Approximating Message Lengths of Hierarchical Bayesian Models Using Posterior Sampling

Daniel F. Schmidt, Enes Makalic and John L. Hopper

Centre for Biostatistics and Epidemiology
The University of Melbourne

29th Australasian Joint Conference on Artificial Intelligence
Hobart, Tasmania
December 8, 2016

Outline

1. Problem Description
   - Hierarchical Bayesian Models
   - Minimum Message Length Inference
2. MML for Hierarchical Models
   - The MML-h Approximation for Hierarchical Models
3. Application: Dense or Sparse?
   - Problem Statement
   - Results

Outline

1. Problem Description
   - Hierarchical Bayesian Models
   - Minimum Message Length Inference
2. MML for Hierarchical Models
   - The MML-h Approximation for Hierarchical Models
3. Application: Dense or Sparse?
   - Problem Statement
   - Results

Hierarchical Bayesian Models (1)

- We have
  - A vector of data $\mathbf{y}^n = (y_1, \ldots, y_n)$
  - A statistical model $p(\mathbf{y}^n|\boldsymbol{\theta})$
  - A prior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha_q})$
- We can write the complete model as

$$\begin{aligned}
\mathbf{y}^n|\boldsymbol{\theta} &\sim p(\mathbf{y}^n|\boldsymbol{\theta})d\mathbf{y}^n, \\
\boldsymbol{\theta}|\boldsymbol{\alpha}_1 &\sim \pi(\boldsymbol{\theta}|\boldsymbol{\alpha}_1)d\boldsymbol{\theta}, \\
\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_2 &\sim \pi(\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_2)d\boldsymbol{\alpha}_1, \\
\boldsymbol{\alpha}_2|\boldsymbol{\alpha}_3 &\sim \pi(\boldsymbol{\alpha}_2|\boldsymbol{\alpha}_3)d\boldsymbol{\alpha}_2, \\
&\ldots \\
\boldsymbol{\alpha}_q &\sim \pi(\boldsymbol{\alpha}_q)d\boldsymbol{\alpha}_q,
\end{aligned}$$

$\Rightarrow$ The $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q$ are called hyperparameters

Hierarchical Bayesian Models (1)

- We have
  - A vector of data $\mathbf{y}^n = (y_1, \ldots, y_n)$
  - A statistical model $p(\mathbf{y}^n|\boldsymbol{\theta})$
  - A prior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha_q})$
- We can write the complete model as

$$\begin{aligned}
\mathbf{y}^n|\boldsymbol{\theta} &\sim p(\mathbf{y}^n|\boldsymbol{\theta})d\mathbf{y}^n, \\
\boldsymbol{\theta}|\boldsymbol{\alpha}_1 &\sim \pi(\boldsymbol{\theta}|\boldsymbol{\alpha}_1)d\boldsymbol{\theta}, \\
\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_2 &\sim \pi(\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_2)d\boldsymbol{\alpha}_1, \\
\boldsymbol{\alpha}_2|\boldsymbol{\alpha}_3 &\sim \pi(\boldsymbol{\alpha}_2|\boldsymbol{\alpha}_3)d\boldsymbol{\alpha}_2, \\
&\ldots \\
\boldsymbol{\alpha}_q &\sim \pi(\boldsymbol{\alpha}_q)d\boldsymbol{\alpha}_q,
\end{aligned}$$

$\Rightarrow$ The $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha_q}$ are called hyperparameters

Hierarchical Bayesian Models (2)

- The choice of statistical model controls distribution of data.
- Important example is the Gaussian linear regression model

$$y_j|\mathbf{x}_j \sim N(\beta_0 + \mathbf{x}_j'\boldsymbol{\beta}, \sigma^2)$$

where

  - $\boldsymbol{\beta}$ are regression coefficients
  - β_0 is the intercept
  - $\mathbf{x}_j$ are predictors
  - σ^2 is the noise variance

- Other common models:
  - Logistic regression for binary data (classifier)
  - Poisson regression for count data, etc.

