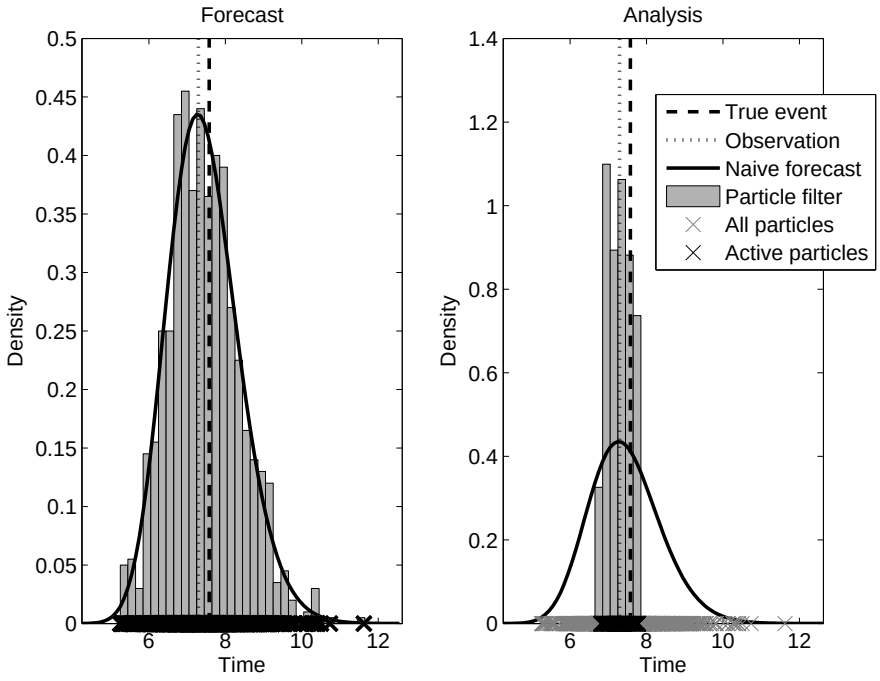


Figure 7.1: Step 1 of the SSIS filter to forecast and estimate the exact occurrence time from a noisy observation and a model prior (forecast). Left panel: "naive" forecast (solid curve) and discrete particle filter approximation using 1000 particles. Black crosses indicate active particles. Right panel: posterior after assimilating the noisy observation into the model prior. Grey crosses indicate particles with zero weight, indicating rapid particle loss.

no resampling to replenish the active particles in the SSIS filter, all particles, whether dead or alive, are propagated by the model forecast, but only active particles (black crosses) contribute to the forecast density. The "naive" forecast is now no longer correct, because it is based on a noisy observed event. However, the discrete particle filter is clearly having difficulty approximating the "true", smooth forecast. Again, increasing the number of particles is feasible but would only lead to a delayed effect. Nevertheless, we can see a marginal improvement of the particle filter over the naive

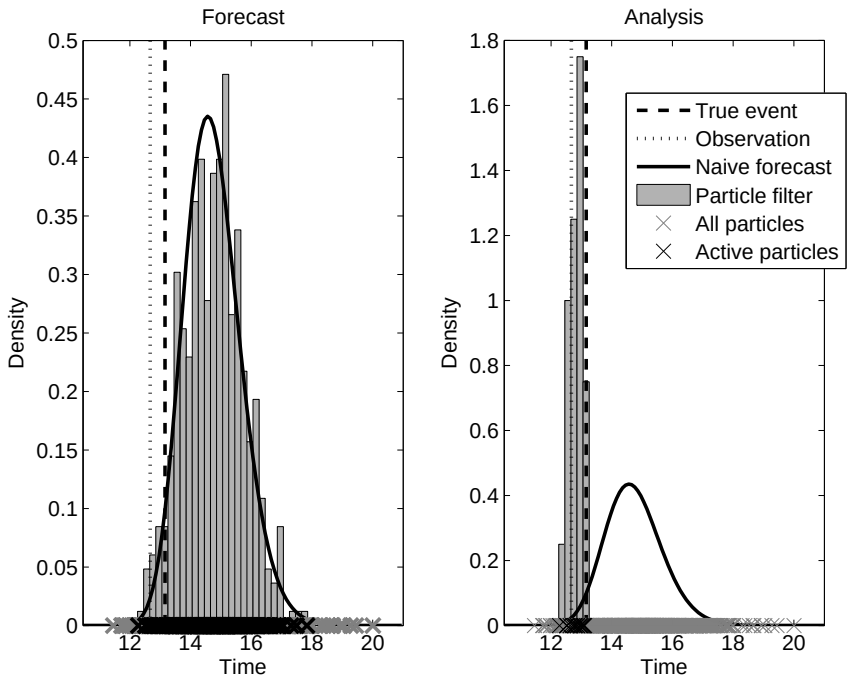


Figure 7.2: Step 2 of the SSIS filter: Only 20 of the original 1000 particles remain after the analysis in step 2.

forecast, because allowing for uncertainty in the last step broadened the particle filter forecast.

The right panel shows the analysis after combing the forecast and the likelihood for the third event. Only about 20 particles remain active. We can conclude that a more sophisticated filter is needed for the scenario of broad forecasts and narrow likelihoods in order to keep the particles alive. Nevertheless, increasing the number of particles to 100,000 can already provide substantial improvement over the naive forecast for a dozen or so steps (not shown).

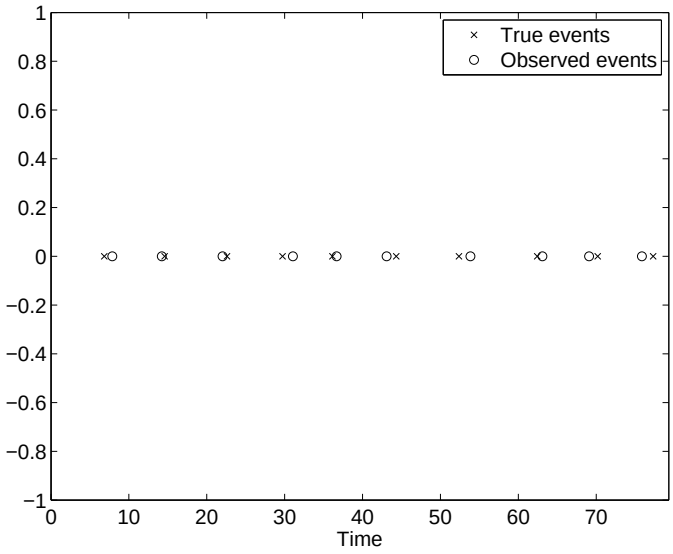


Figure 7.3: Identical twin experiment used as demonstration of the OSIS filter: Realization of the true process (crosses) and the perturbed, noisy observations (circles).

#### 7.10.4.2 OSIS Filter for the Lognormal Renewal Process

We can improve on the SSIS filter by using the optimal importance density to sample from the posterior, i.e. using the OSIS filter defined by Algorithm 7.2 in section 7.7.2. It is usually impossible to sample directly from the optimal importance density (see section 7.6.2.4), but our choice of a truncated, uniform likelihood makes this possible.

To use the optimal density, we need to overcome two problems (see section 7.6.2.4): (i) sampling from the optimal density $p(x_t|x_{t-1}^{(i)}, y_t)$, and (ii) performing the integral $p(y_t|x_{t-1}^{(i)})$.

To overcome (i), in the case of a uniform, truncated likelihood, one can use rejection sampling to sample from the optimal density: sample a candidate $\tilde{x}_t^{(i)}$ from the forecast $p(x_t|x_{t-1}^{(i)}$ but accept the candidate only if it falls into the interval $[y_t - \Delta/2, y_t + \Delta/2]$, as dictated by the likelihood $p(y_t|x_t)$; otherwise reject the candidate

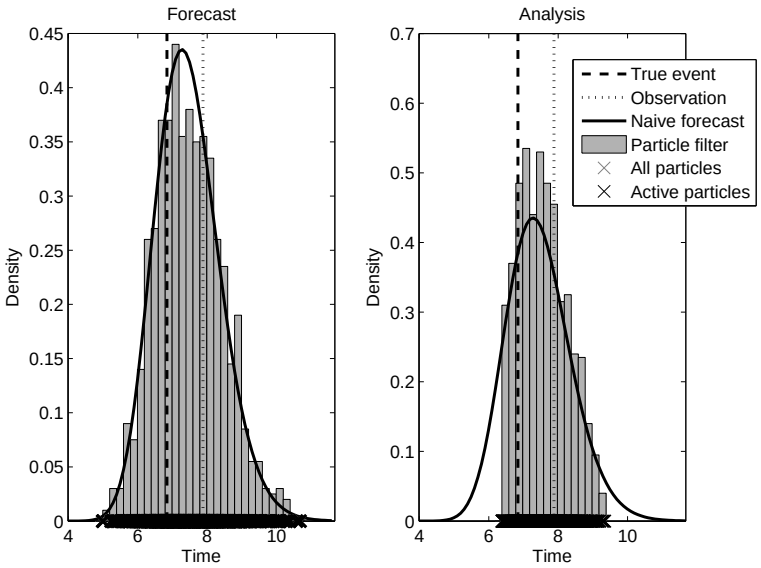


Figure 7.4: Step 1 of the OSIS filter: Same as Figure 7.1, but now all particles remain active (black crosses) during the sampling from the posterior.

and start over.

Moreover, the integral $p(y_t|x_{t-1}^{(i)})$ can be analytically transformed into error functions, which can easily be numerically solved in standard software packages such as Matlab or Mathematica. This opens the path for implementing the OSIS filter.

We again use an identical twin experiment in which we simulate the true process and then perturb it to obtain the observed process. As above, we assume that the first event $t_1 = 0$ is observed exactly, i.e. $t_1^o = t_1^t$. We then simulate 10 additional points $(t_2, \ldots, t_{11})$ according to the lognormal model equation (7.61) with parameters $\mu = 2$ and $\sigma = 1/8$. Next, we perturb these 10 points using $t^o = t^t + \epsilon$ and the uniform, truncated noise distribution (7.66) with $\Delta = 3$ to generate the noisy observations. The resulting true and noisy processes are shown in Figure 7.3.

Figure 7.4 presents the forecast and the analysis of step 1. All particles remain

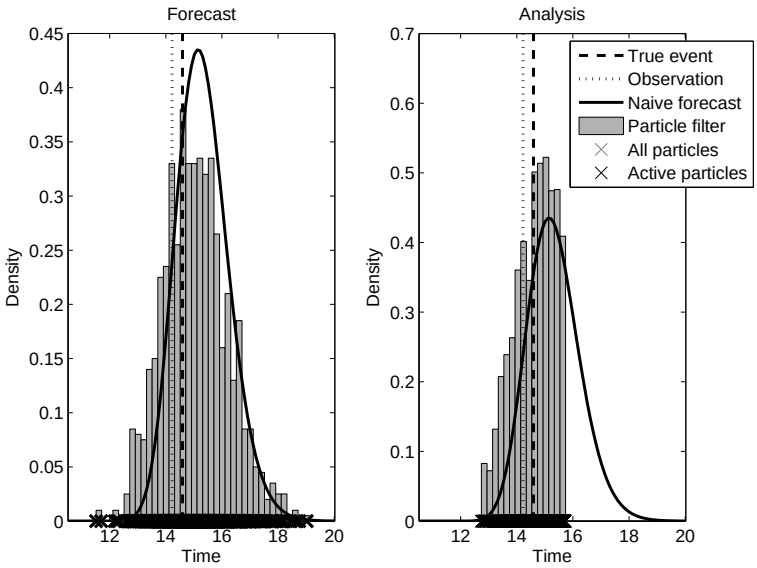


Figure 7.5: Step 2 of the OSIS filter.

active because of the new method of sampling from the posterior. Step 2 already shows a small improvement of the particle filter forecast over the benchmark forecast (see Figure 7.5). Step 5, shown in Figure 7.6, demonstrates that the particle filter forecast can significantly outperform the benchmark forecast, by placing more weight near the true and observed events.

We introduce one potential complication by using the rejection sampling scheme to sample from the posterior: it may be computationally very costly if the rejection rate is high. The rejection rate is high in the following scenario. Say in step $k-1$, the observation $t^o_{k-1}$ was much larger than the true event $t^t_{k-1}$. The support of the posterior is symmetric around the observation, so that a particle of the posterior, say $p_1$, may be as large (late) as $t^t_{k-1}+\Delta$. Now suppose that the mean of the lognormal forecast model is on the order of or even smaller than $\Delta$, then the situation may arise in which $t^t_k < t^t_{k-1}+\Delta$ and perhaps even $t^t_k < t^o_{k-1}$. If, by chance, the observation $t^o_k$ is also small, e.g.

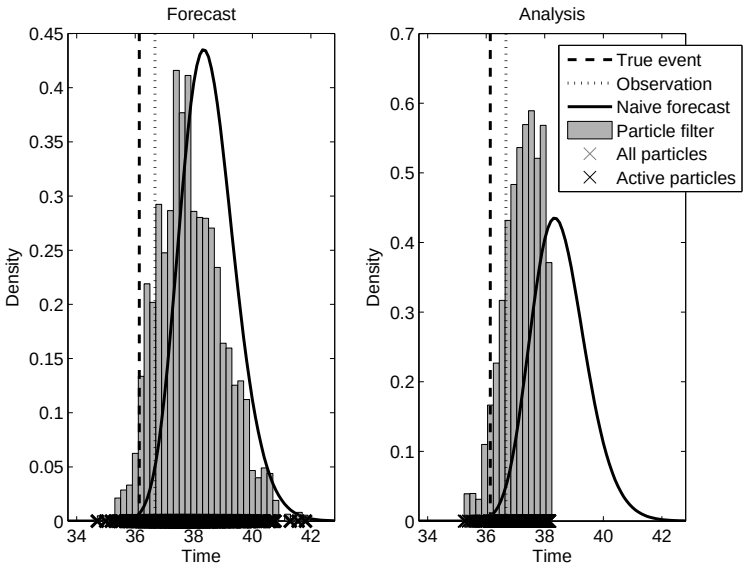


Figure 7.6: Step 5 of the OSIS filter: The particle filter forecast significantly outperforms the benchmark forecast by placing much more weight near the true and observed events.

$t_k^o < t_k^t$, then it may be extremely difficult to propagate particle $p_1$ from $t_{k-1}^t + \Delta$ into the next posterior, which is bounded (in its support) from above by $t_k^o + \Delta/2$, because the probability of propagating $p_1$ from its old position to an allowed new position is nearly zero. In fact, in the case where the observation occurs before the location of particle $p_1$, i.e. when $t_k^o < t_{k-1}^t + \Delta$, it is impossible to propagate the particle. This discussion is intricately connected to the issue of overlapping posteriors, as discussed in section 7.9, and arises from the particular combination of renewal processes and noisy occurrence times in a filter for state-space models. Depending on the parameters of the noise and the model, our rejection sampling scheme may not be feasible in its current form.

Moreover, despite the fact that all particles remain active throughout the procedure, many particles will nevertheless see their weights significantly reduced over time,

while a few lucky survivors dominate the weights. Figure 7.7 shows the cumulative distribution function of the weights of the particles for each of the 10 steps. At first, all particles have equal weight, but after several steps, the distribution becomes more and more skewed. As discussed in section 7.6.2.4, it is impossible to overcome the problem of particle degeneracy, to which Figure 7.7 attests. To counter the problem, we can only resort to resampling.

However, the OSIS filter presented here is already an enormous improvement over the SSIS filter. Depending on the application, the OSIS filter may be more than sufficient and can also easily be upgraded to 100,000 particles on a standard personal computer. Furthermore, one may easily augment the OSIS filter in cases where the rejection rate is high. If a certain time threshold is surpassed attempting to sample from the posterior by using rejection sampling, one may resort to another importance density, such as the model prior or slightly more favorable densities, for the particular step, before resorting back to the optimal importance density. The OSIS filter presented here is thus a powerful, flexible and computationally inexpensive method for sequentially estimating noisy occurrence times in a renewal process.

