Minimum Message Length Analysis of the Behrens–Fisher Problem

Enes Makalic and Daniel F. Schmidt

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

Abstract. Given two sequences of Gaussian data, the Behrens–Fisher problem is to infer whether there exists a difference between the two corresponding population means if the population variances are unknown. This paper examines the Behrens–Fisher-type problem within the minimum message length framework of inductive inference. Using a special bounding on a uniform prior over the population means, a simple Bayesian hypothesis test is derived that does not require computationally expensive numerical integration of the posterior distribution. The minimum message length procedure is then compared against well-known methods on the Behrens–Fisher hypothesis testing problem and the estimation of the common mean problem showing excellent performance in both cases. Extensions to the generalised Behrens–Fisher problem and the multivariate Behrens–Fisher problem are also discussed.

1 Introduction

Consider two mutually independent sequences of i.i.d. data denoted by $\mathbf{y}_1 = (y_{11}, \ldots, y_{1n_1})'$ and $\mathbf{y}_2 = (y_{21}, \ldots, y_{1n_2})'$ and generated by the following Gaussian model:

$$y_{ij} \sim N(\mu_i, \tau_i), \tag{1}$$

where $(i = 1, 2; j = 1, \ldots, n_i)$ and $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ and $\boldsymbol{\tau} = (\tau_1, \tau_2)'$ are the unknown sequence means and variances respectively. The Behrens–Fisher problem is to infer whether there exists a difference between the two population means; that is, whether $\mu_1 = \mu_2$. This paper examines the Behrens–Fisher problem using the minimum message length (MML) principle of inductive inference. The minimum message length approach provides a Bayesian solution that does not require computationally expensive numerical integration of the posterior probability density. The corresponding solution is easily extendable to testing for equality of variances and the generalised Behrens–Fisher problem where the data comprises more than two sequences (that is, $i > 2$).

When the population variances $\boldsymbol{\tau}$ are assumed to be known, or their ratio $\rho = \tau_1/\tau_2$ is specified, a common frequentist solution to the Behrens–Fisher

problem is a hypothesis test based on a Student t pivot. There does not exist a non-randomised frequentist procedure independent of the data for obtaining exact confidence intervals if the population variances are unknown [1]. A common practical solution in this case is to use the Student t pivot with Satterthwaite's approximation for the number of degrees of freedom [2]. An alternative solution is to use a fiducial probability distribution [3, 4] or a fully Bayesian approach. An excellent review of the fiducial and Bayesian solutions to the Behrens–Fisher problem is given in [5].

2 Minimum Message Length (MML)

The minimum message length (MML) principle [6–8] offers a Bayesian framework for inference that is rooted in information theory and is a practical implementation of the theory of inductive inference proposed initially by Solomonoff [9], Kolmogorov [10] and Chaitin [11]. The underlying idea is to view the problem of estimation and model selection as one of data compression. Such an approach naturally leads to criteria that balance a trade-off between the model fit and the model complexity. The model fit is measured by the amount of information, $I(\mathbf{y}|\boldsymbol{\theta})$, required to encode the data using a given model; $I(\mathbf{y}|\boldsymbol{\theta})$ commonly includes the negative log-likelihood function. The model complexity denotes the amount of information, $I(\boldsymbol{\theta})$, needed to encode the selected model relative to some chosen prior beliefs. In this two part decomposition the statement of the chosen model is generally named the *assertion* and the statement of the data using this model is named the *detail*. The model that minimises the sum

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

of the assertion and the detail is accepted as the most *a posteriori* likely explanation of the data in light of the chosen prior beliefs. While the quantity $I(\mathbf{y}, \boldsymbol{\theta})$ can be exactly calculated using the strict MML prescription of [12], this is generally computationally intractable and approximations are used instead. The most commonly used approximation is the MML87 formula of [8], which under suitable regularity conditions gives the joint codelength of model $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$ and data $\mathbf{y}$ as