Hierarchical Bayesian Models (3)

- Prior hierarchy controls beliefs about the model
- Example 1:

$$\begin{aligned} \beta_j|\tau &\sim N(0,\tau), \\ \tau &\sim \pi(\tau)d\tau \end{aligned}$$

leads to Bayesian ridge regression
$\Rightarrow$ expect β_j to be same squared magnitude (β dense)

- Example 2:

$$\begin{aligned} \beta_j|\tau &\sim La(0,\tau), \\ \tau &\sim \pi(\tau)d\tau \end{aligned}$$

leads to Bayesian Lasso ($La(\cdot)$ is Laplace distribution)
$\Rightarrow$ expect β_j's to be same absolute magnitude (β less dense)

Hierarchical Bayesian Models (3)

- Prior hierarchy controls beliefs about the model
- Example 1:

$$\begin{aligned} \beta_j|\tau &\sim N(0,\tau), \\ \tau &\sim \pi(\tau)d\tau \end{aligned}$$

leads to Bayesian ridge regression
$\Rightarrow$ expect β_j to be same squared magnitude ($\boldsymbol{\beta}$ dense)

- Example 2:

$$\begin{aligned} \beta_j|\tau &\sim La(0,\tau), \\ \tau &\sim \pi(\tau)d\tau \end{aligned}$$

leads to Bayesian Lasso ($La(\cdot)$ is Laplace distribution)
$\Rightarrow$ expect β_j's to be same absolute magnitude ($\boldsymbol{\beta}$ less dense)

Hierarchical Bayesian Models (3)

- Prior hierarchy controls beliefs about the model
- Example 1:

$$\begin{aligned} \beta_j|\tau &\sim N(0,\tau), \\ \tau &\sim \pi(\tau)d\tau \end{aligned}$$

leads to Bayesian ridge regression
$\Rightarrow$ expect β_j to be same squared magnitude ($\boldsymbol{\beta}$ dense)

- Example 2:

$$\begin{aligned} \beta_j|\tau &\sim La(0,\tau), \\ \tau &\sim \pi(\tau)d\tau \end{aligned}$$

leads to Bayesian Lasso ($La(\cdot)$ is Laplace distribution)
$\Rightarrow$ expect β_j's to be same absolute magnitude ($\boldsymbol{\beta}$ less dense)

Model Selection for Hierarchical Bayesian Models

- Choice of model and prior have profound effect on estimator
- Standard Bayesian method to choose model is to maximise

$$p(\mathbf{y}^n) = \int p(\mathbf{y}^n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q) d\boldsymbol{\theta} d\boldsymbol{\alpha}_1 \ldots d\boldsymbol{\alpha}_q$$

  $\Rightarrow$ high dimensional integration; difficult problem
- Instead, usually easier to sample from posterior distribution

$$p(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q|\mathbf{y}^n)$$

  - Algorithms exist to compute $p(\mathbf{y}^n)$ from samples
  - These methods either numerically unstable, not suitable for hierarchical models or require multiple sampling passes

Minimum Message Length (1)

- Alternative: minimum message length (MML) model selection
- Practical implementation of theory of inductive inference inspired by Kolmogorov complexity
  - Model that yields the briefest encoding of data in a hypothetical message is optimal
- The message is composed of two-parts
  - assertion, statement describing a particular model $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$
  - detail, encoding of the data $\mathbf{y}$ using the assertion model $\boldsymbol{\theta}$
- Existing MML approximations do not handle general hierarchical models

Minimum Message Length (2)

- Standard MML approximation is Wallace-Freeman (MML87)
- Assigns a score to a model of form:

$$I(\mathbf{y}^n, \boldsymbol{\theta}) = -\log p(\mathbf{y}^n|\boldsymbol{\theta}) + \frac{1}{2}\log|\mathbf{J}(\boldsymbol{\theta})| - \log \pi(\boldsymbol{\theta}) + c(k)$$

  where
  - $\mathbf{J}(\boldsymbol{\theta})$ is the Fisher information matrix,
  - $c(k)$ is a dimensionality constant.

- MML87 has a number of strong properties
  $\Rightarrow$ Aim: extend MML87 to handle hierarchical models

Outline

1. Problem Description
   - Hierarchical Bayesian Models
   - Minimum Message Length Inference
2. MML for Hierarchical Models
   - The MML-h Approximation for Hierarchical Models
3. Application: Dense or Sparse?
   - Problem Statement
   - Results

Fisher Information and Parameter Uncertainty

- The Fisher information matrix has an important interpretation
  - If $\hat{\boldsymbol{\theta}}(\mathbf{y}^n)$ is an unbiased estimator of $\boldsymbol{\theta}$, then

$$|\mathrm{Cov}(\hat{\boldsymbol{\theta}}(\mathbf{y}^n)| \geq |\mathbf{J}^{-1}(\boldsymbol{\theta})|$$

