The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1894–1913
https://doi.org/10.1214/18-AOAS1136

ESTIMATING A COMMON COVARIANCE MATRIX FOR NETWORK META-ANALYSIS OF GENE EXPRESSION DATASETS IN DIFFUSE LARGE B-CELL LYMPHOMA[1]

BY ANDERS ELLERN BILGRAU[*,†,2], RASMUS FROBERG BRØNDUM[†,2], POUL SVANTE ERIKSEN[*], KAREN DYBKÆR[†] AND MARTIN BØGSTED[*,†]

Aalborg University[*] and Aalborg University Hospital[†]

The estimation of covariance matrices of gene expressions has many applications in cancer systems biology. Many gene expression studies, however, are hampered by low sample size and it has therefore become popular to increase sample size by collecting gene expression data across studies. Motivated by the traditional meta-analysis using random effects models, we present a hierarchical random covariance model and use it for the meta-analysis of gene correlation networks across 11 large-scale gene expression studies of diffuse large B-cell lymphoma (DLBCL). We suggest to use a maximum likelihood estimator for the underlying common covariance matrix and introduce an EM algorithm for estimation. By simulation experiments comparing the estimated covariance matrices by cophenetic correlation and Kullback–Leibler divergence the suggested estimator showed to perform better or not worse than a simple pooled estimator. In a posthoc analysis of the estimated common covariance matrix for the DLBCL data we were able to identify novel biologically meaningful gene correlation networks with eigengenes of prognostic value. In conclusion, the method seems to provide a generally applicable framework for meta-analysis, when multiple features are measured and believed to share a common covariance matrix obscured by study dependent noise.

1. Introduction. Human cells carry out their function in concerted interaction via intricate protein signalling networks. These networks are according to the central dogma of molecular biology controlled by expressed genes. It has become popular to perform genome wide measurements of expressed genes and proteins and summarizing the information by huge covariance matrices leading to improved understanding of disease pathology and identification of new drug targets [Agnelli et al. (2011), Clarke et al. (2013)]. Many gene expression studies, however, are hampered by low sample size and it has therefore become of interest to increase

Received June 2016; revised June 2017.

[1]Supported by MSCNET, EU FP6, CHEPRE, the Danish Agency for Science, Technology, and Innovation as well as Karen Elise Jensen Fonden.

[2]Shared first authorship.

Key words and phrases. Covariance estimation, precision estimation, integrative analysis, meta-analysis, network analysis.

sample size by collecting gene expression data across studies. These data are potentially hampered by severe batch effects, and robust methods are therefore required to conduct meta-analysis of covariance matrices.

To the best of our knowledge no approaches exist where meta-analysis of covariance matrices have been addressed explicitly. We acknowledge, however, that a number of indirect methods have been constructed. An immediate and tempting approach is to use one of the many study correcting approaches scattered around in the literature [Irizarry et al. (2003), Johnson, Li and Rabinovic (2007), Lee, Dobbin and Ahn (2014)] followed by estimating the covariance matrix either based on a pooled data set or by pooling covariance matrices estimated from each individual study as suggested by Lee, Dobbin and Ahn (2014). This approach, however, suffers from the same disadvantages as usual meta-analysis based on pooling fixed effects as it puts too much weight on large outliers in the data [Borenstein et al. (2010)].

Motivated by the alternative meta-analysis by random effects [DerSimonian and Laird (1986), Choi et al. (2003)], we suggest a hierarchical model where the covariance for each study is assumed to be drawn from an inverse Wishart distribution with a common mean covariance matrix, and data from each study is then subsequently generated from a multivariate Gaussian distribution with this covariance matrix. We suggest to use a maximum likelihood estimator for the underlying common covariance matrix and introduce an EM algorithm for its estimation. We use the method for the meta-analysis of gene correlation networks across 11 large-scale gene expression studies of diffuse large B-cell lymphoma (DLBCL). It is our expectation that a more suitable handling of the covariance matrix will lead to more adequate estimations of covariance matrices and subsequently inferred gene correlation networks.