## 7.11 Towards General Marked Point Process Models

Section 7.10 addressed noisy occurrence times in renewal processes. Here, we generalize to marked point processes that have an arbitrary temporal structure and unpredictable marks (defined below). We present the (conceptual) Bayesian solution for estimating true magnitudes from noisy magnitude observations but exact occurrence times, with marks that are independently and identically distributed to an invariant distribution $p(m)$. Based on the results of Chapter 6, it is likely that (random) magnitude errors distort forecasts of clustering models significantly more than the (random) temporal errors of section 7.10. Furthermore, we would like to use more realistic clus-

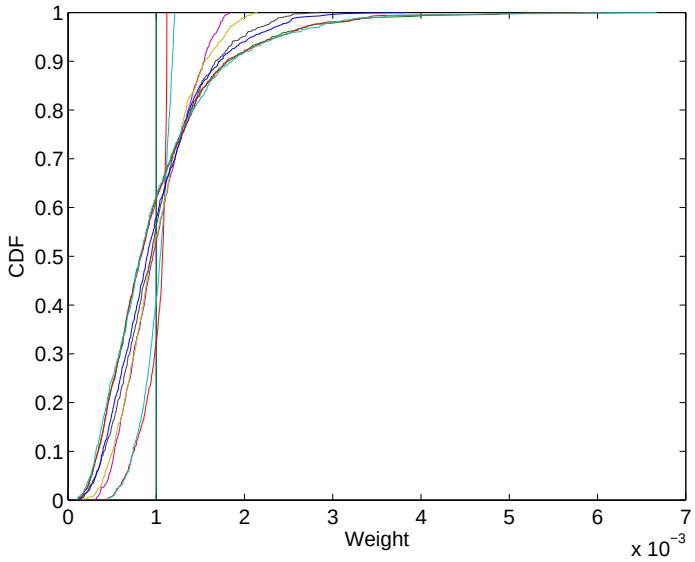


Figure 7.7: Evolution of the cumulative distribution function of the particle weights in the OSIS filter: At first, all particles have equal weight, but as the filter updates the weights at each analysis step, the distribution of the weights becomes more and more skewed at each analysis step.

tering models than the class of renewal processes to produce forecasts. The conceptual solution presented here can be used as the basis for implementing particle filters that estimate posterior distributions of the magnitudes in the ETAS model [*Ogata*, 1988], defined in equation (7.13) in section 7.4.1.

First, we state for clarity the model assumptions, the observational uncertainty assumptions, and the estimation goal. Then we show that estimators without memory are bound to fail. Next, we present the Bayesian solution. We then examine some limiting cases to guide our intuition before concluding.

### 7.11.1 Unpredictably Marked Temporal Point Processes

Point processes are completely defined by their conditional intensity $\lambda(t, m|H)$ which is the instantaneous probability of an event occurring at time $t$ given the history $H$ which contains all information about previous occurrence times and associated marks.

Unpredictable marks are those that do not depend on the history $H$, i.e.

$$\lambda(t, m|H) = p_m(m) \cdot \lambda(t|H) \tag{7.72}$$

independently of the so-called ground process $\lambda(t|H)$. Note, however, that the occurrence times $t_i$ are dependent on previous marks because the time to the next event $t_i$ given the history is equal to:

$$f(t_i|H) = \lambda(t_i|H) e^{-\int_{t_{i-1}}^{t_i} \lambda(u|H) du} H(t_i - t_{i-1}) \tag{7.73}$$

where $H(\cdot)$ is the Heaviside step function.

### 7.11.2 Noisy Marks and Exact Occurrence Times

We assume that marks are noisy according to

$$m^o = m^t + \epsilon \tag{7.74}$$

where $\epsilon$ is an additive noise term drawn independently from some distribution $p_\epsilon(\epsilon)$. At the same time, we assume that occurrence times are exact, such that

$$t_i^o = t_i^t \quad \forall\, i \tag{7.75}$$

### 7.11.3 Goal

The goal of using data assimilation is hence to obtain better estimates (posterior distributions) of the marks $\{m_j^t\}_{1 \leq j \leq i}$ up to the current event $i$, using all available information contained in the entire history and in the model. The outcome of this goal

is further to provide better forecasts for occurrence times based on a more accurate representation of the marks.

### 7.11.4 Estimator Without Memory

We will show in this section that a sequential estimator without memory cannot help us achieve our goal.

The most general updating is achieved via Bayes' theorem:

$$p(t_i^t, m_i^t | t_i^o, m_i^o) = \frac{p(t_i^o, m_i^o | t_i^t, m_i^t) \cdot p(t_i^t, m_i^t)}{p(t_i^o, m_i^o)} \tag{7.76}$$

We know that $t_i^t$ and $m_i^t$ are mutually independent, i.e. the left hand side is a product $p(t_i^t | t_i^o, m_i^o) \cdot p(m_i^t | t_i^o, m_i^o)$. The first term in this product does not depend on $m_i^o$ because of independence and, in any case, equals $\delta(t_i^t - t_i^o)$ as there is no noise. The second term of the product, on the other hand, does not depend on $t_i^o$, i.e. the left hand side of equation (7.76) equals $\delta(t_i^t - t_i^o) \cdot p(m_i^t | m_i^o)$, showing that the problem reduces to estimating the true mark from the observed mark. This estimation problem, however, is completely independent from anything else and trivial: Rewriting $p(m_i^t | m_i^o) \propto p(m_i^o | m_i^t) p_m(m^t) = p_\epsilon(m^t - m^0) p_m(m^t)$ we see that the forecast and likelihood do not depend on previous steps. While this is a perfectly legitimate estimator for marks given observations, it is trivial in that there is no updating.

### 7.11.5 Estimators With Memory

In the previous section, we saw that we need to introduce memory in the estimation problem. There are some choices we need to make. We can estimate the joint pdf of past magnitudes or we can estimate marginal pdfs of each one of the magnitudes. Naturally, the joint pdf determines the marginal. Furthermore, we can use observations up to time $t_k$ to constrain magnitudes up to $m_{k-n}$ where $n$ is not restricted to $1$.

While we may be interested in the marginal pdfs, we will start by determining the joint pdf of magnitudes given occurrence times and observed magnitudes one step ahead of the magnitudes we want to determined (the lag is 1).

Note that the entire history must be estimated, because updating only previous, say, the $(i-1)$st mark does not respect the dependence of the ith observation on the entire history (for the most general temporal point process). This means that the dimensionality of (really, number of entries in) the state vector (the true history $\{t_j^t, m_j^t\}_{1\leq j\leq i}$) increases at each step $i$ by one. Of course, the marginals can be computed after this step.

### 7.11.6 Bayesian Solution for Estimating Magnitude Posteriors from Noisy Magnitudes and Exact Occurrence Times in an Unpredictably Marked, Temporal Point Process

This section presents the solution for estimating $\{m_j^t\}_{1\leq j<i-1}$ from $\{t_j^o, m_j^o\}_{1\leq j\leq i}$.

At time step $i$, we want to estimate $\{m_j^t\}_{1\leq j<i}$ given the observed history $\{t_j, m_j^o\}_{1\leq j\leq i}$. Using Bayes' theorem and neglecting for now the normalization constant, we can write:

$$\begin{aligned} P_i^a &\equiv p_i^a(m_{i-1},\ldots,m_1|t_i,\ldots,t_1,m_{i-1}^o,\ldots,m_1^o) \\ &\propto\ p(t_i|m_{i-1},\ldots,m_1,t_{i-1},\ldots,t_1,m_{i-1}^o,\ldots,m_1^o) \\ &\quad \times p(m_{i-1},\ldots,m_1|t_{i-1},\ldots,t_1,m_{i-1}^o,\ldots,m_1^o) \end{aligned} \tag{7.77}$$

The likelihood can be simplified by noting that conditioning on $m_{i-1}^o,\ldots,m_1^o$ cannot lead to better information about $t_i$ than conditioning on $m_{i-1},\ldots,m_1$, thus making this conditioning redundant. Furthermore, we can rewrite the prior using conditional

probability $P(A,B) = P(A|B)P(B)$ to separate out $m_{i-1}$:

$$\begin{aligned} P_i^a \propto \; & p(t_i|m_{i-1},\ldots,m_1,t_{i-1},\ldots,t_1) \\ & \times p(m_{i-1}|m_{i-2},\ldots,m_1,t_{i-1},\ldots,t_1,m_{i-1}^o,\ldots,m_1^o) \\ & \times p(m_{i-2},\ldots,m_1|t_{i-1},\ldots,t_1,m_{i-1}^o,\ldots,m_1^o) \end{aligned} \tag{7.78}$$

where we can identify in line 1 of equation (7.78) the conditional probability of the ith occurrence time given the true previous history as given by equation (7.73). Furthermore, the only relevant information for $m_{i-1}$ in the second line of equation (7.78) is its observed value $m_{i-1}^o$, because the marks are unpredictable and occurrence times larger than $t_{i-1}$, which could help constrain $m_{i-1}$, are not included in the conditioning. Finally, note that the third line of equation (7.78) is actually $P_{i-1}^a$ as defined in expression (7.77). Thus

$$\begin{aligned} P_i^a \propto \; & \lambda(t_i|H_{i-1})e^{-\int_{t_{i-1}}^{t_i}\lambda(u|H_{i-1})du} \\ & \times p(m_{i-1}|m_{i-1}^o) \\ & \times P_{i-1}^a \end{aligned} \tag{7.79}$$

Now we can use Bayes' theorem again to compute the second line $p(m_{i-1}|m_{i-1}^o)$ to obtain

$$\begin{aligned} P_i^a \propto \; & \lambda(t_i|H_{i-1})e^{-\int_{t_{i-1}}^{t_i}\lambda(u|H_{i-1})du} \\ & \times p(m_{i-1}^o|m_{i-1})p_m(m_{i-1})/p(m_{i-1}^o) \\ & \times P_{i-1}^a \end{aligned} \tag{7.80}$$

where $p_m(m_{i-1})$ is given by the model prior (the exact Gutenberg-Richter law, as it does not depend on the past).

Now we can identify $p(m_{i-1}^o|m_{i-1})$ as the distribution of the noise $p_\epsilon(m_{i-1}^o - m_{i-1})$ introduced below equation (7.74). Furthermore, the distribution of the marks

$p_m(m)$ was introduced in equation (7.72). Therefore, the joint pdf of the magnitudes $m_{i-1}, \ldots, m_1$ given the occurrence times $t_i, \ldots, t_1$ and the observed magnitudes $m^o_{i-1}, \ldots, m^o_1$ is given by

$$\begin{aligned} P^a_i \propto \; & \lambda(t_i|H_{i-1}) e^{-\int_{t_{i-1}}^{t_i} \lambda(u|H_{i-1})du} \\ & \times p_\epsilon(m^o_{i-1} - m_{i-1}) p_m(m_{i-1}) / p(m^o_{i-1}) \\ & \times P^a_{i-1} \end{aligned} \tag{7.81}$$

If we include the normalization constants, we obtain our main result:

$$P^a_i = \frac{\lambda(t_i|H_{i-1}) e^{-\int_{t_{i-1}}^{t_i} \lambda(u|H_{i-1})du} p_\epsilon(m^o_{i-1} - m_{i-1}) p_m(m_{i-1})}{p(t_i|t_{i-1}, \ldots, t_1, m^o_{i-1}, \ldots, m^o_1) p(m^o_{i-1})} \; P^a_{i-1} \tag{7.82}$$

where $P^a_i \equiv p^a_i(m_{i-1}, \ldots, m_1|t_i, \ldots, t_1, m^o_{i-1}, \ldots, m^o_1)$. Equation (7.82) is the exact distribution of the joint pdfs of the magnitudes given the observed magnitudes and the occurrence times up to $t_i$. We thus used the fact that $t_j$ depend on previous marks, to obtain better estimates of those marks than simply $m^j_o$.

### 7.11.7 Limiting Cases and Discussion

Let us examine some limiting cases. First, consider the case where the occurrences times do not depend on past magnitudes, e.g. a Poisson process. Then,

$$P^a_i(\text{Poisson}) = \frac{p_\epsilon(m^o_{i-1} - m_{i-1}) p_m(m_{i-1})}{p(m^o_{i-1})} \; P^a_{i-1} \tag{7.83}$$

where the dependence on occurrence times cancelled with the normalization constant and we are left solely with a product over $p(m^t|t^o)$ for each magnitude. It is clear from expression (7.83) that we have used the information that marks are distributed a priori according to $p_m(m)$. This is the essence of data assimilation.

In the limit of vanishing noise, we obtain simply a product of Dirac functions of

the observed magnitudes, as expected:

$$P_i^a(\text{ Poisson, no noise}) = \frac{\delta(m_{i-1}^o - m_{i-1})p_m(m_{i-1})}{p(m_{i-1}^o)} P_{i-1}^a$$
$$= \frac{\delta(m_{i-1}^o - m_{i-1})p_m(m_{i-1})}{p_m(m_{i-1}^o)} P_{i-1}^a \qquad (7.84)$$

Furthermore, note that even if the occurrence times depend on past marks as in expression (7.82) but there is no noise, then the joint pdf will be zero everywhere except at $m_j = m_j^o$. At those values, the quantitites in the numerator and denominator that depend on $t_i$ cancel. Thus we retrieve the intuition that the joint pdf of marks cannot be influenced by the occurrence times if there is no noise in the observations. On the other hand, when noise is so large that the observations cannot tell us anything about the true marks, e.g. by a flat $p_\epsilon(m^o - m)$, then the best knowledge we have about the system is simply given by the combination of the occurrence times and the a priori distribution of the marks.

In summary, this section presented the (conceptual) Bayesian solution for obtaining the posterior distributions of the exact magnitudes up to time $t_{i-1}$ given noisy magnitudes and exact occurrence times up to time $t_i$. Given the impact that magnitude errors can have on forecasts, as we demonstrated in Chapter 6, an implementation of this solution may drastically improve forecasts. It remains to be implemented using the particle filters described in section 7.7.

## 7.12 Conclusions

This chapter provided a conceptual foundation for basing earthquake forecasting on data assimilation so as to systematically account for uncertainties in forecasting and estimation problems. We began with a discussion motivating the development of this foundation. To introduce non-seismologists to the problem of earthquake forecasting, we briefly reviewed: (i) the observational uncertainties in earthquake catalogs; (ii) the

types of models used for forecasting; and (iii) some basic notions of the classes of point process models relevant to earthquakes. We showed that existing methods in statistical seismology for dealing with uncertainties are limited to simple models and unsuitable for sequential, near-real-time forecasting.

We then proposed the framework of data assimilation as the ideal vehicle for addressing uncertainties and for expanding the toolbox of methods. After reviewing the basic concepts and methods of sequential Bayesian data assimilation, we argued that sequential Monte Carlo methods (particle filters) are a convenient, flexible and easy-to-implement way of estimating posterior distributions of desired quantities. We presented several such particle filters along with algorithms.

The literature on data assimilation methods for point process models is relatively scarce, but progress in developing methods for point processes is being made in spatial statistics, applications to queueing and in situations where the point process is driven by an underlying state-space model. The formulation of state-space models presents some obstacles for the combination of renewal processes and noisy occurrence times, the gravity of which depends on the particular application.

We presented numerical implementations of particle filters for lognormal renewal processes and noisy occurrence times. Using identical twin experiments, we showed that the particle filter outperforms the common method of neglecting uncertainties.

Finally, we generalized the setting to unpredictably marked temporal point processes, which include the class of short-term clustering models popular in time-dependent earthquake forecasting. We presented a conceptual Bayesian solution for estimating magnitude posteriors from exact occurrence times and noisy magnitudes.