  $\Rightarrow$ MML87 penalty inversely proportional to estimator variance

- Fisher of parameters and hyperparameters is difficult to obtain
- Instead use inverse posterior covariance

$$\mathbf{J}(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q) \approx \mathrm{Cov}^{-1}(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q | \mathbf{y}^n)$$

The MML-h Approximation (1)

- The MML-h approximation is then given by

$$\begin{aligned} I_h(\mathbf{y}^n) &= \mathrm{E}\left[-\log p(\mathbf{y}^n|\boldsymbol{\theta}) - \log \pi(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q) \,|\, \mathbf{y}^n\right] \\ &\quad - \frac{1}{2}\log\left|\mathrm{Cov}(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q \,|\, \mathbf{y}^n)\right| + c(k) - \frac{k_{\boldsymbol{\theta}}}{2}, \end{aligned}$$

  where
  - $\mathrm{E}\left[\cdot\,|\mathbf{y}^n\right]$ denotes expectation w.r.t. the posterior distribution
  - k is the total number of parameters and hyperparameters
  - $k_{\boldsymbol{\theta}}$ is the number of parameters
- Evaluating expectations can be difficult

The MML-h Approximation (2)

- Can approximate them using samples from posterior; let

$$\theta^{(j)}, \boldsymbol{\alpha}_1^{(j)}, \ldots, \boldsymbol{\alpha}_q^{(j)}, \quad (j = 1, \ldots, m)$$

  denote m samples from the posterior distribution.
- Then the empirical MML-h approximation is then given by

$$\begin{aligned} I_h(\mathbf{y}^n) \quad \approx \quad & -\left(\frac{1}{m}\right) \sum_{j=1}^{m} \left[\log p(\mathbf{y}^n|\boldsymbol{\theta}^{(j)})\right] \\ & -\left(\frac{1}{m}\right) \sum_{j=1}^{m} \left[\log \pi\left(\boldsymbol{\theta}^{(j)}, \boldsymbol{\alpha}_1^{(j)}, \ldots, \boldsymbol{\alpha}_q^{(j)}\right)\right] \\ & -\frac{1}{2} \log |\mathrm{Cov}(\boldsymbol{\theta}^{(j)}, \boldsymbol{\alpha}_1^{(j)}, \ldots, \boldsymbol{\alpha}_q^{(j)})| + c(k) - \frac{k_{\boldsymbol{\theta}}}{2}, \end{aligned}$$

  $\Rightarrow$ given samples, very simple to apply!

Outline

1. Problem Description
   - Hierarchical Bayesian Models
   - Minimum Message Length Inference
2. MML for Hierarchical Models
   - The MML-h Approximation for Hierarchical Models
3. Application: Dense or Sparse?
   - Problem Statement
   - Results

Problem Statement (1)

- Consider regression model

$$y_j|\mathbf{x}_j \sim N(\beta_0 + \mathbf{x}_j'\boldsymbol{\beta}, \sigma^2)$$

- Usual problem is to estimate $\boldsymbol{\beta} \in \mathbb{R}^p$ from a sample $\mathbf{y}^n$
  - Choice of prior controls behaviour of Bayesian estimator
- Ridge regression

$$\begin{aligned} \beta_j &\sim N(0, \tau^2\sigma^2) \\ \tau &\sim C^+(0,1) \end{aligned}$$

where $C^+(0,1)$ is the half-Cauchy distribution.

  - Implies belief that $\mathrm{E}\left[\beta_j^2\right]/\sigma^2 = \tau$
  - If $\boldsymbol{\beta}$ is sparse (many $\beta_j = 0$), will have high variance

$\Rightarrow$ implies preference for dense models

Problem Statement (2)

- Horseshoe regression (sparse coefficients)

$$\begin{aligned} \beta_j &\sim N(0, \tau^2 \lambda_j^2 \sigma^2) \\ \lambda_j &\sim C^+(0,1) \\ \tau &\sim C^+(0,1) \end{aligned}$$

  - Implies belief that β_j can be very small or very large
  - If $\boldsymbol{\beta}$ is sparse, will have good predictive performance
  - If $\boldsymbol{\beta}$ is dense, can suffer from bias

