Minimum Message Length Analysis of the Behrens–Fisher Problem

Enes Makalic and Daniel F Schmidt

Centre for MEGA Epidemiology
The University of Melbourne

Solomonoff 85th Memorial Conference, 2011

Outline

Outline

Problem Description (1)

- We have two mutually independent sequences of i.i.d. data

$$\mathbf{y}_1 = (y_{11}, \ldots, y_{1n_1})' \text{ and } \mathbf{y}_2 = (y_{21}, \ldots, y_{1n_2})'$$

- Data assumed to be generated by the Gaussian model

$$y_{ij} \sim N(\mu_i, \tau_i) \quad (i = 1, 2; j = 1, \ldots, n_i)$$

- The sequence means and variances are unknown

$$\boldsymbol{\mu} = (\mu_1, \mu_2)' \text{ and } \boldsymbol{\tau} = (\tau_1, \tau_2)'$$

Problem Description (2)

- Task
  - Is there a difference between the two population means?

$$\mu_1 = \mu_2?$$

- Existing solutions
  - Frequentist (based on Student t pivot)
  - Bayes factor
- We will use Minimum Message Length (MML)

Problem Description (2)

- Task
  - Is there a difference between the two population means?

$$\mu_1 = \mu_2?$$

- Existing solutions
  - Frequentist (based on Student t pivot)
  - Bayes factor
- We will use Minimum Message Length (MML)

Problem Description (2)

- Task
  - Is there a difference between the two population means?

$$\mu_1 = \mu_2?$$

- Existing solutions
  - Frequentist (based on Student t pivot)
  - Bayes factor
- We will use Minimum Message Length (MML)

Outline

Introduction (1)

- Practical implementation of theory of inductive inference
  - Initially proposed by Solomonoff
  - Model that yields the briefest encoding of data in a hypothetical message is optimal
- The message comprises
  - the *assertion*, statement describing a particular model $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$
  - the *detail*, encoding of the data $\mathbf{y}$ using the assertion model $\boldsymbol{\theta}$

Introduction (2)

- The total length of the two-part message, $I(\boldsymbol{\theta}, \mathbf{y})$, is sum of the lengths of the assertion and the detail

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

- MML advocates choosing model $\boldsymbol{\theta}$ that minimises the codelength of the two-part message

MML87 (1)

- The Wallace–Freeman, or MML87 codelength, for model $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$ and data $\mathbf{y}$ is

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

  - $p(\mathbf{y}|\boldsymbol{\theta})$ denotes the likelihood function
  - $\pi(\cdot)$ is a prior distribution over the parameter space $\Theta \subset \mathbb{R}^k$
  - $\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is the Fisher information matrix
  - κ_k is the normalised second moment of an optimal quantising lattice in k-dimensions

MML87 (2)

- The model that minimises $I_{87}(\mathbf{y}, \boldsymbol{\theta})$ is (a posteriori) most plausible

$$\hat{\boldsymbol{\theta}}_{87}(\mathbf{y}) = \arg\min_{\boldsymbol{\theta} \in \Theta} \{I_{87}(\mathbf{y}, \boldsymbol{\theta})\}$$

- MML treats both parameter estimation and model class selection on the same footing
- Wallace–Freeman codelengths are invariant under a smooth, one-to-one reparameterisation of the parameters

Outline

MML Solution

- The MML solution to the Behrens–Fisher problem requires codelength of data under
  - Shared mean model ($\mu_1 = \mu_2$)
  - Different means model ($\mu_1 \neq \mu_2$)
- The model resulting in the shortest codelength is preferred
  - Let $\delta = I_{87}(\mathbf{y}, \hat{\mu}, \hat{\boldsymbol{\tau}}) - I_{87}(\mathbf{y}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\tau}})$
    - If $\delta < 0$, single population mean preferred
  - The term $\exp(-\delta)$ is the posterior odds in favour of the model with common population mean

Shared population mean (1)

- Let $\mathbf{y} = (\mathbf{y}_1', \mathbf{y}_2')'$ denote the observed data
- Parameters to be estimated

$$\boldsymbol{\theta} = (\mu, \boldsymbol{\tau}')' \in \mathbb{R}^3, \quad \boldsymbol{\tau} = (\tau_1, \tau_2)'$$

