MML estimation of the multivariate normal distribution

Enes Makalic and Daniel F. Schmidt

July 24, 2014

1 Introduction

Let $\mathcal{D} = (\mathbf{y_1}, \ldots, \mathbf{y}_n)$ denote n i.i.d. data points assumed to be sampled from a p-dimensional Gaussian distribution $N_p(\boldsymbol{\mu}, \Psi)$ with density function

$$p(\mathbf{y}|\boldsymbol{\mu}, \Psi) = \frac{1}{(2\pi)^{p/2}|\Psi|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu})'\Psi^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right), \quad (i = 1, \ldots, n) \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean vector and $\Psi \in \mathbb{R}^{p\times p}$ is the variance-covariance matrix. The aim of this exercise is to derive the message length formula and the corresponding MML estimates for the multivariate Gaussian distribution. The derivation here is similar to that of Wallace [5] (pp. 261–264) and includes additional detail for derivations which were omitted from the original.

Minimal sufficient statistics for the multivariate Gaussian are the sample mean $\bar{\mathbf{y}} \in \mathbb{R}^p$ and sample variance-covariance matrix $\mathbf{S} \in \mathbb{R}^{p\times p}$ which are

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i, \quad \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' = \frac{1}{n}\left(\sum_{i=1}^{n}(\mathbf{y}_i\mathbf{y}_i') - n\bar{\mathbf{y}}'\bar{\mathbf{y}}\right) \tag{2}$$

The negative log-likelihood function of the data $\mathcal{D} = (\mathbf{y_1}, \ldots, \mathbf{y}_n)$ is:

$$-\log p(\mathcal{D}|\boldsymbol{\mu}, \Psi) = \frac{np}{2}\log(2\pi) + \frac{n}{2}\log|\Psi| + \frac{1}{2}\mathrm{tr}(\Psi^{-1}\mathbf{Z}) \tag{3}$$

where

$$\mathbf{Z} = \sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu})(\mathbf{y}_i - \boldsymbol{\mu})' \tag{4}$$

We can write the quadratic form in terms of the sufficient statistics as

$$\mathrm{tr}(\Psi^{-1}\mathbf{Z}) = n(\bar{\mathbf{y}} - \boldsymbol{\mu})'\Psi^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}) + n\mathrm{tr}(\Psi^{-1}\mathbf{S}) \tag{5}$$

The steps required to derive the quadratic form in terms of the mean and variance-covariance matrix are given in Appendix A. In the next section, we derive a message length formula for the multivariate Gaussian using the popular Wallace–Freeman message length approximation [3].

1.1 Minimum message length

Minimum message length (MML) [2, 4, 3, 5] principle of inductive inference states that the best model is one which results in the shortest codelength of the data. The codelength has two parts: (1) the *assertion*, stating a model for the data from a set of candidate models, and (2) the *detail* which encodes the data using the model that was named in the assertion. The assertion and the detail are commonly denoted as $I_{87}(\boldsymbol{\theta})$ and $I_{87}(\mathbf{y}|\boldsymbol{\theta})$ respectively. The most commonly used form of MML is the Wallace–Freeman approximation [3], or the MML87 approximation, which states that the codelength, $I_{87}(\mathbf{y}, \boldsymbol{\theta})$, of model $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$ and data $\mathbf{y} \in \mathbb{R}^n$ is The most popular message length approximation used in practice is due to Wallace and Freeman [3] and is often referred to as MML87. The MML87 approximation states that the codelength, $I_{87}(\mathbf{y}, \boldsymbol{\theta})$, of model $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$ and data $\mathbf{y} \in \mathbb{R}^n$ is

$$I_{87}(\mathbf{y}, \boldsymbol{\theta}) = \underbrace{-\log \pi(\boldsymbol{\theta}) + \frac{1}{2}\log |\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})| + \frac{d}{2}\log \kappa_d}_{I_{87}(\boldsymbol{\theta})} + \underbrace{\frac{d}{2} - \log p(\mathbf{y}|\boldsymbol{\theta})}_{I_{87}(\mathbf{y}|\boldsymbol{\theta})} \tag{6}$$

