# Graphical meta analysis: Estimation of covariance matrices from multiple studies

Anders Ellern Bilgrau
abilgrau@math.aau.dk

Poul Svante Eriksen
svante@math.aau.dk

Martin Bøgsted
m_boegsted@dcm.aau.dk

September 24, 2014

**Abstract**

This paper proposes a model and estimators for a common covariance matrix in cases where multiple datasets are present and thus provide a basis for meta analysis of gaussian graphical models. Our approach is inspired by traditional meta analysis using random effects models. We derive the basic properties and estimators of the model and compare our estimators method to the straight-forward approaches of simple averages or "mega analysis" were the datasets are combined, reprocessed, and plugged into the usual estimators. Though only a modest improvement, explicitly accounting for the inter-study variance is superior to the alternative. The model is also shown to be applicable as an intermediate between linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA).

**Keywords:** *meta analysis, covariance estimation, integrative analysis, network integration, structural meta analysis, gaussian graphical modeling, linear discriminant analysis, quadratic discriminant analysis*

# Contents

# 1 Introduction

The fundamental problem in statistics of accurately and precisely estimating the covariance matrix (or its inverse) is notoriously difficult though computationally easy. The usual bias-corrected maximum likelihood estimator (MLE), the sample covariance matrix, have long been known to perform poorly in general due to high variability [3]. The sample covariance is ill-conditioned when the sample size $n$ is less than the number of variables $p$. Because of its central statistical role the list of statistical methods and applications utilizing the estimated covariance matrix is extraordinarily long. Beside the many standard statistical methods such as principal component analysis (PCA), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), more direct applications include [examples in medicine, biology. genetics, finance, forensics, physics, engineering, signal processing]. Among this expanding list of applications is also an increasing number of high-dimensional applications and datasets publicly available at online repositories.

In high-dimensional datasets the number of features $p$ far exceed the number of samples $n$. Since the number of parameters increase quadratically in $p$ and the sample covariance become singular when $p > n$ a plethora of shrinkage and regularization estimators have been proposed to combat the accompanying problems by effectively increasing the degrees of freedom. Instead of attempting to derive still more sophisticated estimators with diminishing improvements we attempt to alleviate the problem from a different angle by effectively increasing $n$ and using more available data.

We are motivated by gene-gene interaction networks in diffuse large B-cell lymphoma (DLBCL) where the covariance matrix contain all information about the conditional dependencies of the genes. As with many other cancers, a large number of DLBCL studies are now available online and hence we wish to use these studies in combination with our own data to arrive at a good estimate of the covariance matrix and the inter-study variation.

# 2 A graphical random effects model

In a vein similar to regular effect-based meta analysis of Choi et al. [2], we think of the the different studies as related but perturbed experiments. The graphical analog to Choi et al. [2] is the following relatively simple graphical random effects model (GREM) of the observations. Let $p$ be the number of features and $k$ the number of studies. We model the each sample for the $i$th study as a $p$-dimensional zero-mean multivariate gaussian vector with the covariance matrix realized from a inverse Wishart distribution, i.e. the hierarchical model,

$$\begin{aligned} \boldsymbol{\Sigma}_i|\boldsymbol{\Psi},\nu &\sim \mathcal{W}^{-1}(\boldsymbol{\Psi},\nu), \\ \boldsymbol{x}|\boldsymbol{\Sigma}_i &\sim \mathcal{N}_p(\mathbf{0}_p,\boldsymbol{\Sigma}_i), \qquad i=1,...,k, \end{aligned} \tag{1}$$

where $\mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$ denotes an inverse Wishart distribution with probability density function (pdf),

$$f(\boldsymbol{\Sigma}_i|\boldsymbol{\Psi}, \nu) = \frac{|\boldsymbol{\Psi}|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p\big(\frac{\nu}{2}\big)}|\boldsymbol{\Sigma}_i|^{-\frac{\nu+p+1}{2}}e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}_i^{-1})},$$

$\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}_i$ are positive semi-definite, $\nu > p-1$, and $\mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_i)$ is a multivariate gaussian distribution with pdf

$$f(\boldsymbol{x}|\boldsymbol{\Sigma}_i) = (2\pi)^{-\frac{p}{2}}\det(\boldsymbol{\Sigma}_i)^{-\frac{1}{2}}e^{-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}}.$$

Throughout this paper, we use generic notation $f(\cdot|\cdot)$ and $f(\cdot)$ for the conditional and unconditional pdf of random variables. In the model, $\nu$ encodes the inter-study variation where the $\boldsymbol{\Sigma}_i$'s concentrate around $\boldsymbol{\Psi}$ for $\nu \to \infty$. I.e. the inter-study variation goes to zero for larger $\nu$. We wish to infer the parameters $\boldsymbol{\Psi}$ and $\nu$ from the observed data.

## 2.1 The likelihood function

Suppose we have $n_i$ i.i.d. observations, $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$, from the $i = 1, ..., k$ independent studies from the model given in (1). Let $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})^\top$ be the $n_i \times p$ (transposed) expression matrix of the $i$'th study where rows correspond to samples and columns to variables. By the independence assumptions, the log-likelihood is given by

$$\begin{aligned}
&\ell\big(\boldsymbol{\Psi}, \nu\big|\boldsymbol{X}_1, ..., \boldsymbol{X}_k\big) = \log f\big(\boldsymbol{X}_1, ..., \boldsymbol{X}_k\big|\boldsymbol{\Psi}, \nu\big) \\
&= \log\int f(\boldsymbol{X}_1, ..., \boldsymbol{X}_k|\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}, \nu)f(\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_k|\boldsymbol{\Psi}, \nu)d\boldsymbol{\Sigma}_1 \cdots d\boldsymbol{\Sigma}_k \\
&= \log\int \prod_{i=1}^{k} f(\boldsymbol{X}_i|\boldsymbol{\Sigma}_i)f(\boldsymbol{\Sigma}_i|\boldsymbol{\Psi}, \nu)d\boldsymbol{\Sigma}_1 \cdots d\boldsymbol{\Sigma}_k \\
&= \sum_{i=1}^{k}\log\int f(\boldsymbol{X}_i|\boldsymbol{\Sigma}_i)f(\boldsymbol{\Sigma}_i|\boldsymbol{\Psi}, \nu)d\boldsymbol{\Sigma}_i.
\end{aligned}$$