Much work remains to be done on this topic. The particle filters for renewal processes should be applied to (real) earthquake data on faults. The conceptual solution for marked point processes needs to be implemented numerically and tested on

synthetic and real data. Both methods should be compared with existing methods to quantify the improvement. Furthermore, the framework needs to be extended to spatio-temporal marked point processes and noisy locations, especially depth, which is another large source of uncertainty in earthquake catalogs. Thus, a complete ETAS model entirely based on particle filters should be run in near-real-time to provide a gold standard of short-term earthquake forecasting that systematically accounts for uncertainties. Such a model would provide a truly powerful null hypothesis against which other hypotheses can be tested.

According to the literature on particle filters, parameter estimation does not seem to be difficult to include (at least conceptually) in the algorithms. Parameter estimation also opens the path toward introducing additional dependencies of the model on different information, the strength of which can be estimated to test scientific ideas.

On the theoretical side, the restriction to state-space models needs to be overcome. A potential candidate might be the trans-dimensional sequential Monte Carlo method discussed by *Doucet et al.* [2006].

Another exciting avenue of research lies in the adaption of physics-based models, such as rate-and-state friction models or Coulomb stress transfer models, to state-space models in which the observations are a noisy point process and the state corresponds to a physical quantity. Similar existing models in neuroscience need to be generalized to the non-Markovian nature of earthquakes.

Finally, data assimilation provides a framework for combining different types of data, e.g. GPS and InSAR data, into a single forecast model. Such a combination requires the development of models that relate observable seismicity to such data.

In summary, we believe that the framework of data assimilation provides a solid conceptual foundation not just for dealing with uncertainties but also for developing models and testing scientific hypotheses.

# CHAPTER 8

# Conclusions

The characterization of seismicity as a highly stochastic process in which all scales contribute to the fluctuations led us to investigate the role of the smallest length scales, in the form of earthquakes too small to be detected by seismic networks. Much evidence suggests, including some of the points made in this dissertation, that the influence of small, undetected earthquakes on observed seismicity is substantial. The characterization of interacting seismicity via the concept of triggering has so far neglected their effects. We introduced a formalism that explicitly recognized this short-coming. We showed that, as a result of the common neglect of small scales, the geophysical interpretation of clustering and triggering parameters that characterize models of triggered seismicity tied intricately to the ability of instruments to measure small earthquakes, rather than physical fact. As a consequence, declustering methods that use temporal branching models, including state-of-the-art stochastic declustering, cannot even in theory be physically sound. We mentioned the development of magnitude-branching models with similar probability structures as temporal-branching models as a potential way to make declustering more robust.

Furthermore, we noted the failure of the maximum likelihood estimator, proving parameter bias when undetected events are not included, and discussing the violation of assumptions that are required to prove consistency and asymptotic normality for the estimator. It is in principle possible to devise methods that begin to account for the effect of small earthquakes, in an attempt to recover unbiased parameters [*Helmstetter*

*et al.*, 2005c] – but these must necessarily remain approximations. Furthermore, the concept of unbiased parameters may simply not be useful: rather, a more practical interpretation of triggering models, in terms of effective parameters that adjust in a complex fashion to model misfit, clustering and the detection threshold, may be more relevant.

The estimation of effective parameters can, when recovered from the same detection threshold, indicated real relative differences between different regions, but the current direct-search or gradient-based algorithms are neither robust nor do the Hessian-derived confidence limits adequately account for the nonlinearity and non-Gaussianity of the estimation problem. The development of more robust methods, for instance Bayesian or simulation-based, remains an important task for the future [*Werner and Jackson*, 2007].

We discussed the Vere-Jones model [*Vere-Jones*, 2005] because it can be formulated without recourse to a detection threshold. While the model generalizes the class of self-similar random measures, simulation and parameter estimation procedures are currently still lacking. Moreover, undetected events will remain a source of uncertainty in the calibration of the model. However, while the improved development of point process models may indeed provide more robust models, the detection threshold is just one of the observer scales that limit our ability to measure earthquakes. As has been argued by Yan Kagan [*Kagan and Knopoff*, 1981; *Kagan*, 2004; *Kagan and Houston*, 2005], earthquakes seem to be fractal objects which cannot uniquely be defined as events – one first must impose a (temporal, spatial, energy) scale, which is set by our instruments. Rather, the micro-physical modeling of earthquakes needs to extend point processes to continuum stochastic processes in a tensorial stress field, characterized by intermittent but continuous catastrophic deformation flows, which are retrospectively termed events. The emphasis must lie on physical quantities in a more

physics-based modeling approach, e.g. on moment rate rather than number of events.

The search for robust, scale-independent features of seismicity in the statistical physics community continues to stir the debate on the relevance of critical phenomena to the earthquake process. We presented an example which, like many of the recently claimed discoveries of universal scaling laws, did not stand scrutiny. By repeating the analysis, which led to the "discovery", on a set of simulations produced by a null hypothesis in the form of a clustering model, we were able to reject the necessity for such a law. Nevertheless, the discussion raises the question of the role of spatial correlations in self-organized criticality and the role of quenched disorder, heterogeneities and elasticity theory. These are problems that are only beginning to be understood (see e.g. [*Kagan*, 1987a,b,c; *Miltenberger et al.*, 1993; *Sornette et al.*, 1994; *Ouillon and Sornette*, 1996; *Sornette and Ouillon*, 2005]).

Throughout much of this dissertation, we have focused on the Epidemic-Type Aftershock Sequence (ETAS) model, or slightly simpler models of a similar class, as it is rapidly becoming a gold standard in hypothesis testing, parameter estimation, modeling and forecasting. We discussed an example of its use as a benchmark in Chapter 5. Much of the discussion here may therefore be viewed as deeper investigations into the formulation of this null hypothesis, which contributes to our overall understanding and the development of improved benchmarks. Perhaps the most direly needed improvement in the ETAS model is the spatial kernel for aftershocks and its representation of the fault network. Recent papers are beginning to address this issue [*Felzer and Brodsky*, 2005; *Schoenberg*, 2003; *Ogata and Zhuang*, 2006].

In Chapter 6, we saw that our understanding of the effects of catalog uncertainties on the null hypothesis are not yet well understood. There, we focused on the quantification of magnitude uncertainties, on their impact on forecasts and on their evaluation in predictability experiments in a simple aftershock clustering model, which captures

the main ingredients of most popular short-term forecasting models. We proposed that, as a first step, the forecast evaluation methodology must drop the assumption of Poisson likelihoods and allow the model itself to specify the likelihood of observations. Much work remains to be done in this area. There are many other uncertainties and biases that were not taken into account in our study. We did not examine the combined effect of uncertainties and parameter estimation. Our mock RELM/CSEP experiment was not fully to scale and one would like to use the actual models along with actual data to perform synthetic tests for different scenarios. Such large-scale simulations may answer many of the following questions: How much time (and data) is needed before we can reject models, under realistic treatment of the uncertainties? At what confidence limits? Where are the largest uncertainties? Which aspects of the data should be focused on? What is the impact of the many missing aftershocks after large earthquakes in the evaluation of aftershock clustering models? What are optimal tests for time-dependent short-term earthquake forecasts? The science of earthquake predictability is still in a very young stage and much work remains to be done.

Chapter 7 presented a framework that may help address some of these problems. Numerical weather prediction was and is faced with similar questions. Out of this need developed the concept of data assimilation, which, in its general form, is a Bayesian method for estimating as best as possible a desired quantity by using a statistical combination of noisy observations and short-term model forecasts. More generally, data assimilation is a conceptual framework that allows a systematic treatment and discussion of uncertainties. For the strongly nonlinear and non-Gaussian statistics in earthquake seismology, traditional methods based on linear, Gaussian assumptions (e.g. Kalman filtering) must be generalized to Bayesian methods, which characterize the entire distribution function. Sequential Monte Carlo methods, a flexible set of simulation-based techniques for estimating posterior distributions, may be an easy-to-implement, computationally efficient method for Bayesian earthquake forecasting. We discussed some

standard particle filters and presented their algorithms. Using numerical experiments, we demonstrated their use for a lognormal renewal process under noisy occurrence times, motivated by the relevance for seismic risk calculations and model identification based on paleoseismic data. We further presented the conceptual Bayesian solution for the estimation of posterior distributions of magnitudes in an arbitrary temporal point process with unpredictable marks. The implementation of the latter particle filter would provide a first sequential Bayesian method for generating realistic earthquake forecasts in near-real-time that systematically includes (random) magnitude uncertainties. On a theoretical level, the representation of the state-space model is not entirely satisfactory and should be supplanted by a more general one.

But well beyond the goal of including uncertainties in forecasts, the framework of data assimilation allows a focused discussion of states, be they physical quantities such as frictional properties or improved estimates of magnitudes, and observations, which are related to the states through likelihood distributions. The combination of noisy point process models as the observational process of a physical (state) model of the Earth, including additional types of data and models, can be approached conceptually and technically in the general setting of data assimilation. Lastly, it is worth mentioning that progress in seismology will likely stimulate research in mathematical statistics and in other areas of applications.

# APPENDICES

# APPENDIX A

# Apparent Branching Structure Calculations

## A.1 Consistency check: $N_{obs}$ as the sum of "above-water" cascades triggered by the mainshock and by the apparent background events

To complete the calculations and show consistency of the results, we demonstrate that the observed cascades set off by the apparent background events, when added to the original "above-water" cascade, add up to the total observed number of aftershocks of the whole sequence. Each apparent source event will trigger its own cascade above the threshold $m_d$ with branching ratio $n_a$. The total number of events due to the apparent background events and their cascades above the threshold is

$$N_{source} = S_a + S_a\, n_a + S_a\, n_a^2 + ... = \frac{S_a}{1 - n_a}\ . \tag{A.1}$$

Substituting expression (3.22) and using (3.8) gives

$$N_{source} = \begin{cases} N_{obs}\ \frac{(n-n_a)}{1-n_a} & ,\ \ M_1 \geq m_d \\ N_{obs}\ \frac{(n-n_a)}{1-n_a} + \frac{\rho(M_1-m_0)\ f_{obs}}{1-n_a} & ,\ \ M_1 < m_d \end{cases} \tag{A.2}$$

Combining the direct "above-water" cascade (3.18) with the apparent source cascades (A.2) gives the total amount of apparent events observed after the initial event

$$\begin{aligned} N_a &= N_{source} + N_{above} \\ &= \begin{cases} N_{obs}\ \frac{(n-n_a)}{1-n_a} + \frac{\rho(M_1-m_0)\ f_{obs}}{1-n_a} & , \ M_1 \geq m_d \\ N_{obs}\ \frac{(n-n_a)}{1-n_a} + \frac{\rho(M_1-m_0)\ f_{obs}}{1-n_a} + 0 & , \ M_1 < m_d \end{cases} \\ &= N_{obs}, \end{aligned} \tag{A.3}$$

where $N_{obs}$ is given by (3.8). The last equality confirms the consistency of our decomposition into apparently-triggered earthquakes and apparent sources.

## A.2 Generalization to a catalog of an arbitrary number of clusters

In this section, we generalize our analysis of the apparent branching structure of one cluster to that of a whole catalog consisting of an arbitrary number of clusters. The reasoning developed in section 3.1 can be directly applied as follows.

We begin by writing the instantaneous seismicity rate at time $t$:

$$\lambda(t, r) = \mu + \sum_{i|t_i<t} \rho(m_i)\ \psi(t - t_i)\ R(r - r_i) \tag{A.4}$$

We integrate this expression over time to obtain the total number $N$ of earthquakes in the catalog. We restrict this demonstration to the temporal domain. In order to have a finite catalog, we assume that the integral over the background source rate $\mu$ is finite so that the total number $s$ of background events is finite. Stated differently, the integration could also be over a finite but very long period $T$ so that the sources $s = \mu\, T$ are finite but the Omori law decays have effectively ended. We thus obtain

$$N^{tot} = s + \sum_{i|t_i<t} k\ 10^{\alpha(m_i-m_0)} \tag{A.5}$$

where the index $i$ runs over all events.

We now express the total number of earthquakes by grouping each event into one of the $s$ distinct clusters and by averaging over the magnitudes of the indirectly triggered events of the initial background event that set off the cascade. Now the total number is simply the $s$ background events plus the triggered events in their $kth$ cluster, which are averaged over the aftershock magnitudes in the same way as for one cluster in equation (3.6):

$$N^{tot} \simeq s + \sum_{k|t_k<t}^{s} \frac{k\ 10^{\alpha(m_k - m_0)}}{1-n} \tag{A.6}$$

where the index $k$ now only runs over the background events.

The observed number of shocks is expression (A.6) multiplied by the fraction of observed events:

$$N_{obs}^{tot} \simeq s\ f_{obs} + \sum_{k|t_k<t}^{s} \frac{k\ 10^{\alpha(m_k - m_0)}}{1-n}\ \ f_{obs} \tag{A.7}$$

Now we can apply to each cluster the same arguments as we did in section 3.1. Let us denote the number of unobserved background events below the detection threshold $m_d$ by $u = 1, \ldots U$, so that $U = s\ (1 - f_{obs})$, while we call the number of observed background events $l = 1, \ldots, L$, i.e $L = s\ f_{obs}$. Then, the total number of events from all the $L$ above-water sequences due to the observed real sources is

$$N_{above}^{tot} \simeq \sum_{l=1}^{L} \frac{\rho(m_l)}{1-n_a}\ f_{obs} \tag{A.8}$$

where the index $l$ runs over all observed real sources.

The number of apparent sources for each cluster is given by equation (3.22) in section 3.1. For the whole catalog, the total number of apparent sources is thus the sum of observed real sources, the apparent sources in clusters due to observed sources

and the apparent sources from clusters of unobserved sources:

$$\begin{aligned} S_a^{tot} &= L + \sum_{l=1}^{L} \frac{\rho(m_l)}{1-n}\, f_{obs}\, (n - n_a) + \\ &\quad \sum_{u=1}^{U} \left( \frac{\rho(m_u)}{1-n}\, f_{obs}\, (n - n_a) + \rho(m_u)\, f_{obs} \right) \end{aligned} \tag{A.9}$$

This expression shows that the apparent branching structure renormalizes $L = s\ f_{obs}$ observed background events into $S_a^{tot}$ apparent background events. Together with the apparent branching ratio, this completely determines the renormalization of the model when going from $m_0$ to $m_d$.

As for the one cluster case, we can check the consistency of our decomposition by testing whether the apparent number of events $N_a$ is equal to the observed number of events $N_{obs}^{tot}$ given by (A.7). From our decomposition, $N_a$ is given by

$$N_a = L + N_{above}^{tot} + \frac{S_a^{tot} - L}{1 - n_a} \tag{A.10}$$

Substituting the relevant expressions, one can easily show that $N_a = N_{obs}^{tot}$ and that our decomposition is consistent. Note that we have to subtract the real observed sources $L$ from the apparent sources $S_a^{tot}$ because the cascades they set off have already been taken into account in the above-water cascade $N_{above}^{tot}$.

In summary, we have generalized the approach to the case of many cascades. We have shown that analyzing the branching structure above the detection threshold of a complete catalog leads to a renormalized ETAS model (for the averaged rates) with an apparent branching ratio $n_a$ and an apparent number of sources $S_a^{tot}$.

# APPENDIX B

# Proofs and Calculations of Chapter 6

## B.1 The Deviation of the Perturbed Rate from the True Rate as a Sum of Weighted Random Variables

This section shows how the deviation of the perturbed rate due to noisy magnitudes from the true rate can be written as a sum over weighted random variables.

