GWAS IV: Bayesian linear (variance component) models

Dr. Oliver Stegle
Christoh Lippert
Prof. Dr. Karsten Borgwardt

Max-Planck-Institutes Tübingen, Germany

BIOLOGISCHE KYBERNETIK

MAX-PLANCK-GESELLSCHAFT

Tübingen
Summer 2011

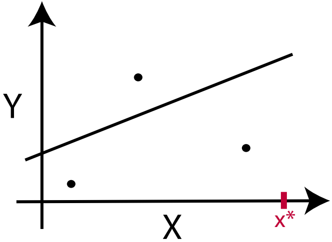
Regression

Lineare regression:

- Making predictions
- Comparison of alternative models

Bayesian and regularized regression:

- Uncertainty in model parameters
- Generalized basis functions



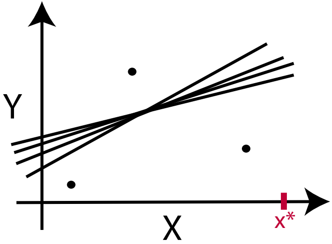
Regression

Lineare regression:

- Making predictions
- Comparison of alternative models

Bayesian and regularized regression:

- Uncertainty in model parameters
- Generalized basis functions



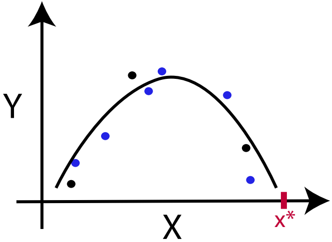
Regression

Lineare regression:

- Making predictions
- Comparison of alternative models

Bayesian and regularized regression:

- Uncertainty in model parameters
- Generalized basis functions



Further reading, useful material

- Christopher M. Bishop: Pattern Recognition and Machine learning [Bishop, 2006]
- Sam Roweis: Gaussian identities [Roweis, 1999]

Outline

Outline

Regression
Noise model and likelihood

- Given a dataset $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^N$, where $\mathbf{x}_n = \{x_{n,1}, \ldots, x_{n,S}\}$ is S dimensional (for example S SNPs), fit parameters $\boldsymbol{\theta}$ of a regressor f with added Gaussian noise:

$$y_n = f(\mathbf{x}_n; \boldsymbol{\theta}) + \epsilon_n \quad \text{where} \quad p(\epsilon \,|\, \sigma^2) = \mathcal{N}\left(\epsilon \,\middle|\, 0, \sigma^2\right).$$

- Equivalent likelihood formulation:

$$p(\mathbf{y} \,|\, \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \,\middle|\, f(\mathbf{x}_n), \sigma^2\right)$$

Regression
Choosing a regressor

- Choose f to be linear:

$$p(\mathbf{y} \,|\, \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \,\middle|\, \mathbf{x}_n \cdot \boldsymbol{\theta} + c, \sigma^2\right)$$

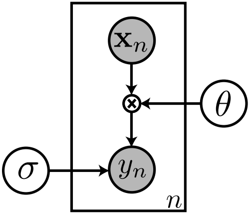
- Consider bias free case, $c = 0$, otherwise inlcude an additional column of ones in each $\mathbf{x}_n$.

Regression
Choosing a regressor

- Choose f to be linear:

$$p(\mathbf{y} \mid \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \,\middle|\, \mathbf{x}_n \cdot \boldsymbol{\theta} + c, \sigma^2\right)$$

- Consider bias free case, $c = 0$, otherwise inlcude an additional column of ones in each $\mathbf{x}_n$.



Equivalent graphical model

Linear Regression
Maximum likelihood

- Taking the logarithm, we obtain

$$\ln p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{X}, \sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}\left(y_n \,\middle|\, \mathbf{x}_n \cdot \boldsymbol{\theta}, \sigma^2\right)$$

$$= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y_n - \mathbf{x}_n \cdot \boldsymbol{\theta})^2}_{\text{Sum of squares}}$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

Linear Regression

Maximum likelihood

- Taking the logarithm, we obtain

$$\ln p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{X}, \sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}\left(y_n \,\middle|\, \mathbf{x}_n \cdot \boldsymbol{\theta}, \sigma^2\right)$$

$$= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y_n - \mathbf{x}_n \cdot \boldsymbol{\theta})^2}_{\text{Sum of squares}}$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

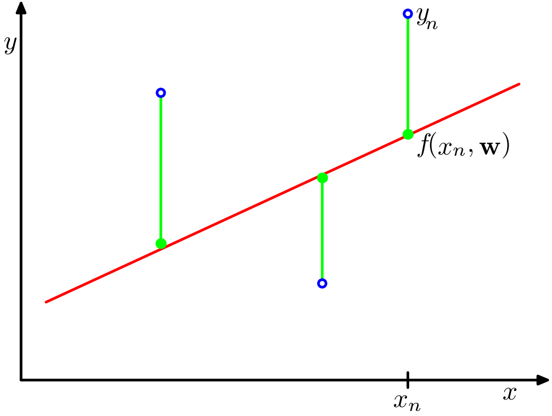
Linear Regression
Maximum likelihood

- Taking the logarithm, we obtain

$$\ln p(\mathbf{y}\,|\,\boldsymbol{\theta},\mathbf{X},\sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}\left(y_n \,\middle|\, \mathbf{x}_n \cdot \boldsymbol{\theta}, \sigma^2\right)$$
$$= -\frac{N}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\underbrace{\sum_{n=1}^{N}(y_n - \mathbf{x}_n\cdot\boldsymbol{\theta})^2}_{\text{Sum of squares}}$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

Linear Regression and Least Squares



(C.M. Bishop, Pattern Recognition and Machine Learning)

$$E(\boldsymbol{\theta}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - \mathbf{x}_n \cdot \boldsymbol{\theta})^2$$

Linear Regression and Least Squares

- Derivative w.r.t a single weight entry θ_i

$$\frac{d}{\mathrm{d}\theta_i} \ln p(\mathbf{y} \,|\, \boldsymbol{\theta}, \sigma^2) = \frac{d}{\mathrm{d}\theta_i}\left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \mathbf{x}_n \cdot \boldsymbol{\theta})^2 \right]$$

$$= \frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \mathbf{x}_n \cdot \boldsymbol{\theta}) x_i$$

- Set gradient w.r.t to $\boldsymbol{\theta}$ to zero

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{y} \,|\, \boldsymbol{\theta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \mathbf{x}_n \cdot \boldsymbol{\theta}) \mathbf{x}_n^{\mathrm{T}} = 0$$

$$\Longrightarrow \boldsymbol{\theta}_{\mathsf{ML}} = \underbrace{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}}_{\text{Pseudo inverse}} \mathbf{y}$$

- Here, the matrix $\mathbf{X}$ is defined as $\mathbf{X} = \begin{bmatrix} x_{1,1} & \dots & x1, D \\ \dots & \dots & \dots \\ x_{N,1} & \dots & x_{N,D} \end{bmatrix}$

Linear Regression and Least Squares

- Derivative w.r.t a single weight entry θ_i

$$\frac{d}{\mathrm{d}\theta_i} \ln p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2) = \frac{d}{\mathrm{d}\theta_i}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{x}_n \cdot \boldsymbol{\theta})^2\right]$$

$$= \frac{1}{\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{x}_n \cdot \boldsymbol{\theta})x_i$$