Since the inverse Wishart distribution is conjugate to the multivariate gaussian distribution the integral, of which the integrand forms a gaussian-inverse-Wishart distribution, can be evaluated. Hence $\boldsymbol{\Sigma}_i$ can be marginalized out, cf. Appendix A, and we arrive at the following expression for the log-likelihood,

$$\begin{aligned}
&\ell\big(\boldsymbol{\Psi}, \nu\big|\boldsymbol{X}_1, ..., \boldsymbol{X}_k\big) = \sum_{i=1}^{k}\log\frac{|\boldsymbol{\Psi}|^{\frac{\nu}{2}}\Gamma_p\big(\frac{\nu+n_i}{2}\big)}{2^{\frac{n_i p}{2}}|\boldsymbol{\Psi} + \boldsymbol{X}_i^\top\boldsymbol{X}_i|^{\frac{\nu+n_i}{2}}\Gamma_p\big(\frac{\nu}{2}\big)} \\
&= c + \sum_{i=1}^{k}\frac{\nu}{2}\log|\boldsymbol{\Psi}| + \log\Gamma_p\Big(\frac{\nu+n_i}{2}\Big) - \frac{\nu+n_i}{2}\log|\boldsymbol{\Psi} + \boldsymbol{X}_i^\top\boldsymbol{X}_i| - \log\Gamma_p\Big(\frac{\nu}{2}\Big), \quad (2)
\end{aligned}$$

where the constant terms are abbreviated by $c$ and $\Gamma_p$ is the multivariate generalization for the gamma function $\Gamma$, which can be seen in equation (10) of

Appendix B.1. As should be expected, the scatter matrix $\boldsymbol{S}_i = \boldsymbol{X}_i^\top \boldsymbol{X}_i$ and $n_i$ are sufficient statistics for each study.

While we are not able to shown general log-concavity of the log-likelihood, we show that it is log-concave in $\nu$ for fixed $\boldsymbol{\Sigma}$, and hence there exists a unique maxima for the marginal $\ell(\nu)$, cf. Appendix B. Reversely, for fixed $\nu$ we can show, that the hessian in any stationary point (where $\frac{\partial \ell}{\partial \boldsymbol{\Psi}} = 0$) is negative semi-definite and hence a local maxima, cf. Appendix C. This combined with the observation that $\ell(\boldsymbol{\Psi}) \to -\infty$ whenever an eigenvalue $\lambda_i(\boldsymbol{\Psi}) \to 0$, we conjecture that any stationary point indeed a global maximum.

In the following, estimation of $\boldsymbol{\Psi}$ for fixed $\nu$ is considered.

## 2.2 Moment estimate

The moment estimate of $\boldsymbol{\Psi}$ is readily available. It is clear, by construction that the expectation of a single scatter matrix is

$$\mathbb{E}[\boldsymbol{S}_i] = \mathbb{E}[\boldsymbol{X}_i^\top \boldsymbol{X}_i] = \mathbb{E}\left[\mathbb{E}[\boldsymbol{X}_i^\top \boldsymbol{X}_i | \boldsymbol{\Sigma}_i]\right] = \mathbb{E}[n_i \boldsymbol{\Sigma}_i] = n_i \mathbb{E}[\boldsymbol{\Sigma}_i] = n_i \frac{\boldsymbol{\Psi}}{\nu - p - 1}.$$

Thus the estimator

$$\hat{\boldsymbol{\Psi}} = \frac{1}{k} \sum_{i=1}^{k} \frac{\nu - p - 1}{n_i} \boldsymbol{X}_i^\top \boldsymbol{X}_i = (\nu - p - 1) \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n_i} \boldsymbol{S}_i \tag{3}$$

of $\boldsymbol{\Psi}$ is unbiased. This estimate is a scaled average of the empirical covariance matrices.

## 2.3 Maximization of the likelihood

To find the maximizing parameters we differentiate (2) w.r.t. $\boldsymbol{\Psi}$ and equate to zero while assuming $\nu$ known and constant. Assuming $\boldsymbol{\Psi}$ unstructured, then $\nabla_{\boldsymbol{Z}} \log |\boldsymbol{Z}| = \boldsymbol{Z}^{-1}$ and

$$\begin{aligned} \boldsymbol{0} &= \frac{k\nu}{2} \boldsymbol{\Psi}^{-1} - \sum_{i=1}^{k} \frac{\nu + n_i}{2} (\boldsymbol{\Psi} + \boldsymbol{S}_i)^{-1} \\ &= \frac{k\nu}{2} \boldsymbol{\Psi}^{-1} - \sum_{i=1}^{k} \frac{\nu + n_i}{2} \left(\boldsymbol{I} + \boldsymbol{\Psi}^{-1} \boldsymbol{S}_i\right)^{-1} \boldsymbol{\Psi}^{-1}. \end{aligned} \tag{4}$$

While assuming $\boldsymbol{\Psi}$ unstructured is not strictly correct it (in this case) leads to a valid equation. The proper 1. order derivative can in alternative notation be seen in equation (13) and can easily be seen to be equivalent to the above. Equation (4) is equivalent to

$$k\nu \boldsymbol{I} - \sum_{i=1}^{k} (\nu + n_i) \left(\boldsymbol{I} - (-\boldsymbol{\Psi}^{-1} \boldsymbol{S}_i)\right)^{-1} = \boldsymbol{0}.$$

which can be rewritten as

$$k\nu\boldsymbol{I} - \sum_{i=1}^{k}(\nu + n_i)\sum_{l=0}^{\infty}\left(-\boldsymbol{\Psi}^{-1}\boldsymbol{S}_i\right)^l = \boldsymbol{0}.$$

by the Neumann series $\left((\boldsymbol{I}+\boldsymbol{A})^{-1} = \sum_{l=0}^{\infty}\boldsymbol{A}^l\right)$ if $\lim_{l\to\infty}(I - \boldsymbol{\Psi}^{-1}\boldsymbol{S}_i)^l = \boldsymbol{0}$ for all $i$, i.e. the eigenvalues of $\boldsymbol{\Psi}^{-1}\boldsymbol{S}_i$ should be less than 1.

