Minimum Message Length Shrinkage Estimation

Enes Makalic *, Daniel F. Schmidt
Faculty of Information Technology, Monash University, Clayton, Australia, 3800

Abstract

This note considers estimation of the mean of a multivariate Gaussian distribution with known variance within the Minimum Message Length (MML) framework. Interestingly, the resulting MML estimator exactly coincides with the positive-part James-Stein estimator under the choice of an uninformative prior. A new approach for estimating parameters and hyperparameters in general hierarchical Bayes models is also presented.

Key words: Hierarchical Bayes, Shrinkage Estimation, Minimum Message Length, Minimum Description Length

1. Introduction

This paper considers the problem of estimating the mean of a multivariate Gaussian distribution with a known variance given a single data sample. Define X as a random variable distributed according to a multivariate Gaussian density $X \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with an unknown mean $\boldsymbol{\mu} \in \mathbb{R}^k$ and a known variance $\boldsymbol{\Sigma} = \mathbf{I}_k$. The accuracy, or risk, of an estimator $\hat{\boldsymbol{\mu}}(\mathbf{x})$ of $\boldsymbol{\mu}$ is defined as (Wald, 1971):

$$R(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}(\mathbf{x})) = \mathrm{E}_{\mathbf{x}} \left[L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}(\mathbf{x})) \right]$$

where $L(\cdot) \geq 0$ is the squared error loss function:

$$L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}(\mathbf{x})) = (\hat{\boldsymbol{\mu}}(\mathbf{x}) - \boldsymbol{\mu})' (\hat{\boldsymbol{\mu}}(\mathbf{x}) - \boldsymbol{\mu})$$

The task is to find an estimator $\hat{\boldsymbol{\mu}}(\mathbf{x})$ which minimises the risk for all values of $\boldsymbol{\mu}$. Specifically, this paper examines the problem of inferring the mean $\boldsymbol{\mu}$ from a single observation $\mathbf{x} \in \mathbb{R}^k$ of the random variable X.

It is well known, that the uniformly minimum variance unbiased (UMVU) estimator of $\boldsymbol{\mu}$ is the least squares estimate (Lehmann and Casella, 2003) given by

$$\hat{\boldsymbol{\mu}}_{\mathrm{LS}}(\mathbf{x}) = \mathbf{x}$$

This estimator is minimax under the squared error loss function and is equivalent to the maximum likelihood estimator. Remarkably, Stein (1956) has demonstrated that for $k \geq 3$, the least-squares estimator is not admissible and is in

* Corresponding author.
Email address: Enes.Makalic@infotech.monash.edu.au (Enes Makalic).

fact dominated by a large class of minimax estimators. The most well known of these dominating estimators is the positive-part James-Stein estimator (James and Stein, 1961):

$$\hat{\boldsymbol{\mu}}_{\mathrm{JS}}(\mathbf{x}) = \left(1 - \frac{k-2}{\mathbf{x}'\mathbf{x}}\right)_+ \mathbf{x}$$

where $(\cdot)_+ = \max(0, \cdot)$. Estimators in the James-Stein class tend to shrink towards some origin (in this case zero) and hence are usually referred to as *shrinkage* estimators. Shrinkage estimators dominate the least-squares estimator by trading some increase in bias for a larger decrease in variance. A common method for deriving the James-Stein estimator is through the *empirical Bayes* method (Robbins, 1964), in which the mean $\boldsymbol{\mu}$ is assumed to be distributed as per $N_k(\mathbf{0}_k, c \cdot \mathbf{I}_k)$ and the hyperparameter $c > 0$ is estimated from the data.

This paper examines the James-Stein problem within the Minimum Message Length framework (see Section 2). Specifically, we derive MML estimators of $\boldsymbol{\mu}$ and c which exactly coincide with the positive-part James-Stein estimator under the choice of an uninformative prior over c (see Section 3). A systematic approach to finding MML estimators for the parameters and hyperparameters in general hierarchical Bayes models is then developed (see Section 4). As a corollary, the new method of hyperparameter estimation appears to provide an information theoretic basis for hierarchical Bayes estimation. Some examples with multiple hyperparameters are discussed in Section 5. Concluding remarks are given in Section 6.