In Section 2, we propose the model for a common covariance matrix across multiple studies, derive estimators thereof, and propose an inter-study homogeneity measure to aid in assessing the variation between studies. We conduct an extensive simulation study in Section 3 comparing the proposed estimator and simple pooling of covariance matrices. We then apply the model in Section 4 to 2046 DLBCL samples across 11 datasets before concluding the manuscript in Section 5.

2. A hierarchical model for the covariance matrix. Let p be the number of features and k the number of studies. We model an observation $\boldsymbol{x}$ from the ith study as a p-dimensional zero-mean multivariate Gaussian vector with covariance matrix realized from an inverse Wishart distribution, that is, $\boldsymbol{x}$ follows the hierarchical model

$$\begin{aligned} \boldsymbol{\Sigma}_i &\sim \mathcal{W}_p^{-1}(\boldsymbol{\Psi}, \nu), \\ \boldsymbol{x}|\boldsymbol{\Sigma}_i &\sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma}_i), \qquad i = 1, \ldots, k, \end{aligned} \tag{2.1}$$

where $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_i)$ denotes a p-dimensional multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and positive definite (p.d.) covariance matrix $\boldsymbol{\Sigma}_i$, and probability density

function (p.d.f.) shown in (B.1) [Bilgrau et al. (2018)], and $\mathcal{W}_p^{-1}(\boldsymbol{\Psi}, \nu)$ denotes a p-dimensional inverse Wishart distribution with $\nu > p - 1$ degrees of freedom, a p.d. $p \times p$ scale matrix $\boldsymbol{\Psi}$, and p.d.f. shown in (B.2) [Bilgrau et al. (2018)]. While the inverse Wishart distribution is defined for all $\nu > p - 1$, the first order moment exists only when $\nu > p + 1$ and is given by

$$\mathbb{E}[\boldsymbol{\Sigma}_i] = \boldsymbol{\Sigma} = \frac{\boldsymbol{\Psi}}{\nu - p - 1} \qquad \text{for } \nu > p + 1. \tag{2.2}$$

Hence, in the Random Covariance Model (RCM) of (2.1), $\boldsymbol{\Sigma}$ can be interpreted as a location-like parameter as it is the expected covariance matrix in each study. The parameter ν inversely controls the inter-study variation and can as such be considered an inter-study homogeneity parameter of the covariance structure. A large ν corresponds to high study homogeneity and vice versa for small ν. This can further be seen as $\boldsymbol{\Sigma}_i$ concentrates around $\boldsymbol{\Sigma}$ for $\nu \to \infty$ which corresponds to a vanishing inter-study variation for increasing ν. This fact is seen directly from variance and covariance expressions for the inverse Wishart [see (F.2) and (F.3), Bilgrau et al. (2018)] where the 4th order denominator grows much faster than the 1st order nominator as polynomials in ν and causing the variance to vanish for $\nu \to \infty$. Thus, the true underlying covariance matrix $\boldsymbol{\Sigma}$ and the homogeneity parameter ν are the effects of interest to be estimated.

2.1. *The likelihood function.* Suppose $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$ are n_i i.i.d. observations from $i = 1, \ldots, k$ independent studies from the model given in (2.1). Let $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})^\top$ be the $n_i \times p$ matrix of observations for the ith study where rows correspond to samples and columns to variables. By the independence assumptions, the log-likelihood for $\boldsymbol{\Psi}$ and ν is given by

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

Throughout, we use the generic notation $f(\cdot|\cdot)$ and $f(\cdot)$ for the conditional and unconditional p.d.f. of random variables, respectively. 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. (B.4) in Appendix B [Bilgrau et al.