Thus, we can approximate the solution by using the first order expansion ($l = 1$) and solve for $\boldsymbol{\Psi}$

$$\begin{aligned}
\boldsymbol{0} &= k\nu\boldsymbol{I} - \sum_{i=1}^{k}(\nu + n_i)(\boldsymbol{I} - \boldsymbol{\Psi}^{-1}\boldsymbol{S}_i) \\
&= k\nu\boldsymbol{I} - k\nu\boldsymbol{I} - n_\bullet\boldsymbol{I} + \boldsymbol{\Psi}^{-1}\sum_{i=1}^{k}\nu\boldsymbol{S}_i + \boldsymbol{\Psi}^{-1}\sum_{i=1}^{k}n_i\boldsymbol{S}_i \\
&= -n_\bullet\boldsymbol{I} + \boldsymbol{\Psi}^{-1}\sum_{i=1}^{k}(\nu + n_i)\boldsymbol{S}_i
\end{aligned}$$

where $n_\bullet = \sum_{i=1}^{k} n_i$ is the total number of observations. This implies

$$\boldsymbol{\Psi}^{-1}\sum_{i=1}^{k}(\nu + n_i)\boldsymbol{S}_i = n_\bullet\boldsymbol{I}$$

which is equivalent to

$$\hat{\boldsymbol{\Psi}} = \frac{\sum_{i=1}^{k}(\nu + n_i)\boldsymbol{S}_i}{n_\bullet} = \frac{\sum_{i=1}^{k}(\nu + n_i)\boldsymbol{X}_i^\top\boldsymbol{X}_i}{\sum_{i=1}^{k} n_i}, \tag{5}$$

corresponding to a weighted sum of the scatter matrices.

## 2.4 Maximization using the EM algorithm

We now derive the updating scheme of the expectation-maximization (EM) algorithm for fixed $\nu$. Suppose $\hat{\boldsymbol{\Psi}}_{(t)}$ is the current estimate of $\boldsymbol{\Psi}$ and that $\boldsymbol{S}_i = \boldsymbol{X}_i^\top\boldsymbol{X}_i = \sum_{j=1}^{n_i}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top$ is the empirical scatter matrix where $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i})^\top$. We now compute the expectation step of the EM-algorithm.

From (1) we have that,

$$\begin{aligned}
\boldsymbol{S}_i|\boldsymbol{\Sigma}_i &\sim \mathcal{W}_p(\boldsymbol{\Sigma}_i, n_i), \qquad i = 1, ..., k \\
\boldsymbol{\Sigma}_i &\sim \mathcal{W}_p^{-1}(\boldsymbol{\Psi}, \nu)
\end{aligned}$$

Let $\boldsymbol{\Delta}_i = \boldsymbol{\Sigma}_i^{-1}$ be the precision matrix (or, concentration matrix) and let $\boldsymbol{\Theta} = \boldsymbol{\Psi}^{-1}$, then

$$\begin{aligned}
\boldsymbol{S}_i|\boldsymbol{\Delta}_i \sim \mathcal{W}_p(\boldsymbol{\Delta}_i^{-1}, n_i) \qquad &\Leftrightarrow \qquad f(\boldsymbol{S}_i|\boldsymbol{\Delta}_i) \propto |\boldsymbol{\Delta}_i|^{\frac{1}{2}}e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Delta}_i\boldsymbol{S}_i)}, \\
\boldsymbol{\Delta}_i \sim \mathcal{W}_p(\boldsymbol{\Theta}, \nu) \qquad &\Leftrightarrow \qquad f(\boldsymbol{\Delta}_i) \propto |\boldsymbol{\Theta}|^{-\frac{\nu}{2}}e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Theta}^{-1}\boldsymbol{\Delta}_i)}.
\end{aligned} \tag{6}$$

From the conjugacy of the inverse Wishart and the Wishart distribution, we have the posterior distribution for the precision matrix,

$$\boldsymbol{\Delta}_i|\boldsymbol{S}_i \sim \mathcal{W}_p\big((\boldsymbol{\Theta}^{-1} + \boldsymbol{S}_i)^{-1}, n_i + \nu\big)$$

Hence, the expectation step is given by

$$\mathbb{E}[\boldsymbol{\Delta}_i|\boldsymbol{S}_i] = (n_i + \nu)(\boldsymbol{\Theta}^{-1} + \boldsymbol{S}_i)^{-1}.$$

The maximization step, in which the log-likelihood $\ell(\boldsymbol{\Theta}|\boldsymbol{\Delta}_1, ..., \boldsymbol{\Delta}_k)$ is maximized, yields the estimate

$$\hat{\boldsymbol{\Theta}} = \frac{1}{k\nu}\sum_{i=1}^{k}\boldsymbol{\Delta}_i,$$

which is the mean of the scaled precision matrices $\frac{1}{\nu}\boldsymbol{\Delta}_i$. The derivation of this estimate can be seen in Appendix D. The above yield the updating scheme

$$\hat{\boldsymbol{\Theta}}_{(t+1)} = \frac{1}{k\nu}\sum_{i=1}^{k}(n_i + \nu)\left(\big(\hat{\boldsymbol{\Theta}}_{(t)}\big)^{-1} + \boldsymbol{S}_i\right)^{-1} \tag{7}$$

for $\boldsymbol{\Theta}$. The connection to the maximum likelihood estimate is immediately seen through equation (4).

### 2.5 Estimation procedure

We propose an alternating procedure between estimating $\nu$ and $\boldsymbol{\Psi}$ while keeping the other fixed. Given parameters $\hat{\nu}_{(t)}$ and $\hat{\boldsymbol{\Psi}}_{(t)}$ at iteration $t$, we we estimate $\hat{\boldsymbol{\Psi}}_{(t+1)}$ using the fixed $\hat{\nu}_{(t+1)}$. Subsequently, find $\hat{\nu}_{(t+1)}$ by standard one-dimensional numerical optimization procedure keeping $\hat{\boldsymbol{\Psi}}_{(t)}$ fixed. This coordinate ascent-like approach is repeated until convergence. More precisely, in pseudo-code, we propose the algorithm seen in Algorithm 1. The update function $U$ in the algorithm is be defined by the derived estimators above. That is, equations (3), (5), and (7) define $U$ to be the moment, approximate MLE, and EM estimate respectively.