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

where k is the number of free parameters, $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function, $\pi(\cdot)$ denotes a prior distribution over the parameter space $\boldsymbol{\Theta}$, $\mathbf{J}(\boldsymbol{\theta})$ is the Fisher information matrix, and κ_k is the normalised second moment of an optimal quantising lattice in k-dimensions. For many dimensions κ_k is not known, and it is common to use the approximation ([6], p. 237)

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

where $\psi(\cdot)$ is the digamma function. In this paper, all codelengths are measured in *nits* (nats), or base-e digits, and as such "log" denotes the natural logarithm. The Wallace–Freeman approximation states that the model $\hat{\boldsymbol{\theta}}_{87}(\mathbf{y})$ that minimises (2) is the most *a posteriori* likely explanation of the data in the light of the chosen priors. Note that the model space $\boldsymbol{\Theta}$ may be enlarged to include models of many different classes if the parameter vector is suitably partitioned into continuous parameters and discrete, structural parameters and these are handled accordingly. In this way, MML treats both parameter estimation and model class selection on the same footing. The Wallace–Freeman approximation provides codelengths (and therefore estimates) that are invariant under smooth, one-to-one reparameterisations of the parameters and has shown to be consistent in difficult inference problems; for example, the Neyman–Scott problem [13].

3 MML and the Behrens–Fisher Problem

Consider the Behrens–Fisher problem from (1) and let $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2')'$ denote the vector comprising two sequences of data with $n = (n_1 + n_2)$ total data points. The solution to the Behrens–Fisher problem within the message length framework requires: (1) the codelength of the data under the assumption that the population means are equal (that is, $\mu_1 = \mu_2$), and (2) the codelength of the data assuming there exist two population means (that is, $\mu_1 \neq \mu_2$). The model resulting in the shortest codelength is then deemed to be *a posteriori* most likely to have generated the data.

3.1 Shared population mean

The model with a single population mean for two data sequences is examined first. The model parameters $\boldsymbol{\theta} = (\mu, \boldsymbol{\tau}')' \in \mathbb{R}^3$ and $\boldsymbol{\tau} = (\tau_1, \tau_2)'$ are considered unknown and must be inferred from the data. Application of the Wallace–Freeman codelength (2) requires a likelihood function, a corresponding Fisher information and prior densities for all parameters. The negative log-likelihood function is

$$-\log p(\mathbf{y}|\boldsymbol{\theta}) = \frac{n}{2}\log 2\pi + \frac{1}{2}\sum_{i=1}^{2}\left(n_i \log \tau_i + \frac{1}{\tau_i}\sum_{j=1}^{n_i}(y_{ij} - \mu)^2\right). \tag{3}$$

The determinant of the Fisher information matrix, $\mathbf{J}(\boldsymbol{\theta})$, is

$$|\mathbf{J}(\boldsymbol{\theta})| = \left(\prod_{i=1}^{2}\frac{n_i}{2\tau_i^2}\right)\left(\frac{n_1}{\tau_1} + \frac{n_2}{\tau_2}\right). \tag{4}$$

It remains to specify the prior densities over the parameters $\boldsymbol{\theta}$. The population variances are considered independent of the mean and given conjugate, scale invariant, prior densities over some compact set Ξ; for example, let $\Xi =$

$(a, b) \times (a, b)$ where $0 < a < b < \infty$. This prior is reasonable as it indicates no preference for any particular measurement scale of the data. Since the two variance parameters are common to both models under consideration, the choice of a prior density and the support Ξ (that is, the choice of a and b) for the variances has no effect on the model selection procedure. Following the procedure in [14], the population mean is given a uniform prior over a special compact support in order to avoid the Jeffreys–Lindley paradox. The chosen support for this prior density can obtained by the following argument. First note that the observed data $\mathbf{y}$ is generated from the model

$$\mathbf{y} = \mathbf{y}_* + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_n)$ and $\mathbf{y}_*$ is the noise-free data (the "signal"). It is clear that

$$\mathrm{E}\left(\mathbf{y}'\mathbf{y}\right) = \mathbf{y}_*'\mathbf{y}_* + \mathrm{tr}\left(\boldsymbol{\Sigma}_n\right). \tag{5}$$