2. Inference by Minimum Message Length

Under the Minimum Message Length (MML) principle (Wallace and Boulton, 1968; Wallace, 2005), inference is performed by seeking the model that admits the briefest encoding (or most compression) of a message transmitted from an imaginary sender to an imaginary receiver. The message is transmitted in two parts; the first part, or *assertion*, states the inferred model, while the second part, or *detail*, states the data using a codebook based on the model stated in the assertion. This two-part message procedure minimises a trade-off between model complexity (the assertion) and model capability (the detail). Simple models tend to yield shorter assertions than more complex models, but possess a reduced ability to explain the data and hence require longer details. The model that minimises this trade-off is chosen by the MML principle as the best explanation of the underlying process.

Let $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$ be an observed data set of $n > 0$ samples and $p(\mathbf{x}|\boldsymbol{\theta})$ a probability density function indexed by a parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$. Here, $\boldsymbol{\Theta}$ is the space of permissible model parameters and $\mathcal{X}$ denotes the complete dataspace. To transmit the message both the receiver and transmitter require a common language and in the MML principle this takes the form of a subjective prior distribution $\pi(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}$. The task is then to construct a codebook over the space $(\boldsymbol{\Theta}, \mathcal{X})$ which minimises the average cost of transmitting a dataset drawn from the marginal distribution $r(\mathbf{x}) = \int \pi(\boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta}$. The Strict Minimum Message Length (SMML) (Wallace and Boulton, 1975) criterion yields an exact solution to this minimisation problem but is computationally intractable (Farr and Wallace, 2002).

The most popular alternative to SMML is Wallace and Freeman's MML87 (Wallace and Freeman, 1987) approximation. Here, under suitable regularity conditions (a discussion of which can be found in Wallace (2005), Chapter 5) the length of the message transmitting data $\mathbf{x}$ using model $\boldsymbol{\theta}$ is approximated by:

$$I_{87}(\mathbf{x}, \boldsymbol{\theta}) = \underbrace{-\log \pi(\boldsymbol{\theta}) + \frac{1}{2}\log|\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})| + \frac{k}{2}\log \kappa_k}_{I_{87}(\boldsymbol{\theta})} + \underbrace{\frac{k}{2} - \log p(\mathbf{x}|\boldsymbol{\theta})}_{I_{87}(\mathbf{x}|\boldsymbol{\theta})} \tag{1}$$

where $I_{87}(\boldsymbol{\theta})$ is the assertion codelength, $I_{87}(\mathbf{x}|\boldsymbol{\theta})$ is the detail codelength, $\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is the Fisher information matrix and κ_k is the normalised second moment of an optimal quantising lattice in k-dimensions (Conway and Sloane, 1998). For the purposes of this paper, the message length is measured in *nits*, where one nit is equal to $\log_2 e$ bits. Inference is then performed by selecting the model that minimises the message length expression (1).

Unlike many other estimators, the SMML and MML87 estimators take into consideration both the data $\mathbf{x}$ that has been observed and expected future data. The SMML criterion achieves this by partitioning the complete dataspace and assigning point estimates to each resulting partition; thus each point estimate is chosen to summarise a set of similar datasets, only one of which is actually observed. In contrast, the MML87 approximation does not construct

a codebook over the entire dataspace. Instead, the Fisher information matrix is used to summarise the expected behaviour of datasets generated by the point estimate under consideration. This approach leads to consistent estimates of nuisance parameters in the case of the Neyman-Scott problem (Dowe and Wallace, 1997) and class labelling in mixture models (Wallace and Dowe, 2000); situations where the method of maximum likelihood (ML) fails. Note that the MML criterion is similar to the Minimum Description Length (MDL) criterion developed by Rissanen (Rissanen, 1978, 1996), but differs on several technical and philosophical details. Perhaps most importantly, under the MML principle a model is required to be *fully specified* in the sense that all parameters must be defined. In contrast, the MDL philosophy generally avoids fully specified models and concerns itself with inference of model classes. For an excellent introduction and discussion of the MDL principle, see Grünwald (2007).

