Approximating Message Lengths of Hierarchical Bayesian Models Using Posterior Sampling

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

The University of Melbourne
Centre for Epidemiology and Biostatistics
Carlton VIC 3053, Australia
{emakalic,dschmidt,j.hopper}@unimelb.edu.au

Abstract. Inference of complex hierarchical models is an increasingly common problem in modern Bayesian data analysis. Unfortunately, there are few computationally efficient and widely applicable methods for selecting between competing hierarchical models. In this paper we adapt ideas from the information theoretic minimum message length principle and propose a powerful yet simple model selection criteria for general hierarchical Bayesian models called MML-h. Computation of this criterion requires only that a set of samples from the posterior distribution be available. The flexibility of this new algorithm is demonstrated by a novel application to state-of-the-art Bayesian hierarchical regression estimation. Simulations show that the MML-h criterion is able to adaptively select between classic ridge regression and sparse horseshoe regression estimators, and the resulting procedure exhibits excellent robustness to the underlying structure of the regression coefficients.

1 Introduction

The Minimum Message Length (MML) [1] principle of inductive inference is a powerful, information theoretic, Bayesian framework for parameter estimation and model selection. The MML principle is based on the connection between statistical inference, algorithmic complexity and information theory, and has been extensively applied to a wide range of statistical and machine learning problems with great success. The basic idea is to quantify the fit of a model to an observed data string by the length of a decodable message, usually measured in *nits*, or base-e digits, that communicates both the model parameters as well as the length of the data string once compressed using the statistical properties of the model. This message length can be decomposed into two components:

$$I(\mathbf{y}^n, \boldsymbol{\theta}) = I(\boldsymbol{\theta}) + I(\mathbf{y}^n|\boldsymbol{\theta}),$$

where $I(\boldsymbol{\theta})$ is the length of the message required to describe the statistical model $\boldsymbol{\theta}$, and $I(\mathbf{y}^n|\boldsymbol{\theta})$ is the message length required to describe the data string $\mathbf{y}^n = (y_1, \ldots, y_n)$ using the properties of the nominated model $\boldsymbol{\theta}$. In this fashion, the message length naturally balances the ability of a model to fit a data string against the complexity of the model, and minimising the message length provides a natural framework for statistical inference. The MML principle is particularly attractive as both continuous model parameters, as well as discrete, structural parameters can be estimated by minimising the message length. A key idea in the MML principle is that estimates should be stated (coded) only to accuracy warranted by the data. This leads to estimators that are invariant under one-to-one reparameterisations, a property not shared by most conventional Bayesian point estimators, and gracefully handle many difficult statistical problems that cause significant issues for procedures such as maximum likelihood and Bayesian posterior mean and mode estimators.

Computing the exact message length is in general an NP problem; however, under suitable regularity conditions, the MML87 approximation [2] provides a simple formula to compute the message length of a statistical model $p(\mathbf{y}^n|\boldsymbol{\theta})$, characterised by continuous parameters $\boldsymbol{\theta}$ and associated prior distribution $\pi(\boldsymbol{\theta})$:

$$I_{87}(\mathbf{y}^n, \boldsymbol{\theta}) = -\log p(\mathbf{y}^n|\boldsymbol{\theta}) + \frac{1}{2}\log|\mathbf{J}(\boldsymbol{\theta})| - \log \pi(\boldsymbol{\theta}) + c(k), \tag{1}$$

where $p(\mathbf{y}^n|\boldsymbol{\theta})$ is the likelihood of the data $\mathbf{y}^n$, $\mathbf{J}(\boldsymbol{\theta})$ is the Fisher information matrix, k is number of free continuous model parameters and

$$c(k) = -\frac{k}{2}\log(2\pi) + \frac{1}{2}\log(k\pi) + \psi(1) \tag{2}$$

are appropriate dimensionality constants. While the MML87 approximation has been applied to derive MML estimators for a wide range of statistical models, there are large classes of problems to which it does not immediately apply, particularly those in which the prior distributions are highly peaked or those in which the Fisher information matrix can be (near) singular.