- Set gradient w.r.t to $\boldsymbol{\theta}$ to zero

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2) = \frac{1}{\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{x}_n \cdot \boldsymbol{\theta})\mathbf{x}_n^{\mathrm{T}} = 0$$

$$\Longrightarrow \boldsymbol{\theta}_{\mathsf{ML}} = \underbrace{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}}_{\text{Pseudo inverse}}\mathbf{y}$$

- Here, the matrix $\mathbf{X}$ is defined as $\mathbf{X} = \begin{bmatrix} x_{1,1} & \dots & x1, D \\ \dots & \dots & \dots \\ x_{N,1} & \dots & x_{N,D} \end{bmatrix}$

Linear Regression and Least Squares

- Derivative w.r.t a single weight entry θ_i

$$\frac{d}{\mathrm{d}\theta_i}\ln p(\mathbf{y}\,|\,\boldsymbol{\theta},\sigma^2) = \frac{d}{\mathrm{d}\theta_i}\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{x}_n\cdot\boldsymbol{\theta})^2\right]$$

$$= \frac{1}{\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{x}_n\cdot\boldsymbol{\theta})x_i$$

- Set gradient w.r.t to $\boldsymbol{\theta}$ to zero

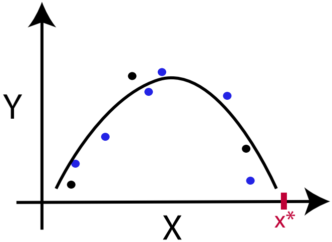
$$\nabla_{\boldsymbol{\theta}}\ln p(\mathbf{y}\,|\,\boldsymbol{\theta},\sigma^2) = \frac{1}{\sigma^2}\sum_{n=1}^{N}(y_n - \mathbf{x}_n\cdot\boldsymbol{\theta})\mathbf{x}_n^{\mathrm{T}} = 0$$

$$\implies \boldsymbol{\theta}_{\mathsf{ML}} = \underbrace{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}}_{\text{Pseudo inverse}}\mathbf{y}$$

- Here, the matrix $\mathbf{X}$ is defined as $\mathbf{X} = \begin{bmatrix} x_{1,1} & \dots & x1,D \\ \dots & \dots & \dots \\ x_{N,1} & \dots & x_{N,D} \end{bmatrix}$

Polynomial Curve Fitting
Motivation

- Non-linear relationships.
- Multiple SNPs playing a role for a particular phenotype.



Polynomial Curve Fitting

Univariate input x

- Use the polynomials up to degree K to construct new features from x

$$\begin{aligned} f(x, \boldsymbol{\theta}) &= \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_K x^K \\ &= \sum_{k=1}^{K} \theta_k \phi_k(x) = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\phi}(x) \end{aligned}$$

  where we defined $\boldsymbol{\phi}(x) = (1, x, x^2, \ldots, x^K)$.
- ϕ can be any feature mapping.
- Possible to show: the feature map ϕ can be expressed in terms of kernels (kernel trick).

Polynomial Curve Fitting

Univariate input x

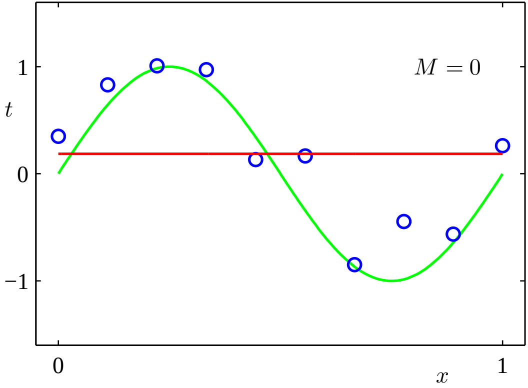
- Use the polynomials up to degree K to construct new features from x

$$
\begin{aligned}
f(x, \boldsymbol{\theta}) &= \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_K x^K \\
&= \sum_{k=1}^{K} \theta_k \phi_k(x) = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\phi}(x)
\end{aligned}
$$

  where we defined $\boldsymbol{\phi}(x) = (1, x, x^2, \ldots, x^K)$.
- ϕ can be any feature mapping.
- Possible to show: the feature map ϕ can be expressed in terms of kernels (kernel trick).

Polynomial Curve Fitting
Overfitting

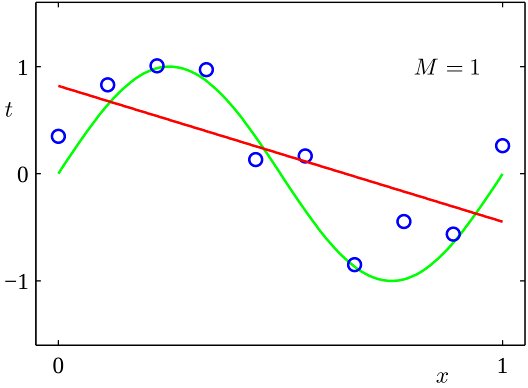
- The degree of the polynomial is crucial to avoid under- and overfitting.



(C.M. Bishop, Pattern Recognition and Machine Learning)

Polynomial Curve Fitting
Overfitting

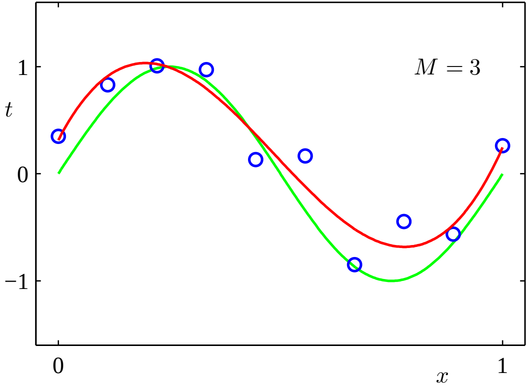
- The degree of the polynomial is crucial to avoid under- and overfitting.



(C.M. Bishop, Pattern Recognition and Machine Learning)

Polynomial Curve Fitting
Overfitting

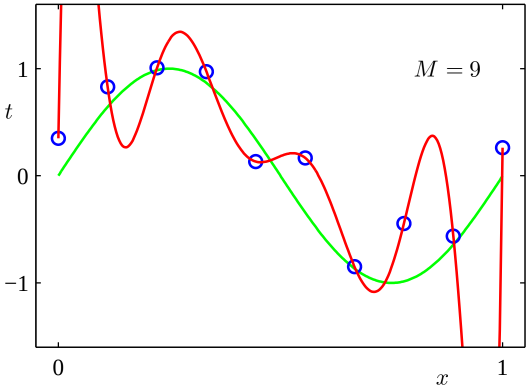
- The degree of the polynomial is crucial to avoid under- and overfitting.



(C.M. Bishop, Pattern Recognition and Machine Learning)

Polynomial Curve Fitting
Overfitting

- The degree of the polynomial is crucial to avoid under- and overfitting.