## 3 Numerical results

To assess the model and the stability of the estimation procedure we generated data from (1) for $p = 10$ variables in $k = 3$ studies each with an equal number of observations, $n = n_1 = n_2 = n_3$. We chose the parameters $\nu = 15$ and

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}.$$

**Algorithm 1** Pseudo-code for the GREM estimation procedure

**procedure** GREM COORDINATE ASCENT
  **Input:**
  *Sufficient data:* $(\boldsymbol{S}_1, n_1), ..., (\boldsymbol{S}_k, n_k)$
  *Initial parameters:* $\hat{\boldsymbol{\Psi}}_{(0)}, \hat{\nu}_{(0)}$
  *Convergence criterion:* $\varepsilon > 0$
  **Output:**
  *Parameters:* $\hat{\boldsymbol{\Psi}}_{(t')}, \hat{\nu}_{(t')}$

  *Initialize*: $l_{(0)} \leftarrow \ell(\hat{\boldsymbol{\Psi}}_{(0)}, \hat{\nu}_{(0)})$
  **for** $t = 1, 2, 3, ...$ **do**
    $\hat{\boldsymbol{\Psi}}_{(t)} \leftarrow U\Big(\hat{\boldsymbol{\Psi}}_{(t-1)}, \hat{\nu}_{(t-1)}\Big)$
    $\nu_{(t)} \leftarrow \arg\max_{\nu} \ell\Big(\hat{\boldsymbol{\Psi}}_{(t)}, \nu\Big)$
    $l_{(t)} \leftarrow \ell\Big(\hat{\boldsymbol{\Psi}}_{(t)}, \hat{\nu}_{(t)}\Big)$
    **if** $l_{(t)} - l_{(t-1)} < \varepsilon$ **then**
      **return** $\Big(\hat{\boldsymbol{\Psi}}_{(t)}, \nu_{(t)}\Big)$

The number of observations is each study, $n_i$, was varied range $[5, 14]$.

We measure the precision of the estimates values against the expected covariance matrix given by

$$\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{\Sigma}_i] = \frac{1}{\nu - p - 1}\boldsymbol{\Psi}.$$

Let $\hat{\boldsymbol{\Psi}}$ and $\hat{\nu}$ be the estimates obtained in the model described and let

$$\hat{\boldsymbol{\Sigma}}_{\text{GREM}} = \frac{1}{\hat{\nu} - p - 1}\hat{\boldsymbol{\Psi}}$$

be an estimate of $\boldsymbol{\Sigma}$. We benchmark this against the pooled covariance matrix,

$$\hat{\boldsymbol{\Sigma}}_{\text{pool}} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i}S_i,$$

as a simple alternative estimate of $\boldsymbol{\Sigma}$. We benchmark the two against each other using the following sum of squared errors,

$$\text{SSE}(\hat{\boldsymbol{\Sigma}}) = \sum_{i \leq j}\frac{(\hat{\Sigma}_{ij} - \Sigma_{ij})^2}{\text{var}(\Sigma_{ij})}$$

where

$$\text{var}(\Sigma_{ij}) = n(\Psi_{ij}^2 + \Psi_{ii}\Psi_{jj}).$$

For each $n$ from 5 to 14, the squared sum of errors for each estimator, $\text{SSE}(\hat{\boldsymbol{\Sigma}}_{\text{GREM}})$, and $\text{SSE}(\hat{\boldsymbol{\Sigma}}_{\text{pool}})$ where computed for 500 datasets and the average of these values seen in figure 1 as function of $n$. Unsurprisingly, the model outperforms the simple pool average covariance matrix.

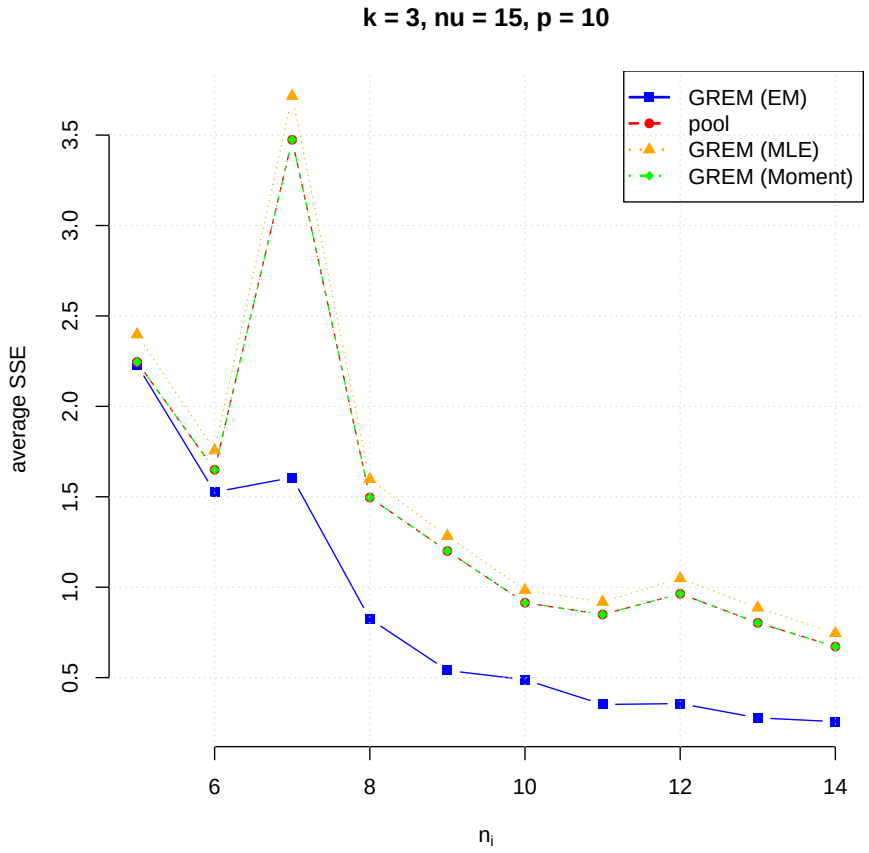



Figure 1: The average sum or squared errors (SSE), of 500 simulations, as a function of the number of samples $n_i$ in each study.

# 4 Applications

## 4.1 Supervised classification

The estimate obtained from the model can be utilized in supervised learning as a intermediate case of linear discriminant analysis (LDA) and quadratic dis-

criminant analysis (QDA) to regularized linear discriminant analysis (RDA).