3. James-Stein Estimation and Minimum Message Length

Minimum Message Length (MML) shrinkage estimation of the mean of a multivariate normal distribution is now considered. The aim is to apply the Wallace and Freeman approximation (1) to inference of the mean parameter, $\boldsymbol{\mu} \in \mathbb{R}^k$, and the hyperparameter $c \in \mathbb{R}$. Recall from Section 2 that MML87 inference requires a negative log-likelihood function, prior densities on all model parameters and the Fisher information matrix. Let $l(\boldsymbol{\mu})$ denote the negative log-likelihood function of $\mathbf{x} \sim N_k(\boldsymbol{\mu}, c \cdot \mathbf{I}_k)$ given $\boldsymbol{\mu}$:

$$l(\boldsymbol{\mu}) = \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'(\mathbf{x} - \boldsymbol{\mu}) + \frac{k}{2}\log(2\pi) \tag{2}$$

The mean parameter is assigned a multivariate Gaussian prior $\boldsymbol{\mu} \sim N_k(\mathbf{0}_k, c \cdot \mathbf{I}_k)$, where the hyperparameter $c > 0$ is considered unknown and must also be estimated from the data. Define the negative log-prior of $\boldsymbol{\mu}$ given hyperparameter c as:

$$-\log \pi(\boldsymbol{\mu}|c) = \frac{1}{2c}\boldsymbol{\mu}'\boldsymbol{\mu} + \frac{k}{2}\log c + \frac{k}{2}\log(2\pi) \tag{3}$$

It is trivial to show that the Fisher Information for $\boldsymbol{\mu}$ is the identity matrix

$$\mathbf{J}_{\boldsymbol{\mu}}(\boldsymbol{\mu}) = \mathbf{I}_k \tag{4}$$

Substituting (2), (3) and (4) into the message length equation (1) and minimising with respect to $\boldsymbol{\mu}$, conditioned on c, yields the MML estimate

$$\hat{\boldsymbol{\mu}}_{87}(\mathbf{x}) = \left(\frac{c}{c+1}\right)\mathbf{x} \tag{5}$$

For small values of c, the prior density $\pi(\boldsymbol{\mu}|c)$ is not locally uniform around the point estimate $\hat{\boldsymbol{\mu}}_{87}(\mathbf{x})$, resulting in a breakdown of the MML87 approximation. Following Wallace and Freeman (1992), this problem is addressed by adding the curvature of the negative log-prior (3) to the information matrix (4) yielding the corrected information matrix:

$$\mathbf{J}_{\boldsymbol{\mu}}(\boldsymbol{\mu}|c) = \left(1 + \frac{1}{c}\right)\mathbf{I}_k$$

The new approximation results in sensible message lengths for all values of $c > 0$. Since the hyperparameter c is not known, it must be inferred from the data and transmitted to the receiver. The receiver then uses the obtained value of c to construct the prior density $\pi(\boldsymbol{\mu}|c)$ over $\boldsymbol{\mu}$. A simple approach to encoding c is to use the asymptotic result derived in (Rissanen, 1989) and state c using $\left(\frac{1}{2}\log k\right)$ nits. This approach was considered by Wallace and Patrick (1993) for MML inference of hyperparameters in decision tree models over an ensemble of binomial distributions. The complete message length is now given by

$$I_{87}(\mathbf{x}, \boldsymbol{\mu}|c) = \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'(\mathbf{x} - \boldsymbol{\mu}) + \frac{k}{2}\log(c+1) + \frac{1}{2c}\boldsymbol{\mu}'\boldsymbol{\mu} + \frac{1}{2}\log k + k\log(2\pi) + \text{const} \tag{6}$$

Simply minimising the message length (6) for $\boldsymbol{\mu}$ and the hyperparameter c yields the estimate

$$\hat{c}_{87}(\mathbf{x}) = \left(\frac{\mathbf{x}'\mathbf{x}}{k} - 1 \right)_+$$

This choice of c results in an estimator that dominates the least-squares estimator for $k \geq 5$ and is inferior to the James-Stein estimator. However, the cardinal rule of the MML principle is that each parameter should only be stated to the accuracy warranted by the data. Since the asymptotic approximation used to code c is inefficient for small amounts of data, a better coding scheme is needed.