1.1 Bayesian Hierarchical Models

Multi-level Bayesian hierarchical models in which the prior distribution $\pi(\boldsymbol{\theta})$ is decomposed into a chain of priors, each potentially depending on further hyperparameters, are a class of models that cannot currently be easily handled within the MML framework. Hierarchical models are becoming increasingly common in machine learning because they can describe complex prior beliefs about the parameters $\boldsymbol{\theta}$, and can also be used as tools to generate estimators with desirable properties; examples of particular importance are the extensive number of Bayesian penalized regression procedures based on scale mixture priors such as the Bayesian horseshoe [3] which represent the state-of-the-art in penalized regression. The general Bayesian hierarchical model can be described by

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

where $(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_q)$ are the hyperparameters of the prior hierarchy. While the original MML87 formula (1) cannot be applied directly to hierarchical models, Makalic and Schmidt [4] extended the formula to this class of problems, and subsequently applied the new approximation to the problem of shrinkage estimation in Gaussian regression models. However, the resulting hierarchical message length approximation depends crucially on modifying the key formulas to account for the highly peaked (non-uniform) prior distributions that can arise when the hyperparameters are estimated from data, rather than being specified *a priori*. As discussed in Section 2, this is only really plausible for specific conjugate likelihood-prior pairings, which greatly restricts the applicability of the method to general hierarchical structures.

1.2 Sampling Approaches to Inference of Hierarchical Models

The standard Bayesian approach to handling complex hierarchical models is to sample from the posterior distribution

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

using Markov Chain Monte Carlo (MCMC) procedures such as Gibbs sampling. While such procedures allow the posterior distribution to be explored, they cannot directly be used to select between different model structures. Standard Bayesian model selection is usually based on the marginal probability of the data $\mathbf{y}^n$ under a given hierarchy, but this is notoriously difficult to compute. Chib's algorithm [5] provides a method for computing marginal probabilities directly from posterior samples, but it is slow, requiring multiple runs of the sampling chain. Further, it requires that all the conditional distributions be fully specified up to normalising constants. Given that this latter restriction is not required to efficiently sample from many posterior distributions, this represents a serious limitation. Other approaches include a proposal to use an empirical Laplace approximation of the posterior distribution [6], and an MML based approach called the message from Monte-Carlo formula [7], which appears to offer a way of efficiently computing message lengths from posterior samples. Unfortunately, it is not immediately clear how, or if, either of these procedures can be adapted to multi-level hierarchies.

What is required is a simple, computationally efficient procedure for performing Bayesian model selection of complex, hierarchical models using posterior samples drawn from a single run of a sampling chain. This paper proposes such a procedure based on the minimum message length principle.

1.3 Our Contribution: Model Selection of Hierarchical Models

Specifically, we propose a simple formula, called MML-h, which is motivated by the MML87 approximation and can be used to compute message lengths of complex Bayesian hierarchical models directly from a single set of posterior samples. In contrast to Chib's [5] algorithm, the method requires only that the *prior* distributions be exactly specified, a condition which is generally satisfied, and which dramatically expands the scope of the procedure.

We demonstrate the simple applicability of the procedure by using it to compute message lengths for two complex Bayesian penalised regression models, the ridge and horseshoe estimators, and show that it can be used to adaptively select between the different hierarchies depending on the properties of the observed data. To further underline the usefulness of our new procedure, we note that prior to this paper, computation of model selection scores for the horseshoe has, to the best of the authors' knowledge, not been undertaken anywhere in the literature due to the complexity of the hierarchy.

2 Curved Message Lengths for Conjugate-Priors

The MML87 approximation was derived under the assumption that the prior distribution, $\pi(\boldsymbol{\theta})$, is approximately uniform in a neighbourhood of the parameter estimates $\boldsymbol{\theta}$ determined by the accuracy to which the parameters are coded. In the case of Bayesian hierarchical models, where the hyperparameters that control the behaviour of the prior distributions are treated as free parameters, this assumption can be violated. In such settings, the approximation fails and minimisation of the resulting message length leads to degenerate estimates.

In the special case that the prior $\pi(\cdot)$ is *conjugate* with the likelihood $p(\cdot)$, C.S. Wallace suggested an ingenious correction to the usual MML87 formula for heavily curved priors that preserves the invariance property of the MML estimator. We discuss this procedure in some detail, as it directly motivates our approximation, and is unfortunately only briefly mentioned by Wallace ([1], pp. 235–236). Consider the two-level hierarchy