(2018)], and we arrive at the following expression for the log-likelihood function,

$$\begin{aligned}
&\ell(\boldsymbol{\Psi}, \nu | \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k) \\
&\quad = \log \prod_{i=1}^{k} \frac{|\boldsymbol{\Psi}|^{\frac{\nu}{2}} \Gamma_p(\frac{\nu+n_i}{2})}{\pi^{\frac{n_i p}{2}} |\boldsymbol{\Psi} + \boldsymbol{X}_i^\top \boldsymbol{X}_i|^{\frac{\nu+n_i}{2}} \Gamma_p(\frac{\nu}{2})} \\
&\quad = \sum_{i=1}^{k} \Big[\frac{\nu}{2} \log |\boldsymbol{\Psi}| - \frac{\nu + n_i}{2} \log |\boldsymbol{\Psi} + \boldsymbol{X}_i^\top \boldsymbol{X}_i| + \log \frac{\Gamma_p(\frac{\nu+n_i}{2})}{\Gamma_p(\frac{\nu}{2})} \Big],
\end{aligned} \tag{2.3}$$

up to an additive constant where Γ_p is the multivariate generalization of the gamma function Γ, see (B.3) [Bilgrau et al. (2018)]. The scatter matrix $\boldsymbol{S}_i = \boldsymbol{X}_i^\top \boldsymbol{X}_i$ and study sample size n_i are sufficient statistics for each study. Note that $\boldsymbol{S}_i$ is conditionally Wishart distributed, $\boldsymbol{S}_i | \boldsymbol{\Sigma}_i \sim \mathcal{W}(\boldsymbol{\Sigma}_i, n_i)$, by construction.

As stated in the following two propositions, the likelihood is not log-concave in general. However, it is log-concave as a function of ν. All proofs have been deferred to Appendix C [Bilgrau et al. (2018)].

PROPOSITION 1 (Non-concavity in $\boldsymbol{\Psi}$). *For a fixed ν, the log-likelihood function (2.3) is not concave in $\boldsymbol{\Psi}$.*

PROPOSITION 2 (Concavity in ν). *For a fixed positive definite $\boldsymbol{\Psi}$, the log-likelihood function (2.3) is concave in ν.*

While the likelihood function is not concave in $\boldsymbol{\Psi}$ we are able to show the existence and uniqueness of a global maximum in $\boldsymbol{\Psi}$.

PROPOSITION 3 (Existence and uniqueness). *The log-likelihood (2.3) has a unique maximum in $\boldsymbol{\Psi}$ for fixed ν and $n_\bullet = \sum_{a=1}^{k} n_a \geq p$.*

In the following section estimators of the parameters are derived using moments and the EM algorithm assuming ν to be fixed.

2.2. *Moment estimator.* The pooled empirical covariance matrix can be viewed as a moment estimator of $\boldsymbol{\Sigma}$. By the model assumptions, the first and second moment of the jth observation in the ith study, $\boldsymbol{x}_{ij}$, is given by $\mathbb{E}[\boldsymbol{x}_{ij}] = \boldsymbol{0}_p$ and

$$\mathbb{E}\big[\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\big] = \mathbb{E}\big[\mathbb{E}\big[\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top | \boldsymbol{\Sigma}_i\big]\big] = \mathbb{E}[\boldsymbol{\Sigma}_i] = \frac{\boldsymbol{\Psi}}{\nu - p - 1} = \boldsymbol{\Sigma}$$

for all $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$. This suggests the estimators

$$\hat{\boldsymbol{\Psi}}_{\text{pool}} = (\nu - p - 1) \frac{\sum_{i=1}^{k} \boldsymbol{S}_i}{\sum_{i=1}^{k} n_i} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_{\text{pool}} = \frac{\sum_{i=1}^{k} \boldsymbol{S}_i}{\sum_{i=1}^{k} n_i}, \qquad \nu > p + 1, \tag{2.4}$$

where the latter is obtained by plugging $\hat{\boldsymbol{\Psi}}_{\text{pool}}$ into (2.2). This is the well-known pooled empirical covariance matrix.

2.3. *Maximization using the EM algorithm.* Here the updating scheme of the expectation-maximization (EM) algorithm [Dempster, Laird and Rubin (1977)] for fixed ν is derived. We now compute the expectation step of the EM-algorithm.