We now derive a more efficient codebook for the hyperparameter c and show that stating c to the optimal accuracy improves the frequentist properties of the resulting estimator. The key step is to note that there is no requirement to transmit all model parameters simultaneously. Instead, one can first send c using an optimal code, then transmit $\boldsymbol{\mu}$ given c, and finally the data $\mathbf{x}$ is transmitted using the previous value of $\boldsymbol{\mu}$. The optimal code for $\boldsymbol{\mu}$ has already been determined; the optimal code for c depends on the sensitivity of $I_{87}(\mathbf{x}, \boldsymbol{\mu}|c)$ to changes in c. As $I_{87}(\mathbf{x}, \boldsymbol{\mu}|c)$ defines a probability distribution over $\mathbf{x}$, one may treat $I_{87}(\cdot)$ as a 'likelihood' and apply the standard MML87 approximation to determine the optimal code lengths. To remove the dependence of $I_{87}(\cdot)$ on the mean $\boldsymbol{\mu}$, we may replace it with an estimator that minimises $I_{87}(\cdot)$, namely (5), resulting in an unnormalised distribution, $I_{87}(\mathbf{x}, \hat{\boldsymbol{\mu}}(\mathbf{x})|c)$, over $\mathbf{x} \in \mathbb{R}^k$ only.

The Fisher information for the hyperparameter c is determined by finding the second derivative of $I_{87}(\mathbf{x}, \hat{\boldsymbol{\mu}}(\mathbf{x})|c)$ with respect to c

$$\frac{d^2 I_{87}(\mathbf{x}, \hat{\boldsymbol{\mu}}(\mathbf{x})|c)}{dc^2} = \frac{2\mathbf{x}'\mathbf{x} - k(c+1)}{2(c+1)^3}$$

and taking the required expectations. Note that $\mathrm{E}\left[\mathbf{x}'\mathbf{x}\right] = k(c+1)$ as $\mathbf{x}$ is distributed according to the marginal distribution $r(\mathbf{x}|c) = \int \pi(\boldsymbol{\mu}|c) p(\mathbf{x}|\boldsymbol{\mu}) d\boldsymbol{\mu} = N_k(\mathbf{0}_k, (c+1)\cdot\mathbf{I}_k)$ (Lehmann and Casella, 2003). The Fisher information for c is therefore

$$J_c(c) = \frac{k}{2(c+1)^2} \tag{7}$$

Clearly, the coding accuracy for c increases as c approaches zero ($c \to 0$), which is unsurprising when one considers the effect of varying c on the estimate $\hat{\boldsymbol{\mu}}_{87}(\mathbf{x})$:

$$\frac{d\hat{\boldsymbol{\mu}}_{87}(\mathbf{x})}{dc} = -\frac{1}{c^2}$$

When c is large, minor perturbations of c leave the estimate $\boldsymbol{\mu}$ relatively unchanged; in contrast, when c is small, minor perturbations of c result in large changes of the estimate.

It remains to determine a suitable prior density over c. Lacking prior knowledge of the distribution of $\boldsymbol{\mu}$, the authors opt for the traditionally uninformative uniform prior $\pi(c) \propto 1$ (Berger and Strawderman, 1996) (also suggested by Stein (1962), given that c is essentially a parameter describing $\boldsymbol{\mu}'\boldsymbol{\mu}$) over some suitable support. The complete message length is now

$$I_{87}(\mathbf{x}, \boldsymbol{\mu}, c) = \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'(\mathbf{x} - \boldsymbol{\mu}) + \frac{k-2}{2}\log(c+1) + \frac{1}{2c}\boldsymbol{\mu}'\boldsymbol{\mu} + k\log(2\pi) + \text{const} \tag{8}$$

Minimising (8) for $\boldsymbol{\mu}$ and c yields the estimate

$$\hat{c}_{87}(\mathbf{x}) = \left(\frac{\mathbf{x}'\mathbf{x}}{k-2} - 1 \right)_+$$

since c is assumed positive. Under this choice of c, the MML estimate for the mean, $\hat{\boldsymbol{\mu}}_{87}(\mathbf{x})$, coincides exactly with the James-Stein estimator.