(C.M. Bishop, Pattern Recognition and Machine Learning)

Multivariate regression

Polynomial curve fitting

$$
\begin{aligned}
f(x, \boldsymbol{\theta}) &= \theta_0 + \theta_1 x + \cdots + \theta_K x^K \\
&= \sum_{k=1}^{K} \theta_k \phi_k(x) \\
&= \boldsymbol{\phi}(x) \cdot \boldsymbol{\theta},
\end{aligned}
$$

Multivariate regression (SNPs)

$$
\begin{aligned}
f(x, \boldsymbol{\theta}) &= \sum_{s=1}^{S} \theta_s x_s \\
&= \mathbf{x} \cdot \boldsymbol{\theta}
\end{aligned}
$$

- Note: When fitting a single binary SNP genotype $\mathbf{x}_i$, a linear model is most general!

Multivariate regression

Polynomial curve fitting

$$\begin{aligned} f(x, \boldsymbol{\theta}) &= \theta_0 + \theta_1 x + \dots + \theta_K x^K \\ &= \sum_{k=1}^{K} \theta_k \phi_k(x) \\ &= \boldsymbol{\phi}(x) \cdot \boldsymbol{\theta}, \end{aligned}$$

Multivariate regression (SNPs)

$$\begin{aligned} f(x, \boldsymbol{\theta}) &= \sum_{s=1}^{S} \theta_s x_s \\ &= \mathbf{x} \cdot \boldsymbol{\theta} \end{aligned}$$

- Note: When fitting a single binary SNP genotype $\mathbf{x}_i$, a linear model is most general!

Regularized Least Squares

- Solutions to avoid overfitting:
  1. Intelligently choose number of dimensions
  2. Regularize the regression weights $\boldsymbol{\theta}$
- Quadratically regularized objective function

$$E(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta})^2}_{\text{Squared error}} + \underbrace{\frac{\lambda}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}}_{\text{Regularizer}}$$

Regularized Least Squares

- Solutions to avoid overfitting:
  1. Intelligently choose number of dimensions
  2. Regularize the regression weights $\boldsymbol{\theta}$
- Quadratically regularized objective function

$$E(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta})^2}_{\text{Squared error}} + \underbrace{\frac{\lambda}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}}_{\text{Regularizer}}$$

Regularized Least Squares
More general regularizers

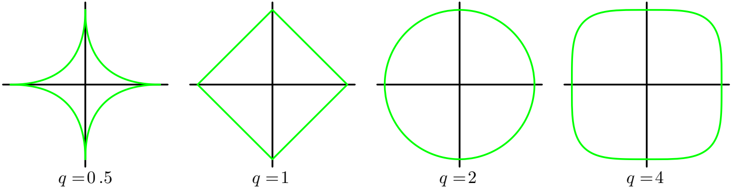
- More general regularization:

$$E(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta})^2}_{\text{Squared error}} + \underbrace{\frac{\lambda}{2}\sum_{d=1}^{D}|\theta_d|^{q}}_{\text{Regularizer}}$$

Regularized Least Squares
More general regularizers

- More general regularization:

$$E(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta})^2}_{\text{Squared error}} + \underbrace{\frac{\lambda}{2}\sum_{d=1}^{D}|\theta_d|^{q}}_{\text{Regularizer}}$$

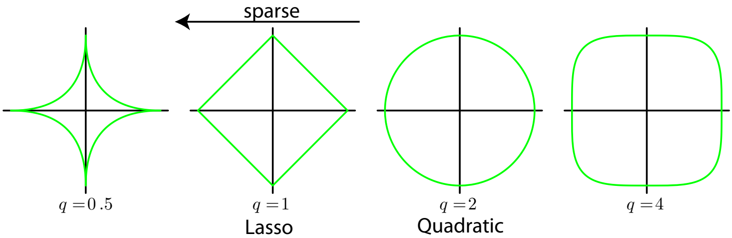


(C.M. Bishop, Pattern Recognition and Machine Learning)

Regularized Least Squares
More general regularizers

- More general regularization:

$$E(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta})^2}_{\text{Squared error}} + \underbrace{\frac{\lambda}{2}\sum_{d=1}^{D}|\theta_d|^{q}}_{\text{Regularizer}}$$



(C.M. Bishop, Pattern Recognition and Machine Learning)

Loss functions and related methods

- Even more general: general loss function

$$E(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}\mathcal{L}(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta})}_{\text{Loss}} + \underbrace{\frac{\lambda}{2}\sum_{d=1}^{D}|\theta_d|^q}_{\text{Regularizer}}$$

- Many state-of-the-art machine learning methods can be expressed within this framework.
  - Linear Regression: squared loss, squared regularizer.
  - Support Vector Machine: hinge loss, squared regularizer.
  - Lasso: squared loss, L1 regularizer.
- Inference: minimize the cost function $E(\boldsymbol{\theta})$, yielding a point estimate for $\boldsymbol{\theta}$.
- Q: How to determine q and the a suitable loss function?

Loss functions and related methods

- Even more general: general loss function

$$E(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}\mathcal{L}(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta})}_{\text{Loss}} + \underbrace{\frac{\lambda}{2}\sum_{d=1}^{D}|\theta_d|^q}_{\text{Regularizer}}$$

- Many state-of-the-art machine learning methods can be expressed within this framework.
  - Linear Regression: squared loss, squared regularizer.
  - Support Vector Machine: hinge loss, squared regularizer.
  - Lasso: squared loss, L1 regularizer.
- Inference: minimize the cost function $E(\boldsymbol{\theta})$, yielding a point estimate for $\boldsymbol{\theta}$.
- Q: How to determine q and the a suitable loss function?

Loss functions and related methods

- Even more general: general loss function

$$E(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}\mathcal{L}(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta})}_{\text{Loss}} + \underbrace{\frac{\lambda}{2}\sum_{d=1}^{D}|\theta_d|^{q}}_{\text{Regularizer}}$$

- Many state-of-the-art machine learning methods can be expressed within this framework.
  - Linear Regression: squared loss, squared regularizer.
  - Support Vector Machine: hinge loss, squared regularizer.
  - Lasso: squared loss, L1 regularizer.
- Inference: minimize the cost function $E(\boldsymbol{\theta})$, yielding a point estimate for $\boldsymbol{\theta}$.
- Q: How to determine q and the a suitable loss function?

Loss functions and related methods

- Even more general: general loss function

$$E(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}\mathcal{L}(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta})}_{\text{Loss}} + \underbrace{\frac{\lambda}{2}\sum_{d=1}^{D}|\theta_d|^q}_{\text{Regularizer}}$$

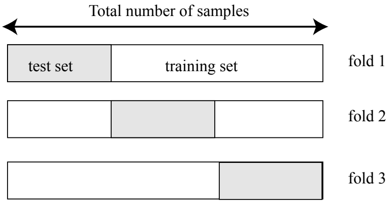
- Many state-of-the-art machine learning methods can be expressed within this framework.
  - Linear Regression: squared loss, squared regularizer.
  - Support Vector Machine: hinge loss, squared regularizer.
  - Lasso: squared loss, L1 regularizer.
- Inference: minimize the cost function $E(\boldsymbol{\theta})$, yielding a point estimate for $\boldsymbol{\theta}$.
- Q: How to determine q and the a suitable loss function?

Loss functions and related methods
Cross validation: minimization of expected loss

For each candidate model $\mathcal{H}$:

- Split data into K folds
- Training-test evaluation for each fold
- Assess average loss on test set

$$E_{\mathcal{H}} = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_k^{\text{test}}$$



Probabilistic interpretation

- So far: minimization of error functions.
- Back to probabilities?

$$E(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}\left(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta}\right)^2}_{\text{Squared error}} + \underbrace{\frac{\lambda}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}}_{\text{Regularizer}}$$

- Most alternative choices of regularizers and loss functions can be mapped to an equivalent probabilistic representation in a similar way.