The perturbed rate is given by

$$\lambda^p(t|H_t^c) = \lambda_c + \sum_{i_c|t_{i_c}<t} \frac{k\ e^{a(m_{i_c}^o - m_d)}}{(t - t_{i_c} + c)^p} \tag{B.1}$$

while the true rate is given by

$$\lambda^t(t|H_t^c) = \lambda_c + \sum_{i_c|t_{i_c}<t} \frac{k\ e^{a(m_{i_c}^t - m_d)}}{(t - t_{i_c} + c)^p} \tag{B.2}$$

where $m^o = m^t + \epsilon$ and $p_\epsilon(\epsilon)$ is the distribution of the noise given by

$$p_\epsilon(\epsilon) = \frac{1}{2\nu_c} e^{\left(-\frac{|\epsilon|}{\nu_c}\right)} \tag{B.3}$$

Hence, given any catalog realization $H_t^c$ (of cluster centers), the deviation of the per-

turbed rate from the true rate is:

$$\begin{aligned}
\Delta\lambda(t|H_t^c,\theta) &= \lambda^p(t|H_t^c,\theta) - \lambda(t|H_t^c,\theta) \\
&= \sum_{i_c|t_{i_c}<t} \frac{k\, e^{a(m_{i_c}^o - m_d)}}{(t-t_{i_c}+c)^p} - \frac{k\, e^{a(m_{i_c}^t - m_d)}}{(t-t_{i_c}+c)^p} \\
&= \sum_{i_c|t_{i_c}<t} \frac{k\, e^{a(m_{i_c}^t + \epsilon_i - m_d)}}{(t-t_{i_c}+c)^p} - \frac{k\, e^{a(m_{i_c}^t - m_d)}}{(t-t_{i_c}+c)^p} \\
&= \sum_{i_c|t_{i_c}<t} \frac{k\, e^{a(m_{i_c}^t - m_d)}}{(t-t_{i_c}+c)^p} \cdot (e^{a\epsilon_i} - 1) \\
&= \sum_{i_c|t_{i_c}<t} w_i \cdot z_i
\end{aligned} \tag{B.4}$$

where the last equality expresses the deviation as a sum over a product of two terms: a quenched weight $w_i$ (i.e.. which is fixed for a given catalog but unknowable)

$$w_i = \frac{k\, e^{a(m_{i_c}^t - m_d)}}{(t-t_{i_c}+c)^p} \tag{B.5}$$

and a random variable $z_i$

$$z_i = e^{a\epsilon_i} - 1 \tag{B.6}$$

The weight $w_i$ measures the influence of the ith cluster center according to its magnitude $m_{i_c}^t$ through the productivity law $\rho(m_{i_c}) = \exp(a(m_{i_c}^t - m_d))$ and its occurrence time according to the Omori-Utsu law $\phi(t - t_{i_c}) = (t - t_i + c)^{-p}$. The weights $w_i$ thus depend sensitively on the specific catalog realization $\{m_{i_c}^t, t_{i_c}^t\}_{1\le i_c \le N_c}$ and the parameters $\theta$.

The weights $w_i$ are "quenched" or "frozen" because they are fixed for a realization but result from a random process. In statistical physics of spin glasses (a similar situation because of the frozen random variables), there are two types of disorder that are treated differently: quenched disorder, where the average is taken over the logarithm of the partition function; and annealed disorder, where the average is taken directly over the partition function. The latter case would correspond in our context to looking

at the full distribution of weights, rather than assuming they are fixed. However, we are interested in the fluctuations of the perturbed rate given a fixed catalog.

The random variables $z_i = (\exp(a\epsilon_i) - 1)$ modulate the weights due to the magnitude noise $\epsilon_i$. Without noise, $\epsilon = 0$ and hence $z_i = 0$ so that $\Delta\lambda = 0$. Their distribution is the subject of the next section.

## B.2 The Distribution of the Random Variables $z$

Using the distribution of $\epsilon$ from equation (6.1), we can determine the distribution of the $z_i$:

$$
\begin{aligned}
p_z(z) \quad &= p_\epsilon(\epsilon) \left|\frac{d\epsilon}{dz}\right| = \begin{cases} \frac{1}{2\nu_c}\, e^{-\epsilon/\nu_c} \left|\frac{d\epsilon}{dz}\right|, & 0 \le \epsilon < \infty \\ \frac{1}{2\nu_c}\, e^{+\epsilon/\nu_c} \left|\frac{d\epsilon}{dz}\right|, & -\infty < \epsilon < 0 \end{cases} \\
&= \begin{cases} \frac{1}{2a\nu_c(z+1)}\, e^{(-\log(z+1)/a\nu_c)}, & 0 \le z < \infty \\ \frac{1}{2a\nu_c(z+1)}\, e^{(\log(z+1)/a\nu_c)}, & -1 < z < 0 \end{cases} \\
&= \begin{cases} \frac{\alpha}{2}\frac{1}{(z+1)^{1+\alpha}}, & 0 \le z < \infty \\ \frac{\alpha}{2}\frac{1}{(z+1)^{1-\alpha}}, & -1 < z < 0 \end{cases} \quad \text{where } \alpha = \frac{1}{a\nu_c} \\
&= \frac{\alpha/2}{(z+1)^{1\pm\alpha}} \quad \begin{matrix} (+): & 0 \le z < \infty \\ (-): & -1 < z < 0 \end{matrix}
\end{aligned}
\tag{B.7}
$$

Figure 6.10 shows a double logarithmic plot of the pdf of the random variable z for several choices of the noise scale parameter $\nu_c = (0.1, 0.2, 0.3, 0.4, 0.5)$. We assumed $a = \ln(10) = 2.3$ so that the exponent $\alpha = (4.34, 2.17, 1.45, 1.09, 0.87)$, respectively.

## B.3 Proof of Proposition 1

In this section, we prove that the deviation $\Delta\lambda(t|H_t)$ of the perturbed seismic rates from the true seismic rate due to magnitude noise is a random variable with a distri-

bution having a power law tail with exponent $\alpha$ and scale factor $C_{\Delta\lambda}$. Equation (B.4) shows that $\Delta\lambda$ can be written as a finite sum of weighted random variables $z$, where the $z$ are distributed according to (B.7). The proof follows in two steps. First, we show that $z$ is regularly varying. Second, we invoke the result that the sum of weighted, regularly varying variables is equally regularly varying in the tail with the same exponent. We will frequently refer to the rigorous *Jessen and Mikosch* [2006] (hereafter JM), but the definitions and proofs can equally be found in other sources. *Sornette* [2004] provides a heuristic and intuitive development of the results we use.

DEFINITION 2.1 of JM: *One-dimensional regularly varying random variables $X$ with distribution function $P(X > x)$ are defined by*

$$P(X > x) \sim q' x^{-\alpha} L(x) \qquad \text{and} \qquad P(X \leq -x) \sim q'' x^{-\alpha} L(x)$$
$$q' + q'' = 1 \tag{B.8}$$

*where $L$ is a slowly varying function, i.e. $L(cx)/L(c) \to 1$ as $x \to \infty$ for every $c > 0$. Condition (B.8) is also referred to as a tail balance condition. The cases $q' = 0$ or $q'' = 0$ are not excluded. Here and in what follows we write $f(x) \sim g(x)$ as $x \to \infty$ if $f(x)/g(x) \to 1$.*

$z$ is power-law distributed with exponent $\alpha$ for $z > 0$ so that the slowly varying function $L(z)$ is simply a constant. For $z < 0$, the power law is truncated at $-1$, so that $q'' = 0$. Hence the tail balance condition is easily verified for the positive tail of $z$. Therefore, the tail of $z$ is regularly varying with index $\alpha$. This concludes the first part of the proof.

To prove that the finite, weighted sum of regularly varying variables $z$ is also a regularly varying function with the same exponent, we invoke Lemma 3.3 of JM:

LEMMA 3.3 of JM: *Let $(Z_i)$ be an independently and identically distributed sequence of regularly varying random variables satisfying the tail balance condition (B.8). Then for any real constants $\psi_i$ and $m \geq 1$,*

$$\begin{aligned} P(\psi_1 Z_1 + \quad & \cdots + \psi_m Z_m > x) \\ \sim \quad & P(|Z_1| > x) \sum_i^m \left[ q'(\psi_i^+)^\alpha + q''(\psi_i^-)^\alpha \right] \end{aligned} \tag{B.9}$$

*where $\psi_i^+$ and $\psi_i^-$ are defined by $P(\psi_i Z_i > x) = P(\psi_i^+ Z_i^+ > x) + P(\psi_i^- Z_i^- > x)$ where $x^\pm = 0 \vee (\pm x)$ (where $\vee$ means "or").*

In our case, the constants $\psi_i$ are given by the weights $w_i$. Plugging in $q'' = 0$ (i.e. $q' = 1$) from above and using our notation, we have shown that

$$p(\Delta\lambda) \sim p_z(z) \cdot \sum_i^m (w_i)^\alpha \tag{B.10}$$

Denoting $C_{\Delta\lambda} = \sum_i^{N_c} (w_i)^\alpha$, we have shown that

$$p(\Delta\lambda) \sim \frac{C_{\Delta\lambda}}{(\Delta\lambda)^{1+\alpha}} \qquad \text{for} \quad \Delta\lambda \to \infty \tag{B.11}$$

which completes the proof of Proposition 1.

## B.4 Proof of Proposition 2

We first state well-established results for a slightly different definition of the scale factor (denoted by $C(t)$, where the time $t$ of evaluation is fixed) for which Proposition 2 can be easily proven (section A4.1). In section A4.2, we consider the more difficult case for our definition of the scale factor ($C(N)$ where the number of cluster centers is fixed).

### B.4.1 Results for Fixed-Time Scale Factor $C(t)$

Recall that the scale factor is defined by:

$$C_{\Delta\lambda} = \sum_{i=1}^{N} \frac{k^{\alpha} e^{a\alpha(m_i^t - m_d)}}{(t - t_i + c)^{p\alpha}} \ . \tag{B.12}$$

When we simulate catalogs, perturb them and calculate the differences between the perturbed rates and the true rate, we are interested in the deviations from the true rate given a fixed number of events. We did not constrain the time $t$ to a fixed value because a more useful and practical result would be the deviations for a fixed number of events. We therefore let $t$ adjust according to when the Nth main shock happened. We then evaluated the rates at time $t = t_N + dt$ just after the Nth main shock. Since the Nth occurrence time is random, $t$ is therefore also random (for different catalogs or realizations of the Poisson process). Let us denote our scale factor by $C(N)$ to stress the fact that the number of main shocks is fixed, not the time $t$.

On the other hand, a wealth of results is available for the scale factor $C(t)$, for which the time $t$ is fixed and $N$ fluctuates. The scale factor is then defined by:

$$C(t) = \sum_{i=1}^{N(t)} \frac{k^{\alpha} e^{a\alpha(m_i^t - m_d)}}{(t - t_i + c)^{p\alpha}} \ . \tag{B.13}$$

where $N(t)$ is a now random variable for fixed $t$. Now we make the crucial identification of the scale factor $C(t)$ as the intensity of power law shot noise [e.g. *Lowen and Teich* [1990]], defined in general by

$$I(t) = \sum_{j|t_j<t} K_j h(t - t_j) \tag{B.14}$$

where $I(t)$ is the "current" or noise, $t_j$ are Poisson occurrence times with rate $\lambda < \infty$, the $K_j$ are i.i.d. stochastic amplitudes and the "impulse" function $h(t) = t^{-\delta}$ is an inverse power law function on the interval $[A, B]$ and zero otherwise. These correspond exactly to the fixed-time scale factor $C(t)$, the main shock occurrence times

at rate $\lambda_c$, the productivities $k^\alpha \exp(-a\alpha(m_j - m_d))$ and the Omori-like decay $(t - t_j + c)^{-p\alpha}$, respectively. Note that in our case $A = c$, $B = \infty$ and $\delta = \alpha p$.

The cumulants $c_n$ of $I(t)$, which determine the moments of the shot noise, are given by the following equations [Rice, 1945]. For $\delta \neq 1/n$:

$$c_n = \lambda \langle K^n \rangle \times \frac{A^{1-n\delta} - B^{1-n\delta}}{n\delta - 1} \tag{B.15}$$

while for $\delta = 1/n$:

$$c_n = \lambda \langle K^n \rangle \times \ln(B/A) \tag{B.16}$$

The nth cumulant is hence infinite if the nth moment $\langle K^n \rangle$ of the stochastic amplitudes is infinite, if $A = 0$ and $\delta \geq 1/n$, or if $B = \infty$ and $\delta \leq 1/n$.

Since earthquakes cannot have infinite moment (magnitude), their distribution is truncated and hence all moments of the stochastic amplitudes are finite: $\langle K^n \rangle < \infty \ \forall n$. However, the productivities are power law distributed up to the truncation with an exponent $\beta/a$ which lies in the range $1 < \beta/a < 2$. Therefore, the amplitudes fluctuate as if power-law distributed until the sampling actually "feels" the corner magnitudes. For specific regions of the world, it may take millenia for these corner magnitudes to occur. Therefore, while mathematically all moments of the stochastic amplitude are finite, fluctuations will be power-law like with infinite variance until the upper truncation is actually felt.

Using results from *Rice* [1945] and *Lowen and Teich* [1990], we now prove the three elements of Proposition 2 for the fixed-time scale factor $C(t)$:

1. First, we show that $C(t) < \infty$ almost surely (a.s.) even as $t \to \infty$ if $p\alpha > 1$. For this, we only need to demonstrate that all cumulants $c_n$ of $C(t)$ given by (B.15) or (B.16) are finite. We already stated that the moments of the stochastic amplitudes are mathematically finite because the Gutenberg-Richter distribution

is truncated. Furthermore, by definition $\delta = p\alpha > 1 \geq 1/n \; \forall \; n$, and $A = c > 0$. Therefore, $c_n < \infty \; \forall \; n$, which in turn implies that $C(t) < \infty$ a.s. $\forall \; t$.

2. Second, we show that $C(t)$ diverges a.s. as $t \rightarrow \infty$ if $p\alpha < 1$. To prove this, we will bound $C(t)$ from below and show that this lower bound diverges a.s. As noted above, only magnitudes down to $m_d$ are included in the process. Therefore, the smallest productivity is given by $k^\alpha$. We create a lower bound for $C(t)$ by replacing all productivities by their lower bound $k^\alpha$, i.e.

$$C(t) \geq D(t) = \sum_{i=1}^{N_c(t)} \frac{k^\alpha}{(t - t_i + c)^{p\alpha}} \tag{B.17}$$

Now the process $D(t)$ has moment generating function $Q(s)$ that is equal to equation (A1) in *Lowen and Teich* [1990] [their Appendix A]. There, the authors show that, for $\delta \leq 1$, $Q(s) = 1$ for $s = 0$ and zero otherwise, so that

$$\Pr\{D_t < x\} = 0, \text{ for all } x < \infty$$

which proves the a.s. divergence of $C(t)$ as $t \rightarrow \infty$.

3. Third, in the regime $p\alpha > 1$ for which $C(t) < \infty$ almost surely, we show that $C(t)$ remains a random variable with a non-degenerate distribution. For this, we only need to state that the variance of $C(t)$ is non-zero as $t \rightarrow \infty$. We can actually calculate the variance explicitly, being equal to the second cumulant $c_2$ given by equation (B.15):

$$\text{Var}(C_t) = \frac{\lambda_c \langle K^2 \rangle}{(2\alpha p - 1) c^{2\alpha p - 1}} \tag{B.18}$$

As stated above, the second moment of the amplitude is mathematically finite. However, if the earthquake catalog under study does not actually sample the upper magnitude cut-off, the variance will behave as if infinite. Thus, not only is $C(t)$ a random variable, it fluctuates wildly. Sampling the corner magnitude

may take hundreds to thousands of years even in relatively active regions like California. To get a sense of the numbers, set $\lambda_c = 0.01$, $\langle K^2 \rangle = 1$ and $\alpha p = 2$ for simplicity, neglecting for a moment the large second moment of $K$. For these values, the variance is on the order of $10^6$, far larger than typical earthquake rates of e.g. 1 per day above $m_d = 4$ in California. This completes the proof of Proposition 2 for the fixed-time scale factor $C(t)$.