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

where $\boldsymbol{\alpha}$ are specified hyperparameters controlling the behaviour of the prior distribution for $\boldsymbol{\theta}$. The key idea is to first propose some imaginary "prior data" $\mathbf{y}_0^m$ whose properties depend only on the prior hyperparameters $\boldsymbol{\alpha}$. It is then possible to view the prior $\pi(\boldsymbol{\theta}|\boldsymbol{\alpha})$ as a *posterior* of the likelihood of this prior data $\mathbf{y}_0^m$ and some initial uninformative prior $\pi_0(\cdot)$ that does not depend on the hyperparameters $\boldsymbol{\alpha}$. Formally, we seek the decomposition

$$\pi(\boldsymbol{\theta}|\boldsymbol{\alpha}) = C_0\pi_0(\boldsymbol{\theta})p(\mathbf{y}_0^m|\boldsymbol{\theta}),$$

where $p(\mathbf{y}_0^m|\boldsymbol{\theta})$ is the likelihood of m imaginary prior samples, $\pi_0(\boldsymbol{\theta})$ is an uninformative prior and C_0 is a suitable normalisation constant not dependent on $\boldsymbol{\theta}$. The corrected Fisher information $\mathbf{J}^*(\boldsymbol{\theta})$ is then constructed from the new combined likelihood $p(\mathbf{y}^n, \mathbf{y}_0^m|\boldsymbol{\theta}) = p(\mathbf{y}^n|\boldsymbol{\theta})p(\mathbf{y}_0^n|\boldsymbol{\theta})$, and the corrected MML87 approximation is simply

$$\begin{aligned}I_{87}^*(\mathbf{y}^n, \boldsymbol{\theta}) &= -\log \pi(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \frac{1}{2}\log|\mathbf{J}^*(\boldsymbol{\theta})| - \log p(\mathbf{y}^n|\boldsymbol{\theta}) + c(k),\\ &= -\log \pi_0(\boldsymbol{\theta})C_0 + \frac{1}{2}\log|\mathbf{J}^*(\boldsymbol{\theta})| - \log p(\mathbf{y}^n, \mathbf{y}_0^m|\boldsymbol{\theta}) + c(k),\end{aligned}$$

where $c(k)$ are terms depending only on k. This correction has the interesting interpretation of being based on the asymptotic lower bound of the inverse of the variance of the maximum likelihood estimator for the combined data $(\mathbf{y}^n, \mathbf{y}_0^m)$ rather than for simply the observed data $\mathbf{y}^n$, and in this sense, it naturally incorporates the effects of the curvature of the prior on the variance of the resulting estimator. While this is clearly a very neat solution, it does have two drawbacks. The first is that it can only be used if the prior is conjugate with the likelihood, a situation that will often not be the case in many Bayesian models. The second is that it cannot be applied when the prior distribution is itself a multi-level hierarchy with additional hyperparameters that require estimation.

2.1 Poisson-Exponential Hierarchy

We now demonstrate the above procedure on a simple, but commonly used, statistical model. Consider the following Bayesian hierarchy:

$$\begin{aligned} y_i|\lambda &\sim \mathrm{Poi}(\lambda), \\ \lambda|\alpha &\sim \mathrm{Exp}(\alpha), \end{aligned}$$

where $\mathrm{Poi}(\lambda)$ denotes a Poisson distribution with with rate parameter λ, and $\mathrm{Exp}(\alpha)$ is an exponential distribution with scale parameter α and probability density

$$p(\lambda|\alpha) = \frac{1}{\alpha} \exp\left(-\frac{\lambda}{\alpha}\right). \tag{3}$$

The hyperparameter α controls how concentrated the prior density is around $\lambda = 0$; for smaller α the curvature of the prior distribution becomes very large relative to the width of the coding quantum, and the usual condition of a "roughly uniform" prior under which MML87 was derived becomes increasingly violated as $\alpha \to 0$. As the exponential distribution is conjugate to the Poisson distribution, we can use the procedure discussed in Section 2 to derive an appropriately corrected Fisher information.