Probabilistic interpretation

- So far: minimization of error functions.
- Back to probabilities?

$$E(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}\left(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta}\right)^2}_{\text{Squared error}} + \underbrace{\frac{\lambda}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}}_{\text{Regularizer}}$$

$$= -\sum_{n=1}^{N}\ln\mathcal{N}\left(y_n \,\middle|\, \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta}, \sigma^2\right) - \ln\mathcal{N}\left(\boldsymbol{\theta}\,\middle|\,\mathbf{0}, \frac{1}{\lambda}\mathbf{I}\right)$$

- Most alternative choices of regularizers and loss functions can be mapped to an equivalent probabilistic representation in a similar way.

Probabilistic interpretation

- So far: minimization of error functions.
- Back to probabilities?

$$E(\boldsymbol{\theta}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}\left(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta}\right)^2}_{\text{Squared error}} + \underbrace{\frac{\lambda}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}}_{\text{Regularizer}}$$

$$= -\sum_{n=1}^{N}\ln\mathcal{N}\left(y_n \,\middle|\, \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta}, \sigma^2\right) - \ln\mathcal{N}\left(\boldsymbol{\theta}\,\middle|\,\mathbf{0}, \frac{1}{\lambda}\mathbf{I}\right)$$

$$= -\ln p(\mathbf{y}\,|\,\boldsymbol{\theta}, \boldsymbol{\Phi}(\mathbf{X}), \sigma^2) - \ln p(\boldsymbol{\theta})$$

- Most alternative choices of regularizers and loss functions can be mapped to an equivalent probabilistic representation in a similar way.

Probabilistic interpretation

- So far: minimization of error functions.
- Back to probabilities?

$$
\begin{aligned}
E(\boldsymbol{\theta}) &= \underbrace{\frac{1}{2}\sum_{n=1}^{N}\left(y_n - \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta}\right)^2}_{\text{Squared error}} & &+ \underbrace{\frac{\lambda}{2}\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}}_{\text{Regularizer}} \\
&= -\sum_{n=1}^{N}\ln\mathcal{N}\left(y_n \,\middle|\, \boldsymbol{\phi}(\mathbf{x}_n)\cdot\boldsymbol{\theta}, \sigma^2\right) & &- \ln\mathcal{N}\left(\boldsymbol{\theta}\,\middle|\,\mathbf{0}, \frac{1}{\lambda}\mathbf{I}\right) \\
&= -\ln p(\mathbf{y}\,|\,\boldsymbol{\theta}, \boldsymbol{\Phi}(\mathbf{X}), \sigma^2) & &- \ln p(\boldsymbol{\theta})
\end{aligned}
$$

- Most alternative choices of regularizers and loss functions can be mapped to an equivalent probabilistic representation in a similar way.

Outline

Bayesian linear regression

- Likelihood as before

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \,\middle|\, \boldsymbol{\phi}(\mathbf{x}_n) \cdot \boldsymbol{\theta}, \sigma^2\right)$$

- Define a conjugate prior over $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta} \mid \mathbf{m}_0, \mathbf{S}_0\right)$$

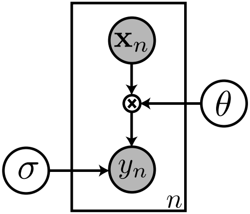
Bayesian linear regression

- Likelihood as before

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \,\middle|\, \boldsymbol{\phi}(\mathbf{x}_n) \cdot \boldsymbol{\theta}, \sigma^2\right)$$

- Define a conjugate prior over $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta} \mid \mathbf{m}_0, \mathbf{S}_0\right)$$



Bayesian linear regression

- Posterior probability of $\boldsymbol{\theta}$

$$
\begin{aligned}
p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{X}, \sigma^2) &\propto \prod_{n=1}^{N} \mathcal{N}\left(y_n \,\middle|\, \boldsymbol{\phi}(\mathbf{x}_n) \cdot \boldsymbol{\theta}, \sigma^2\right) \cdot \mathcal{N}\left(\boldsymbol{\theta} \,|\, \mathbf{m}_0, \mathbf{S}_0\right) \\
&= \mathcal{N}\left(\mathbf{y} \,\middle|\, \boldsymbol{\Phi}(\mathbf{X}) \cdot \boldsymbol{\theta}, \sigma^2 \mathbf{I}\right) \cdot \mathcal{N}\left(\boldsymbol{\theta} \,|\, \mathbf{m}_0, \mathbf{S}_0\right) \\
&= \mathcal{N}\left(\boldsymbol{\theta} \,\middle|\, \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)
\end{aligned}
$$

- where

$$
\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{\theta}} &= \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \frac{1}{\sigma^2} \boldsymbol{\Phi}(\mathbf{X})^{\mathrm{T}} \mathbf{y} \right) \\
\boldsymbol{\Sigma}_{\boldsymbol{\theta}} &= \left[\mathbf{S}_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{\Phi}(\mathbf{X})^{\mathrm{T}} \boldsymbol{\Phi}(\mathbf{X}) \right]^{-1}
\end{aligned}
$$

Bayesian linear regression
Prior choice

- Choice of prior: regularized (ridge) regression

$$p(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta} \,|\, \mathbf{m}_0, \mathbf{S}_0\right).$$

- In this case

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{X}, \sigma^2) \propto \mathcal{N}\left(\boldsymbol{\theta} \,|\, \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \frac{1}{\sigma^2}\boldsymbol{\Phi}(\mathbf{X})^{\mathrm{T}}\mathbf{y} \right)$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \left[\mathbf{S}_0^{-1} + \frac{1}{\sigma^2}\boldsymbol{\Phi}(\mathbf{X})^{\mathrm{T}}\boldsymbol{\Phi}(\mathbf{X}) \right]^{-1}$$

- Equivalent to maximum likelihood estimate for $\lambda \rightarrow 0$!

Bayesian linear regression

Prior choice

- Choice of prior: regularized (ridge) regression

$$p(\boldsymbol{\theta}) = \mathcal{N}\big(\boldsymbol{\theta} \,|\, \mathbf{0}, \frac{1}{\lambda}\mathbf{I}\big).$$

- In this case

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{X}, \sigma^2) \propto \mathcal{N}\left(\boldsymbol{\theta} \,\middle|\, \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \left(\qquad \frac{1}{\sigma^2}\boldsymbol{\Phi}(\mathbf{X})^{\mathrm{T}}\mathbf{y} \right)$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \left[\lambda\mathbf{I} + \frac{1}{\sigma^2}\boldsymbol{\Phi}(\mathbf{X})^{\mathrm{T}}\boldsymbol{\Phi}(\mathbf{X})\right]^{-1}$$

- Equivalent to maximum likelihood estimate for $\lambda \to 0$!