where $\pi(\cdot)$ denotes a prior distribution over the support $\boldsymbol{\Theta}$, $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function, $\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is the $(d \times d)$ Fisher information matrix and $\kappa_d > 0$ is a dimensionality constant. Following Wallace (p. 237, [5]), the dimensionality constant is well approximated by

$$\frac{d}{2}(\log \kappa_d + 1) \approx -\frac{d}{2}\log(2\pi) + \frac{1}{2}\log(d\pi) + \psi(1) \tag{7}$$

where $\psi(\cdot)$ is the digamma function. MML is a Bayesian principle and requires stating a prior density over the free model parameters. Under MML87, the model $\hat{\boldsymbol{\theta}}_{87}(\mathbf{y})$ which minimises (6) is chosen as the most a posteriori likely explanation of the data $\mathbf{y}$, in view of the chosen prior density $\pi(\cdot)$.

MML treatment of the Gaussian distribution requires suitable priors for all parameters and determinant of the Fisher information matrix. Our first derivation of MML formulae is with respect to the variance-covariance matrix Ψ. This is in contrast to Wallace [5] where the precision matrix was used instead. In Section 1.6, we present additional derivations with respect to the precision matrix and using the prior distributions for model parameters that were employed by Wallace.

1.2 Fisher information

The Fisher information matrix for a multivariate Gaussian is

$$n \begin{pmatrix} \Psi^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2} D_p'(\Psi^{-1} \otimes \Psi^{-1}) D_p \end{pmatrix} \tag{8}$$

where D_p is a $p^2 \times p(p+1)/2$ duplication matrix [1] (pp. 311–317). Mathematical details of the Fisher information derivation can be found in [1] (pp. 387–391). Since

$$|D_p'(\Psi^{-1} \otimes \Psi^{-1}) D_p| = 2^{p(p-1)/2} |\Psi|^{-(p+1)} \tag{9}$$

(see, e.g., [1], pp. 317), the determinant of the Fisher information matrix is

$$\mathcal{J}(\boldsymbol{\mu}, \Psi) = \left(n^p |\Psi|^{-1}\right) \left(\left(\frac{n}{2}\right)^{p(p+1)/2} 2^{p(p-1)/2} |\Psi|^{-(p+1)} \right) = \frac{n^p n^{p(p+1)/2}}{2^p |\Psi|^{(p+2)}} \tag{10}$$

In the case of a univariate ($p = 1$) Gaussian $N_1(\mu, \tau)$ with mean $\mu \in \mathbb{R}$ and variance $\tau > 0$, equation (10) simplifies to

$$\mathcal{J}(\mu, \tau) = \frac{n^2}{2\tau^3}. \tag{11}$$

1.3 Prior distributions

A possible uninformative prior for the parameters $(\boldsymbol{\mu}, \Psi)$ is

$$\pi_0(\boldsymbol{\mu}, \Psi) \propto |\Psi|^{-1} \tag{12}$$

which is improper and must be normalised over some range. Following Wallace [5] (pp. 261–264), we construct an invariant conjugate prior for the parameters as follows. Consider prior data $\{(m_0, \boldsymbol{\mu}_0), (m_1, \Psi_1)\}$ comprising m_0 data vectors with sample mean $\boldsymbol{\mu}_0$ and m_1 data vectors having sample variance-covariance matrix Ψ_1 about their mean. Combining the prior *data*, with the uninformative prior (12), we have the conjugate prior

$$\pi(\boldsymbol{\mu}, \Psi | m_0, \boldsymbol{\mu}_0, m_1, \Psi_1) \propto \pi_0(\boldsymbol{\mu}, \Psi) \pi_\mu(\boldsymbol{\mu}_0 | \boldsymbol{\mu}, \Psi) \pi_\Psi(\Psi_1 | \Psi) \tag{13}$$

which is the (unnormalised) posterior distribution corresponding to the uninformative prior (12) combined with the likelihood of the prior data $\{(m_0, \boldsymbol{\mu}_0), (m_1, \Psi_1)\}$. The prior mean and variance-covariance matrix are selected as follows