Remark 1: In its current form, the MML coding scheme does not handle the special case where $\hat{c}_{87}(\mathbf{x}) = 0$. The problem is easily rectified if one instead uses an alternate model with no free parameters that assumes the data is distributed as per $N_k(\mathbf{0}_k, \mathbf{I}_k)$ and prepends the code with an indicator variable selecting the appropriate coding scheme.

This solution inflates the codelength in both cases by $(\log 2)$ nits (Hansen and Yu, 2001).

Remark 2: Behaviour for $k < 3$. For $k < 3$, it is straightforward to show that $\partial I_{87}(\cdot)/\partial c < 0$ for $0 < c < \infty$, yielding an estimate $\hat{c}_{87}(\mathbf{x}) = \infty$ in the limit. Under this diffuse choice of hyperparameter, the MML estimates coincide exactly with the least squares estimates. Interestingly, this breakdown in the message length approximation yields estimates that coincide with least squares and are thus admissible when $k < 3$. Unfortunately, the message lengths themselves become nonsensical and lose a strict coding interpretation (the message lengths are negative for $k = 1$). Although the breakdown has little effect in the case where k is fixed, it may be particularly problematic in situations where the message lengths are used to select between candidate models of varying complexity.

Remark 3: Model selection. One of the strengths of the MML principle is that it unifies parameter estimation and model selection within the same framework by assigning a posterior probability mass to fully specified models. Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{R}^k$ be a partition of the data into two groups; $\mathbf{x}_1 \in \mathbb{R}^p$ (with $p \geq 3$) is assumed to have non-zero means and $\mathbf{x}_2 \in \mathbb{R}^m$ zero means. Using the approximation (see Wallace (2005), p. 257–258)

$$\frac{p}{2}\left(\log \kappa_p + 1\right) \approx -\frac{p}{2}\log(2\pi) + \frac{1}{2}\log(p\pi) - \gamma$$

where γ is Euler's constant, the full message length up to terms not depending on p is now given by

$$I_{87}(\mathbf{x}, \boldsymbol{\mu}, c) = \frac{1}{2}(\mathbf{x}_1 - \boldsymbol{\mu})'(\mathbf{x}_1 - \boldsymbol{\mu}) + \frac{1}{2}\mathbf{x}_2'\mathbf{x}_2 + \frac{p-2}{2}\log(c+1) + \frac{1}{2c}\boldsymbol{\mu}'\boldsymbol{\mu} + \log p \tag{9}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ are the estimated means for the non-zero components. In order to perform model selection we select the partition that minimises (9). This has an interesting extension to subset selection in orthonormal regression which is the focus of future research.

4. Message Lengths of Hierarchical Bayes Structures

The previous section suggests a general procedure for estimating parameters in a hierarchical Bayes structure. Given a parametrised probability density $p(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \sim \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ and $\boldsymbol{\alpha} \sim \pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$, first find the message length of $\mathbf{x}$ given $\boldsymbol{\theta}$ conditioned on $\boldsymbol{\alpha}$, i.e.

$$I_{87}(\mathbf{x}, \boldsymbol{\theta}|\boldsymbol{\alpha}) = l(\boldsymbol{\theta}) - \log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\alpha}) + \frac{1}{2}\log|\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})| + \text{const} \tag{10}$$

Next find the estimates $\hat{\boldsymbol{\theta}}_{87}(\mathbf{x}|\alpha)$ that minimise the message length (10). In most cases, it appears necessary to apply the curved prior correction (see Wallace (2005), p. 236–237) to the MML87 message length formula (see Section 3); for conjugate priors this may be done in a fashion that perserves invariance of the estimators. Form a new 'profile message length' of $\mathbf{x}$ as $I_{87}(\mathbf{x}, \hat{\boldsymbol{\theta}}_{87}(\mathbf{x}|\alpha)|\boldsymbol{\alpha})$ by substituting the parameters $\boldsymbol{\theta}$ by their estimates. The accuracy to which $\boldsymbol{\alpha}$ must be stated is determined by finding the Fisher Information of the hyperparameters