  $\Rightarrow$ Ridge and horseshoe both good for different settings

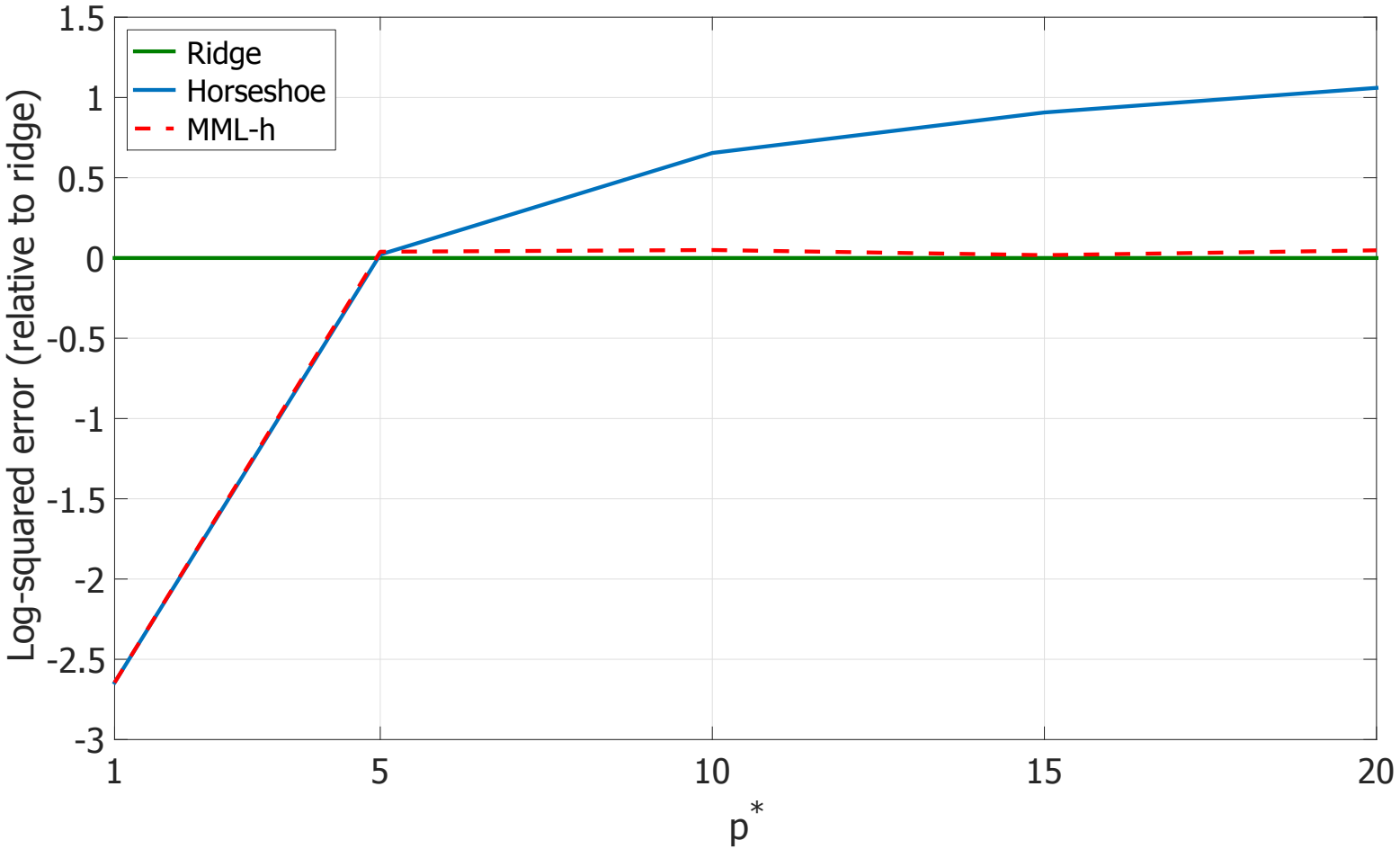
- Use MML-h to select between ridge and HS using data $\mathbf{y}^n$
  $\Rightarrow$ only model selection criterion currently for HS!