Shared population mean (2)

- The negative log-likelihood function

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

- The determinant of the Fisher information matrix

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

Shared population mean (3)

- Prior densities over the parameters $\boldsymbol{\theta}$

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

  - Population variances

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

  - Population mean

$$\begin{aligned} \pi(\mu) &= \frac{1}{\mathrm{vol}(\Lambda_1)} = \left(\frac{n}{4\mathbf{y}'\mathbf{y}}\right)^{1/2}, \quad \mu \in \Lambda_1 \\ \Lambda_1 &= \{\mu : n\mu^2 \leq \mathbf{y}'\mathbf{y}\} \end{aligned}$$

Shared population mean (4)

- Prior density for the population mean
  - Observed data $\mathbf{y}$ is generated from the model

$$\mathbf{y} = \mathbf{y}_* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_n)$$

  - One can show that

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

  - An estimate $(\mathbf{1}_n\hat{\mu})$ of μ of $\mathbf{y}_*$ should then satisfy

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

Shared population mean (5)

- Total Wallace–Freeman codelength, $I_{87}(\mathbf{y}, \mu, \boldsymbol{\tau})$

$$\frac{n}{2}\log 2\pi + \frac{1}{2}\sum_{i=1}^{2}\left(n_i \log \tau_i + \frac{1}{\tau_i}\sum_{j=1}^{n_i}(y_{ij}-\mu)^2\right) + \frac{1}{2}\log\left(\frac{n_1}{\tau_1}+\frac{n_2}{\tau_2}\right) + \frac{1}{2}\log\left(\frac{\Omega^2(\mathbf{y}'\mathbf{y})}{n}\prod_{i=1}^{2} n_i\right) - 2{\cdot}32$$

- Wallace–Freeman parameter estimates

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

Different population means (1)

- Parameters to be estimated

$$\boldsymbol{\theta} = (\boldsymbol{\mu}', \boldsymbol{\tau}')' \in \mathbb{R}^4 \text{ where } \boldsymbol{\mu} = (\mu_1, \mu_2)', \boldsymbol{\tau} = (\tau_1, \tau_2)'$$

- The negative log-likelihood function

$$-\log p(\mathbf{y}|\boldsymbol{\theta}) = \frac{n_1}{2} \log 2\pi\tau_1 + \frac{n_2}{2} \log 2\pi\tau_2 + \sum_{i=1}^{2} \sum_{j=1}^{n_i} \frac{(y_{ij} - \mu_i)^2}{2\tau_i}$$

- The determinant of the Fisher information matrix is

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

Different population means (2)

- Prior densities over the parameters $\boldsymbol{\theta}$

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

  - Population variances

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

  - Population means

$$\begin{aligned} \pi(\boldsymbol{\mu}) &= \frac{1}{\mathrm{vol}(\Lambda_2)} = \frac{\sqrt{n_1 n_2}}{\pi \mathbf{y}'\mathbf{y}}, \quad \boldsymbol{\mu} \in \Lambda_2 \\ \Lambda_2 &= \left\{ (\mu_1, \mu_2) : \sum_{i=1}^{2} n_i \mu_i^2 \leq \mathbf{y}'\mathbf{y} \right\} \end{aligned}$$

Different population means (3)

- Total Wallace–Freeman codelength, $I_{87}(\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\tau})$

$$\frac{n}{2}\log 2\pi + \frac{1}{2}\left(\sum_{i=1}^{2}(n_i - 1)\log\hat{\tau}_i\right) + \frac{n-2}{2} + \log\left(\mathbf{y}'\mathbf{y}\sqrt{n_1 n_2}\Omega\pi/2\right) - 3{\cdot}14$$

- Wallace–Freeman parameter estimates

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

Outline

Simulation and Discussion (1)

- MML approach empirically compared to standard procedures using artificial data
  - Hypothesis testing
  - Parameter estimation

Simulation and Discussion (2)

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

Simulation and Discussion (3)

- 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