$$\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \mathrm{E}_r\left[\frac{\partial^2 I_{87}(\mathbf{x}, \hat{\boldsymbol{\theta}}_{87}(\mathbf{x}|\alpha)|\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}\right]$$

where the expectation is taken with respect to the marginal distribution $r(\mathbf{x}|\boldsymbol{\alpha}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\mathbf{x}|\boldsymbol{\theta})d\boldsymbol{\theta}$. The complete message length for $\mathbf{x}$, $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ is given by

$$\begin{aligned} I_{87}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\alpha}) &= I_{87}(\boldsymbol{\alpha}) + I_{87}(\boldsymbol{\theta}|\boldsymbol{\alpha}) + I_{87}(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\alpha}) \\ &= l(\boldsymbol{\theta}) - \log \pi_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\alpha})\pi_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) + \frac{1}{2}\log|\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})||\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| + \text{const} \end{aligned} \tag{11}$$

Finally, one may minimise $I_{87}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\alpha})$ to jointly estimate both parameters $\boldsymbol{\theta}$ and the hyperparameters $\boldsymbol{\alpha}$. This procedure may be trivially extended to multiple hierarchies of parameters. This general procedure may be used to extend the shrinkage results in this paper to the more complex problem of regularised regression, with the message length

score allowing combined model selection (i.e. selection of important regressors) and estimation of the regularisation parameter. This has interesting applications in the analysis of microarray data in which there are generally many more potential features than datapoints, and is a focus of future research. The hierarchical message length procedure is applied to several problems with multiple hyperparameters in the next section.

5. Shrinkage Towards a Grand Mean

The following extension to the basic JS shrinkage estimator was proposed by Lindley in the discussion of (Stein, 1962) and has been applied to several problems by Efron and Morris (1973, 1975). Lindley suggests that instead of shrinking to the origin, one may wish to shrink the parameters to another point in the parameter space. Under this modification, the parameters μ_i $(i = 1, \ldots, k)$ are assumed to be normally distributed, $\mu_i \sim N(a, c)$, where both $a \in \mathbb{R}$ and $c \in \mathbb{R}^+$ are unknown parameters and must be inferred from the data. The combined message length of data $\mathbf{x}$ and parameters $\boldsymbol{\mu}$ conditioned on hyperparameters $\boldsymbol{\alpha} \equiv (a, c)$, is

$$I(\mathbf{x}, \boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'(\mathbf{x} - \boldsymbol{\mu}) + \frac{k}{2}\log(c+1) + \frac{1}{2c}(\boldsymbol{\mu} - a)'(\boldsymbol{\mu} - a) + k\log(2\pi) + \text{const} \tag{12}$$

The MML estimates of $\hat{\boldsymbol{\mu}}_{87}(\mathbf{x})$ that minimise (12), conditioned on $\boldsymbol{\alpha}$, are

$$\hat{\boldsymbol{\mu}}_{87}(\mathbf{x}|\boldsymbol{\alpha}) = \frac{c}{c+1}\left(\mathbf{x} + \frac{a}{c}\right) \tag{13}$$

It remains to determine the accuracy to which $\boldsymbol{\alpha}$ should be stated. Following the general procedure detailed in Section 4, begin by substituting (13) into (12) and finding the second derivatives of $I(\mathbf{x}, \hat{\boldsymbol{\mu}}_{87}(\mathbf{x}|\boldsymbol{\alpha})|\boldsymbol{\alpha})$. Expectations are taken by noting that the marginal distribution of $\mathbf{x}$ is $N_k(a\mathbf{1}_k, (c+1)\mathbf{I}_k)$, yielding the Information Matrix

$$\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \begin{bmatrix} \dfrac{k}{c+1} & 0 \\ 0 & \dfrac{k}{2(c+1)^2} \end{bmatrix}$$

and $|\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| = k^2/2/(c+1)^3$. The accuracy to which the hyperparameters must be stated is independent of a; this is not unexpected as a is a location parameter. A uniform prior $\pi(a) \propto 1$ is chosen as nothing is known *a priori* about the possible locations of the grand mean. This leads to a combined message length of