From (2.1) we have that,

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

Let $\boldsymbol{\Delta}_i = \boldsymbol{\Sigma}_i^{-1}$ be the precision matrix and let $\boldsymbol{\Theta} = \boldsymbol{\Psi}^{-1}$, then we equivalently have that

$$\begin{aligned} \boldsymbol{\Delta}_i &\sim \mathcal{W}_p(\boldsymbol{\Theta}, \nu), \\ \boldsymbol{S}_i|\boldsymbol{\Delta}_i &\sim \mathcal{W}_p\big(\boldsymbol{\Delta}_i^{-1}, n_i\big). \end{aligned} \tag{2.5}$$

From the conjugacy of the inverse Wishart and the Wishart distribution, the posterior distribution of the precision matrix is

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

Hence, by the expectation of the Wishart distribution,

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

The maximization step, in which the log-likelihood $\ell(\boldsymbol{\Theta}|\boldsymbol{\Delta}_1, \dots, \boldsymbol{\Delta}_k)$ 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 $\frac{1}{\nu}\boldsymbol{\Delta}_i$ [derived in Appendix D, Bilgrau et al. (2018)]. Let $\hat{\boldsymbol{\Theta}}_{(t)}$ be the current estimate of $\boldsymbol{\Theta}$. This yields the updating scheme

$$\hat{\boldsymbol{\Theta}}_{(t+1)} = \frac{1}{k\nu}\sum_{i=1}^k (n_i + \nu)\big(\hat{\boldsymbol{\Theta}}_{(t)}^{-1} + \boldsymbol{S}_i\big)^{-1} \tag{2.6}$$

for $\boldsymbol{\Theta}_{(t)}$. We denote the inverse of the estimate obtained by repeated iteration of (2.6) by $\hat{\boldsymbol{\Psi}}_{\mathrm{EM}}$. The EM algorithm can be sensitive to starting values. Hence, starting the algorithm in different starting values can help assesing if a global maximum has been reached.

An approximate maximum likelihood estimator using a first order approximation is also possible [derived in Appendix E, Bilgrau et al. (2018)].

2.4. *Estimation procedure.* We propose a procedure alternating between estimating ν and $\boldsymbol{\Psi}$ while keeping the other fixed. Given parameters $\hat{\nu}_{(t)}$ and $\hat{\boldsymbol{\Psi}}_{(t)}$ at iteration t, we estimate $\hat{\boldsymbol{\Psi}}_{(t+1)}$ using fixed $\hat{\nu}_{(t)}$. Subsequently, we find $\hat{\nu}_{(t+1)}$ by a standard one-dimensional numerical optimization procedure using the fixed $\hat{\boldsymbol{\Psi}}_{(t+1)}$. This coordinate ascent approach is repeated until convergence as described in Algorithm 1. The update function U in the algorithm is defined by the derived

Algorithm 1 RCM coordinate ascent estimation procedure

Input:
Sufficient data: $(\boldsymbol{S}_1, n_1), \ldots, (\boldsymbol{S}_k, n_k)$
Initial parameters: $\hat{\boldsymbol{\Psi}}_{(0)}, \hat{\nu}_{(0)}$
Convergence criterion: $\varepsilon > 0$
Output:
Parameter estimates: $\hat{\boldsymbol{\Psi}}, \hat{\nu}$
procedure FITRCM($\boldsymbol{S}_1, \ldots, \boldsymbol{S}_k, n_1, \ldots, n_k, \hat{\boldsymbol{\Psi}}_{(0)}, \hat{\nu}_{(0)}, \varepsilon$)
    Initialize: $l_{(0)} \leftarrow \ell(\hat{\boldsymbol{\Psi}}_{(0)}, \hat{\nu}_{(0)})$
    for $t = 1, 2, 3, \ldots$ **do**
        $\hat{\boldsymbol{\Psi}}_{(t)} \leftarrow U(\hat{\boldsymbol{\Psi}}_{(t-1)}, \hat{\nu}_{(t-1)})$
        $\hat{\nu}_{(t)} \leftarrow \arg\max_{\nu} \ell(\hat{\boldsymbol{\Psi}}_{(t)}, \nu)$
        $l_{(t)} \leftarrow \ell(\hat{\boldsymbol{\Psi}}_{(t)}, \hat{\nu}_{(t)})$
        if $l_{(t)} - l_{(t-1)} < \varepsilon$ **then**
            return $(\hat{\boldsymbol{\Psi}}_{(t)}, \nu_{(t)})$
        end if
    end for
end procedure