$$\begin{aligned} \boldsymbol{\mu}_0 | \boldsymbol{\mu}, \Psi &\sim \mathrm{N}_p\left(\boldsymbol{\mu}, \Psi / m_0\right) && (14) \\ \Psi_1 | \Psi &\sim \mathcal{W}^{-1}(m_1 - 1, \Psi^{-1}) && (15) \end{aligned}$$

where $\mathcal{W}^{-1}(m_1 - 1, \Psi^{-1})$ denotes an inverse Wishart distribution with $(m_1 - 1)$ degrees of freedom and positive-definite scale matrix Ψ^{-1}. This is a proper distribution provided $(m_1 > p)$ and is slightly different to that used by Wallace. The unnormalised prior $\pi(\boldsymbol{\mu}, \Psi | m_0, \boldsymbol{\mu}_0, m_1, \Psi_1)$ is

$$\underbrace{|\Psi|^{-1}}_{\pi_0(\boldsymbol{\mu}, \Psi)} \underbrace{|\Psi / m_0|^{-1/2} \exp\left(-\frac{m_0}{2}(\boldsymbol{\mu}_0 - \boldsymbol{\mu})' \Psi^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu})\right)}_{\pi_\mu(\boldsymbol{\mu}_0 | \boldsymbol{\mu}, \Psi)} \underbrace{|\Psi|^{-(m_1 - 1)/2} |\Psi_1|^{-(m_1 + p)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\Psi^{-1} \Psi_1^{-1})\right)}_{\pi_\Psi(\Psi_1 | \Psi)}$$

which, after simplifying, becomes

$$|\Psi / m_0|^{-1/2} \exp\left(-\frac{m_0}{2}(\boldsymbol{\mu}_0 - \boldsymbol{\mu})' \Psi^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu})\right) |\Psi|^{-(m_1 + 1)/2} |\Psi_1|^{(m_1 + p)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\Psi^{-1} \Psi_1^{-1})\right).$$

The first factor can be written as

$$|\Psi / m_0|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)' (\Psi / m_0)^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right)$$

which is a Gaussian distribution with mean $\boldsymbol{\mu}_0 \in \mathbb{R}^p$ and variance-covariance matrix Ψ / m_0. The second factor,

$$|\Psi|^{-(m_1 + 1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\Psi_1^{-1} \Psi^{-1})\right)$$

implies that the prior for Ψ is an inverse Wishart distribution with $(m_1 - p)$ degrees of freedom. Hence, the prior distribution for the parameters is

$$\pi(\boldsymbol{\mu}, \Psi | m_0, \boldsymbol{\mu}_0, m_1, \Psi_1) \sim \mathrm{N}_p(\boldsymbol{\mu}_0, \Psi / m_0) \times \mathcal{W}^{-1}(m_1 - p, \Psi_1^{-1}). \tag{16}$$

This is a proper distribution provided $m_0 > 0$ and $m_1 > (2p - 1)$.

1.4 MML Estimates

The Fisher information may be adjusted to include the effects of the prior data, yielding

$$\mathcal{J}(\boldsymbol{\mu}, \Psi) = \frac{(n+m_0)^p (n+m_1)^{p(p+1)/2}}{2^p |\Psi|^{(p+2)}} \tag{17}$$

The message length, omitting constants, is

$$\frac{n}{2}\log|\Psi| + \frac{1}{2}\mathrm{tr}(\Psi^{-1}\mathbf{Z}) - \frac{p+2}{2}\log|\Psi| + \frac{m_1+1}{2}\log|\Psi| + \frac{1}{2}\mathrm{tr}(\Psi^{-1}\Psi_1^{-1}) + \frac{1}{2}\mathrm{tr}\left(m_0\Psi^{-1}\mathbf{Z}_0\right) + \frac{1}{2}\log|\Psi| \tag{18}$$

which simplifies to

$$A = \left(\frac{n+m_1-p}{2}\right)\log|\Psi| + \frac{1}{2}\mathrm{tr}\left[\Psi^{-1}\left(\mathbf{Z} + m_0\mathbf{Z}_0 + \Psi_1^{-1}\right)\right] \tag{19}$$

where

$$\mathbf{Z}_0 = (\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)'. \tag{20}$$