Let $Y$ be a random variable denoting the class $1, ..., k$ and suppose $\boldsymbol{x}$ is a random vector of the explanatory variables. Recall, that QDA (and LDA) finds the class $y$ maximizing

$$P(Y = y|\boldsymbol{X} = \boldsymbol{x}) = \frac{\pi_y f(\boldsymbol{x}|Y = y)}{\sum_{y'=1}^{k} \pi_{y'} f(\boldsymbol{x}|Y = y')}$$

where $\boldsymbol{x}|Y = y$ is assumed multivariate normal, i.e.

$$(\boldsymbol{x}|Y = y) \sim \mathcal{N}_p(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y).$$

LDA differs from QDA only by the additional assumption that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_y$ for all classes $y$. An intermediate classifier of LDA and QDA can thus be constructed by assuming the $\boldsymbol{\Sigma}_y$'s inversely Wishart distributed as in (1), i.e. $\boldsymbol{\Sigma}_y \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}, \nu)$. This hierarchical discriminant analysis (HDA) is thus straight-forward to implement given that

$$\begin{aligned} f(\boldsymbol{x}|Y = y) &= \int f(\boldsymbol{x}|\boldsymbol{\Sigma}_y, Y = y) f(\boldsymbol{\Sigma}_y|Y = y) d\boldsymbol{\Sigma}_y \\ &= \frac{|\boldsymbol{\Psi}|^{\frac{\nu}{2}} \Gamma_p\left(\frac{\nu+1}{2}\right)}{\pi^{-\frac{n}{2}} |\boldsymbol{\Psi} + (\boldsymbol{x} - \mu_k)(\boldsymbol{x} - \mu_k)^\top|^{\frac{\nu+1}{2}} \Gamma_p\left(\frac{\nu}{2}\right)}, \end{aligned}$$

analogous to the computation done in appendix A, using the matrix determinant lemma, $|\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^\top| = (1 + \boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{u})|\boldsymbol{A}|$, to simplify and speed up the computations [].

### 4.2 DLBCL graphical meta analysis

## 5 Concluding remarks

While the improvements are modest, the results above demonstrate an advantage of modelling the inter-study variance as a hierachical random effects model. However, the virtue of such a model is not from improvement in accuracy alone. Also desireable is the explicit quantification of the inter-study variance. If $\hat{\nu}$ is estimated to be large, the studies exhibit a largely common covariance structure, and vice-versa when $\hat{\nu}$ is small.

The generalization of the model to $n \gg p$ is extremely interesting though out of scope for this article.

# References


[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004. ISBN 9780511804441. doi: 10.1017/CBO9780511804441.

[2] J. K. Choi, U. Yu, S. Kim, and O. J. Yoo. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19(Suppl 1):i84–i90, July 2003. ISSN 1367-4803. doi: 10.1093/bioinformatics/btg1010.

[3] AP Dempster. Covariance Selection. *Biometrics*, 28(1):157–175, 1972.

[4] KB Petersen and MS Pedersen. The Matrix Cookbook. *Technical University of Denmark*, 2008. URL `http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3274`.

# A Marginalization of $\boldsymbol{\Sigma}$

This sections shows the marginalization out of $\boldsymbol{\Sigma}$ in (2). For ease of notation we drop the subscript $i$ used in $\boldsymbol{\Sigma}_i$, $\boldsymbol{X}_i$, $\boldsymbol{S}_i = \boldsymbol{X}_i \boldsymbol{X}_i^\top$, and $n_i$ in the above text. We do the computation straight-forwardly by the assumptions of the model,

$$
\begin{aligned}
&f(\boldsymbol{X}|\boldsymbol{\Psi}, \nu) \\
&= \int f(\boldsymbol{X}|\boldsymbol{\Sigma}) f(\boldsymbol{\Sigma}|\boldsymbol{\Psi}, \nu) d\boldsymbol{\Sigma} \\
&= \int \left[ \prod_{j=1}^{n} (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\Sigma}^{-1})} \right] \frac{|\boldsymbol{\Psi}|^{-\frac{\nu}{2}}}{2^{\frac{\nu p}{2}} \Gamma_p\left(\frac{\nu}{2}\right)} |\boldsymbol{\Sigma}|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})} \, d\boldsymbol{\Sigma} \\
&= (2\pi)^{-\frac{np}{2}} \frac{|\boldsymbol{\Psi}|^{-\frac{\nu}{2}}}{2^{\frac{\nu p}{2}} \Gamma_p\left(\frac{\nu}{2}\right)} \int |\boldsymbol{\Sigma}|^{-\frac{n}{2}} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{S}\boldsymbol{\Sigma}^{-1})} |\boldsymbol{\Sigma}|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Psi}\boldsymbol{\Sigma}^{-1})} \, d\boldsymbol{\Sigma} \\
&= \frac{|\boldsymbol{\Psi}|^{-\frac{\nu}{2}}}{\pi^{-\frac{np}{2}} 2^{\frac{(\nu+n)p}{2}} \Gamma_p\left(\frac{\nu}{2}\right)} \int |\boldsymbol{\Sigma}|^{-\frac{(\nu+n)+p+1}{2}} e^{-\frac{1}{2}\operatorname{tr}\left((\boldsymbol{\Psi}+\boldsymbol{S})\boldsymbol{\Sigma}^{-1}\right)} \, d\boldsymbol{\Sigma}.
\end{aligned}
$$

The integrand here can be recognized as a unnormalized inverse Wishart pdf, $\mathcal{W}^{-1}(\boldsymbol{\Psi} + \boldsymbol{S}, \nu + n)$, and so integral evaluates to the reciprocal value of the normalizing constant in that density. Thus,

$$
\begin{aligned}
f(\boldsymbol{X}|\boldsymbol{\Psi}, \nu) &= \frac{|\boldsymbol{\Psi}|^{-\frac{\nu}{2}}}{\pi^{-\frac{np}{2}} 2^{\frac{(\nu+n)p}{2}} \Gamma_p\left(\frac{\nu}{2}\right)} \frac{2^{\frac{(v+n)p}{2}} \Gamma\left(\frac{\nu+n}{2}\right)}{|\boldsymbol{\Psi} + \boldsymbol{S}|^{\frac{\nu+n}{2}}} \\
&= \frac{|\boldsymbol{\Psi}|^{\frac{\nu}{2}} \Gamma\left(\frac{\nu+n}{2}\right)}{\pi^{-\frac{np}{2}} |\boldsymbol{\Psi} + \boldsymbol{S}|^{\frac{\nu+n}{2}} \Gamma_p\left(\frac{\nu}{2}\right)},
\end{aligned}
$$

which was what was wanted.