estimators. That is, equations (2.4), (2.6), or (E.2) [Bilgrau et al. (2018)] define U as the pooled, EM, or approximate MLE estimates, respectively.

The procedure using the EM step utilizes the results about the RCM log-likelihood and thus provides a guarantee of convergence along with the advantage of a very simple implementation. Both the EM step and the ν update will always yield an increase in the likelihood. The disadvantage is that the identified stationary point might be a local maximum or saddle-point when considering the log-likelihood function jointly in $(\boldsymbol{\Psi}, \nu)$. Intuitively, the latter possibility happens with zero probability, but it cannot be excluded that the maximum found is not global.

Variations on the convergence criterion can also be considered, such as (a) using the difference in successive parameter estimates, or (b) using relative rather than absolute differences.

2.5. *Interpretation and inference.*

Intra-study correlation coefficient. The heterogeneity parameter ν has no straightforward interpretation partly because the values of ν which corresponds to a large study heterogeneity is dependent on the dimension p. We therefore introduce a descriptive statistic analogous to the intra-study correlation coefficient (ICC) [Shrout and Fleiss (1979)] well known from ordinary meta-analysis. For the RCM this follows from the definition of the ICC which is defined to be the ratio of

the between-study variation $\mathrm{Var}(\Sigma_{ij})$ and the total variation $\mathrm{Var}(S_{ij})$ of any single pair of variables. In Appendix F [Bilgrau et al. (2018)] it is shown that the ICC is given by:

$$\mathrm{ICC}(\nu) = \frac{1}{\nu - p}. \tag{2.7}$$

The ICC might in this sense be utilized in better quantifying the reproducibility of the covariance across studies. A straight-forward plug-in estimator $\widehat{\mathrm{ICC}}(\nu)$ of the ICC of some gene-gene interaction is then $\mathrm{ICC}(\hat{\nu})$.

Though $\nu > p+3$ is required for the variances to exist, it is clear that $\mathrm{ICC}(\nu) \to 1$ for $\nu \to (p+1)^+$ and $\mathrm{ICC}(\nu) \to 0$ for $\nu \to \infty$ as should be expected.

Test for no study heterogeneity. By the RCM ν parameterizes an inter-study variance where the size of ν corresponds to the homogeneity between the studies. A large ν yields high study homogeneity while a small ν yields low homogeneity. Thus, it might be of interest to test if the estimated homogeneity $\hat{\nu}$ is extreme under the null-hypothesis of no heterogeneity (i.e., infinite homogeneity). That is, a test for the hypothesis $H_0 : \nu = \infty$ which is equivalent to

$$H_0 : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}.$$

The two are equivalent since sampling the covariance matrix from the inverse Wishart distribution becomes deterministic for $\nu = \infty$. Therefore, testing this hypothesis can also be interpreted as testing whether the data is adequately explained when leaving out the hierarchical structure.