In fact, more results are known about the statistical properties of $C(t)$ [e.g. *Lowen and Teich* [1990], their Figure 3, and references therein]. If $A > 0$ and $\delta > 1$ so that cumulants exist (assuming the stochastic amplitudes have finite moments), then in the limit of infinite Poisson driving rate $\lambda \to \infty$, the intensity $C(t)$ is distributed according to a Gaussian with mean equal to the first cumulant and variance equal to the second cumulant. Even in this limit (which is not directly relevant to earthquakes since there $\lambda$ is small) the variance remains huge. Furthermore, if $A = 0$ and $\delta > 1$, the distribution of $C(t)$ is Levy-stable for all Poisson rates with exponent $1/\delta$. Since for earthquakes, $A = c$ is very small, we expect the distribution of $C(t)$ to be close to Levy with exponent $1/\alpha p = a\nu_c/p$. For reasonable values $a = 2.3$, $\nu_c = 0.2$ and $p = 1.2$, this results in an extremely small Levy exponent 0.4.

### B.4.2 Results for Fixed-Number Scale Factor $C(N)$

Results do not seem widely established for the fixed-number scale factor. Note, however, that both the mean number and the variance of the number of events in a Poisson process diverge as $t \to \infty$. The higher moments of degree $n$ of the Poisson process are Touchard polynomials of degree $n$ of the variable $\lambda t$. For $t \to \infty$, they also diverge. Furthermore, the probability of having a finite number k of events in an infinite interval $\Pr\{N(0,T] = k\} = (\lambda T)^k \exp(-\lambda T)/k!$ is zero for $T \to \infty$. And vice versa, the probability of having an infinite number of events $k = \infty$ in a finite interval $T$ is

equally zero. Therefore, the limit $t \to \infty$ and $N \to \infty$ are equivalent so that we expect the same results to hold for $C(N)$ as for $C(t)$. Without a more formal statement of this equivalence, however, we proceed to consider separately $C(N)$ as $N \to \infty$.

Let us rewrite expression (6.16) for $C(N)$ explicitly as

$$C(N) = \sum_{i=1}^{N} \frac{k^{\alpha} e^{a\alpha(m_i^t - m_d)}}{(t_N - t_i + c')^{p\alpha}} \,. \tag{B.19}$$

where $c' = c + dt$ is a constant since $t = t_N + dt$. We bound $C(N)$ from below and from above by noting that

$$k^{\alpha} \leq k^{\alpha} e^{a\alpha(m_i^t - m_d)} \leq k^{\alpha} e^{a\alpha(M - m_d)} \,. \tag{B.20}$$

since $m_d \leq m_i^t \leq M$, where $M$ is an upper magnitude bound, which always exists due to the finiteness of the Earth. Thus,

$$k^{\alpha} \sum_{i=1}^{N_c} \frac{1}{(t_N - t_i + c')^{p\alpha}} \leq C(N) \leq k^{\alpha} e^{a\alpha(M - m_d)} \sum_{i=1}^{N_c} \frac{1}{(t_N - t_i + c')^{p\alpha}} \,. \tag{B.21}$$

1. First, we show that, when $p\alpha > 1$, the scale factor $C_{\Delta\lambda}$ converges to a finite value as $N_c \to \infty$ under the assumption of an arbitrary small but finite minimum time interval $0 < \tau_{min} << 1/\lambda$ between events. In this case, we can further bound the right-hand-side of equation (B.20) from above by replacing the intervals $t_N - t_i$ by $(N - i)\tau_{min}$:

$$C(N) \leq k^{\alpha} e^{a\alpha(M - m_d)} \sum_{i=1}^{N_c} \frac{1}{((N - i) \cdot \tau_{min} + c')^{p\alpha}} \tag{B.22}$$

   The sum in (B.22) is in turn bounded from above by the Riemann zeta function $\zeta(\alpha p) = \sum_j^{\infty} 1/j^{\alpha p}$, which converges absolutely for $\alpha p > 1$. This completes the proof that the scale factor $C_{\Delta\lambda}$ converges to a finite value as $N_c \to \infty$ if $p\alpha > 1$.

2. Second, we show, when $p\alpha \leq 1$, that (i) the expectation of the scale factor $C(N)$ diverges as $N_c \to \infty$ using Jensen's inequality and (ii) $C(N)$ diverges under the assumption of an arbitrarily large but finite maximum interval $\tau_{max}$ between events. In the latter case (ii), we can bound $C(N)$ from below by replacing the intervals $t_N - t_i$ by $(N-i)\tau_{max}$:

$$C(N) \geq k^\alpha \sum_{i=1}^{N_c} \frac{1}{((N-i)\cdot\tau_{max} + c')^{p\alpha}} \tag{B.23}$$

for which the right-hand-side diverges for $p\alpha < 1$ so that $C(N)$ diverges. In the other part (i), we show that the expectation of $C_{\Delta\lambda}$ diverges using Jensen's inequality [e.g. *Durrett* [2005]]. Jensen's inequality theorem states that for any convex function $g(x)$ (i.e., with non-negative second derivative $g''(x) \geq 0$, $\forall\, x$, if the second derivative exists) and for any random variable $\xi$ with finite expectation, the following inequality holds true:

$$\mathrm{E}[g(\xi)] \geq g(\mathrm{E}[\xi])\ , \tag{B.24}$$

where $\mathrm{E}[x]$ denotes the expectation of the random variable $x$. The equality sign holds true in (B.24) only for a degenerate distribution of $\xi$. Now we use $g(x) = 1/x^{\alpha p}$, having checked that its second derivative is positive $g''(x) > 0$, and let the random variable be $\xi = t - t_i + c$. Using Jensen's inequality:

$$\mathrm{E}\left[\frac{1}{(t-t_i+c)^{\alpha p}}\right] \geq \frac{1}{(\mathrm{E}[t-t_i+c])^{\alpha p}}\ . \tag{B.25}$$

Now if the $t_i$'s are assumed to be a stationary sequence (not even necessarily Poissonian), then $E[t - t_i + c] = const \cdot (N-i) + c$. Using this argument on each term in the sum leads to

$$\mathrm{E}\left[\sum_{j=1}^{N} \frac{1}{(t-t_i+c)^{\alpha p}}\right] \geq \sum_{j=1}^{N} \frac{1}{(const\cdot(N-j)+c)^{\alpha p}} \tag{B.26}$$

But the term on the right hand side diverges $\to \infty$ as $N \to \infty$ for $\alpha p < 1$ so that the expectation of the scale factor also diverges. This proves the second element of Proposition 2.

3. Finally, in the regime $p\alpha > 1$ for which $C(N)$ converges almost surely to a finite value as $N \to \infty$, we show that $C(N)$ remains a random variable dependent on the specific catalog distributed according to a non-degenerate distribution. We rewrite (B.19) as

$$C(N) = \sum_{i=1}^{N} \omega_i X_i \ , \tag{B.27}$$

where $\omega_i \equiv 1/(t - t_i + c)^{p\alpha}$ and $X_i \equiv k^{\alpha} e^{a\alpha(m_i^t - m_d)}$. The scale factor $C(N)$ is rewritten in (B.27) as a randomly weighted sum of i.i.d. random variables $X_i$, where the random weights are functions of the random occurrence times and the random variables are the magnitude-dependent productivities. The weights are non-identically distributed and dependent while the $X_i$ are i.i.d. We will show that for any fixed configuration of occurrence times and random $X_i$, $C(N)$ remains distributed. For $p\alpha > 1$, we have shown in 1. that $W_N \equiv \sum_{i=1}^{N} \omega_i < \infty$ for $N \to \infty$. We then use the result quoted from *Jamison et al.* [1965]: "[if the sum $W_N$ of the weights converges, then] $C(N)/W_N$ [the normalized weighted sum] either fails to converge in probability or converges almost surely to a non-degenerate limit." In plain words, in the latter case, this means that the random variable $S_n/W_n$ remains distributed according to a non-degenerate probability distribution, even in the limit $N \to \infty$. Thus, in both cases, the variance of the scale factor remains non-zero. The intuition behind this result is that the convergence of the weights ensures that there are only a finite number of terms in the infinite sum that contribute to it. This implies that, notwithstanding the existence of an infinite number of contributions, the sum remains a random variable controlled by a finite number of them. This completes the proof that the

scale factor does not converge to a unique constant (a degenerate limit) when the exponent $p/(a \cdot \nu_c) > 1$.

# BIBLIOGRAPHY


Abercrombie, R. E. (1995a), Earthquake locations using single-station deep borehole recordings: Implications for microseismicity on the san andreas fault in southern california, *J. Geophys. Res.,*, *100(B12)*, 24,003–24,014.

Abercrombie, R. E. (1995b), Earthquake source scaling relationships from $-1$ to $5$ $m_l$ using seismograms recorded at 2.5 km depth, *J. Geophys. Res.,*, *100(B12)*, 24,015 – 24,036.

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Automatic Control*, *AC-19*(6), 716–723.

Aki, K. (1996), Scale dependence in earthquake phenomena and its relevance to earthquake prediction, *Proc. Nat. Acad. Sci. USA*, *93*, 3740–3747.

Aki, K. (2000), Scale-dependence in earthquake processes and seismogenic structures,, *Pure and Appl. Geophys.*, *157*, 2249–2258.

Alexander, F., G. Eyink, and J. Restrepo (2005), Accelerated Monte Carlo for optimal estimation of time series, *J. Stat. Phys.*, *119*, 1331–1345(15), doi:doi:10.1007/s10955-005-3770-1.

Allègre, C. J., J. L. Le Mouel, and A. Provost (1982), Scaling rules in rock fracture and possible implications for earthquake prediction, *Nature*, *297*, 47–49, doi:10.1038/297047a0.

Anderson, D. R., and K. P. Burnham (2002), Avoiding pitfalls when using information-theoretic methods, *The Journal of Wildlife Management*, *66*(3), 912–918.

Andrieu, C., E. E. Barat, and A. Doucet (2001), Bayesian deconvolution of noisy filtered point processes, *Signal Processing, IEEE Transactions on [see also Acous-*

*tics, Speech, and Signal Processing, IEEE Transactions on]*, *49*(1), 134–146, doi:10.1109/78.890355.

Arulampalam, M., S. Maskell, N. Gordon, and T. Clapp (2002), A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, *Signal Processing, IEEE Transactions on*, *50*(2), 174–188, doi:10.1109/78.978374.

Athreya, K. B., and P. E. Ney (1972), *Branching Processes*, Springer Verlag.

Bak, P. (1996), *How Nature Works: The Science of Self-Organized Criticality*, 212 pp., Springer, New York.

Bak, P., and C. Tang (1989), Earthquakes as a self-organized critical phenomena, *J. Geophys. Res.*, *94*(B11), 15,635–15,637.

Bak, P., C. Tang, and K. Wiesenfeld (1987), Self-organized criticality: An explanation of the 1/f noise, *Phys. Rev. Lett.*, *59*(4), 381–384, doi:10.1103/PhysRevLett.59.381.

Bak, P., K. Christensen, L. Danon, and T. Scanlon (2002), Unified scaling law for earthquakes, *Phys. Rev. Lett.*, *88*(17), 178,501, doi:10.1103/PhysRevLett.88.178501.

Bakun, W. H., et al. (2005), Implications for prediction and hazard assessment from the 2004 Parkfield earthquake, *nature*, *437*, 969–974, doi:10.1038/nature04067.

Barbieri, R., M. C. Quirk, L. M. Frank, M. A. Wilson, and E. N. Brown (2001), Construction and analysis of non-poisson stimulus-response models in neural spiking activity, *Journal of Neuroscience Methods*, *105*, 25–37.

Båth, M. (1965), Lateral inhomogeneities in the upper mantle, *Tectonophysics*, *2*, 483–514.

Bebbington, M., and D. Harte (2003), The linked stress release model for spatio-temporal seismicity: formulations, procedures and applications, *Geophysical Journal International*, *154*, 925–946, doi:10.1046/j.1365-246X.2003.02015.x.

Bebbington, M. S. (2005), Information Gains for Stress Release Models, *Pure and Applied Geophysics*, *162*, 2299–2319, doi:10.1007/s00024-005-2777-5.

Ben-Zion, Y. (2003), Appendix 2: Key formulas in earthquake seismology, in *International Handbook of Earthquake and Engineering Seismology, Part B*, edited by W. H. K. Lee, H. Kanamori, P. C. Jennings, and C. Kisslinger, pp. 1857–1875, Academic Press.

Benz, H., J. Filson, W. Arabasz, L. Gee, and L. Wald (2005), ANSS-Advanced National Seismic System, U.S. Geological Survey Fact Sheet 075-00, Online version 1.0.

Bird, P. (2003), An updated digital model of plate boundaries, *Geochemistry Geophysics Geosystems*, *4*(3), 1027, doi:10.1029/2001GC000252.

Bird, P., and Y. Y. Kagan (2004), Plate-tectonic analysis of shallow seismicity: Apparent boundary width, beta, corner magnitude, coupled lithosphere thickness, and coupling in seven tectonic settings, *Bull. Seismol. Soc. Am.*, *94*(6), 2380–2399.

Bird, P., Y. Y. Kagan, D. D. Jackson, F. P. Schoenberg, and M. J. Werner (2007), Earthquake production by subduction zones is not linear in relative plate velocity, *Eos Trans. AGU*, *88*, Fall Meet. Suppl., Abstract S32B-02.

Bowman, D. D., G. Ouillon, C. G. Sammis, A. Sornette, and D. Sornette (1998), An observational test of the critical earthquake concept, *J. Geophys. Res.*, *103*, 24,359–24,372, doi:10.1029/98JB00792.

Brémaud, P., and L. Massoulié (2001), Hawkes branching processes without ancestors, *J. Appl. Proba.*, *38*, 122–135.

Brown, E., L. Frank, D. Tang, M. C. Quirk, and M. Wilson (1998), A Statistical Paradigm for Neural Spike Train Decoding Applied to Position Prediction from Ensemble Firing Patterns of Rat Hippocampal Place Cells, *J. Neurosci.*, *18*(18), 7411–7425.

Brown, E. N., D. Nguyen, L. Frank, M. Wilson, and V. Solo (2001), An analysis of neural receptive field plasticity by point process adaptive filtering, *Proceedings of the National Academy of Sciences*, *98*(21), 12,261–12,266, doi:10.1073/pnas.201409398.

Bufe, C. G., and D. J. Varnes (1993), Predictive modeling of the seismic cycle of the greater San Francisco Bay region, *J. Geophys. Res.*, *98*, 9871–9883.

Burnham, K. P., and D. R. Anderson (2004), Multimodel Inference: Understanding AIC and BIC in Model Selection, *Sociological Methods Research*, *33*(2), 261–304, doi:10.1177/0049124104268644.

Cardy, J. L. (1996), *Scaling and Renormalization in Statistical Physics*, Cambridge University Press, Cambridge, UK.

Christensen, K., N. Farid, G. Pruessner, and M. Stapleton (2007), On the finite-size scaling of probability density functions, *submitted to Phys. Rev. Lett.*

Console, R., A. M. Lombardi, M. Murru, and D. Rhoades (2003a), Båth's law and the self-similarity of earthquakes, *J. Geophys. Res.*, *108*(B2), 2128, doi:10.1029/2001JB001651.

Console, R., M. Murru, and A. M. Lombardi (2003b), Refining earthquake clustering models, *J. Geophys. Res.*, *108*(B10), 2468, doi:10.1029/2002JB002123.

Console, R., M. Murru, F. Catalli, and G. Falcone (2007), Real time forecasts through an earthquake clustering model constrained by the rate-and-state constitutive law: Comparison with the purely stochastic etas model, *Seismol. Res. Lett.*, *78*(1), 49.

Consoli, M., and P. Stevenson (2000), Physical mechanisms generating spontaneous symmetry breaking and a hierarchy of scales, *Int. J. Mod. Phys. A*, *15*, 133–157.

Corral, A. (2003), Local distributions and rate fluctuations in a unified scaling law for earthquakes, *Phys. Rev. E*, *68*(3), 035,102, doi:10.1103/PhysRevE.68.035102.