The Wallace–Freeman estimates for $(\boldsymbol{\mu}, \Psi)$ are

$$\hat{\boldsymbol{\mu}}_{87} = \frac{m_0\boldsymbol{\mu}_0 + \sum_{i=1}^n \mathbf{y}_i}{n+m_0} \tag{21}$$

$$\hat{\Psi}_{87} = \frac{m_0\hat{\mathbf{Z}}_0 + \Psi_1^{-1} + \hat{\mathbf{Z}}}{n+m_1-p} \tag{22}$$

1.5 Message length

The number of free parameters in the model is $d = p(p+3)/2$. The total message length is

$$I_{87}(\mathcal{D}) = \left(\frac{n+m_1-p}{2}\right)\left(p + \log|\hat{\Psi}_{87}|\right) + \text{const.} \tag{23}$$

where

$$\begin{aligned}\text{const.} = \quad & \frac{(n+1)p}{2}\log(2\pi) + \frac{1}{2}\log\left(m_0^{-p}(n+m_0)^p(n+m_1)^{p(p+1)/2}\right) + \frac{p(m_1-p-1)}{2}\log 2 \\ & + \frac{m_1-p}{2}\log|\Psi_1| + \log\Gamma_p((m_1-p)/2) + \frac{d}{2}(1+\log\kappa_d)\end{aligned}$$

and $m_0 > 0$ and $m_1 > (2p-1)$.

1.6 Precision matrix parametarisation

The message length and parameter estimates derived in Wallace [5] are given in terms of the precision matrix $\Lambda \equiv \Psi^{-1}$ instead of the variance-covariance matrix . The prior distribution suggested by Wallace is also slightly different to our selection (16). In order to match Wallace's derivation, we now re-parameterise the problem and derive MML estimates for $(\boldsymbol{\mu}, \Lambda)$ using an equivalent prior distribution for the parameters. We also assume the availability of prior data $\{(m_0, \boldsymbol{\mu}_0), (m_1, \Lambda_1)\}$ comprising m_0 data vectors with sample mean $\boldsymbol{\mu}_0$ and m_1 data vectors having sample precision matrix Λ_1 about their mean. The negative log-likelihood function in terms of the precision matrix is

$$-\log p(\mathcal{D}|\boldsymbol{\mu}, \Lambda) = \frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Lambda| + \frac{1}{2}\mathrm{tr}(\Lambda\mathbf{Z}). \tag{24}$$

The determinant of the Fisher information matrix in the new parametarisation is:

$$\mathcal{J}_\Lambda(\boldsymbol{\mu}, \Lambda) = |(\mathbf{J}^{-1})'||\mathcal{J}(\boldsymbol{\mu}, \Lambda^{-1})||(\mathbf{J}^{-1})| = \mathcal{J}(\boldsymbol{\mu}, \Lambda^{-1})|\mathbf{J}|^{-2} \tag{25}$$

where $\mathbf{J}$ is the Jacobian of the transformation $g(\mathbf{M}) = \mathbf{M}^{-1}$ evaluated at Λ. Assuming that Ψ is symmetric, the Jacobian is

$$\partial \mathrm{vec}(\Lambda) = -D_p^+(\Psi^{-1} \otimes \Psi^{-1})D_p \partial \mathrm{vech}(\Psi)$$

where D_p is a $p^2 \times p(p+1)/2$ duplication matrix. The determinant of the Jacobian is

$$|\mathbf{J}| = (-1)^{n(n+1)/2}|\Lambda|^{p+1}, \quad |\mathbf{J}|^{-2} = |\Lambda|^{-2(p+1)}$$

which yields the Fisher information

$$\mathcal{J}_\Lambda(\boldsymbol{\mu}, \Lambda) = \frac{(n+m_0)^p (n+m_1)^{p(p+1)/2}}{2^p |\Lambda|^p} \tag{26}$$

in terms of the precision matrix, adjusted for the effects of prior data. In the case of a univariate ($p = 1$) Gaussian $N_1(\mu, \lambda)$ with mean $\mu \in \mathbb{R}$ and precision $\lambda > 0$, the determinant of the Fisher information simplifies to

$$\mathcal{J}(\mu, \lambda) = \frac{n^2}{2\lambda}. \tag{27}$$

where the prior data has not been taken into account.