where $\mathrm{E}(\cdot)$ denotes the expectation operator. For any value of the population mean μ, one can construct an estimate of $\mathbf{y}_*$, say $(\mathbf{1}_n\hat{\mu})$, and since $\mathrm{tr}\left(\boldsymbol{\Sigma}_n\right)$ is unknown and strictly positive, by (5), this estimate should satisfy

$$\mathbf{y}'\mathbf{y} \geq \left(\mathbf{1}_n\hat{\mu}\right)'\left(\mathbf{1}_n\hat{\mu}\right) = n\hat{\mu}^2, \tag{6}$$

where $\mathbf{1}_n$ is a $(n \times 1)$ vector of ones. From (6), the feasible parameter set $\Lambda_1 \subset \mathbb{R}$ is

$$\Lambda_1 = \left\{\mu : n\mu^2 \leq \mathbf{y}'\mathbf{y}\right\}.$$

A suitable prior for the population mean is then a uniform density defined over the support Λ_1. Within the context of MML, this prior is perfectly acceptable as the data component of the prior (that is, $\mathbf{y}'\mathbf{y}/n$) can be encoded, with codelength $O(\log n)$, prior to encoding the parameters, and the data given the parameters (see below). Further arguments for this choice of prior density are given in Appendix A. The complete prior density for all parameters is

$$\pi(\boldsymbol{\theta}) = \pi_\mu(\mu)\pi_{\boldsymbol{\tau}}(\boldsymbol{\tau}), \tag{7}$$

$$\pi(\mu) = \frac{1}{\mathrm{vol}(\Lambda_1)} = \left(\frac{n}{4\mathbf{y}'\mathbf{y}}\right)^{1/2}, \quad \mu \in \Lambda_1, \tag{8}$$

$$\pi_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \left(\Omega\tau_1\tau_2\right)^{-1}, \quad \tau_1, \tau_2 \in \Xi, \tag{9}$$

where $\Omega > 0$ is a suitable normalisation constant. Substituting (3), (4) and (7) into (2) yields the total codelength $I_{87}(\mathbf{y}, \mu, \boldsymbol{\tau})$

$$\frac{n}{2}\log 2\pi + \frac{1}{2}\sum_{i=1}^{2}\left(n_i \log \tau_i + \frac{1}{\tau_i}\sum_{j=1}^{n_i}(y_{ij} - \mu)^2\right) + \frac{1}{2}\log\left(\frac{n_1}{\tau_1} + \frac{n_2}{\tau_2}\right) + \frac{1}{2}\log\left(\frac{\Omega^2(\mathbf{y}'\mathbf{y})}{n}\prod_{i=1}^{2} n_i\right) + c(3), \tag{10}$$

where $c(3) = -2{\cdot}32$. Minimising (10) numerically yields the Wallace–Freeman parameter estimates

$$(\hat{\mu}, \hat{\boldsymbol{\tau}}) = \arg\min_{\mu,\boldsymbol{\tau}} \{I_{87}(\mathbf{y}, \mu, \boldsymbol{\tau})\}. \tag{11}$$

The Wallace–Freeman estimate of μ is equal to the maximum likelihood estimate only when the variances are known. The optimal Wallace–Freeman model under the assumption the data shares a common mean has codelength $I_{87}(\mathbf{y}, \hat{\mu}, \hat{\boldsymbol{\tau}})$.

In its current form, the total codelength (10) is not strictly valid as the prior density for the population mean is a function of the observed data $\mathbf{y}'\mathbf{y}$. This is easily rectified if one assumes existence of a suitable preamble code stating the data constant (that is, $\mathbf{y}'\mathbf{y}/n$) prior to transmitting the data itself. The length of this code can be shown to be approximately $\log(n)/2$ nits. As the preamble code is now common to both models under consideration (that is, both codelengths are extended by $\log(n)/2$ nits) it has no effect on the choice of model made by MML and is omitted from further discussion.