Experimental Procedure

- Generate $p = 20$ predictors
- Number of nonzero elements $p^* = \{1, 5, 10, 15, 20\}$ of $\boldsymbol{\beta}$
  $\Rightarrow$ covers varying levels of sparsity
- Generated the non-zero elements of $\boldsymbol{\beta}$ from:
  - Constant model, $\beta_j = 1$
  - Normal model, $\beta_j \sim N(0, 1)$
  - Cauchy model, $\beta_j \sim C(0, 1)$
- For each of 100 iterations
  - Generate data according to model with SNR $= 5$
  - Get posterior samples from ridge and horseshoe models
  - Use MML-h to select best hierarchy

Results (1)

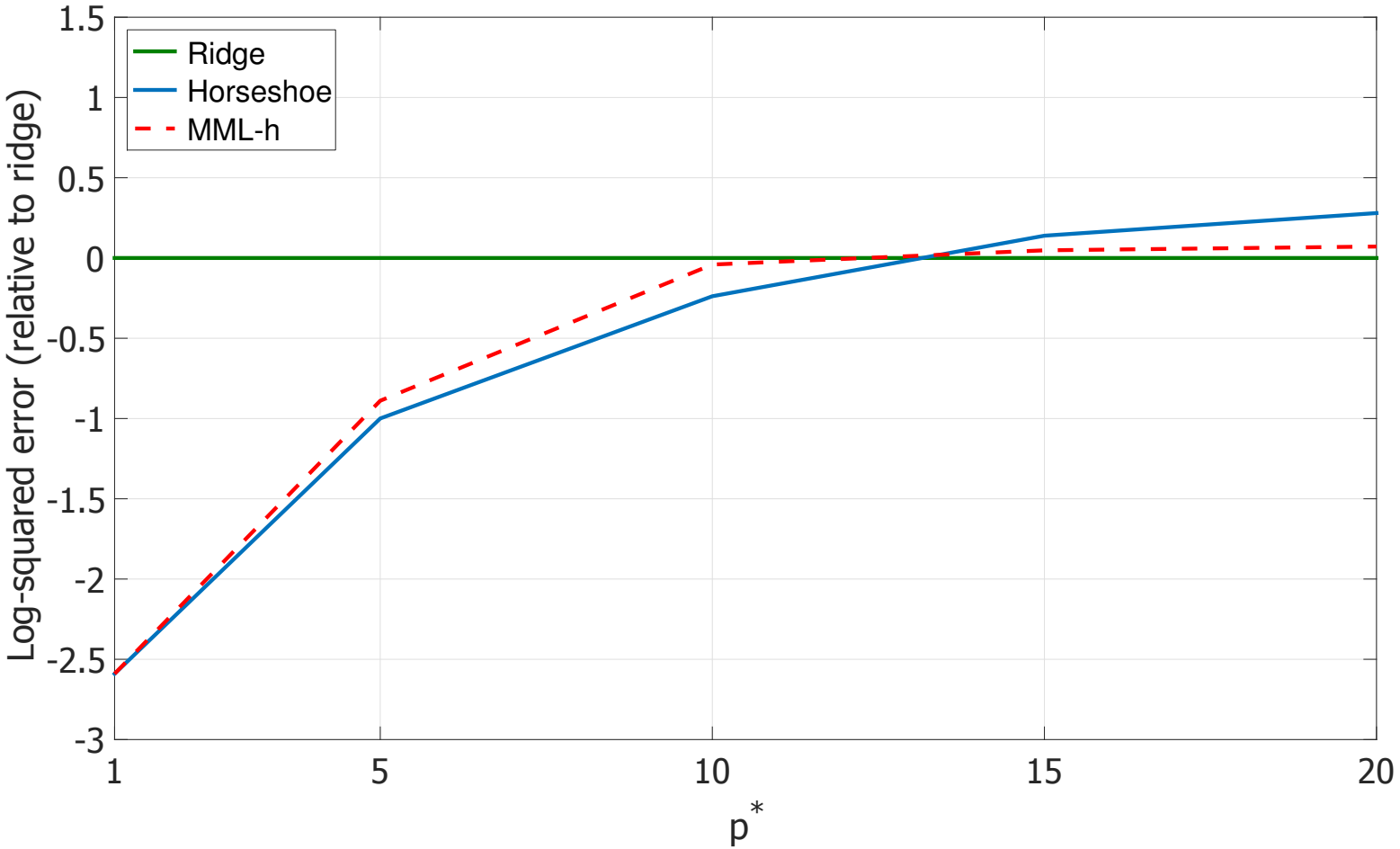
- Coefficient model: $\beta_j = 1$



- Squared errors are calculated relative to ridge regression

Results (2)