Corral, A. (2004a), Universal local versus unified global scaling laws in the statistics of seismicity, *Physica A*, *340*, 590–597.

Corral, A. (2004b), Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes, *Phys. Rev. Lett.*, *92*, 108,501.

Corral, A. (2005a), Mixing of rescaled data and bayesian inference for earthquake recurrence times, *Nonlinear Processes in Geophysics*, *12*, 89–100.

Corral, A. (2005b), Renormalization-group transformations and correlations of seismicity, *Phys. Rev. Lett.*, *95*, 028,501.

Corral, A. (2006), Universal earthquake-occurrence jumps, correlations with time, and anomalous diffusion, *Phys. Rev. Lett.*, *97*, 178,501.

Corral, A., and K. Christensen (2006), Comment on "earthquakes descaled: On waiting time distributions and scaling laws", *Phys. Rev. Lett.*, *96*, 109,801.

Daley, D. J., and D. Vere-Jones (2003), *An Introduction to the Theory of Point Processes*, vol. I, Springer, New York, USA.

Daley, R. (1991), *Atmospheric Data Analysis*, Cambridge University Press.

Davidsen, J., and C. Goltz (2004), Are seismic waiting time distributions universal?, *Geophys. Res. Lett.*, *31*, doi:10.1029/2004GL020892.

Davidsen, J., and M. Paczuski (2005), Analysis of the spatial distribution between successive earthquakes, *Phys. Rev. Lett.*, *94*, 048,501, doi:10.1103/PhysRevLett.94.048501.

Davidsen, J., and M. Paczuski (2007), Reply to comment on "analysis of the spatial distribution between successive earthquakes, *Phys. Rev. Lett.*, *99*, 179,802, doi:10.1103/PhysRevLett.99.179802.

Davidsen, J., P. Grassberger, and M. Paczuski (2006), Earthquake recurrence as a record breaking process, *Geophys. Res. Lett.*, *33*, L11304, doi:10.1029/2006GL026122.

Davidsen, J., P. Grassberger, and M. Paczuski (2007), Networks of recurrent events, a theory of records, and an application to finding causal signatures in seismicity, *submitted Phys. Rev. Lett.*

Davis, P. M., D. D. Jackson, and Y. Y. Kagan (1989), The longer it has been since the last earthquake, the longer the expected time till the next?, *Bull. Seismol. Soc. Am.*, *79*(5), 1439–1456.

Davis, S. D., and C. Frohlich (1991), Single-link cluster analysis of earthquake aftershocks: Decay laws and regional variations, *J. Geophys. Res.*, *96*, 6335–6350.

de Freitas, N. (1999), Bayesian methods for neural networks, Ph.D. thesis, Cambridge University, available from www.cs.ubc.ca/nando/ publications.html.

Dieterich, J. (1992), Earthquake nucleation on faults with rate- and state-dependent strength, *Tectonophysics*, *211*, 115–134.

Dieterich, J. (1994), A constitutive law for rate of earthquake production and its application to earthquake clustering, *J. Geophys. Res.*, *99*, 2601–2618.

Doucet, A., S. Godsill, and C. Andrieu (2000), On sequential Monte Carlo sampling methods for Bayesian filtering, *Statistics and Computing*, *10*, 197–208.

Doucet, A., N. de Freitas, and N. Gordon (Eds.) (2001), *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, New York.

Doucet, A., L. Montesano, and A. Jasra (2006), Optimal filtering for partially observed point processes using trans-dimensional sequential monte carlo, *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, *5*, doi:10.1109/ICASSP.2006.1661346.

Durrett, R. (2005), *Probability: Theory and Examples*, 3rd ed., 497 pp., Brooks/Cole, Belmont, CA, USA.

Dziewonski, A. M., and J. H. Woodhouse (1983), An experiment in systematic study of global seismicity: Centroid-moment tensor solutions fro 201 moderate and large earthquakes of 1981, *J. Geophys. Res.*, *88*, 3247–3271.

Dziewonski, A. M., T.-A. Chou, and J. H. Woodhouse (1981), Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, *J. Geophys. Res.*, *86*, 2825–2852.

Ehrendorfer, M. (1994), The liouville equation and its potential usefulness for the prediction of forecast skill. part i: Theory, *Mon. Wea. Rev.*, *122*, 703713.

Ekstrom, G., A. M. Dziewonski, N. N. Maternovskaya, and M. Nettles (2005), Global seismicity of 2003: Centroid-moment-tensor solutions for 1087 earthquakes, *Physics of the Earth and Planetary Inter.*, *148*, 327–351.

Englert, F. (2004), A brief course in spontaneous symmetry breaking ii. modern times: The beh mechanism, Talk presented at the 2001 Corfu Summer Institute on Elementary Particle Physics. hep-tp/0203097.

Evison, F. F., and D. A. Rhoades (2004), Demarcation and Scaling of Long-term Seismogenesis, *Pure Appl. Geophys.*, *161*, 21–45, doi:10.1007/s00024-003-2435-8.

Felzer, K. R., and E. E. Brodsky (2005), Testing the stress shadow hypothesis, *J. Geophys. Res.*, *110, B05S09*.

Felzer, K. R., and E. E. Brodsky (2006), Decay of aftershock density with distance indicates triggering by dynamic stress, *Nature*, *441*(7094), 735–738.

Felzer, K. R., T. W. Becker, R. E. Abercrombie, G. Ekstrom, and J. R. Rice (2002), Triggering of the 1999 Mw 7.1 Hector Mine earthquake by aftershocks of the 1992 Mw 7.3 Landers earthquake, *J. Geophys. Res.*, *107*(B09, 2190).

Felzer, K. R., R. E. Abercrombie, and G. Ekstrom (2003), Secondary aftershocks and their importance for aftershock forecasting, *Bull. Seism. Soc. Am.*, *93*(4), 1433–1448.

Felzer, K. R., R. E. Abercrombie, and G. Ekstrom (2004), A Common Origin for Aftershocks, Foreshocks, and Multiplets, *Bull. Seismol. Soc. Am.*, *94*(1), 88–98, doi:10.1785/0120030069.

Field, E. H. (2007), A Summary of Previous Working Groups on California Earthquake Probabilities, *Bull. Seismol. Soc. Am.*, *97*(4), 1033–1053, doi:10.1785/0120060048.

Frisch, U. (1995), *Turbulence. The legacy of A.N. Kolmogorov*, Cambridge University Press, Cambridge, UK.

Frohlich, C., and S. D. Davis (1990), Single-link cluster analysis as a method to evaluate spatial and temporal properties of earthquake catalogs, *Geophys. J. Int.*, *100*, 19–32.

Gardner, J. K., and L. Knopoff (1974), Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian?, *Bull. Seis. Soc. Am.*, *64*(5), 1363–1367.

Geilikman, M. B., and V. F. Pisarenko (2000), About the self-similarity in geophysical phenomena, in *Discrete Properties of Geophysical Media*, edited by M. Sadovskii, pp. 109–130, Nauka, Moscow, in Russian.

Geller, R. J. (1997), Earthquake prediction: a critical review, *Geophysical Journal International*, *131*(3), 425–450, doi:10.1111/j.1365-246X.1997.tb06588.x.

Geller, R. J., D. D. Jackson, Y. Y. Kagan, and F. Mulargia (1997a), Earthquakes cannot be predicted, *Science*, *275*, 1616–1617.

Geller, R. J., D. D. Jackson, Y. Y. Kagan, and F. Mulargia (1997b), Response - cannot earthquakes be predicted, *Science*, *278*, 488–490.

Gerstenberger, M. C., S. Wiemer, L. M. Jones, and P. A. Reasenberg (2005), Real-time forecasts of tomorrow's earthquakes in California, *Nature*, *435*(7040), 328–331.

Giardini, D. (1988), Frequency distribution and quantification of deep earthquakes, *J. Geophys. Res.*, *93*, 2095–2105.

Guo, Z., and Y. Ogata (1997), Statistical relations between parameters of aftershocks in time, space, and magnitude, *J. Geophys. Res.*, *102, B2*, 2857–2873.

Gutenberg, B., and C. F. Richter (1944), Frequency of earthquakes in California, *Bull. Seis. Soc. Am.*, *34*, 184–188.

Hainzl, S., and Y. Ogata (2005), Detecting fluid signals in seismicity data through statistical earthquake modeling, *J. Geophys. Res.*, *110, B05S07*, doi:10.1029/2004JB003247.

Hainzl, S., F. Scherbaum, and C. Beauval (2006), Estimating Background Activity Based on Interevent-Time Distribution, *Bull. Seismol. Soc. Am.*, *96*(1), 313–320, doi:10.1785/0120050053.

Hanks, T. C., and H. Kanamori (1979), A moment magnitude scale, *J. Geophys. Res.*, *84*, 2348–2350.

Harris, R. A., and J. R. Arrowsmith (2006), Introduction to the Special Issue on the 2004 Parkfield Earthquake and the Parkfield Earthquake Prediction Experiment, *Bull. Seismol. Soc. Am.*, *96*(4B), S1–10, doi:10.1785/0120050831.

Harte, D., and D. Vere-Jones (1999), Differences in coverage between the PDE and New Zealand local earthquake catalogs, *New Zealand J. of Geol. Geophys.*, *42*, 237–253.

Harte, D., and D. Vere-Jones (2005), The Entropy Score and its Uses in Earthquake Forecasting, *Pure and Applied Geophysics*, *162*, 1229–1253, doi:10.1007/s00024-004-2667-2.

Hawkes, A. G. (1971a), Spectra of some self-exciting and mutually exciting point processes, *Biometrika*, *58*(1), 83–90, doi:10.1093/biomet/58.1.83.

Hawkes, A. G. (1971b), Point spectra of some mutually exciting point processes, *J. Royal Stat. Soc. Series B (Meth.)*, *33*(3), 438–443.

Hawkes, A. G., and D. Oakes (1974), A cluster process representation of a self-exciting process, *J. of Appl. Prob.*, *11*(3), 493–503.

Heimpel, M., and P. Malin (1998), Aseismic slip in earthquake nucleation and self-similarity: evidence from parkfield, california, *Earth and Planet. Sci. Lett.*, *157*, 249–254.

Helffrich, G. R. (1997), How good are routinely determined focal mechanisms? empirical statistics based on a comparison of harvard, usgs and eri moment tensors, *Geophys. J. Int.*, *131*, 741–750.

Helmstetter, A. (2003), Is earthquake triggering driven by small earthquakes?, *Phys. Rev. Lett.*, *91*(5), 058,501, doi:10.1103/PhysRevLett.91.058501.

Helmstetter, A., and D. Sornette (2002), Subcritical and supercritical regimes in epidemic models of earthquake aftershocks, *J. Geophys. Res.*, *107*(B10), 2237, doi:10.1029/2001JB001580.

Helmstetter, A., and D. Sornette (2003a), Båth's law derived from the gutenberg-richter law and from aftershock properties, *Geophys. Res. Lett.*, *30*(11), 2069, doi:10.1029/2003GL018186.

Helmstetter, A., and D. Sornette (2003b), Importance of direct and indirect triggered seismicity, *Geophys. Res. Lett.*, *30*(11), 1576, doi:10.1029/2003GL017670.

Helmstetter, A., and D. Sornette (2003c), Foreshocks explained by cascades of triggered seismicity, *J. Geophys. Res.*, *108*(B10), 2457, doi:10.1029/2003JB002409.

Helmstetter, A., and D. Sornette (2003d), Predictability in the etas model of interacting triggered seismicity, *J. Geophys. Res.*, *108*(B10), 2482, doi:10.1029/2003JB002485.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2005a), Importance of small earthquakes for stress transfers and earthquake triggering, *J. Geophys. Res.*, *110*, doi:10.1029/2004JB003286.

Helmstetter, A., G. Ouillon, and D. Sornette (2005b), Are aftershocks of large californian earthquakes diffusing?, *J. Geophys. Res.*, *108*(B10), 2483, doi:10.1029/2003JB002503.

Helmstetter, A., D. Sornette, and M. J. Werner (2005c), Unbiased likelihood analysis of earthquake catalogs, unpublished.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2006), Comparison of short-term and time-independent earthquake forecast models for southern california, *Bull. Seismol. Soc. Am.*, *96*(1), doi:10.1785/0120050067.

Hergarten, S. (2002), *Self-Organized Criticality in Earth Systems*, Springer, Berlin.

Hero, I., A.O. (1991), Timing estimation for a filtered Poisson process in Gaussian noise, *Information Theory, IEEE Transactions on*, *37*(1), 92–106, doi:10.1109/18.61107.

Hill, D. P., et al. (1993), Seismicity remotely triggered by the magnitude 7.3 Landers, California, earthquake, *Science*, *260*(5114), 1617–1623.

Hinrichsen, H. (2000), Non-equilibrium critical phenomena and phase transitions into absorbing states, *Advances In Physics*, *49*, 815–958(144).

Huang, Y., H. Saleur, C. Sammis, and D. Sornette (1998), Precursors, aftershocks, criticality and self-organized criticality, *Europhysics Letters*, *41*, 43–48, doi:10.1209/epl/i1998-00113-x.

Ide, K., A. Bennett, P. Courtier, M. Ghil, and A. Lorenc (1997), Unified notation for data assimilation: Operational, sequential and variational, in Data Assimilation, Meteorology and Oceanography: Theory and Practice, *J. Meteor. Soc. Japan*, *75*(1B), 71–79.

Ide, S., and M. Takeo (1997), Determination of constitutive relations of fault slip based on seismic wave analysis, *J. Geophys. Res.*, *102*(B12), 27,379.

Iio, Y. (1991), Minimum size of earthquakes and minimum value of dynamic rupture velocity, *Tectonophysics*, *197*, 19–25.

Izenman, A. J. (1991), Recent developments in nonparametric density estimation, *J. Am. Stat. Assoc.*, *86*(413), 205–224.

Jackson, D. D. (1996), Hypothesis testing and earthquake prediction, *Proc. Natl. Acad. Sci. USA*, *93*, 3772–3775.

Jackson, D. D., and Y. Kagan (1999), Testable earthquake forecasts for 1999, *Seismol. Res. Lett.*, *70*(4), 393–403.

Jackson, D. D., and Y. Y. Kagan (2006), The 2004 Parkfield Earthquake, the 1985 Prediction, and Characteristic Earthquakes: Lessons for the Future, *Bull. Seismol. Soc. Am.*, *96*(4B), S397–409, doi:10.1785/0120050821.

Jamison, B., S. Orey, and W. Pruitt (1965), Convergence of weighted averages of independent random variables, *Z. Wahrscheinlichkeitstheorie*, *4*, 40–44.

Jensen, H. J. (1998), *Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems*, Cambridge University Press, Cambridge, UK.

Jessen, A. H., and T. Mikosch (2006), Regularly varying functions, *Publications de l'Institut Mathématique, Nouvelle Série*, *79*(93).

Johansen, A., and D. Sornette (1998), Evidence of discrete scale invariance by canonical averaging, *Int. J. Mod. Phys. C*, *9*(3), 433–447.

Johansen, A., D. Sornette, H. Wakita, U. Tsunogai, W. I. Newman, and H. Saleur (1996), Discrete Scaling in Earthquake Precursory Phenomena: Evidence in the Kobe Earthquake, Japan, *Journal de Physique I*, *6*, 1391–1402.

Jordan, T. H. (2006), Earthquake predictability: Brick by brick, *Seismol. Res. Lett.*, *77*(1), 3–6.

Kagan, Y. Y. (1973), Statistical methods in the study of seismic processes, *Bull. Int. Stat. Inst.*, *45*, 437–453.

Kagan, Y. Y. (1982), Stochastic model of earthquake fault geometry, *Geophys. J. Roy. astr. Soc.*, *71*, 659–691.

Kagan, Y. Y. (1987a), Point sources of elastic deformation: Elementary sources, static displacements, *Geophys. J. Roy. astr. Soc.*, *90*, 1–34.