Bayesian linear regression

Prior choice

- Choice of prior: regularized (ridge) regression

- In this case

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{X}, \sigma^2) \propto \mathcal{N}\left(\boldsymbol{\theta} \,\middle|\, \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)$$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \left(\qquad \frac{1}{\sigma^2} \boldsymbol{\Phi}(\mathbf{X})^{\mathrm{T}} \mathbf{y} \right)$$

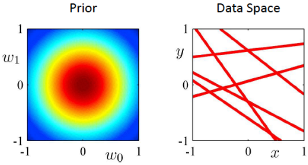
$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \left[\lambda \mathbf{I} + \frac{1}{\sigma^2} \boldsymbol{\Phi}(\mathbf{X})^{\mathrm{T}} \boldsymbol{\Phi}(\mathbf{X}) \right]^{-1}$$

- Equivalent to maximum likelihood estimate for $\lambda \to 0$!

Bayesian linear regression

Example

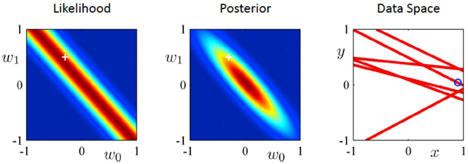
0 Data points



(C.M. Bishop, Pattern Recognition and Machine Learning)

Bayesian linear regression
Example

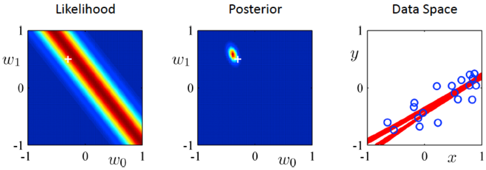
1 Data point



(C.M. Bishop, Pattern Recognition and Machine Learning)

Bayesian linear regression
Example

20 Data points



(C.M. Bishop, Pattern Recognition and Machine Learning)

Making predictions

- Prediction for fixed weight $\hat{\boldsymbol{\theta}}$ at input $\mathbf{x}^\star$ trivial:

$$p(y^\star \mid \mathbf{x}^\star, \hat{\boldsymbol{\theta}}, \sigma^2) = \mathcal{N}\left(y^\star \,\middle|\, \phi(\mathbf{x}^\star)\hat{\boldsymbol{\theta}}, \sigma^2\right)$$

- Integrate over $\boldsymbol{\theta}$ to take the posterior uncertainty into account

$$\begin{aligned} p(y^\star \mid \mathbf{x}^\star, \mathcal{D}) &= \int_{\boldsymbol{\theta}} p(y^\star \mid \mathbf{x}^\star, \boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}, \sigma^2) \\ &= \int_{\boldsymbol{\theta}} \mathcal{N}\left(y^\star \mid \phi(\mathbf{x}^\star)\boldsymbol{\theta}, \sigma^2\right) \mathcal{N}\left(\boldsymbol{\theta} \mid \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right) \\ &= \mathcal{N}\left(y^\star \,\middle|\, \phi(\mathbf{x}^\star) \cdot \boldsymbol{\mu}_{\boldsymbol{\theta}}, \sigma^2 + \phi(\mathbf{x}^\star)^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \phi(\mathbf{x}^\star)\right) \end{aligned}$$

- Key:
  - prediction is again Gaussian
  - Predictive variance is increase due to the posterior uncertainty in $\boldsymbol{\theta}$.

Making predictions

- Prediction for fixed weight $\hat{\boldsymbol{\theta}}$ at input $\mathbf{x}^\star$ trivial:

$$p(y^\star \mid \mathbf{x}^\star, \hat{\boldsymbol{\theta}}, \sigma^2) = \mathcal{N}\left(y^\star \,\middle|\, \phi(\mathbf{x}^\star)\hat{\boldsymbol{\theta}}, \sigma^2\right)$$

- Integrate over $\boldsymbol{\theta}$ to take the posterior uncertainty into account

$$\begin{aligned} p(y^\star \mid \mathbf{x}^\star, \mathcal{D}) &= \int_{\boldsymbol{\theta}} p(y^\star \mid \mathbf{x}^\star, \boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y}, \sigma^2) \\ &= \int_{\boldsymbol{\theta}} \mathcal{N}\left(y^\star \mid \phi(\mathbf{x}^\star)\boldsymbol{\theta}, \sigma^2\right) \mathcal{N}\left(\boldsymbol{\theta} \mid \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right) \\ &= \mathcal{N}\left(y^\star \,\middle|\, \phi(\mathbf{x}^\star) \cdot \boldsymbol{\mu}_{\boldsymbol{\theta}}, \sigma^2 + \phi(\mathbf{x}^\star)^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \phi(\mathbf{x}^\star)\right) \end{aligned}$$

- Key:
  - prediction is again Gaussian
  - Predictive variance is increase due to the posterior uncertainty in θ.

Making predictions

- Prediction for fixed weight $\hat{\boldsymbol{\theta}}$ at input $\mathbf{x}^\star$ trivial:

$$p(y^\star \,|\, \mathbf{x}^\star, \hat{\boldsymbol{\theta}}, \sigma^2) = \mathcal{N}\left(y^\star \,\middle|\, \boldsymbol{\phi}(\mathbf{x}^\star)\hat{\boldsymbol{\theta}}, \sigma^2\right)$$

- Integrate over $\boldsymbol{\theta}$ to take the posterior uncertainty into account

$$\begin{aligned} p(y^\star \,|\, \mathbf{x}^\star, \mathcal{D}) &= \int_{\boldsymbol{\theta}} p(y^\star \,|\, \mathbf{x}^\star, \boldsymbol{\theta}, \sigma^2) p(\boldsymbol{\theta} \,|\, \mathbf{X}, \mathbf{y}, \sigma^2) \\ &= \int_{\boldsymbol{\theta}} \mathcal{N}\left(y^\star \,\middle|\, \boldsymbol{\phi}(\mathbf{x}^\star)\boldsymbol{\theta}, \sigma^2\right) \mathcal{N}\left(\boldsymbol{\theta} \,\middle|\, \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right) \\ &= \mathcal{N}\left(y^\star \,\middle|\, \boldsymbol{\phi}(\mathbf{x}^\star) \cdot \boldsymbol{\mu}_{\boldsymbol{\theta}}, \sigma^2 + \boldsymbol{\phi}(\mathbf{x}^\star)^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{\phi}(\mathbf{x}^\star)\right) \end{aligned}$$

- Key:
  - prediction is again Gaussian
  - Predictive variance is increase due to the posterior uncertainty in $\boldsymbol{\theta}$.

Outline

Model comparison

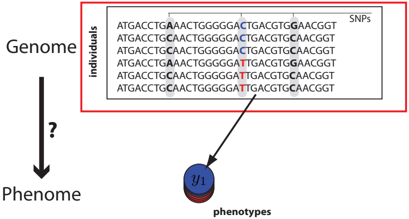
Motivation

- What degree of polynomials describes the data best?
- Is the linear model at all appropriate?
- Association testing.

Model comparison

Motivation

- What degree of polynomials describes the data best?
- Is the linear model at all appropriate?
- Association testing.