Let $\mathbf{y}^n = (y_1, \ldots, y^n)$ be n samples from an unknown Poisson distribution. The probability of $\mathbf{y}^n$ for a rate parameter λ is

$$p(\mathbf{y}^n|\lambda) = \frac{\lambda^{\bar{y}} \exp(-n\lambda)}{\prod_{i=1}^n \Gamma(y_i + 1)}, \tag{4}$$

where $\bar{y} = \sum_{i=1}^n y_i$ is the sufficient statistic. Following the procedure in Section 2 we can write the likelihood of the "prior data" $\mathbf{y}_0^m$ as

$$p(\mathbf{y}_0^m|\lambda) = \frac{\lambda^{\bar{y}_0} \exp(-m\lambda)}{K}.$$

Setting $m = 1/\alpha$, $\bar{y}_0 = 0$, $\pi_0(\lambda) \propto 1$ and $C_0 = K/\alpha$ yields

$$\pi(\lambda|\alpha) = \underbrace{(1)}_{\pi_0(\lambda)} \cdot \underbrace{\left(\frac{K}{\alpha}\right)}_{C_0} \cdot \underbrace{\left[\frac{\exp(-\lambda/\alpha)}{K}\right]}_{p(\mathbf{y}_0^m|\lambda)} = \frac{1}{\alpha} \exp\left(-\frac{\lambda}{\alpha}\right).$$

Thus, the effect of the prior is to augment the data with $1/\alpha$ additional samples. The augmented negative log-likelihood (up to constants) is then

$$l^*(\lambda) = -(\bar{y} + \bar{y}_0)\log\lambda + \lambda(n + 1/\alpha).$$

The corrected Fisher information, using the expectation $\mathrm{E}\left[\bar{y} + \bar{y}_0\right] = \lambda(n+1/\alpha)$, is given by

$$J^*(\lambda) = \frac{n + 1/\alpha}{\lambda} \tag{5}$$

The effect of the correction is to inflate the amount of data by $1/\alpha$; as $\alpha \to 0$ and the prior becomes highly curved, the Fisher information increases because the coding quantum must be correspondingly smaller.

3 Approximate Message Lengths from Posterior Samples

A well known property of the Fisher information matrix is that it provides a lower bound on the covariance of unbiased estimates. In particular, if $\hat{\boldsymbol{\theta}}(\mathbf{y}^n)$ is an unbiased estimator of $\boldsymbol{\theta}$ for some parametric model, then the Cramer-Rao lower bound states that

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

Therefore, the accuracy to which parameters are encoded in the MML framework, which is determined by the determinant of the Fisher information matrix, is approximately inversely proportional to the (generalized) variance of the maximum likelihood estimates. The corrected Fisher information matrix $J^*(\boldsymbol{\theta})$ discussed in Section 2 can also be interpreted in a similar fashion by treating it as the Fisher information matrix for both real data and the additional Fisher information contributed by the prior distributions, so that the accuracy to which the parameters are encoded is approximately inversely proportional to the generalized variance of the resulting *maximum a posteriori* (MAP) estimates.

3.1 Approximating Message Lengths of Hierarchical Models

This observation motivates the main idea of this paper, which is to note that while the covariance matrix of the MAP estimates may be very difficult to compute, particularly for hierarchical models, we can approximate it by the covariance matrix of the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q | \mathbf{y}^n)$. This choice is particularly attractive because the posterior covariance automatically takes into account the effects of the prior distribution in a similar way to the curved conjugate-prior corrected Fisher information, and immediately extends to the complete set of parameters and hyperparameters in any hierarchy.

We propose to approximate the minimised message length of a complex, potentially non-conjugate hierarchical Bayesian model by

$$\begin{aligned} I(\mathbf{y}^n, \hat{\boldsymbol{\theta}}(\mathbf{y}^n)) \approx I_h(\mathbf{y}^n) = \mathrm{E}\left[-\log p(\mathbf{y}^n|\boldsymbol{\theta}) - \log\pi(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q) \,|\, \mathbf{y}^n\right] \\ -\frac{1}{2}\log\left|\mathrm{Cov}(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q \,|\, \mathbf{y}^n)\right| + c(k) - \frac{k_{\boldsymbol{\theta}}}{2}, \end{aligned} \tag{6}$$

where the expectation is taken with respect to the posterior distribution, k is the total number of free parameters and hyperparameters in the model, $c(k)$ is given by (2), $k_{\boldsymbol{\theta}}$ is the dimensionality of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}(\mathbf{y}^n)$ are the estimates that minimise the exact message length. We call (6) the MML-h approximation. The last term in (6) accounts for the fact that the usual increase in the negative log-likelihood of $\mathbf{y}^n$ due to "rounding off" the model parameters $\boldsymbol{\theta}$ is already taken into account through the posterior expectation of the negative log-likelihood.