Kagan, Y. Y. (1987b), Point sources of elastic deformation: Elementary sources, dynamic displacements, *Geophys. J. Roy. astr. Soc.*, *91*, 891–912.

Kagan, Y. Y. (1987c), Multipole expansions of extended sources of elastic deformation, *Geophys. J. Roy. astr. Soc.*, *93*, 101–114.

Kagan, Y. Y. (1989), Earthquakes and fractals, *Annual Review of Materials Science: Fractal Phenomena in Disordered Systems*, *19*, 520–522.

Kagan, Y. Y. (1991), Likelihood analysis of earthquake catalogs, *Geophys. J. Intern.*, *106*, 135–148.

Kagan, Y. Y. (1992), Seismicity: Turbulence of solids, *Nonlinear Sci. Today*, *2*, 1–13.

Kagan, Y. Y. (1993), Statistics of characteristic earthquakes, *Bull. Seismol. Soc. Am.*, *83*(1), 7–24.

Kagan, Y. Y. (1994), Observational evidence for earthquakes as a nonlinear dynamic process, *Physica D*, *77*, 160–192.

Kagan, Y. Y. (1997a), Statistical aspects of parkfield earthquake sequence and parkfield prediction experiment, *Tectonophysics*, *270*, 207–219.

Kagan, Y. Y. (1997b), Are earthquakes predictable?, *Geophys. J. Int.*, *131*, 505–525.

Kagan, Y. Y. (1999a), Universality of the seismic moment-frequency relation, *Pure and Appl. Geophys.*, *155*, 537–573.

Kagan, Y. Y. (1999b), Is earthquake seismology a hard, quantitative science?, *Pure Appl. Geoph.*, *155*, 233–258.

Kagan, Y. Y. (2002a), Aftershock Zone Scaling, *Bull. Seismol. Soc. Am.*, *92*(2), 641–655, doi:10.1785/0120010172.

Kagan, Y. Y. (2002b), Modern California Earthquake Catalogs and Their Comparison, *Seismol. Res. Lett.*, *73*(6), 921–929.

Kagan, Y. Y. (2003), Accuracy of modern global earthquake catalogs, *Phys. Earth Plan. Inter.*, *135*, 173–209.

Kagan, Y. Y. (2004), Short-term properties of earthquake catalogs and models of earthquake source, *Bull. Seismol. Soc. Am.*, *94*(4), 1207–1228.

Kagan, Y. Y. (2006), Why does theoretical physics fail to explain and predict earthquake occurrence?, in *Modelling Critical and Catastrophic Phenomena in Geoscience: A Statistical Physics Approach, Lecture Notes in Physics, 705*, edited by P. Bhattacharyya and B. K. Chakrabarti, pp. 303–359, Springer Verlag, Berlin.

Kagan, Y. Y. (2007), Earthquake spatial distribution: the correlation dimension, *Geophys. J. Intern.*, *168*, 1175–1194, doi:10.1111/j.1365-246X.2006.03251.x.

Kagan, Y. Y., and H. Houston (2005), Relation between mainshock rupture process and omori's law for aftershock moment release rate, *Geophys. J. Int.*, *163*, 1039–1048, doi:10.1111/j.1365-246X.2005.02772.x.

Kagan, Y. Y., and D. D. Jackson (1991a), Long term earthquake clustering, *Geophys. J. Int.*, *104*(1), 117–133.

Kagan, Y. Y., and D. D. Jackson (1991b), Seismic gap hypothesis: Ten years after, *J. Geophys. Res.*, *96*, 21,419–21,431.

Kagan, Y. Y., and D. D. Jackson (1994), Long-term probabilistic forecasting of earthquakes, *J. Geophys. Res.*, *99*(B7), 13,685–13,700.

Kagan, Y. Y., and D. D. Jackson (1995), New seismic gap hypothesis: Five years after, *J. Geophys. Res.*, *100*(B3), 3943–3959.

Kagan, Y. Y., and D. D. Jackson (2000), Probabilistic forecasting of earthquakes, *Geophys. J. Int.*, *143*, 483–453.

Kagan, Y. Y., and L. Knopoff (1981), Stochastic synthesis of earthquake catalogs, *J. Geophys. Res.*, *86*(B4), 2853–2862.

Kagan, Y. Y., and L. Knopoff (1987), Statistical short-term earthquake prediction, *Science*, *236*(4808), 1563–1567.

Kalnay, E. (2003), *Atmospheric Modeling, Data Assimilation and Predictability*, 364 pp., Cambridge Univ. Press, Cambridge, UK.

Kanamori, H. (1977), The energy release in great earthquakes, *J. Geophys. Res.*, *82*, 2981–2988.

Kanamori, H., and E. E. Brodsky (2004), The physics of earthquakes, *Rep. Prog. Phys.*, *67*, 1429.

Karr, A. (1986), *Point Processes and Their Statistical Inference*, Marcel Dekker, Inc., New York and Basel.

Kay, S. M. (1993), *Fundamentals of Statistical Signal Processing: Estimation Theory*, *Prentice-Hall Signal Processing Series*, vol. I, 595 pp., Prentice Hall, Inc., Upper Saddle River, New Jersey.

Keilis-Borok, V. I., and A. A. Soloviev (Eds.) (2003), *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*, 337 pp., Springer-Verlag, Berlin Heidelberg.

Klein, F. W. (2002), User's Guide to HYPOINVERSE-2000, a Fortran Program to Solve for Earthquake Locations and Magnitudes, USGS Open File Report 02-171, Version 1.0.

Knopoff, L. (2000), The magnitude distribution of declustered earthquakes in Southern California, *Proceedings of the National Academy of Sciences*, *97*(22), 11,880–11,884.

Knopoff, L., and Y. Kagan (1977), Analysis of the theory of extremes as applied to earthquake problems, *J. Geophys. Res.*, *82*, 5647–5657.

Kong, A., J. Liu, and W. Wong (1994), Sequential imputations and Bayesian missing data problems, *J. Am. Stat. Assoc.*, *89*(425), 278–288.

Kuge, K. (1992), Systematic difference in the isc body-wave magnitude-seismic moment relationship between intermediate and deep earthquakes, *Bull. Seismol. Soc. Am.*, *82*(2), 819–835.

Kwakernaak, H. (1980), Estimation of pulse heights and arrival times, *Automatica*, *16*, 367–377.

Landau, L. D., and E. M. Lifshitz (1980), *Statistical Physics*, *Course on Theoretical Physics*, vol. 5, 3rd ed., Butterworth-Heinemann, Oxford, UK.

Lapusta, N., and J. R. Rice (2003), Nucleation and early seismic propagation of small and large events in a crustal earthquake model, *J. Geophys. Res.*, *108*(B4), 2205, doi:10.1029/2001JB000793.

Li, Y. G., K. Aki, D. Adams, A. Hasemi, and W. H. K. Lee (1994), Seismic guided waves trapped in the fault zone of the landers, california, earthquake of 1992, *J. Geophys. Res.*, *99*, 11,705–11,722.

Lindman, M., K. Jonsdottir, R. Roberts, B. Lund, and R. Bdvarsson (2005), Earthquakes descaled: On waiting time distributions and scaling laws, *Phys. Rev. Lett.*, *94*, 108,501.

Lindman, M., K. Jonsdottir, R. Roberts, B. Lund, and R. Bdvarsson (2006), Reply to comment by a. corral and k. christensen, *Phys. Rev. Lett.*, *96*, 109,802.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, 343 pp., Springer-Verlag, New York.

Liu, J. S., and R. Chen (1998), Sequential monte carlo methods for dynamic systems, *J. Am. Stat. Assoc.*, *93*(443), 1032–1044.

Livina, V. N., S. Havlin, and A. Bunde (2006), Memory in the occurrence of earthquakes, *Phys. Rev. Lett.*, *95*, 208,501.

Lowen, S. B., and M. C. Teich (1990), Power-law shot noise, *IEEE Transactions on Information Theory*, *36*(6), 1302.

Luebeck, S. (2004), Universal scaling behavior of non-equilbrium phase transitions, *Intern. J. Mod. Phys. B*, *18*, 3977.

Lund, J., and M. Rudemo (2000), Models for point processes observed with noise, *Biometrika*, *87*(2), 235–249.

Marone, C., and B. Kilgore (1993), Scaling of the critical slip distance for seismic faulting with shear strain in fault zones, *Nature*, *362*, 618–621.

Matsu'ura, M. (1999), Physical modeling and simulation of the earthquake cycles, in *1-st ACES Workshop Proc.*, edited by P. Mora, pp. 159–160, APEC Cooperation Earthquake Simulation, Brisbane, Australia.

McCann, W., S. Nishenko, L. Sykes, and J. Krause (1979), Seismic gaps and plate tectonics: Seismic potential for major boundaries, *Pure Appl. Geophys.*, *117*, 1082–1147.

Mézard, M., G. Parisi, and M. Virasoro (1987), *Spin glass theory and beyond*, World Scientific Lecture Notes in Physics Vol. 9, Cambridge Univ. Press, Cambridge, UK.

Mignan, A., D. D. Bowman, and G. C. P. King (2006), An observational test of the origin of accelerating moment release before large earthquakes, *Journal of Geophysical Research (Solid Earth)*, *111*(B10), 11,304, doi:10.1029/2006JB004374.

Mikumo, T., K. B. Olsen, E. Fukuyama, and Y. Yagi (2003), Stress-breakdown time and slip-weakening distance inferred from slip-velocity functions on earthquake faults, *Bull. Seismol. Soc. Am.*, *93*, 264–282.

Miller, R. N., E. F. Carter, and S. T. Blue (1999), Data assimilation into nonlinear stochastic models, *Tellus*, *51A*(2), 167–194.

Miltenberger, P., D. Sornette, and C. Vanneste (1993), Fault self-organization as optimal random paths selected by critical spatiotemporal dynamics of earthquakes, *Physical Review Letters*, *71*, 3604–3607, doi:10.1103/PhysRevLett.71.3604.

Molchan, G. (2005), Interevent time distribution in seismicity: A theoretical approach, *Pure and Appl. Geophys.*, *162*, 1135–1150, doi:10.1007/s00024-004-2664-5.

Molchan, G., and T. Kronrod (2005), On the spatial scaling of seismicity rate, *Geophys. J. Int.*, *162*(3), 899–909, doi:10.1111/j.1365-246X.2005.02693.x.

Molchan, G. M. (1990), Strategies in strong earthquake prediction, *Physics of the Earth and Planetary Interiors*, *61*, 84–98, doi:10.1016/0031-9201(90)90097-H.

Molchan, G. M., and Y. Y. Kagan (1992), Earthquake prediction and its optimization, *J. Geophys. Res.*, *97*, 4823–4838.

Nature Debates (1999), Nature debates: Is the reliable prediction of individual earthquakes a realistic scientific goal?, available from http://www.nature.com/nature/debates/.

Nechad, H., A. Helmstetter, R. E. Guerjouma, and D. Sornette (2004), Andrade and critical time-to-failure laws in fiber-matrix composites: Experiments and model, *J. Mech. Phys. Solids*, *53*, 1099–1127.

Nishenko, S. P. (1991), Circum-pacific seismic potential 1989-1999, *Pure Appl. Geophys.*, *135*, 169–259.

Ogata, Y. (1978), The asymptotic behaviour of maximum likelihood estimators for stationary point processes, *Ann. Inst. Stat. Math.*, *5*(2), 379–402.

Ogata, Y. (1988), Statistical models for earthquake occurrence and residual analysis for point processes, *J. Am. Stat. Assoc.*, *83*, 9–27.

Ogata, Y. (1989), Statistical model for standard seismicity and detection of anomalies by residual analysis, *Tectonophysics*, *169*, 159–174.

Ogata, Y. (1992), Detection of precursory relative quiescence before great earthquakes through a statistical model, *J. Geophys. Res.*, *97*, 19,845–+.

Ogata, Y. (1998), Space-time point-process models for earthquake occurrences, *Ann. Inst. Stat. Math.*, *5*(2), 379–402.

Ogata, Y. (1999a), Estimating the hazard of rupture using uncertain occurrence times of paleoearthquakes, *J. Geophys. Res.*, *104*(B8), 17,995–18,014.

Ogata, Y. (1999b), Seismicity Analysis through Point-process Modeling: A Review, *Pure and Applied Geophysics*, *155*, 471–507.

Ogata, Y. (2001), Increased probability of large earthquakes near aftershock regions with relative quiescence, *J. Geophys. Res.*, *106*(B5), 8729–8744.

Ogata, Y. (2002), Slip-size-dependent renewal processes and bayesian inferences for uncertainties, *J. Geophys. Res.*, *107*(B11), 2268, doi:10.1029/2001JB000668.

Ogata, Y. (2004), Space-time model for regional seismicity and detection of crustal stress changes, *J. Geophys. Res.*, *109*(B03308), doi:10.1029/2003JB002621.

Ogata, Y. (2005), Synchronous seismicity changes in and around the northern Japan preceding the 2003 Tokachi-ocki earthquake of M8.0, *J. Geophys. Res.*, *110*(B08305), doi:10.1029/2004JB003323.

Ogata, Y. (2006a), Monitoring of anomaly in the aftershock sequence of the 2005 earthquake of M7.0 off coast of the western Fukuoka, Japan, by the ETAS model, *Geophys. Res. Lett.*, *33*, 1303–+, doi:10.1029/2005GL024405.

Ogata, Y. (2006b), Seismicity anomaly scenario prior to the major recurrent earthquakes off the east coast of Miyagi Prefecture, northern Japan, *Tectonophysics*, *424*, 291–306.

Ogata, Y., and K. Katsura (2006), Immediate and updated forecasting of aftershock hazard, *Geophys. Res. Lett.*, *33*, 10,305–+, doi:10.1029/2006GL025888.

Ogata, Y., and J. Zhuang (2006), Space-time ETAS models and an improved extension, *Tectonophysics*, *413*, 13–23.

Ogata, Y., R. S. Matsu'ura, and K. Katsura (1993), Fast likelihood computation of epidemic-type aftershock-sequence model, *Geophys. Res. Lett.*, *20*(19), 2143–2146.

Ogata, Y., L. M. Jones, and S. Toda (2003), When and where the aftershock activity was depressed: Contrasting decay patterns of the proximate large earthquakes in southern california, *J. Geophys. Res.*, *108*(B6), 2318, doi:10.1029/2002JB002009.

Olam, Z., H. J. S. Feder, and K. Christensen (1992), Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes, *Phys. Rev. Lett.*, *68*(8), 1244–1247.

Omori, F. (1894), On after-shocks of earthquakes, *J. Coll. Sci. Imp. Univ. Tokyo*, *7*, 111–200.

Oppenheimer, D. (2007), Catalog production in Northern CA and by the ANSS, Talk presented at the CSEP April 2007 meeting at USC, available from `http://scecdata.usc.edu/csep/`.

Ouillon, C. C., G., and D. Sornette (1996), Hierarchical geometry of faulting, *J. Geophys. Res.*, *101*(B3), 5477–5487.

Ouillon, G., and D. Sornette (2005), Magnitude-dependent Omori law: Empirical study and theory, *J. Geophys. Res.*, *110*(B04306), doi:10.1029/2004JB003311.

Ozaki, T. (1979), Maximum likelihood estimation of hawkes' self-exciting point processes, *Ann. Inst. Stat. Math.*, *31*(Part B), 145–155.

Papangelou, F. (1972), Integrability of expected increments of point processes and a related random change of scale, *Trans. Amer. Math. Soc.*, *164*, 438–506.

Patton, H. J. (2001), Regional magnitude scaling, transportability, and $m_s$ : $m_b$ discrimination at small magnitudes, *Pure and Applied Geophys.*, *158*, 1951–2015.

Peng, Z., J. E. Vidale, M. Ishii, and A. Helmstetter (2007), Seismicity rate immediately before and after main shock rupture from high-frequency waveforms in japan, *J. Geophys. Res.*, *112*(B03306), doi:10.1029/2006JB004386.