Following Wallace, an uninformative prior for the parameters $(\boldsymbol{\mu}, \Lambda)$ is

$$\pi_0(\boldsymbol{\mu}, \Lambda) \propto |\Lambda|^{-1} \tag{28}$$

which is improper and must be normalised over some range. As in Section 1.3, combining the prior data with the uninformative prior (28) yields

$$\pi(\boldsymbol{\mu}, \Lambda | m_0, \boldsymbol{\mu}_0, m_1, \Lambda_1) \propto \pi_0(\boldsymbol{\mu}, \Lambda)\pi_\mu(\boldsymbol{\mu}_0 | \boldsymbol{\mu}, \Lambda)\pi_\Lambda(\Lambda_1 | \Lambda) \tag{29}$$

The prior mean and variance-covariance matrix are distributed as

$$\boldsymbol{\mu}_0 | \boldsymbol{\mu}, \Lambda \sim \mathrm{N}_p\left(\boldsymbol{\mu}, (m_0\Lambda)^{-1}\right) \tag{30}$$

$$\Psi_1 | \Psi \sim \mathcal{W}(m_1 - 1, \Lambda^{-1}) \tag{31}$$

where $\mathcal{W}(m_1 - 1, \Lambda - 1)$ denotes a Wishart distribution with $(m_1 - 1)$ degrees of freedom and positive-definite scale matrix Λ^{-1}. After normalisation, the prior distribution is

$$\pi(\boldsymbol{\mu}, \Lambda | m_0, \boldsymbol{\mu}_0, m_1, \Lambda_1) \sim \mathrm{N}_p(\boldsymbol{\mu}_0, (m_0\Lambda)^{-1}) \times \mathcal{W}(m_1 + p - 2, \Psi_1^{-1}). \tag{32}$$

This is a proper distribution provided $m_1 > 1$.

The message length, omitting constants, is

$$-\frac{n}{2}\log|\Lambda| + \frac{1}{2}\mathrm{tr}(\Lambda\mathbf{Z}) - \frac{p}{2}\log|\Lambda| - \frac{m_1 - 3}{2}\log|\Lambda| + \frac{1}{2}\mathrm{tr}(\Lambda_1\Lambda) + \frac{1}{2}\mathrm{tr}\left(m_0\Lambda\mathbf{Z}_0\right) - \frac{1}{2}\log|\Lambda| \tag{33}$$

which simplifies to

$$-\left(\frac{n + m_1 + p - 2}{2}\right)\log|\Lambda| + \frac{1}{2}\mathrm{tr}\left[\Lambda\left(\mathbf{Z} + m_0\mathbf{Z}_0 + \Lambda_1\right)\right] \tag{34}$$

The Wallace–Freeman estimates for $(\boldsymbol{\mu}, \Lambda)$ are

$$\hat{\boldsymbol{\mu}}_{87} = \frac{m_0\boldsymbol{\mu}_0 + \sum_{i=1}^n \mathbf{y}_i}{n + m_0} \tag{35}$$

$$\hat{\Lambda}_{87} = \frac{m_0\mathbf{Z}_0 + \Lambda_1 + \mathbf{Z}}{n + m_1 + p - 2} \tag{36}$$

A Sufficient statistics

We want to show that

$$\text{tr}(\Psi^{-1}\mathbf{Z}) = n(\bar{\mathbf{y}}-\boldsymbol{\mu})'\Psi^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu}) + n\text{tr}(\Psi^{-1}\mathbf{S}) \tag{37}$$

where

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i, \quad \mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i-\bar{\mathbf{y}})(\mathbf{y}_i-\bar{\mathbf{y}})' = \frac{1}{n}\left(\sum_{i=1}^{n}(\mathbf{y}_i\mathbf{y}_i') - n\bar{\mathbf{y}}'\bar{\mathbf{y}}\right) \tag{38}$$