3.2 Different population means

Consider now the model where the population mean differs between the two data sequences. The model parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}', \boldsymbol{\tau}')' \in \mathbb{R}^4$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)'$, are again considered unknown and must be inferred from the data. Following the same argument as in Section 3.1 the feasible parameter set for the population means is now the ellipsoid

$$\Lambda_2 = \left\{ (\mu_1, \mu_2) : \sum_{i=1}^{2} n_i \mu_i^2 \leq \mathbf{y}'\mathbf{y} \right\}$$

with volume $\mathrm{vol}(\Lambda_2) = \pi \mathbf{y}'\mathbf{y}/\sqrt{n_1 n_2}$. The prior densities for the population variances are taken to be equivalent to (9). The determinant of the Fisher information matrix is

$$|\mathbf{J}(\boldsymbol{\theta})| = \prod_{i=1}^{2} \left(\frac{n_i^2}{2\tau_i^3} \right).$$

Following the procedure in Section 3.1, the total Wallace–Freeman codelength $I_{87}(\mathbf{y}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\tau}})$ is

$$\frac{n}{2}\log 2\pi + \frac{1}{2}\left(\sum_{i=1}^{2} (n_i - 1) \log \hat{\tau}_i \right) + \frac{n-2}{2} + \log\left(\mathbf{y}'\mathbf{y}\sqrt{n_1 n_2}\Omega\pi/2\right) + c(4), \tag{12}$$

where

$$\hat{\mu_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \quad \hat{\tau}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2, \quad (i = 1, 2) \tag{13}$$

are the Wallace–Freeman parameter estimates $\hat{\boldsymbol{\theta}}_{87}(\mathbf{y})$, and $c(4) = -3{\cdot}14$. In this case, the Wallace–Freeman parameter estimates are the same as the regular unbiased estimates.

3.3 MML Hypothesis Testing

Let $\delta = (I_{87}(\mathbf{y}, \hat{\mu}, \hat{\boldsymbol{\tau}}) - I_{87}(\mathbf{y}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\tau}}))$ denote the difference in Wallace–Freeman codelengths between the model with a shared population mean (10) and a model with two population means (12). Within the minimum message length framework, the optimal hypothesis is the one resulting in the briefest encoding. Thus, if $\delta < 0$, the hypothesis of a single population mean for the Behrens–Fisher problem is deemed optimal, and vice versa. The term $\exp(-\delta)$ can be directly interpreted as the posterior odds in favour of the model with a common population mean. Large values of $\exp(-\delta)$, perhaps ten or greater, indicate strong preference to the simpler model with a single population mean.

4 Simulation and Discussion

The minimum message length procedure was compared against well-known methods on the Behrens–Fisher hypothesis testing problem (see Examples 1 and 2) and the estimation of the common mean problem (see Example 3).

Example 1. Hypothesis testing. The minimum message length solution to the Behrens–Fisher problem was compared against two alternative approaches: (1) a popular frequentist method, and (2) a Bayesian approach [15]. For the frequentist procedure, the null distribution of $\Delta = (\mu_2 - \mu_1)$ was approximated by a Student t density with Satterthwaite's approximation for the relevant degrees of freedom. In the Bayesian procedure, the prior density for the parameters was taken to be the vague reference prior distribution $\pi(\boldsymbol{\theta}) \propto (\tau_1 \tau_2)^{-1}$ for $\boldsymbol{\mu}$ and $\log \boldsymbol{\tau}$ and the Cochran and Cox method was used to approximate the posterior density of Δ (see equation (33) in [5]). Interestingly, a Bayesian procedure with these prior distributions is numerically equivalent to Fisher's fiducial inference procedure for the Behrens–Fisher problem [5].