Bayesian model comparison

- How do we choose among alternative models?
- Assume we want to choose among models $\mathcal{H}_0, \ldots, \mathcal{H}_M$ for a dataset $\mathcal{D}$.
- Posterior probability for a particular model i

$$p(\mathcal{H}_i \,|\, \mathcal{D}) \propto \underbrace{p(\mathcal{D} \,|\, \mathcal{H}_i)}_{\text{Evidence}} \underbrace{p(\mathcal{H}_i)}_{\text{Prior}}$$

Bayesian model comparison

- How do we choose among alternative models?
- Assume we want to choose among models $\mathcal{H}_0, \ldots, \mathcal{H}_M$ for a dataset $\mathcal{D}$.
- Posterior probability for a particular model i

$$p(\mathcal{H}_i \,|\, \mathcal{D}) \propto \underbrace{p(\mathcal{D} \,|\, \mathcal{H}_i)}_{\text{Evidence}} \underbrace{p(\mathcal{H}_i)}_{\text{Prior}}$$

Bayesian model comparison

How to calculate the evidence

- The evidence is not the model likelihood!

$$p(\mathcal{D} \,|\, \mathcal{H}_i) = \int_{\boldsymbol{\Theta}} \mathrm{d}\boldsymbol{\Theta} p(\mathcal{D} \,|\, \boldsymbol{\Theta}) p(\boldsymbol{\Theta}) \text{ for model parameters } \boldsymbol{\Theta}.$$

- Remember:

$$p(\boldsymbol{\Theta} \,|\, \mathcal{H}_i, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \mathcal{H}_i, \boldsymbol{\Theta}) p(\boldsymbol{\Theta})}{p(\mathcal{D} \,|\, \mathcal{H}_i)}$$

Bayesian model comparison

How to calculate the evidence

- The evidence is not the model likelihood!

$$p(\mathcal{D} \,|\, \mathcal{H}_i) = \int_{\boldsymbol{\Theta}} \mathrm{d}\boldsymbol{\Theta} p(\mathcal{D} \,|\, \boldsymbol{\Theta}) p(\boldsymbol{\Theta}) \text{ for model parameters } \boldsymbol{\Theta}.$$

- Remember:

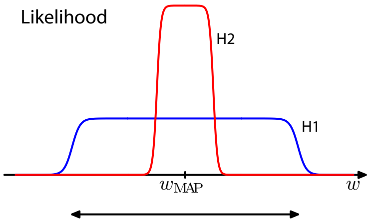
$$p(\boldsymbol{\Theta} \,|\, \mathcal{H}_i, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \mathcal{H}_i, \boldsymbol{\Theta}) p(\boldsymbol{\Theta})}{p(\mathcal{D} \,|\, \mathcal{H}_i)}$$

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{Evidence}}$$

Bayesian model comparison

Ocam's razor

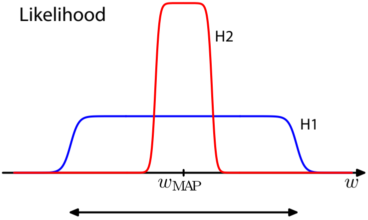
- The evidence integral penalizes overly complex models.
- A model with few parameters and lower maximum likelihood ($\mathcal{H}_1$) may win over a model with a peaked likelihood that requires many more parameters ($\mathcal{H}_2$).



(C.M. Bishop, Pattern Recognition and Machine Learning)

Bayesian model comparison
Ocam's razor

- The evidence integral penalizes overly complex models.
- A model with few parameters and lower maximum likelihood ($\mathcal{H}_1$) may win over a model with a peaked likelihood that requires many more parameters ($\mathcal{H}_2$).



(C.M. Bishop, Pattern Recognition and Machine Learning)

Application to GWA

Relevance of a single SNP

- Consider an association study.
  - $\mathcal{H}_0$: no association

$$p(\mathbf{y} \,|\, \mathcal{H}_0, \mathbf{X}, \boldsymbol{\Theta}_0) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{0}, \sigma^2\mathbf{I}\right)$$
$$p(\mathcal{D} \,|\, \mathcal{H}_0) = \int_{\sigma^2} \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{0}, \sigma^2\mathbf{I}\right) p(\sigma^2)$$

  - $\mathcal{H}_1$: linear association

$$p(\mathbf{y} \,|\, \mathcal{H}_1, \mathbf{x}_i, \boldsymbol{\Theta}_1) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{x}_i \cdot \theta, \sigma^2\mathbf{I}\right)$$
$$p(\mathcal{D} \,|\, \mathcal{H}_1) = \int_{\sigma^2, \theta} \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{x}_i \cdot \theta, \sigma^2\mathbf{I}\right) p(\sigma^2) p(\theta)$$

- Depending on the choice of priors, $p(\sigma^2)$ and $p(\theta)$, the required integrals are often tractable in closed form.

Application to GWA

Relevance of a single SNP

- Consider an association study.
  - $\mathcal{H}_0$: no association

$$p(\mathbf{y} \,|\, \mathcal{H}_0, \mathbf{X}, \mathbf{\Theta}_0) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{0}, \sigma^2\mathbf{I}\right)$$
$$p(\mathcal{D} \,|\, \mathcal{H}_0) = \int_{\sigma^2} \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{0}, \sigma^2\mathbf{I}\right) p(\sigma^2)$$

  - $\mathcal{H}_1$: linear association

$$p(\mathbf{y} \,|\, \mathcal{H}_1, \mathbf{x}_i, \mathbf{\Theta}_1) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{x}_i \cdot \theta, \sigma^2\mathbf{I}\right)$$
$$p(\mathcal{D} \,|\, \mathcal{H}_1) = \int_{\sigma^2, \theta} \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{x}_i \cdot \theta, \sigma^2\mathbf{I}\right) p(\sigma^2) p(\theta)$$

- Depending on the choice of priors, $p(\sigma^2)$ and $p(\theta)$, the required integrals are often tractable in closed form.

Application to GWA

Relevance of a single SNP

- Consider an association study.
  - $\mathcal{H}_0$: no association

$$p(\mathbf{y} \,|\, \mathcal{H}_0, \mathbf{X}, \boldsymbol{\Theta}_0) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{0}, \sigma^2\mathbf{I}\right)$$
$$p(\mathcal{D} \,|\, \mathcal{H}_0) = \int_{\sigma^2} \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{0}, \sigma^2\mathbf{I}\right) p(\sigma^2)$$

  - $\mathcal{H}_1$: linear association

$$p(\mathbf{y} \,|\, \mathcal{H}_1, \mathbf{x}_i, \boldsymbol{\Theta}_1) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{x}_i \cdot \theta, \sigma^2\mathbf{I}\right)$$
$$p(\mathcal{D} \,|\, \mathcal{H}_1) = \int_{\sigma^2, \theta} \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{x}_i \cdot \theta, \sigma^2\mathbf{I}\right) p(\sigma^2) p(\theta)$$

- Depending on the choice of priors, $p(\sigma^2)$ and $p(\theta)$, the required integrals are often tractable in closed form.