We start by expanding

$$\begin{aligned}
\text{tr}(\Psi^{-1}\mathbf{Z}) &= \text{tr}\left[\Psi^{-1}\left(\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i' - \boldsymbol{\mu}\mathbf{y}_i' - \mathbf{y}_i\boldsymbol{\mu}' + \boldsymbol{\mu}\boldsymbol{\mu}'\right)\right] \\
&= \text{tr}\left[\Psi^{-1}\left(\left(\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i'\right) - \boldsymbol{\mu}\left(\sum_{i=1}^{n}\mathbf{y}_i\right)' - \left(\sum_{i=1}^{n}\mathbf{y}_i\right)\boldsymbol{\mu}' + n\boldsymbol{\mu}\boldsymbol{\mu}'\right)\right] \\
&= \text{tr}\left[\Psi^{-1}\left(\left(\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i'\right) - n\boldsymbol{\mu}\bar{\mathbf{y}}' - n\bar{\mathbf{y}}\boldsymbol{\mu}' + n\boldsymbol{\mu}\boldsymbol{\mu}'\right)\right] \\
&= \text{tr}\left[\Psi^{-1}\left(\left(\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i'\right) - n\bar{\mathbf{y}}'\bar{\mathbf{y}} + n\bar{\mathbf{y}}'\bar{\mathbf{y}} - n\boldsymbol{\mu}\bar{\mathbf{y}}' - n\bar{\mathbf{y}}\boldsymbol{\mu}' + n\boldsymbol{\mu}\boldsymbol{\mu}'\right)\right] \\
&= n\text{tr}\left(\Psi^{-1}\mathbf{S}\right) + n\text{tr}\left[\Psi^{-1}\left(\bar{\mathbf{y}}'\bar{\mathbf{y}} - \boldsymbol{\mu}\bar{\mathbf{y}}' - \bar{\mathbf{y}}\boldsymbol{\mu}' + \boldsymbol{\mu}\boldsymbol{\mu}'\right)\right] \\
&= n\text{tr}\left(\Psi^{-1}\mathbf{S}\right) + n\text{tr}\left(\Psi^{-1}(\boldsymbol{\mu}-\bar{\mathbf{y}})(\boldsymbol{\mu}-\bar{\mathbf{y}})'\right) \\
&= n(\bar{\mathbf{y}}-\boldsymbol{\mu})'\Psi^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu}) + n\text{tr}\left(\Psi^{-1}\mathbf{S}\right)
\end{aligned} \tag{39}$$

B Derivations

Differentiating the message length A w.r.t. to the mean, we get

$$\begin{aligned}
\partial A &= \frac{1}{2}\text{tr}\left[\Psi^{-1}\partial\left(\mathbf{Z} + m_0\mathbf{Z}_0\right)\right] \\
&= \frac{1}{2}\text{tr}\left[\Psi^{-1}\left(\partial\mathbf{Z} + m_0\partial\mathbf{Z}_0\right)\right] \\
&= \frac{1}{2}\text{tr}\left[\Psi^{-1}\left(m_0(\partial\boldsymbol{\mu}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)' + (\boldsymbol{\mu}-\boldsymbol{\mu}_0)\partial\boldsymbol{\mu}') - (\partial\boldsymbol{\mu})\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})' - \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})\partial\boldsymbol{\mu}'\right)\right] \\
&= \frac{1}{2}\text{tr}\left[\Psi^{-1}\partial\boldsymbol{\mu}\left(m_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})\right)'\right] + \frac{1}{2}\text{tr}\left[\Psi^{-1}\left(m_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})\right)\partial\boldsymbol{\mu}'\right] \\
&= \frac{1}{2}\left(m_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})\right)'\Psi^{-1}(\partial\boldsymbol{\mu}) + \frac{1}{2}(\partial\boldsymbol{\mu})'\Psi^{-1}\left(m_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})\right) \\
&= \frac{1}{2}(\partial\boldsymbol{\mu})'\Psi^{-1}\left(m_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})\right) + \frac{1}{2}(\partial\boldsymbol{\mu})'\Psi^{-1}\left(m_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})\right) \\
&= \Psi^{-1}\left(m_0(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})\right)(\partial\boldsymbol{\mu})
\end{aligned}$$