In general, analytical evaluation of (6) is a non-trivial proposition. However, as discussed in Section 1.2, a standard approach to Bayesian analysis is to characterise posterior distributions by pseudo-random samples generated from an MCMC sampling procedure. If a chain of samples drawn from the posterior distribution is available it is straightforward to approximate the posterior covariance matrix by the empirical posterior covariance matrix. Let

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

denote the chain of m parameter and hyperparameter samples drawn from the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q | \mathbf{y}^n)$ of a Bayesian hierarchy. We can then approximate the minimised message length of the hierarchical model by the following simple formula:

$$\begin{aligned} I_h(\mathbf{y}^n) \approx \left(\frac{1}{m}\right) \sum_{j=1}^{m} \Big[-\log p(\mathbf{y}^n | \boldsymbol{\theta}^{(j)}) - \log \pi \left(\boldsymbol{\theta}^{(j)}, \boldsymbol{\alpha}_1^{(j)}, \ldots, \boldsymbol{\alpha}_q^{(j)} \right) \Big] \\ - \frac{1}{2} \log |\mathrm{Cov}(\boldsymbol{\theta}^{(j)}, \boldsymbol{\alpha}_1^{(j)}, \ldots, \boldsymbol{\alpha}_q^{(j)})| + c(k) - \frac{k_{\boldsymbol{\theta}}}{2}, \end{aligned} \quad (7)$$

where k is the total number of free parameters and hyperparameters in the complete hierarchy and $k_{\boldsymbol{\theta}}$ is the dimensionality of $\boldsymbol{\theta}$. To use the approximation (7) we require only the ability to sample from the posterior distribution, which makes it widely applicable. By taking m sufficiently large, we can approximate the formula (6) by (7) to any desired degree of accuracy.

3.2 Discussion

In comparison with the usual MML87 formula (1), the proposed MML-h approximation (7) uses the empirical covariance matrix of the posterior samples in place of the Fisher information matrix to determine the accuracy to which all parameters and hyperparameters are to be encoded. It also uses the posterior expected negative-log data-prior probabilities to approximate the maximised product of data and prior probabilities. Such an approximation is expected to work even if a hierarchy contains many hyperparameters, because the posterior variances of these hyperparameters will be generally large, and they will contribute very little to the total message length.

It is straightforward to establish that under suitable regularity conditions the approximation (6) converges to the minimised MML87 message length in the case that the hierarchy contains no adjustable hyperparameters and the number of parameters remains fixed. Convergence in the general setting is more difficult to establish, and is a topic of future research. The criterion is similar to the Laplace approach to approximate the marginal distribution of non-hierarchical Bayesian models used in [6], and the similarities may offer an approach to derive more general properties.

The empirical covariance matrix can potentially be broken into a series of block-wise independent covariance matrices for each level of the hierarchy to improve numerical stability and reduce computational burden if the dimensionality of the complete parameter and hyperparameter space is large; in this case the determinant of the posterior covariance matrix can then be replaced by

$$|\mathrm{Cov}(\boldsymbol{\theta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_q \,|\, \mathbf{y}^n)| \approx |\mathrm{Cov}(\boldsymbol{\theta}|\mathbf{y}^n)| \cdot \prod_{j=1}^{q} |\mathrm{Cov}\left(\boldsymbol{\alpha}_j|\mathbf{y}^n\right)|.$$

Finally, it is important to note that while the approximation (6) allows us to assign message lengths to potentially complex hierarchical models, it is in general not invariant under reparameterisations, and the regular MML87 approximation should be preferred in those situations in which it can be applied. The lack of invariance is expected to be a minor issue, particularly for moderate to large sample sizes, as the differences in message lengths between parameterisations will generally be small, and will decrease with increasing sample size.

4 Example 1: Poisson-Exponential

As a simple example, we apply the MML-h approximation to the Poisson-exponential hierarchy discussed in Section 2.1. In this case, the posterior distribution of the Poisson parameter, and its associated posterior variance can be exactly computed, allowing for straightforward comparison of the MML-h message length (7) with the MML87 message length, without needing to approximate quantities via MCMC sampling. The resulting expressions are very close, which lends further confidence in the validity of the approximation (7) when applied to significantly more complex problems, such as those discussed in Section 5. Recall the Poisson-exponential hierarchy discussed in Section 2.1:

$$\begin{aligned} y_i|\lambda &\sim \mathrm{Poi}(\lambda), \\ \lambda|\alpha &\sim \mathrm{Exp}(\alpha). \end{aligned}$$