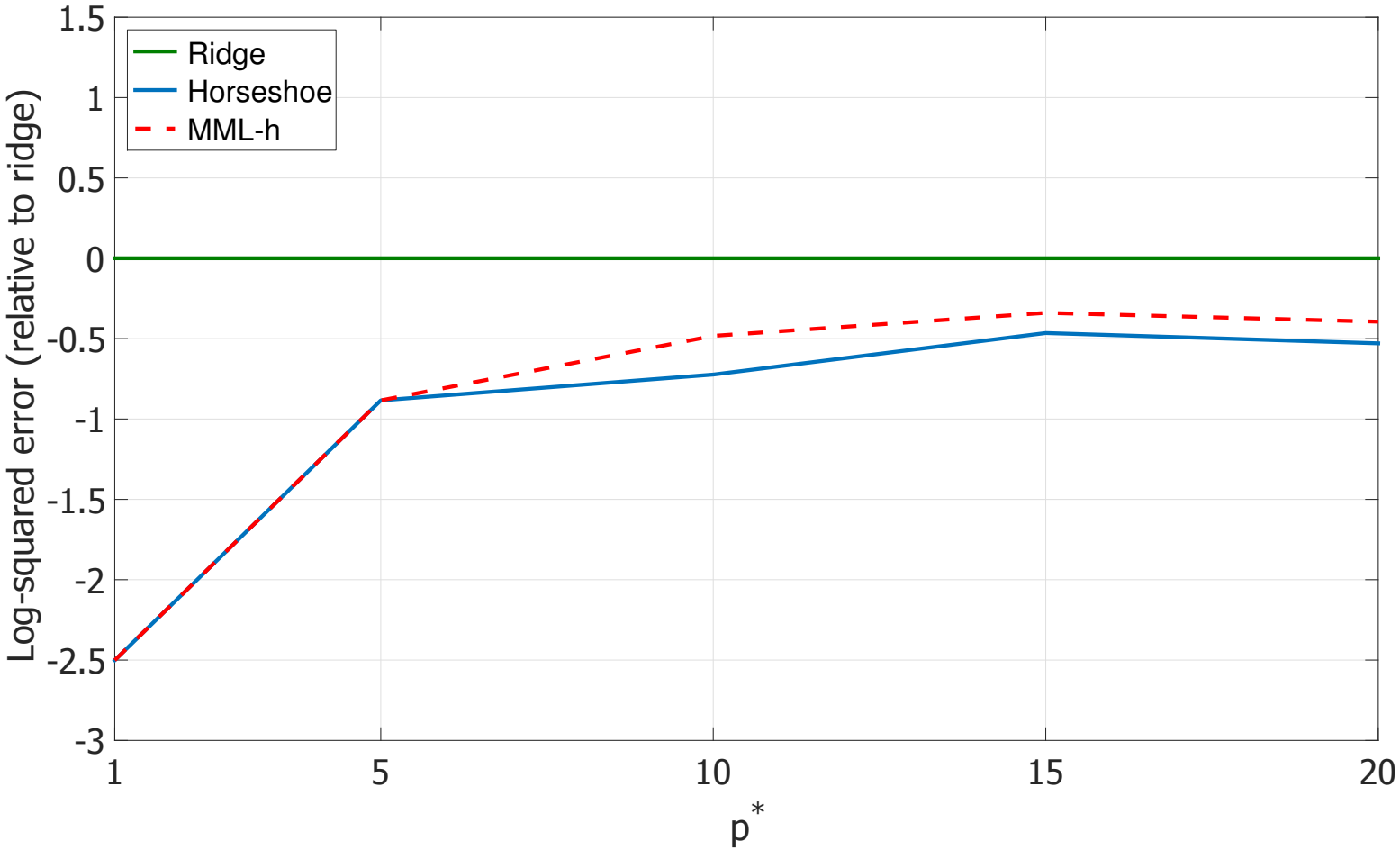
- Coefficient model: $\beta_j \sim N(0,1)$



- Squared errors are calculated relative to ridge regression

Results (3)

- Coefficient model: $\beta_j \sim C(0, 1)$



- Squared errors are calculated relative to ridge regression

Conclusion

- The MML-h approximation provides a simple approach to model selection in complex hierarchical Bayesian models
- Improved Bayesian regression by selecting priors
  $\Rightarrow$ showed very promising performance

- MATLAB software for Bayesian regression available from:

  http://au.mathworks.com/matlabcentral/fileexchange/60335-bayesian-regularized-linear-and-logistic-regression

- R package available as package "bayesreg" from CRAN
- Paper describing the bayesreg package available at:

  https://arxiv.org/pdf/1611.06649v1/

- Thank you – questions?