# B Non-concavity of log-likelihood

The log-likelihood is not concave. This section analyse the (non)-concavity of the log-likelihood function,

$$
\begin{aligned}
&\ell\big(\boldsymbol{\Psi}, \nu \big| \boldsymbol{X}_1, ..., \boldsymbol{X}_k\big) \\
&= c + \sum_{i=1}^{k} \frac{\nu}{2} \log|\boldsymbol{\Psi}| + \log\Gamma_p\Big(\frac{\nu + n_i}{2}\Big) - \frac{\nu + n_i}{2} \log|\boldsymbol{\Psi} + \boldsymbol{S}_i| - \log\Gamma_p\Big(\frac{\nu}{2}\Big). \qquad (8)
\end{aligned}
$$

We first analyze the terms involving $\boldsymbol{\Psi}$ in (8). Not counting the constant, the first term, $\frac{\nu}{2}\log|\boldsymbol{\Psi}|$, is concave by the well known result that the log-determinant of positive semi-definite matrix is concave [See e.g. 1, pp. 73-74]. The third term in (8), $-\frac{\nu+n_i}{2}\log|\boldsymbol{\Psi} + \boldsymbol{S}_i|$, is convex by the same argument however negated by the sign. The sum $\boldsymbol{\Psi} + \boldsymbol{S}_i$ is positive semi-definite since both summands are.

Next, we look at the terms involving $\nu$. Cleary, the mixed terms involving both $\nu$ and $\boldsymbol{\Psi}$ are log-linear in $\nu$ and hence log-concave. We thus restrict our attention to the remaining terms not dependent on $\boldsymbol{\Psi}$. By themselves,

the second term in (8), $\log\Gamma_p\left(\frac{\nu+n_i}{2}\right)$ is convex since the multivariate gamma function is log-convex, cf. section B.1. In the same manner, the fourth term, $-\log\Gamma_p\left(\frac{\nu}{2}\right)$, is concave by the negative sign. The sum however of these terms involving $\Gamma_p$ are concave in $\nu$, since

$$\log\Gamma_p\left(\frac{\nu+n_i}{2}\right) - \log\Gamma_p\left(\frac{\nu}{2}\right) = \log\frac{\Gamma_p\left(\frac{\nu+n_i}{2}\right)}{\Gamma_p\left(\frac{\nu}{2}\right)} = \log\prod_{j=1}^{p}\frac{\Gamma\left(\frac{\nu+1-j}{2}+\frac{n_i}{2}\right)}{\Gamma\left(\frac{\nu+1-j}{2}\right)}.$$

which can be seen to be concave since $n_i \geq 2$ for all $i$ and

$$x \mapsto \log\left(\frac{\Gamma(x+a)}{\Gamma(x)}\right) \tag{9}$$

is concave for all $x > 0$ and $a > 0$. While we do not provide a proof of this the mapping is plotted in figure 2. Hence, the log-likelihood is log-concave in $\nu$.

```
logGammaRatio <- function(x, a) {
  log(gamma(x + a)/gamma(x))
}
xs <- seq(0.01, 2, by = 0.01)
par(mfrow = c(1,2), mar = c(2,2,0,0)+ 0.5)
plot(xs, logGammaRatio(xs, a = 1e-3), type = "l", xlab = "", ylab = "")
```

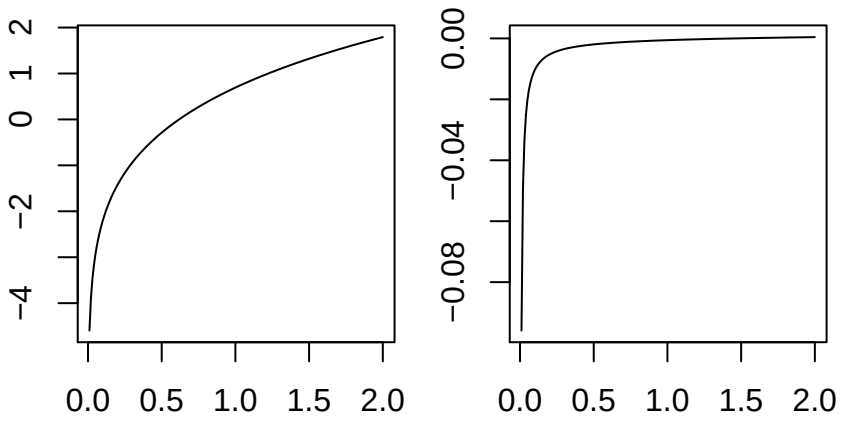



Figure 2: Plots of the mapping given in (9) for diffent values of $a$.

## B.1 log-convexity of the multivariate gamma function

The log-convexity of the multivariate gamma function can seen using the following characterization of $\Gamma_p$,

$$\Gamma_p(t) = \pi^{\frac{1}{2}\binom{p}{2}}\prod_{j=1}^{p}\Gamma\left(t+\frac{1-j}{2}\right) \text{ where } \Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx. \tag{10}$$

From this

$$\log \Gamma_p(t) = \frac{1}{2}\binom{p}{2} \log \pi + \sum_{j=1}^{p} \Gamma\left(t + \frac{1-j}{2}\right), \tag{11}$$

which is convex since $\Gamma$ is log-convex and sums of convex functions are convex. Hence $\Gamma_p$ is log-convex.

## B.2 One-dimensional case

In the one-dimensional case, we have

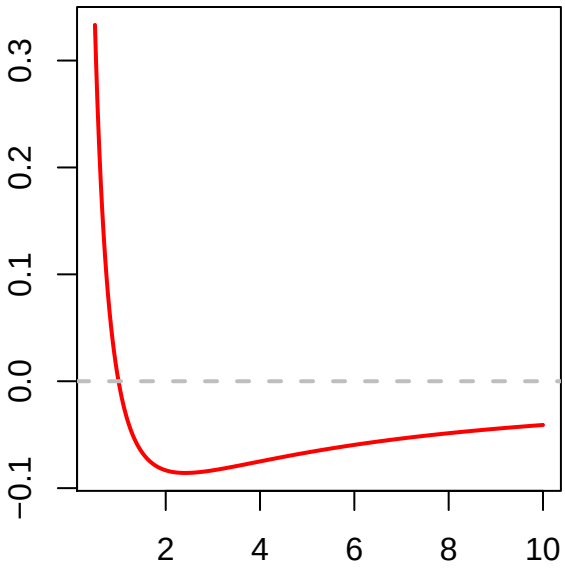
$$\ell'(\phi) = \frac{k\nu}{2}\frac{1}{\phi} + \sum_{i=1}^{k} \frac{\nu + n_i}{2} \frac{1}{\phi + x_i^2}$$

which is not clearly convex. We see, that

$$\lim_{\phi \to 0} \ell'(\phi) = \infty \text{ and } \lim_{\phi \to \infty} \ell'(\phi) = 0$$

We draw this derivative with $k = 1$ and other appropriately chosen parameters:

```
dl <- function(phi, k = 1, nu = 1, ni = 1, xi = 1)  {
  k*nu/2*1/phi - (nu + ni)/2 * 1/(phi + xi^2)
}
phi <- seq(0.5, 10, by = 0.01)
plot(phi, dl(phi), type = "l", col = "red", lwd = 2)
abline(h = 0, col = "grey", lty = 2, lwd = 2)
```




Which is be seen to have a unique root. However, the log-likelihood $\ell$ is not log-concave since a differentiable function of one variable is concave if and only if its derivative is monotonically non-increasing.