The testing setup was as follows: (1) randomly choose the true hypothesis ($\mu_1 = \mu_2$) or ($\mu_1 \neq \mu_2$) with equal probability, (2) sample all parameters from their respective prior densities, and (3) sample data $\mathbf{y}$ from the resulting model with $(n_1, n_2) \in \{5, 10, 25, 50, 100, 500\}^2$. The population variances were sampled from the compact set $\tau_i \in \Xi = [0{\cdot}01, 20]$ for $(i = 1, 2)$; the normalisation constant is then $\Omega \approx 57{\cdot}77$. The population mean(s) were uniformly sampled from $(-5 \leq \mu_i \leq 5)$. For each data set, the Wallace–Freeman, frequentist and Bayesian procedures were asked to nominate which of the two possible hypotheses was used to generate the data. To aid in comparison, the minimum message length tests were completed first and the resultant empirical type I error rate was chosen as the significance level for the frequentist and Bayesian procedures. This is necessary as the MML principle has no in-built notion of type I and type II error rates. By controlling the type I error rate, the performance of the three methods can be compared solely on the number of type II errors. The number of times each criterion selected the generating hypothesis was then recorded (see Table 1). The entire procedure was repeated for 10^4 iterations. As an alternative, the probability of choosing the hypothesis ($\mu_1 = \mu_2$) was set to the observed type

I error rate of the MML procedure, and the experiments repeated as before. This did not result in any significant changes to results and their interpretation.

The MML criterion obtained superior scores when there was an imbalance in the generated data; for example, $(n_1, n_2) = (5, 500)$. When the sample size was small, $(n_1, n_2 < 25)$, the MML criterion obtained a higher proportion of correct classifications compared to both the frequentist and the Bayesian approaches. These differences in performance may potentially be attributed to the accuracies of the various approximations used in the three procedures as well as the choice of prior density over the population means. As expected, all tested criteria performed well for moderate and large samples sizes.

Table 1. Proportion of times each criterion correctly selected the data generating hypothesis

Criterion	n_1	n_2					
		5	10	25	50	100	500
MML	5	82·9	84·8	86·4	86·6	86·4	85·9
	10	85·0	86·9	87·8	89·4	89·8	90·0
	25	85·9	89·2	90·7	92·3	92·5	93·2
	50	86·9	89·3	91·8	93·4	93·6	94·8
	100	86·8	90·2	92·5	93·8	95·0	96·1
	500	86·5	89·8	93·7	95·1	96·0	97·3
Student t	5	81·4	83·2	84·7	84·3	83·7	82·6
	10	83·5	86·3	87·4	88·7	88·9	89·3
	25	84·1	88·3	90·5	91·5	92·1	92·6
	50	84·9	88·3	91·6	93·1	93·3	94·5
	100	83·9	88·7	92·1	93·7	95·0	95·9
	500	82·7	88·0	93·2	94·8	96·1	97·2
Bayesian	5	81·3	83·2	84·7	84·2	83·6	82·4
	10	83·2	86·4	87·4	88·7	88·9	89·2
	25	83·9	88·3	90·5	91·5	92·2	92·6
	50	84·8	88·4	91·6	93·1	93·3	94·5
	100	83·6	88·7	92·0	93·7	95·0	95·9
	500	82·5	88·0	93·2	94·8	96·1	97·2

Example 2. Driving time data ([16], p. 83; [5]). The driving times along two different routes from a person's house to work were measured; there were $n_1 = 5$ trips for the first route and $n_2 = 11$ trips for the second route. The complete data set is given below

$$\mathbf{y}_1 = (6{\cdot}5, 6{\cdot}8, 7{\cdot}1, 7{\cdot}3, 10{\cdot}2),$$
$$\mathbf{y}_2 = (5{\cdot}8, 5{\cdot}8, 5{\cdot}9, 6{\cdot}0, 6{\cdot}0, 6{\cdot}0, 6{\cdot}3, 6{\cdot}3, 6{\cdot}4, 6{\cdot}5, 6{\cdot}5).$$

The task is to determine whether there is a difference in the average travel times for the two routes. For this problem, both the frequentist and Bayesian procedures find that the difference between the two means is not significant at a significance level of $\alpha = 0{\cdot}05$. The Wallace–Freeman codelengths for the model with a common population mean and the model with two different population means were 19·30 nits and 20·14 nits respectively. Thus, the MML approach prefers the model with one population mean with a posterior odds of 2·3.