Application to GWA
Scoring models

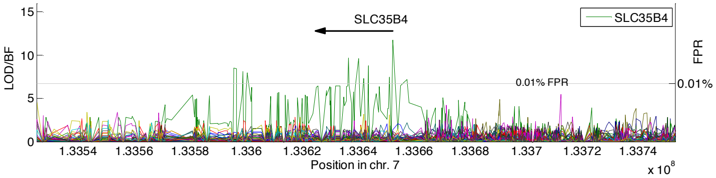
- Similar to likelihood ratios, the ratio of the evidences, the Bayes factor can be used to score alternative models:

$$BF = \ln \frac{p(\mathcal{D} \,|\, \mathcal{H}_1)}{p(\mathcal{D} \,|\, \mathcal{H}_0)}.$$

Application to GWA
Scoring models

- Similar to likelihood ratios, the ratio of the evidences, the Bayes factor can be used to score alternative models:

$$BF = \ln \frac{p(\mathcal{D} \,|\, \mathcal{H}_1)}{p(\mathcal{D} \,|\, \mathcal{H}_0)}.$$



Application to GWA

Posterior probability of an association

- Bayes factors are useful, however we would like a probabilistic answer how certain an association really is.
- Posterior probability of $\mathcal{H}_1$

$$p(\mathcal{H}_1 \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{H}_1)p(\mathcal{H}_1)}{p(\mathcal{D})}$$

$$= \frac{p(\mathcal{D} \mid \mathcal{H}_1)p(\mathcal{H}_1)}{p(\mathcal{D} \mid \mathcal{H}_1)p(\mathcal{H}_1) + p(\mathcal{D} \mid \mathcal{H}_0)p(\mathcal{H}_0)}$$

- $p(\mathcal{H}_1 \mid \mathcal{D}) + p(\mathcal{H}_0 \mid \mathcal{D}) = 1$, prior probability of observing a real association.

Application to GWA

Posterior probability of an association

- Bayes factors are useful, however we would like a probabilistic answer how certain an association really is.
- Posterior probability of $\mathcal{H}_1$

$$p(\mathcal{H}_1 \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \mathcal{H}_1) p(\mathcal{H}_1)}{p(\mathcal{D})}$$
$$= \frac{p(\mathcal{D} \,|\, \mathcal{H}_1) p(\mathcal{H}_1)}{p(\mathcal{D} \,|\, \mathcal{H}_1) p(\mathcal{H}_1) + p(\mathcal{D} \,|\, \mathcal{H}_0) p(\mathcal{H}_0)}$$

- $p(\mathcal{H}_1 \,|\, \mathcal{D}) + p(\mathcal{H}_0 \,|\, \mathcal{D}) = 1$, prior probability of observing a real association.

Application to GWA
Posterior probability of an association

- Bayes factors are useful, however we would like a probabilistic answer how certain an association really is.
- Posterior probability of $\mathcal{H}_1$

$$p(\mathcal{H}_1 \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{H}_1) p(\mathcal{H}_1)}{p(\mathcal{D})}$$
$$= \frac{p(\mathcal{D} \mid \mathcal{H}_1) p(\mathcal{H}_1)}{p(\mathcal{D} \mid \mathcal{H}_1) p(\mathcal{H}_1) + p(\mathcal{D} \mid \mathcal{H}_0) p(\mathcal{H}_0)}$$

- $p(\mathcal{H}_1 \mid \mathcal{D}) + p(\mathcal{H}_0 \mid \mathcal{D}) = 1$, prior probability of observing a real association.

Bayes factor verus likelihood ratio

Bayes factor

- Models of different complexity can be objectively compared.
- Statistical significance as posterior probability of a model.
- Typically hard to compute.

Likelihood ratio

- Likelihood ratio scales with the number of parameters.
- Likelihood ratios have known null distribution, yielding p-values.
- Often easy to compute.

Bayes factor verus likelihood ratio

Bayes factor

- Models of different complexity can be objectively compared.
- Statistical significance as posterior probability of a model.
- Typically hard to compute.

Likelihood ratio

- Likelihood ratio scales with the number of parameters.
- Likelihood ratios have known null distribution, yielding p-values.
- Often easy to compute.

Marginal likelihood of variance component models

- Consider a linear model, accounting for a set of measured SNPs $\mathbf{X}$

$$p(\mathbf{y}\,|\,\mathbf{X},\boldsymbol{\theta},\sigma^2) = \mathcal{N}\left(\mathbf{y}\,\middle|\,\sum_{s=1}^{S}\mathbf{x}_s\theta_s,\sigma^2\mathbf{I}\right)$$

- Choose identical Gaussian prior for all weights

$$p(\boldsymbol{\theta}) = \prod_{s=1}^{S}\mathcal{N}\left(\theta_s\,\middle|\,0,\sigma_g^2\right)$$

- Marginal likelihood

$$p(\mathbf{y}\,|\,\mathbf{X},) = \int_{\boldsymbol{\theta}}\mathcal{N}\left(\mathbf{y}\,|\,\mathbf{X}\boldsymbol{\theta},\sigma^2\mathbf{I}\right)\mathcal{N}\left(\boldsymbol{\theta}\,|\,\mathbf{0},\sigma_g^2\mathbf{I}\right)$$

- Number of hyperparameters independent of number of SNPs

Marginal likelihood of variance component models

- Consider a linear model, accounting for a set of measured SNPs $\mathbf{X}$

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma^2) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \sum_{s=1}^{S} \mathbf{x}_s \theta_s, \sigma^2 \mathbf{I}\right)$$

- Choose identical Gaussian prior for all weights

$$p(\boldsymbol{\theta}) = \prod_{s=1}^{S} \mathcal{N}\left(\theta_s \,\middle|\, 0, \sigma_g^2\right)$$

- Marginal likelihood

$$p(\mathbf{y} \mid \mathbf{X},) = \int_{\boldsymbol{\theta}} \mathcal{N}\left(\mathbf{y} \,\middle|\, \mathbf{X}\boldsymbol{\theta}, \sigma^2 \mathbf{I}\right) \mathcal{N}\left(\boldsymbol{\theta} \,\middle|\, \mathbf{0}, \sigma_g^2 \mathbf{I}\right)$$

- Number of hyperparameters independent of number of SNPs

Marginal likelihood of variance component models

- Consider a linear model, accounting for a set of measured SNPs $\mathbf{X}$

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma^2) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \sum_{s=1}^{S} \mathbf{x}_s \theta_s, \sigma^2 \mathbf{I}\right)$$

- Choose identical Gaussian prior for all weights

$$p(\boldsymbol{\theta}) = \prod_{s=1}^{S} \mathcal{N}\left(\theta_s \,\middle|\, 0, \sigma_g^2\right)$$

- Marginal likelihood

$$p(\mathbf{y} \mid \mathbf{X}, \sigma^2, \sigma_g^2) = \int_{\boldsymbol{\theta}} \mathcal{N}\left(\mathbf{y} \mid \mathbf{X}\boldsymbol{\theta}, \sigma^2 \mathbf{I}\right) \mathcal{N}\left(\boldsymbol{\theta} \mid \mathbf{0}, \sigma_g^2 \mathbf{I}\right)$$
$$= \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma_g^2 \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma^2 \mathbf{I}\right)$$

- Number of hyperparameters independent of number of SNPs

Marginal likelihood of variance component models

- Consider a linear model, accounting for a set of measured SNPs $\mathbf{X}$

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma^2) = \mathcal{N}\left(\mathbf{y} \,\middle|\, \sum_{s=1}^{S} \mathbf{x}_s \theta_s, \sigma^2 \mathbf{I}\right)$$

- Choose identical Gaussian prior for all weights

$$p(\boldsymbol{\theta}) = \prod_{s=1}^{S} \mathcal{N}\left(\theta_s \mid 0, \sigma_g^2\right)$$

- Marginal likelihood

$$p(\mathbf{y} \mid \mathbf{X}, \sigma^2, \sigma_g^2) = \int_{\boldsymbol{\theta}} \mathcal{N}\left(\mathbf{y} \mid \mathbf{X}\boldsymbol{\theta}, \sigma^2 \mathbf{I}\right) \mathcal{N}\left(\boldsymbol{\theta} \mid \mathbf{0}, \sigma_g^2 \mathbf{I}\right)$$
$$= \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma_g^2 \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma^2 \mathbf{I}\right)$$

- Number of hyperparameters independent of number of SNPs

Marginal likelihood of variance component models
Application to GWAs

The missing heritability paradox

- Complex traits are regulated by a large number of small effects
  - Human height: the best single SNP explains little variance.
  - But: the parents are highly predictive for the height of the child!