$$I(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\alpha}) = \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'(\mathbf{x} - \boldsymbol{\mu}) + \frac{k-3}{2}\log(c+1) + \frac{1}{2c}(\boldsymbol{\mu} - a)'(\boldsymbol{\mu} - a) + k\log(2\pi) + \text{const}$$

resulting in the following estimators for a and c:

$$\hat{a}_{87}(\mathbf{x}) = \frac{1}{k}\sum_{i=1}^{k} x_i$$

$$\hat{c}_{87}(\mathbf{x}) = \left(\frac{(\mathbf{x} - \hat{a}_{87}(\mathbf{x}))'(\mathbf{x} - \hat{a}_{87}(\mathbf{x}))}{k-3} - 1\right)_+$$

This is instantly recognisable as the positive part James-Stein estimator proposed by Lindley that is known to dominate LS for $k \geq 4$.

Remark 4: Asymptotic Codelengths. Replacing the finite sample codelength with the asymptotic codelength as suggested by Wallace and Patrick (1993) and Hansen and Yu (2001) leads to the estimate

$$\hat{c}_{87}(\mathbf{x}) = \left(\frac{(\mathbf{x} - \hat{a}_{87}(\mathbf{x}))'(\mathbf{x} - \hat{a}_{87}(\mathbf{x}))}{k} - 1\right)_+$$

which only dominates least-squares for $k \geq 6$. While this does not make a significant difference when inferring the mean of a multivariate normal, the effect is drastic in scenarios where the number of hyperparameters grows proportionally to the number of parameters. This problem is illustrated in the following example.

5.1. *Inference of 'Nuisance' Hyperparameters*

The following toy example illustrates the importance of encoding the hyperparameters efficiently. Suppose the data now consists of m groups of p points where each data group is assigned a separate mean hyperparameter

$$\mathbf{x}_i \sim N(\mathbf{0}_p, \mathbf{I}_p)$$

$$\boldsymbol{\mu}_i \sim N(a_i \mathbf{1}_p, c\mathbf{I}_p)$$

where $i = 1, \ldots, m$. The mean and variance hyperparameters $\boldsymbol{\alpha} = (a_1, \ldots, a_m, c)$ are given uniform priors, and the MML analysis proceeds in a similar fashion to the previous section. The Fisher information for $\boldsymbol{\alpha}$ is now given by $|\mathbf{J}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})| = (c+1)^{(-m-2)}$. The estimates of the hyperparameter means are $\hat{a}_i(\mathbf{x}_i) = \frac{1}{p}\sum_{j=1}^{p} x_{(i,j)}$ which are simply the arithmetic means of the each data group. The variance estimate is given by

$$\hat{c}_{87}(\mathbf{x}) = \left(\frac{\sum_{i=1}^{m} (\mathbf{x}_i - \hat{a}_i(\mathbf{x}_i))'(\mathbf{x}_i - \hat{a}_i(\mathbf{x}_i))}{n - m - 2} - 1 \right)_+ \tag{14}$$

where $n = mp$. It is straightforward to show that (14) is a consistent estimate of c. Replacing the finite sample codelength with the asymptotic codelength leads to an inconsistent estimate. Using an inconsistent estimate of c (for example, the marginal maximum likelihood estimator (Maritz, 1970)), leads to a greater squared-error risk. This is analogous to the MML solution for the conventional Neyman-Scott problem (Dowe and Wallace, 1997). While this problem may appear rather artificial, similar problems of inconsistency in estimation of hyperparameters may be expected to arise if complex forms of priors based on mixture models are employed. In this case, in a similar fashion to the regular mixture modelling problem, it is expected that estimation based on the hierarchical message length formulation should lead to consistent estimates.