Example 3. Parameter estimation. The performance of the Wallace–Freeman estimator (11) is now compared against the maximum likelihood (ML) estimator on the problem of inferring the common mean of two normal populations with unknown variances; that is, the true hypothesis is assumed to be $(\mu_1 = \mu_2)$ and the MML and ML methods are compared solely on their parameter estimation performance. The testing setup was as follows: (1) sample all parameters $\boldsymbol{\theta} = (\mu, \tau_1, \tau_2)' \in \mathbb{R}^3$ from their respective prior densities, (2) sample data $\mathbf{y}$ from the resulting model with $(n_1, n_2) \in \{5, 10, 25, 50, 100, 500\}^2$. The population variances were sampled from the compact set $\tau_i \in \Xi = [0{\cdot}1, 5]$ for $(i = 1, 2)$; the normalisation constant is then $\Omega \approx 15{\cdot}30$. The common population mean was uniformly sampled from $(-5 \leq \mu \leq 5)$. For each data set, the Wallace–Freeman (11) estimator and the maximum likelihood estimator were used to infer the parameters $\boldsymbol{\theta}$. The entire procedure was repeated for 10^5 iterations. Following each iteration, the Kullback–Leibler (KL) divergence [17] of the two estimators from the data generating distribution was computed. The results expressed in terms of the median KL divergence are presented in Table 2. The Wallace–Freeman estimator is clearly superior to the maximum likelihood estimator for small samples sizes $(n_1, n_2 \leq 25)$. The two criteria performed similarly when there was a large imbalance in the sample sizes which agrees with the results presented in [18]. Both the Wallace–Freeman and ML estimators performed identically for all samples sizes $(n_1, n_2 \geq 50)$.

Table 2. The median Kullback–Leibler divergence computed over 10^5 iterations between the data generating distribution and the MML and ML estimators

Estimator	n_1	n_2					
		5	10	25	50	100	500
MML	5	0·329	0·208	0·126	0·094	0·074	0·055
	10	0·207	0·137	0·082	0·059	0·045	0·029
	25	0·127	0·082	0·050	0·035	0·025	0·014
	50	0·095	0·060	0·035	0·024	0·017	0·009
	100	0·074	0·045	0·025	0·017	0·012	0·006
	500	0·055	0·029	0·014	0·009	0·006	0·002
ML	5	0·416	0·239	0·136	0·098	0·077	0·055
	10	0·237	0·149	0·086	0·061	0·046	0·029
	25	0·137	0·086	0·051	0·036	0·025	0·014
	50	0·099	0·062	0·036	0·025	0·017	0·009
	100	0·077	0·046	0·026	0·017	0·012	0·006
	500	0·056	0·029	0·014	0·009	0·006	0·002

5 Extensions

It is relatively straightforward to extend the Wallace–Freeman codelength formulae from Section 3 to the generalised Behrens–Fisher problem. The data now comprises $(d > 2)$ mutually independent samples of i.i.d. sequences generated

by the following Gaussian model:

$$y_{ij} \sim N(\mu_i, \tau_i), \tag{14}$$

where $(i = 1, \ldots, d; j = 1, \ldots, n_i)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)'$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_d)'$ are the unknown sequence means and variances respectively. The complete data set is denoted by $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_d')'$ and comprises $n = (n_1 + \cdots + n_d)$ samples. The generalised Behrens–Fisher problem is testing whether or not there exists a difference between the population means; that is, whether $(\mu_1 = \mu_2 = \cdots = \mu_d)$.

Consider first the Wallace–Freeman codelength under the assumption that there exists a common population mean across the d data sequences. The parameter vector is $\boldsymbol{\theta} = (\mu, \boldsymbol{\tau}')' \in \mathbb{R}^{d+1}$, and assuming the uniform prior density for the population mean (8) and conjugate scale invariant prior densities for the population variances $\pi_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = (\Omega_d \tau_1 \tau_2 \ldots \tau_d)^{-1}$, the total Wallace–Freeman codelength is

$$\frac{n}{2}\log 2\pi + \frac{1}{2}\sum_{i=1}^{d}\left(n_i \log \tau_i + \frac{1}{\tau_i}\sum_{j=1}^{n_i}(y_{ij}-\mu)^2\right) + \frac{1}{2}\log\left(\sum_{i=1}^{d}\frac{n_i}{\tau_i}\right) + \frac{1}{2}\log\left(\frac{\Omega_d^2(\mathbf{y}'\mathbf{y})}{n}\prod_{i=1}^{d} n_i\right) + c(d+1),$$

where Ω_d is a suitable normalisation constant. As in Section 3, the total codelength must be numerically minimised for $(\mu, \boldsymbol{\tau}')'$.