Peters, O., and J. D. Neelin (2006), Critical phenomena in atmospheric precipitation, *Nature Physics*, *2*, 393–396, doi:10.1038/nphys314.

Pham, D. T. (2001), Stochastic methods for sequential data assimilation in strongly nonlinear systems, *Monthly Weather Review*, *129*, 1194–1207.

Pisarenko, V. F., and D. Sornette (2003), Characterization of the frequency of extreme events by the generalized pareto distribution, *Pure and Appl. Geophys.*, *160*, 2343–2364.

Pisarenko, V. F., D. Sornette, and M. V. Rodkin (2004), Deviations of the distributions of seismic energies from the gutenberg-richter law, *Computational Seismology*, *35*, 138–159.

Pisarenko, V. F., A. Sornette, D. Sornette, and M. V. Rodkin (2008), New approach to the characterization of mmax and of the tail of the distribution of earthquake magnitudes, *in press in Pure and Applied Geophysics*, e-print at arxiv.org/abs/physics/0703010.

Preliminary Determination of Epicenters (PDE) (2001), Monthly Listings, US Department of Interior / Geological Survey, National Earthquake Information Center, April-June 2000, Open File Report 2000-600-B.

Pruessner, G. (2004), Studies in self-organized criticality, Ph.D. thesis, Imperial College London, available from http://www.ma.imperial.ac.uk/ pruess/publications/.

Rao, C. R. (1965), *Linear statistical inference and its applications*, John Wiley & Sons, Inc., New York, USA.

Rathbun, S. L. (1996), Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes, *J. Stat. Planning and Inference*, *51*, 55–74.

Reasenberg, P. (1985), Second-order moment of central california seismicity, 1969-82, *J. Geophys. Res.*, *90*, 5479–5495.

Reasenberg, P., and L. M. Jones (1989), Earthquake hazard after a mainshock in california, *Science*, *243*(4895), 1173–1176.

Reasenberg, P., and L. M. Jones (1994), Earthquake aftershocks: Update, *Science*, *265*(4895), 1251–1252.

Rhoades, D. A., and F. F. Evison (2004), Long-range Earthquake Forecasting with Every Earthquake a Precursor According to Scale, *Pure Appl. Geophys.*, *161*, 47–72, doi:10.1007/s00024-003-2434-9.

Rhoades, D. A., R. V. Dissen, and D. Dowrick (1994), On the handling of uncertainties in estimating the hazard of rupture on a fault segment, *J. Geophys. Res.*, *99*(B7), 13,701 – 13,712.

Rice, S. O. (1945), Mathematical analysis of random noise, in *Selected Papers on Noise and Stochastic Processes (1954)*, edited by N. Wax, pp. 133–294, Dover, New York, USA.

Richardson, E., and T. H. Jordan (2002), Seismicity in deep gold mines of south africa: Implications for tectonic earthquakes, *Bull. Seismo. Soc. Am.*, *92*, 1766–1782.

Rong, Y., D. D. Jackson, and Y. Y. Kagan (2003), Seismic gaps and earthquakes, *Journal of Geophysical Research (Solid Earth)*, *108*, 2471, doi:10.1029/2002JB002334.

Sadovskii, M. A. (1999), *Geophysics and Physics of Explosion (selected works)*, 335 pp., Nauka, Moscow, Russia.

Sahimi, S., and S. Arbabi (1996), Scaling laws for fracture of heterogeneous materials and rock, *Phys. Rev. Lett.*, *77*, 3689–3692.

Saichev, A., and D. Sornette (2004), Anomalous power law distribution of total lifetimes of branching processes: Application to earthquake aftershock sequences, *Phys. Rev. E*, *70*, 046,123, doi:10.1103/PhysRevE.70.046123.

Saichev, A., and D. Sornette (2005a), Distribution of the largest aftershocks in branching models of triggered seismicity: Theory of the universal båth's law, *Phys. Rev. E*, *71*, 056,127.

Saichev, A., and D. Sornette (2005b), Vere-Jones' self-similar branching model, *Phys. Rev. E*, *72*, 056,122.

Saichev, A., and D. Sornette (2006a), Renormalization of branching models of triggered seismicity from total to observable seismicity, *Eur. Phys. J. B*, *51*, 443–459.

Saichev, A., and D. Sornette (2006b), "universal" distribution of interearthquake times explained, *Phys. Rev. Lett.*, *97*, 078,501.

Saichev, A., and D. Sornette (2007), Theory of earthquake recurrence times, *J. Geophys. Res.*, *112*(B04313), doi:10.1029/2006JB004536.

Saichev, A., A. Helmstetter, and D. Sornette (2005), Anomalous scaling of offspring and generation numbers in branching processes, *Pure and App. Geophys.*, *162*, 1113–1134.

Saleur, H., C. G. Sammis, and D. Sornette (1996a), Renormalization group theory of earthquakes, *Nonlinear Processes in Geophysics*, *3*, 102–109.

Saleur, H., C. G. Sammis, and D. Sornette (1996b), Discrete scale invariance, complex fractal dimensions, and log-periodic fluctuations in seismicity, *J. Geophys. Res.*, *101*, 17,661–17,678, doi:10.1029/96JB00876.

Sammis, S., and D. Sornette (2002), Positive feedback, memory and the predictability of earthquakes, *Proceedings of the National Academy of Sciences*, *99*, 2501–2508.

Schoenberg, F. P. (2002), On rescaled Poisson processes and the Brownian bridge, *Ann. Int. Stat. Math.*, *54*(2), 445–457.

Schoenberg, F. P. (2003), Multidimensional residual analysis of point process models for earthquake occurrences, *J. Am. Stat. Assoc.*, *98*, 789–795(7), doi:doi:10.1198/016214503000000710.

Schoenberg, F. P., A. Chu, and A. Veen (2007), On the relationship between lower magnitude thresholds and bias in ETAS parameters, unpublished.

Scholz, C. H. (1998), Earthquakes and friction laws, *Nature*, *391*, 37–42.

Scholz, C. H. (2002), *The Mechanics of Earthquakes and Faulting*, 2nd ed., Cambridge University Press, Cambridge.

Schorlemmer, D., and M. Gerstenberger (2007), RELM testing center, in *Special Issue on Working Group on Regional Earthquake Likelihood Models (RELM)*, *Seismol. Res. Lett.*, *78*(1), 30.

Schorlemmer, D., M. Gerstenberger, S. Wiemer, D. Jackson, and D. Rhoades (2007), Earthquake likelihood model testing, in *Special Issue on Working Group on Regional Earthquake Likelihood Models (RELM)*, *Seismol. Res. Lett.*, *78*(1), 17.

Sellers, E. J., M. O. Kataka, and L. M. Linzer (2003), Source parameters of acoustic emission events and scaling with mining-induced seismicity, *J. Geophys. Res.*, *108*(B9), doi:10.1029/2001JB000670.

Shcherbakov, R., D. L. Turcotte, and J. Rundle (2004), A generalized omori's law for earthquake aftershock decay, *Geophys. Res. Lett.*, *31*(L11613), doi:10.1029/2004GL019808.

Shcherbakov, R., D. L. Turcotte, and J. Rundle (2005), A generalized omori's law for earthquake aftershock decay, *Pure Appl. Geophys.*, *162*, 1051.

Sipkin, S. A. (1986), Estimation of earthquake source parameters by the inversion of waveform data: Global seismicity, 1981-1983, *Bull. Seismol. Soc. Am.*, *76*(6), 1515–1541.

Sipkin, S. A. (1994), Rapid determination of global moment-tensor solutions, *Geophys. Res. Lett.*, *21*(16), 1667–1670.

Smalley, R. F., Jr., D. L. Turcotte, and S. A. Solla (1985), A renormalization group approach to the stick-slip behavior of faults, *J. Geophys. Res.*, *90*, 1894–1900.

Sornette, A., and D. Sornette (1989), Self-organized criticality and earthquakes, *Europhysics Letters*, *9*, 197–+.

Sornette, A., and D. Sornette (1999a), Earthquake rupture as a critical point: Consequences for telluric precursors, *Tectonophysics*, *179*, 327–334.

Sornette, A., and D. Sornette (1999b), Renormalization of earthquake aftershocks, *Geophys. Res. Lett.*, *6*, 1981–1984.

Sornette, D. (1992), Critical phase transitions made self-organized : a dynamical system feedback mechanism for self-organized criticality, *J. Phys. I France*, *2*, 2065–2073, doi:10.1051/jp1:1992267.

Sornette, D. (2002), Predictability of catastrophic events: material rupture, earthquakes, turbulence, financial crashes and human birth, *Proc. Nat. Acad. Sci. USA*, *99*, 2522–2529.

Sornette, D. (2004), *Critical Phenomena in Natural Sciences: Chaos, Fractals, Self-organization and Disorder: Concepts and Tools*, 2nd ed., 529 pp., Springer, Berlin.

Sornette, D., and A. Helmstetter (2002), Occurrence of finite-time singularities in epidemic models of rupture, earthquakes and starquakes, *Phys. Rev. Lett.*, *89*(15), 158,501.

Sornette, D., and K. Ide (2001), The Kalman-Lévy filter, *Physica D Nonlinear Phenomena*, *151*, 142–174.

Sornette, D., and L. Knopoff (1997), The paradox of the expected time until the next earthquake, *Bull. Seismol. Soc. Am.*, *87*(4), 789–798.

Sornette, D., and G. Ouillon (2005), Multifractal scaling of thermally-activated rupture processes, *Phys. Rev. Lett.*, *94*, 038,501.

Sornette, D., and C. G. Sammis (1995), Complex critical exponents from renormalization group theory of earthquakes : Implications for earthquake predictions, *J.Phys.I France*, *5*, 607–619.

Sornette, D., and M. J. Werner (2005a), Constraints on the size of the smallest triggering earthquake from the epidemic-type aftershock sequence model, båth's law, and observed aftershock sequences, *J. Geophys. Res.*, *110*(B08304), doi: 10.1029/2004JB003535.

Sornette, D., and M. J. Werner (2005b), Apparent clustering and apparent background earthquakes biased by undetected seismicity, *J. Geophys. Res.*, *110*(B09303), doi: 10.1029/2005JB003621.

Sornette, D., P. Miltenberger, and C. Vanneste (1994), Statistical physics of fault patterns self-organized by repeated earthquakes, *Pure and Applied Geophysics*, *142*, 491–527, doi:10.1007/BF00876052.

Sornette, D., L. Knopoff, Y. Kagan, and C. Vanneste (1996), Rank-ordering statistics of extreme events : application to the distribution of large earthquakes, *J. Geophys. Res.*, *101*, 13,883–13,893.

Stanley, H. E. (1999), Scaling, universality, and renormalization: Three pillars of modern critical phenomena, *Rev. Mod. Phys.*, *71*(2), S358–S366, doi:10.1103/ RevModPhys.71.S358.

Stein, R. S. (1999), The role of stress transfer in earthquake occurrence, *Nature*, *402*(6762), 605–609.

Stein, R. S., A. A. Barka, and J. H. Dieterich (1997), Progressive failure on the Northen Anatolian fault since 1939 by earthquake stress triggering, *Geophys. J. Int.*, *128*(3), 594–604.

Suteanu, C., D. Zugravescu, and F. Munteanu (2000), Fractal approach of structuring by fragmentation, *Pure and App. Geophys.*, *157*(4), 539–557.

Sykes, L. R., and S. C. Jaumé (1990), Seismic activity on neighbouring faults as a long-term precursor to large earthquakes in the San Francisco Bay area, *Nature*, *348*, 595–599, doi:10.1038/348595a0.

Turcotte, D. L. (1997), *Fractals and Chaos in Geology and Geophysics*, 2nd ed., 398 pp., Cambridge University Press, Cambridge, UK.

Turcotte, D. L., W. I. Newman, and A. Gabrielov (2000), A statistical physics approach to earthquakes, in *GeoComplexity and the Physics of Earthquake*, edited by J. B. Rundle, D. L. Turcotte, and W. Klein, pp. 83–96, American Geophysical Union.

Utsu, T. (1969), Aftershock and earthquake statistics (i): Some parameters which characterize an aftershock sequence and their interrelations, *J. Faculty of Science, Hokkaido Univ.*, *3*(Ser. VII (Geophysics)), 129–195.

Utsu, T., Y. Ogata, and R. S. Matsu'ura (1995), The centenary of the Omori formula for a decay law of aftershock activity, *J. Phys. Earth*, *43*, 1–33.

Veen, A., and F. P. Schoenberg (2007), Estimation of space-time branching process models in seismology using an EM-type algorithm, *in press in J. Am. Stat. Assoc.*

Vere-Jones, D. (1970), Stochastic models for earthquake occurrence, *J. Roy. Stat. Soc. Series B (Methodological)*, *32*(1), 1–62.

Vere-Jones, D. (1977), Statistical theories of crack propagation, *Mathematical Geology*, *9*, 455–481.

Vere-Jones, D. (1978), Earthquake prediction - a statistician's view, *J. Phys. Earth.*, *26*, 129–146.

Vere-Jones, D. (1995), Forecasting earthquakes and earthquake risk, *Intern. J. Forecasting*, *11*, 503–538.

Vere-Jones, D. (2000), Seismology - a statistical vignette, *J. Am. Stat. Assoc.*, *95*(451), 975–978.

Vere-Jones, D. (2005), A class of self-similar random measure, *Adv. in Appl. Probab.*, *37*(4), 908–914.

Vere-Jones, D. (2006), The development of statistical seismology: A personal experience, *Tectonophysics*, *413*, 5–12.

Vere-Jones, D., R. Robinson, and W. Yang (2001), Remarks on the accelerated moment release model: problems of model formulation, simulation and estimation, *Geophys. J. Intern.*, *144*, 517–531, doi:10.1046/j.1365-246X.2001.01348.x.

Voight, B. (1988), A method for prediction of volcanic eruptions, *Nature*, *332*, 125–130, doi:10.1038/332125a0.

Wang, Y., A. R. C. Paiva, and J. C. Principe (2006), A monte carlo sequential estimation for point process optimum filtering, *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, pp. 1846–1850.

Werner, M. J., and D. D. Jackson (2007), ETES: The inverse problem, in preparation.

Werner, M. J., and D. Sornette (2007a), Comment on "analysis of the spatial distribution between successive earthquakes", *Phys. Rev. Lett.*, *99*, 179,801.

Werner, M. J., and D. Sornette (2007b), Magnitude uncertainties impact seismic rate estimates, forecasts and predictability experiments, *submitted to J. Geophys. Res.*

Werner, M. J., K. Ide, and D. Sornette (2007), Earthquake forecasting based on data assimilation, in preparation.

Wesnousky, S. G. (1994), The Gutenberg-Richter or characteristic earthquake distribution, which is it?, *Bulletin of the Seismological Society of America*, *84*(6), 1940–1959.

Wilson, K. (1979), Problems in physics with many scales of length, *Sci. Am.*, *241*, 140–157.

Yeomans, J. M. (1992), *Statistical Mechanics of Phase Transitions*, Oxford University Press Inc., New York.

Zechar, J. D., and T. H. Jordan (2007), Testing alarm-based earthquake predictions, *accepted for publication in Geophys. J. Int.*

Zee, A. (2003), *Quantum Field Theory in a Nutshell*, Princeton University Press, Princeton, New Jersey.

Zheng, X.-G., and D. Vere-Jones (1991), Application of stress release models to historical earthquakes from North China, *Pure and Applied Geophysics*, *135*, 559–576, doi:10.1007/BF01772406.

Zhuang, J., Y. Ogata, and D. Vere-Jones (2002), Stochastic declustering of space-time earthquake occurrences, *J. Am. Stat. Assoc.*, *97*, 369–380.

Zhuang, J., Y. Ogata, and D. Vere-Jones (2004), Analyzing earthquake clustering features by using stochastic reconstruction, *J. Geophys. Res.*, *109*(B05301), doi: 10.1029/2003JB002879.