6. Conclusion

This paper has examined the task of estimating the mean of multivariate normal distribution with known variance given a single data sample within the Minimum Message Length framework. We considered this problem in a hierarchical Bayes setting where the prior distribution on the mean depends on an unknown hyperparameter that must be estimated from the data. We show that if the hyperparameter is stated suboptimally, the resulting solution is inferior to the James-Stein estimator. Once the message length is extended to incorporate an efficient code for the hyperparameter, the resulting MML estimator coincides exactly with the James-Stein estimator. This procedure is subsequently generalised to provide a framework for joint estimation of both the parameters and hyperparameters, and applied to the problem of estimating multiple hyperparameters with success. Essentially, the new approach is akin to hierarchical Bayes estimation by compact coding. Furthermore, the results obtained for the mean estimation problem suggest that hierarchical Bayes MML estimators may be a fruitful approach to systematically constructing shrinkage estimators for arbitrary problems.

References

Berger, J. O., Strawderman, W. E., June 1996. Choice of hierarchical priors: Admissibility in estimation of normal means. The Annals of Statistics 24 (3), 931–951.

Conway, J. H., Sloane, N. J. A., December 1998. Sphere Packing, Lattices and Groups, 3rd Edition. Springer-Verlag.

Dowe, D. L., Wallace, C. S., 1997. Resolving the Neyman-Scott problem by Minimum Message Length. In: Proc. Computing Science and Statistics - 28th Symposium on the interface. Vol. 28. pp. 614–618.

Efron, B., Morris, C., 1973. Combining possibly related estimation problems. Journal of the Royal Statistical Society (Series B) 35 (3), 379–421.

Efron, B., Morris, C., June 1975. Data analysis using Stein's estimator and its generalizations. Journal of the American Statistical Association 70 (350), 311–319.

Farr, G. E., Wallace, C. S., 2002. The complexity of strict minimum message length inference. Computer Journal 45 (3), 285–292.

Grünwald, P. D., 2007. The Minimum Description Length Principle. Adaptive Communication and Machine Learning. The MIT Press.

Hansen, M. H., Yu, B., 2001. Model selection and the principle of minimum description length. Journal of the American Statistical Association 96 (454), 746–774.

James, W., Stein, C. M., 1961. Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium. Vol. 1. University of California Press, pp. 361–379.

Lehmann, E. L., Casella, G., 2003. Theory of Point Estimation, 4th Edition. Springer Texts in Statistics. Springer.

Maritz, J. S., 1970. Empirical Bayes Methods. Methuen.

Rissanen, J., September 1978. Modeling by shortest data description. Automatica 14 (5), 465–471.

Rissanen, J., 1989. Stochastic Complexity in Statistical Inquiry. World Scientific.

Rissanen, J., January 1996. Fisher information and stochastic complexity. IEEE Transactions on Information Theory 42 (1), 40–47.

Robbins, H., March 1964. The empirical bayes approach to statistical decision problems. The Annals of Mathematical Statistics 35 (1), 1–20.

Stein, C. M., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. Berkeley, University of California Press, pp. 197–206.

Stein, C. M., 1962. Confidence sets for the mean of a multivariate normal distribution. Journal of the Royal Statistical Society (Series B) 24 (2), 265–296.

Wald, A., 1971. Statistical Decision Functions, 2nd Edition. Chelsea Pub Co.

Wallace, C., Boulton, D., 1975. An invariant Bayes method for point estimation. Classification Society Bulletin 3 (3), 11–34.

Wallace, C. S., 2005. Statistical and Inductive Inference by Minimum Message Length, 1st Edition. Information Science and Statistics. Springer.

Wallace, C. S., Boulton, D. M., August 1968. An information measure for classification. Computer Journal 11 (2), 185–194.
URL http://www.allisons.org/ll/MML/Structured/1968-WB-CJ/

Wallace, C. S., Dowe, D. L., January 2000. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. Statistics and Computing 10 (1), 73–83.

Wallace, C. S., Freeman, P. R., 1987. Estimation and inference by compact coding. Journal of the Royal Statistical Society (Series B) 49 (3), 240–252.

Wallace, C. S., Freeman, P. R., 1992. Single-factor analysis by minimum message length estimation. Journal of the Royal Statistical Society (Series B) 54 (1), 195–209.

Wallace, C. S., Patrick, J. D., April 1993. Coding decision trees. Machine Learning 11 (1), 7–22.