MML estimate of $\boldsymbol{\mu}$ is then

$$\hat{\boldsymbol{\mu}}_{87} = \frac{m_0\boldsymbol{\mu}_0 + \sum_{i=1}^{n}\mathbf{y}_i}{n + m_0} \tag{40}$$

Differentiating the message length w.r.t. Ψ we get

$$\begin{aligned}
\partial A &= \left(\frac{n + m_1 - p}{2}\right)\partial \log|\Psi| + \frac{1}{2}\mathrm{tr}\partial\left[\Psi^{-1}\left(\mathbf{Z} + m_0\mathbf{Z}_0 + \Psi_1^{-1}\right)\right] \\
&= \left(\frac{n + m_1 - p}{2}\right)\mathrm{tr}(\Psi^{-1}\partial\Psi) + \frac{1}{2}\mathrm{tr}\left[\partial\Psi^{-1}\left(\mathbf{Z} + m_0\mathbf{Z}_0 + \Psi_1^{-1}\right)\right] + \frac{1}{2}\mathrm{tr}\left[\Psi^{-1}\partial\left(\mathbf{Z} + m_0\mathbf{Z}_0 + \Psi_1^{-1}\right)\right] \\
&= \left(\frac{n + m_1 - p}{2}\right)\mathrm{tr}(\Psi^{-1}\partial\Psi) - \frac{1}{2}\mathrm{tr}\left[\Psi^{-1}(\partial\Psi)\Psi^{-1}\left(\mathbf{Z} + m_0\mathbf{Z}_0 + \Psi_1^{-1}\right)\right] \\
&= \frac{1}{2}\mathrm{tr}((\partial\Psi)(n + m_1 - p)\Psi^{-1}) - \frac{1}{2}\mathrm{tr}\left[(\partial\Psi)\Psi^{-1}\left(\mathbf{Z} + m_0\mathbf{Z}_0 + \Psi_1^{-1}\right)\Psi^{-1}\right] \\
&= \frac{1}{2}\mathrm{tr}(\partial\Psi)\left((n + m_1 - p)\Psi^{-1} - \Psi^{-1}\left(\mathbf{Z} + m_0\mathbf{Z}_0 + \Psi_1^{-1}\right)\Psi^{-1}\right)
\end{aligned}$$

Solving

$$\begin{aligned}
(n + m_1 - p)\Psi^{-1} - \Psi^{-1}\left(\mathbf{Z} + m_0\mathbf{Z}_0 + \Psi_1^{-1}\right)\Psi^{-1} &= \mathbf{0} \\
\Psi^{-1}\left((n + m_1 - p)\Psi - (\mathbf{Z} + m_0\mathbf{Z}_0 + \Psi_1^{-1})\right)\Psi^{-1} &= \mathbf{0} \\
(n + m_1 - p)\Psi &= (m_0\mathbf{Z}_0 + \Psi_1^{-1} + \mathbf{Z})
\end{aligned}$$

yields

$$\hat{\Psi}_{87} = \frac{m_0\mathbf{Z}_0 + \Psi_1^{-1} + \mathbf{Z}}{n + m_1 - p} \tag{41}$$

References

[1] K. M. Abadir and J. R. Magnus. *Matrix Algebra*. Econometric Exercises. Cambridge University Press, August 2005.

[2] C. S. Wallace and D. M. Boulton. An information measure for classification. *Computer Journal*, 11(2):185–194, August 1968.

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

[4] Chris Wallace and David Boulton. An invariant Bayes method for point estimation. *Classification Society Bulletin*, 3(3):11–34, 1975.

[5] Chris S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Information Science and Statistics. Springer, first edition, 2005.