The Wallace–Freeman codelength for the model in which $(\mu_1, \ldots, \mu_d)$ are free parameters is easily derived from (12). The feasible parameter set for the d population means is now a hyper-ellipsoid

$$\Lambda_d = \left\{(\mu_1, \mu_2, \ldots, \mu_d) : \sum_{i=1}^{d} n_i \mu_i^2 \leq \mathbf{y}'\mathbf{y}\right\},$$

resulting in the uniform prior density

$$\pi_{\boldsymbol{\mu}}(\boldsymbol{\mu}) = \frac{1}{\mathrm{vol}(\Lambda_d)} = \frac{\Gamma(d/2+1)}{(\pi \mathbf{y}'\mathbf{y})^{(d/2)}}\left(\prod_{i=1}^{d} n_i\right)^{(1/2)}, \quad \boldsymbol{\mu} \in \Lambda_d.$$

where $\Gamma(\cdot)$ is the gamma function. Invariant conjugate scale prior densities are again used for the population variances. The total Wallace–Freeman codelength for the model with d population means is then

$$\frac{n}{2}\log 2\pi + \frac{1}{2}\left(\sum_{i=1}^{d}(n_i - 1)\log \hat{\tau}_i\right) + \frac{n-d}{2} + \frac{d}{2}\log(\pi \mathbf{y}'\mathbf{y}) + \frac{1}{2}\sum_{i=1}^{d}\log\left(\frac{n_i}{2}\right) - \log\Gamma(d/2+1) + \log\Omega_d + c(2d)$$

where the Wallace–Freeman estimates of $\boldsymbol{\tau}$ are equivalent to those in (13).

Testing the hypothesis of the existence of a common mean across the d data sequences follows the same procedure as per Section 3.3. This process can also be extended to other hypothesis tests, such as testing for a common population variance. Furthermore, an analysis of the multivariate Behrens–Fisher problem under the minimum message length framework is possible given Wallace–Freeman codelengths for a multivariate normal distribution ([6], pp. 261–264).

A Prior distribution over the population means

Consider the standard linear regression model for data $\mathbf{y} \in \mathbb{R}^n$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{X}$ is a $(n \times p)$ design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ denotes the coefficient vector and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ are zero mean i.i.d. Gaussian variates with covariance matrix $\boldsymbol{\Sigma}$. The Behrens–Fisher problem is then a special case of the linear regression model for a suitable choice of design matrix $\mathbf{X}$ and noise covariance matrix $\boldsymbol{\Sigma}$. The aim here is to derive a prior density $\pi(\cdot)$ for the coefficients $\boldsymbol{\beta}$ which can be used in the absence of any subjective knowledge.

Ideally, the prior density should give each combination of regression coefficients the same probability. A possible choice is to use an independent uniform prior for each coefficient, however this requires arbitrary bounding of the parameter space and the resulting model selection criteria would be highly dependent on the chosen support. An alternative approach is to exploit the fact that the observed data $\mathbf{y}$ are generated by the model

$$\mathbf{y} = \mathbf{y}_* + \boldsymbol{\varepsilon},$$

where $\mathbf{y}_*$ denotes the "true" signal. Note that

$$\mathrm{E}\left[\mathbf{y}'\mathbf{y}\right] = \mathbf{y}_*'\mathbf{y}_* + \mathrm{tr}\left(\boldsymbol{\Sigma}\right).$$

where $\mathrm{E}[\cdot]$ denotes the expectation operator. Having observed $\mathbf{y}$, one can form an estimate, say $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, of the true signal. Since the covariance matrix $\boldsymbol{\Sigma}$ is strictly positive definite, it is expected that the estimate, $\hat{\mathbf{y}}'\hat{\mathbf{y}}$, should satisfy

$$\mathbf{y}'\mathbf{y} \geq \hat{\mathbf{y}}'\hat{\mathbf{y}} = \hat{\boldsymbol{\beta}}'\left(\mathbf{X}'\mathbf{X}\right)\hat{\boldsymbol{\beta}}. \tag{15}$$