Using the likelihood (4), the prior distribution (3) and the corrected Fisher information (5) derived in Section 2.1 in the MML87 formula (1), and minimising with respect to λ, yields the minimised MML87 message length

$$\begin{aligned} I_{87}(\mathbf{y}^n, \hat{\lambda}(\mathbf{y}^n)) = -(\bar{y}+1/2)\log\left(\frac{\bar{y}+1/2}{n+1/\alpha}\right) + (\bar{y}+1/2) + \sum_{i=1}^{n} \log \mathit{\Gamma}(y_i+1) \\ + \log\alpha + \frac{1}{2}\log(n+1/\alpha) + c(1). \end{aligned} \tag{8}$$

Under the specified hierarchy, the posterior distribution of λ is well known to be

$$\lambda \,|\, \mathbf{y}^n, \alpha \sim \mathrm{Ga}(\bar{y}+1, n+1/\alpha),$$

where $\mathrm{Ga}(a, b)$ is a Gamma distribution with shape parameter a and inverse-scale parameter b. We can use this to compute the approximate minimised message length using approximation (6). The posterior variance is given by

$$\mathrm{Var}(\lambda \,|\, \mathbf{y}^n, \alpha) = \frac{\bar{y}+1}{(n+1/\alpha)^2}$$

and the required expectations are

$$\mathrm{E}\left[\lambda \,|\, \mathbf{y}^n, \alpha\right] = \frac{\bar{y}+1}{n+1/\alpha}, \quad \mathrm{E}\left[\log \lambda \,|\, \mathbf{y}^n\right] = \psi(\bar{y}+1) - \log(n+1/\alpha).$$

Using the approximation $\psi(z) = \log(z) - 1/2/z + O(1/z^2)$, the MML-$h$ message length is given by

$$\begin{aligned} I_h(\mathbf{y}^n) = {} & -(\bar{y}+1/2)\log\left(\frac{\bar{y}+1}{n+1/\alpha}\right) + (\bar{y}+1) + \sum_{i=1}^{n} \log \mathit{\Gamma}(y_i+1) \\ & + \log\alpha + \frac{1}{2}\log(n+1/\alpha) + c(1) + O(1/n). \end{aligned} \tag{9}$$

Comparing (8) with (9) reveals close similarities between the two minimised message lengths. For finite n, the MML-h message length is slightly longer, and for large n the two message lengths converge.

5 Example 2: Ridge versus Horseshoe Regression

Estimation of potentially high-dimensional regression models using Bayesian shrinkage techniques is an important and active area of research, with important applications in a wide range of areas such as statistical genetics. In this setting, the prior distribution expresses a belief about the relative magnitudes of the underlying regression coefficients. Consider the following local-global hierarchy for linear regresson:

$$\begin{aligned} \mathbf{y}^n|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \\ \beta_j|\lambda_j^2, \tau^2, \sigma^2 &\sim \mathcal{N}(0, \lambda_j^2\tau^2\sigma^2), \\ \sigma^2 &\sim \sigma^{-2}d\sigma^2, \\ \lambda_j &\sim \pi(\lambda_j)d\lambda_j, \\ \tau &\sim \mathcal{C}^+(0,1), \end{aligned} \tag{10}$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a matrix of predictor variables (not necessarily full rank), $\mathcal{N}_k(\cdot,\cdot)$ is the k-variate Gaussian distribution, and $\mathcal{C}^+(0,1)$ is the standard half-Cauchy distribution with probability density function

$$p(z) = \frac{2}{\pi(1+z^2)}, \quad z > 0.$$

The parameter τ is a *global shrinkage* parameter that controls the overall level of regularisation applied to the estimated regression coefficients; in contrast, the λ_j hyperparameters are *local shrinkage* parameters that control the level of regularisation applied to individual regression coefficients. The choice of prior distributions $\pi(\lambda_j)$ associated with the local shrinkage parameters can lead to prior specifications with very different models of the underlying regression coefficients. The two priors that we consider are

$$\lambda_j \sim \begin{cases} \delta_1(\lambda_j) d\lambda_j & \text{(ridge regression)} \\ \mathcal{C}^+(0,1) & \text{(horseshoe regression)} \end{cases}$$

where $\delta_1(x)$ is a distribution with a point-mass at $x = 1$. In the case of ridge regression, there is no local shrinkage, which implies a belief that the underlying vector $\boldsymbol{\beta}$ is dense (most coefficients are non-zero), and that the majority of the coefficients have similar magnitude of association with the outcome y. In contrast, the powerful horseshoe hierarchy [3], which enjoys a number of favourable theoretical properties, allows each regression coefficient to be shrunk individually, and the half-Cauchy prior distribution over the λ_j hyperparameters implies a prior belief that the coefficients will either be close to zero, and unimportant, or large and relatively unaffected by shrinkage. These are two very different prior beliefs about $\boldsymbol{\beta}$, and both ridge regression and horseshoe estimation can yield excellent parameter estimates when applied in the appropriate setting.