# C Negative semi-definite hessian in stationary points

This section proves that the hessian is negative semi-definite in all stationary points. The log-likelihood in (2), assuming $\nu$ fixed, obey

$$2\ell(\boldsymbol{\Psi}) = k\nu \log|\boldsymbol{\Psi}| - \sum_{a=1}^{k}(n_a+\nu)\log|\boldsymbol{\Psi}+\boldsymbol{S}_a| \tag{12}$$

up to an addition of a constant. The matrix cookbook by Petersen and Pedersen [4] has been a useful reference here[1].

## C.1 1. order derivatives

From the log-likelihood expression, we compute the 1. order derivative $\nabla_{\boldsymbol{\Psi}} 2\ell(\boldsymbol{\Psi})$ which is the matrix-valued function where each entry is given by

$$\frac{\partial 2\ell}{\partial \Psi_{ij}} = k\nu \operatorname{tr}\left(\boldsymbol{E}^{ij}\boldsymbol{\Psi}^{-1}\right) - \sum_{a=1}^{k}(\nu+n_a)\operatorname{tr}\left(\boldsymbol{E}^{ij}\left(\boldsymbol{\Psi}+\boldsymbol{S}_a\right)^{-1}\right). \tag{13}$$

where $\boldsymbol{E}^{ij}$ is a matrix with ones at entries $(i,j)$ and $(j,i)$ and zeros elsewhere. This $\boldsymbol{E}^{ij}$ is introduced as the derivative is not straight-forward because of the symmetric structure of $\boldsymbol{\Psi}$. Had $\boldsymbol{\Psi}$ been unstructured, then $\frac{\partial}{\partial \boldsymbol{\Psi}}\log|\boldsymbol{\Psi}| = \boldsymbol{\Psi}^{-1}$. However, when $\boldsymbol{\Psi}$ is symmetric we have that $\frac{\partial}{\partial \Psi_{ij}}\log|\boldsymbol{\Psi}| = \operatorname{tr}(\boldsymbol{E}^{ij}\boldsymbol{\Psi}^{-1})$ which is to say $\frac{\partial}{\partial \boldsymbol{\Psi}}\log|\boldsymbol{\Psi}| = 2\boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\circ\boldsymbol{I}$ where $\circ$ denotes the Hadamard product [4, eq. (43) and (141)].

The first order derivative lives in a $\binom{p+1}{2}$-dimensional vector space indexed by $(i,j)$, $i \le j$, with basis vectors $\boldsymbol{E}^{ij}$.

## C.2 2. order derivatives

We proceed with the second order derivative $\nabla^2_{\boldsymbol{\Psi}} 2\ell(\boldsymbol{\Psi})$ with entries given by

$$\begin{aligned}\frac{\partial^2 2\ell}{\partial \Psi_{kl}\partial \Psi_{ij}} &= -k\nu \operatorname{tr}\left(\boldsymbol{E}^{ij}\boldsymbol{\Psi}^{-1}\boldsymbol{E}^{kl}\boldsymbol{\Psi}^{-1}\right) \\ &\quad + \sum_{a=1}^{k}(\nu+n_a)\operatorname{tr}\left(\boldsymbol{E}^{ij}\left(\boldsymbol{\Psi}+\boldsymbol{S}_a\right)^{-1}\boldsymbol{E}^{kl}\left(\boldsymbol{\Psi}+\boldsymbol{S}_a\right)^{-1}\right),\end{aligned}$$

obtained by differentiation of (13) combined with $\frac{\partial}{\partial \Psi_{ij}}\boldsymbol{\Psi}^{-1} = -\boldsymbol{\Psi}^{-1}\boldsymbol{E}^{ij}\boldsymbol{\Psi}^{-1}$ for symmetric matrices [4, eq. (40)] and the linearity of the trace operator.

The second order derivative is a $\binom{p+1}{2} \times \binom{p+1}{2}$-dimensional matrix indexed by $(i,j)$ and $(k,l)$, $i \le j$, $k \le l$.

[1] See equation (41, p. 8) and (59, p. 9) and pages 14 and 52-53 in `http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3274`

## C.3 Negative semi-definiteness in stationary points

With the above expressions we now show that the hessian is negative semi-definite in all stationary points (or, extrema). Let $\boldsymbol{Y} = \sum_{(i,j)} y_{ij} E_{ij}$ be a symmetric matrix in the vector space where $\boldsymbol{Y} \neq \boldsymbol{0}$. The analog to $\boldsymbol{y}^\top \boldsymbol{A}\boldsymbol{y} = y_i \sum_{ij} A_{ij} y_j \leq 0$ in our vector space then becomes

$$\sum_{i \leq j, k \leq j} Y_{ij} \left( \nabla^2_{\boldsymbol{\Psi}} 2\ell(\boldsymbol{\Psi}) \right)_{(i,j),(k,l)} Y_{kl} \leq 0$$

which amounts to showing that

$$-k\nu \operatorname{tr}\big(\boldsymbol{Y}\boldsymbol{\Psi}^{-1}\boldsymbol{Y}\boldsymbol{\Psi}^{-1}\big) + \sum_{a=1}^{k} (\nu + n_a) \operatorname{tr}\big(\boldsymbol{Y} \left(\boldsymbol{\Psi} + \boldsymbol{S}_a\right)^{-1} \boldsymbol{Y} \left(\boldsymbol{\Psi} + \boldsymbol{S}_a\right)^{-1}\big) \leq 0. \tag{14}$$