Marginal likelihood of variance component models

Application to GWAs

Multivariate additive models for complex traits

- Multivariate model over causal SNPs

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma^2) = \mathcal{N}\Big(\mathbf{y} \mid \sum_{s \in \text{causal}} \mathbf{x}_s \theta_s, \sigma^2 \mathbf{I}\Big)$$

- Common variance prior for causal SNPs $p(\theta_s) = \mathcal{N}\left(\theta_s \,\middle|\, 0, \sigma_g^2\right)$
- Marinalize out weights

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_g^2, \sigma_e^2) = \mathcal{N}\Big(\mathbf{y} \mid \mathbf{0}, \sigma_g^2 \sum_{s \in \text{causal}} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\Big)$$

- Which SNPs are causal ?
  Approximation: consider all SNPs [Yang et al., 2011]

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_g^2, \sigma_e^2) = \mathcal{N}\Big(\mathbf{y} \mid \mathbf{0}, \sigma_g^2 \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\Big)$$

Marginal likelihood of variance component models

Application to GWAs

Multivariate additive models for complex traits

- Multivariate model over causal SNPs

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma^2) = \mathcal{N}\left(\mathbf{y} \mid \sum_{s \in \text{causal}} \mathbf{x}_s \theta_s, \sigma^2 \mathbf{I}\right)$$

- Common variance prior for causal SNPs $p(\theta_s) = \mathcal{N}\left(\theta_s \,\middle|\, 0, \sigma_g^2\right)$
- Marinalize out weights

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_g^2, \sigma_e^2) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma_g^2 \sum_{s \in \text{causal}} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\right)$$

- Which SNPs are causal ?
  Approximation: consider all SNPs [Yang et al., 2011]

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_g^2, \sigma_e^2) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma_g^2 \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\right)$$

Marginal likelihood of variance component models

Application to GWAs

Multivariate additive models for complex traits

- Multivariate model over causal SNPs

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma^2) = \mathcal{N}\Big(\mathbf{y} \mid \sum_{s \in \text{causal}} \mathbf{x}_s \theta_s, \sigma^2 \mathbf{I}\Big)$$

- Common variance prior for causal SNPs $p(\theta_s) = \mathcal{N}\left(\theta_s \,\middle|\, 0, \sigma_g^2\right)$
- Marinalize out weights

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_g^2, \sigma_e^2) = \mathcal{N}\Big(\mathbf{y} \mid \mathbf{0}, \sigma_g^2 \sum_{s \in \text{causal}} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\Big)$$

- Which SNPs are causal ?
  Approximation: consider all SNPs [Yang et al., 2011]

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_g^2, \sigma_e^2) = \mathcal{N}\big(\mathbf{y} \mid \mathbf{0}, \sigma_g^2 \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\big)$$

Marginal likelihood of variance component models

Application to GWAs

Multivariate additive models for complex traits

- Multivariate model over causal SNPs

$$p(\mathbf{y} \,|\, \mathbf{X}, \boldsymbol{\theta}, \sigma^2) = \mathcal{N}\Big(\mathbf{y} \,|\, \sum_{s \in \text{causal}} \mathbf{x}_s \theta_s, \sigma^2 \mathbf{I}\Big)$$

- Common variance prior for causal SNPs $p(\theta_s) = \mathcal{N}\left(\theta_s \,\middle|\, 0, \sigma_g^2\right)$
- Marinalize out weights

$$p(\mathbf{y} \,|\, \mathbf{X}, \sigma_g^2, \sigma_e^2) = \mathcal{N}\Big(\mathbf{y} \,|\, \mathbf{0}, \sigma_g^2 \sum_{s \in \text{causal}} \mathbf{x}_s \mathbf{x}_s^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\Big)$$

- Which SNPs are causal ?
  Approximation: consider all SNPs [Yang et al., 2011]

$$p(\mathbf{y} \,|\, \mathbf{X}, \sigma_g^2, \sigma_e^2) = \mathcal{N}\Big(\mathbf{y} \,|\, \mathbf{0}, \sigma_g^2 \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\Big)$$

Marginal likelihood of variance component models

Application to GWAs

- Approximate variance model

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_g^2, \sigma_e^2) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma_g^2 \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\right)$$

- Genetic variance σ_g^2 across chromosomes
- Heritability $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$

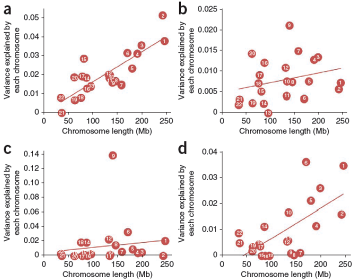
Marginal likelihood of variance component models

Application to GWAs

- Approximate variance model

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_g^2, \sigma_e^2) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma_g^2 \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\right)$$

- Genetic variance σ_g^2 across chromosomes
- Heritability $h^2 = \dfrac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$



[Yang et al., 2011]

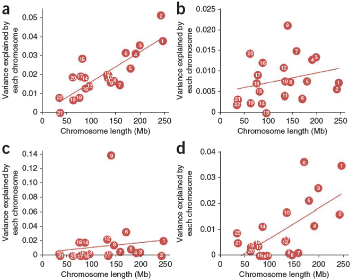
Marginal likelihood of variance component models

Application to GWAs

- Approximate variance model

$$p(\mathbf{y} \mid \mathbf{X}, \sigma_g^2, \sigma_e^2) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{0}, \sigma_g^2 \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma_e^2 \mathbf{I}\right)$$

- Genetic variance σ_g^2 across chromosomes
- Heritability $h^2 = \dfrac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$



[Yang et al., 2011]

Outline

Summary

- Generalized linear models for Curve fitting and multivariate regression.
- Maximum likelihood and least squares regression are identical.
- Construction of features using a mapping ϕ.
- Regularized least squares and other models that correspond to different choices of loss functions.
- Bayesian linear regression.
- Model comparison and ocam's razor.
- Variance component models in GWAs.

Tasks

- Prove that the product of two Gaussians is Gaussian distributed.
- Try to understand the convolution formula of Gaussian random variables.

References I

C. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

S. Roweis. Gaussian identities. *technical report*, 1999. URL `http://www.cs.nyu.edu/~roweis/notes/gaussid.pdf`.

J. Yang, T. Manolio, L. Pasquale, E. Boerwinkle, N. Caporaso, J. Cunningham, M. de Andrade, B. Feenstra, E. Feingold, M. Hayes, et al. Genome partitioning of genetic variation for complex traits using common snps. *Nature Genetics*, 43(6):519–525, 2011.