An obvious question is whether we can use the data to decide which hierarchy we should be using to estimate the regression coefficients? In this example, we use the MML-h approximation to compute message lengths for both ridge and horseshoe models, and select the hierarchy that results in the shortest message length. This is an interesting and novel application of model selection to penalised regression that we believe has the potential to make significant improvements by *mitigating the weaknesses* of individual prior distributions, and is a topic for future research. This example also neatly highlights the ease of applicability of the MML-h approximation to highly complex hierarchical models such as the horseshoe, for which there are no software packages to compute model selection quantities such as marginal probabilities.

5.1 Message Lengths for Ridge and Horseshoe Regression

Sampling from the horseshoe, and ridge posterior distributions, can be efficiently done using the algorithm presented in [8]. The horseshoe and ridge hierarchies are nested, in the sense that the ridge is simply a special case of the horseshoe, and this makes computation of the approximate message lengths for both models relatively straightforward. In the case of ridge regression, we can simply set $(\lambda_j = 1)$ for all $j = (1, \ldots, p)$, and ignore these hyperparameters when computing (7). The covariance quantities used in the ridge regression case were

$$|\mathrm{Cov}(\boldsymbol{\beta}, \tau, \sigma^2 \,|\, \mathbf{y}^n)| = |\mathrm{Cov}(\boldsymbol{\beta} \,|\, \mathbf{y}^n)| \cdot \mathrm{Var}(\tau \,|\, \mathbf{y}^n) \cdot \mathrm{Var}(\sigma^2 \,|\, \mathbf{y}^n).$$

In the case of the horseshoe, the covariance quantities were

$$|\mathrm{Cov}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \tau, \sigma^2 \,|\, \mathbf{y}^n)| = |\mathrm{Cov}(\boldsymbol{\beta}, \tau, \sigma^2 \,|\, \mathbf{y}^n)| \cdot |\mathrm{Cov}(\boldsymbol{\lambda} \,|\, \mathbf{y}^n)|.$$

Coef. Model	p^*	Pairwise Correlation			Töplitz Correlation		
		HS	MML-h	MML Avg	HS	MML-h	MML Avg
$\beta_j^* = 1$	1	0.071	0.071	0.071	0.067	0.067	0.067
	5	1.022	1.040	0.971	0.731	1.079	0.996
	10	1.924	1.051	1.104	1.379	1.000	1.001
	15	2.476	1.018	1.090	1.710	1.000	1.002
	20	2.886	1.049	1.144	2.112	1.000	1.000
$\mathcal{N}(0,1)$	1	0.075	0.075	0.075	0.093	0.093	0.093
	5	0.368	0.411	0.396	0.373	0.390	0.411
	10	0.788	0.960	0.953	0.778	0.932	0.932
	15	1.149	1.049	1.035	1.090	1.006	1.006
	20	1.323	1.074	1.049	1.348	1.021	1.020
$\mathcal{C}(0,1)$	1	0.082	0.082	0.082	0.060	0.060	0.060
	5	0.413	0.413	0.413	0.427	0.432	0.432
	10	0.485	0.617	0.630	0.467	0.550	0.550
	15	0.628	0.712	0.713	0.639	0.724	0.731
	20	0.589	0.674	0.652	0.691	0.767	0.734

Table 1. Median squared prediction errors, relative to ridge regression, for the horseshoe (HS), the model with smallest MML-h message length, and the MML-h weighted mixture of ridge and horseshoe regression, computed over 100 test iterations. For each test, the number of non-zero coefficients, out of $p = 20$, is given by p^*. The sample size of the training data was $n = 50$.