Now, the positive-definiteness of $\boldsymbol{\Psi}$, let

$$\begin{aligned} \boldsymbol{Y} &:= \boldsymbol{\Psi}^{-\frac{1}{2}} \boldsymbol{Y} \boldsymbol{\Psi}^{-\frac{1}{2}} \text{ and} \\ \boldsymbol{S}_a &:= \boldsymbol{\Psi}^{-\frac{1}{2}} \boldsymbol{S}_a \boldsymbol{\Psi}^{-\frac{1}{2}}. \end{aligned}$$

we can assume that $\boldsymbol{\Psi} = \boldsymbol{I}$. Hence, the likelihood equation (12) equated to zero, becomes

$$k\nu \boldsymbol{I} = \sum_a (n_a + \nu)(\boldsymbol{I} + \boldsymbol{S}_a)$$

which implies (by multiplication by $\boldsymbol{Y}^2$)

$$\begin{aligned} k\nu \operatorname{tr}(\boldsymbol{Y}^2) &= \sum_a (n_a + \nu) \operatorname{tr}\big(\boldsymbol{Y}^2(\boldsymbol{I} + \boldsymbol{S}_a)\big) \\ &= \sum_a (n_a + \nu) \operatorname{tr}(\boldsymbol{Y}(\boldsymbol{I} + \boldsymbol{S}_a)\boldsymbol{Y}) \,. \end{aligned} \tag{15}$$

We substitute (15) into (14) to get

$$\begin{aligned} &\sum_a (n_a + \nu) \operatorname{tr}\left(\boldsymbol{Y}(\boldsymbol{I} + \boldsymbol{S}_a)\boldsymbol{Y}(\boldsymbol{I} + \boldsymbol{S}_a) - \boldsymbol{Y}(\boldsymbol{I} + \boldsymbol{S}_a)\boldsymbol{Y}\right) \\ &= \sum_a (n_a + \nu) \operatorname{tr}\Big(\boldsymbol{Y}(\boldsymbol{I} + \boldsymbol{S}_a)\boldsymbol{Y}\left[(\boldsymbol{I} + \boldsymbol{S}_a) - \boldsymbol{I}\right]\Big) \leq 0 \end{aligned}$$

Since $\boldsymbol{S}_a$ is positive semi-definite, it can be diagonalized $\boldsymbol{S}_a = \boldsymbol{U}_a \boldsymbol{D}_a \boldsymbol{U}_a^\top$ by the spectral theorem where the diagonal matrix $\boldsymbol{D}_a \succeq 0$ (all non-zero entries are positive) and $\boldsymbol{U}_a$ is orthonormal (i.e. $\boldsymbol{U}_a \boldsymbol{U}_a^\top = \boldsymbol{U}_a^\top \boldsymbol{U}_a = \boldsymbol{I}$). Using the diagonalization,

$$\sum_a (n_a + \nu) \operatorname{tr}\Big(\underbrace{\boldsymbol{U}^\top \boldsymbol{Y} (\boldsymbol{I} + \boldsymbol{S}_a)^{-1} \boldsymbol{Y} \boldsymbol{U}} \left[(\boldsymbol{I} + \boldsymbol{D}_a)^{-1} - \boldsymbol{I}\right]\Big) \leq 0 \tag{16}$$

where the underbraced matrix is positive semi-definite and hence have non-negative diagonal elements. Furthermore, since $\boldsymbol{D}_a$ is diagonal also with non-negative elements, the diagonal matrix $(\boldsymbol{I}+\boldsymbol{D}_a)^{-1}-\boldsymbol{I}$ clearly have non-positive entries and is thus negative semi-definite.

Since the trace of a matrix product is the sum of the element-wise products, the trace (and thus the sum) will always be non-positive and hence (16) will always hold.

# D log-likelihood of the precision

Suppose we have $k$ i.i.d. realizations, $\boldsymbol{\Delta}_1, ..., \boldsymbol{\Delta}_k$, from the Wishart distribution given in equation (6). The corresponding log-likelihood can be computed straight-forwardly:

$$\begin{aligned}
\ell(\boldsymbol{\Theta}|\boldsymbol{\Delta}_1, ..., \boldsymbol{\Delta}_k) &= \sum_{i=1}^{k} \log f(\boldsymbol{\Delta}_i|\boldsymbol{\Theta}) \\
&= \sum_{i=1}^{k} \log \frac{|\boldsymbol{\Theta}|^{-\frac{\nu}{2}}}{2^{-\frac{\nu p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)} |\boldsymbol{\Delta}_i|^{\frac{\nu-p-1}{2}} e^{-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Theta}^{-1}\boldsymbol{\Delta}_i)} \\
&= c + \sum_{i=1}^{k} -\frac{\nu}{2}\log|\boldsymbol{\Theta}| - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Theta}^{-1}\boldsymbol{\Delta}_i) \\
&= c + \frac{\nu k}{2}\left(\log|\boldsymbol{\Theta}| + \operatorname{tr}\left(\boldsymbol{\Theta}^{-1}\frac{1}{\nu k}\sum_{i=1}^{k}\boldsymbol{\Delta}_i\right)\right).
\end{aligned}$$

The last expression is to be maximized with respect to $\boldsymbol{\Theta}$ and can be recognized as the MLE problem in a multivariate Gaussian distribution. Hence,

$$\boldsymbol{\Theta} = \frac{1}{k\nu}\sum_{i=1}^{k}\boldsymbol{\Delta}_i,$$

is the MLE in this model.

# E Model and notation overview

Let $\boldsymbol{\Delta}_i = \boldsymbol{\Sigma}_i^{-1}$, $\boldsymbol{\Theta} = \boldsymbol{\Psi}^{-1}$, and $\boldsymbol{S}_i = \boldsymbol{X}_i^\top \boldsymbol{X}_i$. Then the following equivalences hold.

$$\begin{array}{ccc}
\boldsymbol{\Sigma}_i \sim \mathcal{W}_p^{-1}(\boldsymbol{\Psi}, \nu) & \iff & \boldsymbol{\Delta}_i \sim \mathcal{W}_p(\Theta, \nu) \\
\boldsymbol{X}_i|\boldsymbol{\Sigma}_i \sim \mathcal{N}_p(\boldsymbol{0}_p, \boldsymbol{\Sigma}_i) & \iff & \boldsymbol{X}_i|\boldsymbol{\Delta}_i \sim \mathcal{N}_p(\boldsymbol{0}_p, \boldsymbol{\Delta}_i^{-1}) \\
\Updownarrow & & \Updownarrow \\
\boldsymbol{\Sigma}_i|X_i \sim \mathcal{W}_p^{-1}(\boldsymbol{\Psi}+\boldsymbol{S}_i, \nu+n_i) & \iff & \boldsymbol{\Delta}_i|X_i \sim \mathcal{W}_p((\boldsymbol{\Theta}^{-1}+\boldsymbol{S}_i)^{-1}, \nu+n_i)
\end{array}$$

# List of Corrections