The least-squares estimates, the James-Stein shrunken least squares estimates [19], and other estimates that obtain minimax squared error risk satisfy restriction (15), which offers strong support for this choice of prior. Hence, the feasible parameter space for the regression coefficients is given by the hyper-ellipsoid

$$\Lambda = \left\{\boldsymbol{\beta} : \boldsymbol{\beta}'\left(\mathbf{X}'\mathbf{X}\right)\boldsymbol{\beta} \leq \mathbf{y}'\mathbf{y}\right\}.$$

A suitable joint prior density for the regression coefficients is then

$$\pi(\boldsymbol{\beta}) = \frac{1}{\mathrm{vol}(\Lambda)} = \frac{\Gamma(p/2+1)\sqrt{|\mathbf{X}'\mathbf{X}|}}{(\pi\mathbf{y}'\mathbf{y})^{p/2}}, \quad \boldsymbol{\beta} \in \Lambda \tag{16}$$

where $\Gamma(\cdot)$ is the gamma function. This is a uniform prior over the volume of the feasible set Λ and is equivalent to assigning the same probability mass to each possible combination of regressors. The prior density (16) has been used to derive an MML model selection criterion for linear regression models that has the desirable property of being invariant under full-rank affine transformations of the design matrix [14].

References

1. Scheffé, H.: On solutions of the Behrens–Fisher problem, based on the t-distribution. The Annals of Mathematical Statistics **14**(1) (March 1943) 35–44
2. Satterthwaite, F.E.: An approximate distribution of estimates of variance components. Biometrics Bulletin **2**(6) (December 1946) 110–114
3. Fisher, R.A.: Inverse probability. Proceedings of the Cambridge Philosophical Society **26** (1930) 528–535
4. Fisher, R.A.: The fiducial argument in statistical inference. Annals of Eugenics **6** (1935) 391–398
5. Kim, S.H., Cohen, A.S.: On the Behrens–Fisher problem: A review. Journal of Educational and Behavioral Statistics **23**(4) (1998) 356–377
6. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length. First edn. Information Science and Statistics. Springer (2005)
7. Wallace, C.S., Boulton, D.M.: An information measure for classification. Computer Journal **11**(2) (August 1968) 185–194
8. 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
9. Solomonoff, R.J.: A formal theory of inductive inference. Information and Control **7**(2) (1964) 1–22, 224–254
10. Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Problems of Information Transmission **1**(1) (1965) 1–7
11. Chaitin, G.J.: A theory of program size formally identical to information theory. Journal of the Association for Computing Machinery **22**(3) (1975) 329–340
12. Wallace, C., Boulton, D.: An invariant Bayes method for point estimation. Classification Society Bulletin **3**(3) (1975) 11–34
13. Dowe, D.L., Wallace, C.S.: Resolving the Neyman-Scott problem by minimum message length. In: Proc. 28th Symposium on the interface. Volume 28 of Computing Science and Statistics., Sydney, Australia (1997) 614–618
14. Schmidt, D., Makalic, E.: MML invariant linear regression. In: The 22nd Australasian Joint Conference on Artificial Intelligence, Melbourne, Australia (2009) 312–321
15. Jeffreys, H.: Note on the Behrens–Fisher formula. Annals of Eugenics **10** (1940) 48–51
16. Lehmann, E.L.: Nonparametrics: Statistical methods based on ranks. Mcgraw–Hill (1974)
17. Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics **22**(1) (March 1951) 79–86
18. Pal, N., Lin, J.J., Chang, C.H., Kumar, S.: A revisit to the common mean problem: Comparing the maximum likelihood estimator with the Graybill-Deal estimator. Computational Statistics & Data Analysis **51**(12) (2007) 5673–5681
19. Sclove, S.L.: Improved estimators for coefficients in linear regression. Journal of the American Statistical Association **63**(322) (June 1968) 596–606