We chose to use the block-diagonal covariance structure between $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ because these parameters are largely uncorrelated. Both models have dimensionality $k_{\boldsymbol{\theta}} = p+1$. The total number of parameters and hyperparameters for the ridge model is $k = p+2$, and for the horseshoe is $k = 2p+2$

5.2 Simulations

To test the ability of MML-h to discriminate between the horseshoe and ridge prior hierarchies we undertook a small simulation study. For each test iteration, ($p = 20$) covariates were generated from a multivariate normal distribution with zero mean, and either: (i) pair-wise correlations of 1/2 between each covariate, or (ii) a Töplitz correlation structure with $\mathrm{corr}(X_i, X_j) = (1/2)^{|i-j|}$. Then, a true coefficient vector $\boldsymbol{\beta}^*$ was generated according to three different models: (i) $\beta_j^* \sim \delta_1(\beta_j^*) d\beta_j^*$ (i.e., $\beta_j^* = 1$), (ii) $\beta_j^* \sim \mathcal{N}(0,1)$ and (ii) $\beta_j^* \sim \mathcal{C}(0,1)$, where $\mathcal{C}(\cdot)$ denotes the Cauchy distribution. These models cover a wide range of true coefficient patterns. The first $p^* \leq 20$ coefficients were retained, and the remaining $(p - p^*)$ coefficients were set to zero to simulate different levels of sparsity.

Finally, ($n = 50$) datapoints were generated from the model $y_i = \mathbf{x}_i \boldsymbol{\beta}^* + \varepsilon_i$, where $\mathrm{Var}(\varepsilon_i)$ was chosen to attain a signal-to-noise ratio of five. Once the data was generated, $m = 2,000$ samples were drawn from the posterior distribution of the ridge and horseshoe model, and the coordinatewise medians of the samples were used as representative point estimates. The MML-h method was then applied, using the posterior samples, to estimate message lengths for the two different prior models and the prior model with the smaller message length was selected. Additionally, a posterior weighted average of the two coefficient estimates was produced, using the exponentiated negative message lengths as unnormalized weights. Finally, the expected squared prediction error for all four methods was calculated and recorded. This process was repeated 100 times for all combinations of sparsity level $p^* = \{1, 5, 10, 15, 20\}$, correlation structure and coefficient models. The results are presented in Table 1 as median prediction errors attained by the horseshoe, the model with smallest MML-h message length (ridge or horseshoe), and the MML-h weighted posterior average of the ridge and horseshoe models, all relative to the ridge regression model. Numbers less than one indicate a performance increase relative to ridge regression, while numbers greater than one indicate poorer performance relative to ridge regression.

5.3 Results and Discussion

The results demonstrate that for specific types of underlying true coefficients both ridge regression and horseshoe regression can have significant differences in performance relative to each other. For high to moderate sparsity, the horseshoe outperformed ridge regression, while for dense models, ridge regression generally performed better. The exception was when coefficients were generated from the Cauchy distribution. The heavy tails of this distribution lead to a mixture of large and small coefficients that ridge regression priors cannot adequately model.

In contrast, both MML-h based methods exhibit excellent *robustness* to the underlying structure. While they rarely achieve the outright smallest prediction errors, there are no situations in which they perform substantially worse than the best performing method. The most difficult situation in which to identify the best fitting hierarchy appears to be the grey region between dense and sparse, which is expected. These results clearly demonstrate that the MML-h message lengths, computed from only a small number of posterior samples, are able to discriminate well between complex prior hierarchies

References

1. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length. First edn. Information Science and Statistics. Springer (2005)
2. Wallace, C.S., Freeman, P.R.: Estimation and inference by compact coding. Journal of the Royal Statistical Society (Series B) **49**(3) (1987) 240–252
3. Carvalho, C.M., Polson, N.G., Scott, J.G.: The horseshoe estimator for sparse signals. Biometrika **97**(2) (2010) 465–480
4. Makalic, E., Schmidt, D.F.: Minimum message length shrinkage estimation. Statistics & Probability Letters **79**(9) (2009) 1155–1161
5. Chib, S.: Marginal likelihood from the Gibbs output. Journal of the American Statistical Association **90**(432) (December 1995) 1313–1321
6. Lewis, S.M., Raftery, A.E.: Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. Journal of the Amer. Stat. Assoc. (1997)
7. Fitzgibbon, L.J., Dowe, D.L., Allison, L.: Univariate polynomial inference by Monte Carlo message length approximation. In: Proceedings of the Nineteenth International Conference on Machine Learning (ICML'02). (2002) 147–154
8. Makalic, E., Schmidt, D.F.: A simple sampler for the horseshoe estimator. IEEE Signal Processing Letters **23**(1) (2016) 179–182