The distribution of $\hat{\nu}$ under the null hypothesis is not tractable. However, in practice under H_0 or when ν is extremely large the estimated $\hat{\nu}_{\mathrm{obs}}$ will be finite as the intra-study variance dominates the total variance. We note that the null distribution of $\hat{\nu}$ does not depend on $\boldsymbol{\Sigma}$. We propose approximating the distribution of $\hat{\nu}$ under H_0 by resampling. To do this, the model is simply fitted a large number of times N on datasets re-sampled under H_0 mimicked by permuted study labels to get $\hat{\nu}_0^{(1)}, \ldots, \hat{\nu}_0^{(N)}$. As *small* values of $\hat{\nu}$ are critical for H_0 approximate acceptance regions can be constructed from $\hat{\nu}_0^{(j)}$, $j = 1, \ldots, N$. Likewise, an approximation of the p-value testing H_0 can be obtained by

$$P = \frac{1}{N+1}\left(1 + \sum_{j=1}^{N} \mathbb{1}\big[\hat{\nu}_0^{(j)} < \hat{\nu}_{\mathrm{obs}}\big]\right), \tag{2.8}$$

where $\mathbb{1}[\cdot]$ is the indicator function. The addition of one to both nominator and denominator adds a positive bias to the approximate p-value and is considered minimally needed according to Phipson and Smyth (2010). This is approximately the fraction of $\hat{\nu}_0^{(j)}$'s smaller than $\hat{\nu}_{\mathrm{obs}}$.

2.6. *Implementation and availability.* Algorithm 1 and the different estimators are implemented in the statistical programming language R [R Core Team (2012)] with core functions in C++ using packages Rcpp and RcppArmadillo [Eddelbuettel and François (2011), François, Eddelbuettel and Bates (2012)]. They are incorporated in the open-source R-package `correlateR` freely available for forking and editing [Bilgrau (2014)]. We refer to the information here for further details and installation instructions. This document was prepared with `knitr` [Xie (2013)] and LaTeX. To reproduce this document see http://github.com/AEBilgrau/RCM.

3. Simulation experiments.

3.1. *Evalutation of network estimation.* To assess the estimation procedures ability to estimate Σ we generated data from the hierarchical model (2.1) in two different scenarios. In the first scenario we define a simple block matrix of dimension $p = 40$ with four blocks of size 10. Each block has an internal pairwise correlation of 0.5, blocks 1 and 2 and 3 and 4 have a correlation of 0.3 between all pairs, and the remaing correlations are set at 0.1. In the second scenario we select the top 100 genes, ranked by variance, from the IDRC dataset (see Table 2) and used the scatter matrix of these genes, scaled as a correlation matrix, as the Σ matrix for simulation. For both scenarios we performed agglomerative hierchacical clustering using Ward-linkage and 1 minus the absolute correlation as a distance measure. Heatmaps with associated hierarchical clustering of both Σ matrices are shown in Supplementary Figure A.1 [Bilgrau et al. (2018)].

For both scenarios we simulate data with $k = 3$ and a range of values for n_i and ν. Each simulation was repeated 100 times, and the correlation matrix was estimated using the EM, MLE, and Pool approaches as outlined in Section 2. The similarity of the estimated and true Σ matrices and associated networks were evaluated using respectively the Kullback–Leibler divergence [Mattiussi et al. (2011)] and the cophenetic correlation [Sokal and Rohlf (1962)]. The cophenetic correlation is defined as the correlation of cophenetic distances of all pairwise distances in a tree, where the cophenetic distance is the height of the lowest point on the tree where two points merge. Results from the first scenario [EM and Pool method in Table 1, full results in Supplementary Table A.1, Bilgrau et al. (2018)] show that for heterogenous data ($\nu = 50, 100$) and $n_i \geq p$ the EM estimator outperforms the Pool and MLE estimators using both measures. Examples of tanglegrams comparing networks estimated with the EM and Pool method and the true Σ matrix are shown in Supplementary Figure A.2 [Bilgrau et al. (2018)]. Tanglegrams were constructed using the R-package `dendextend` [Galili (2015)]. Increasing the ν parameter, thereby making the data more homogeneous across groups diminishes the advantage of the EM estimator. Similar results were found in the second scenario using a Σ matrix based on the IDRC dataset [Table A.2, Bilgrau et al. (2018)]. Results furthermore showed that the estimates in terms of cophenetic correlation for the MLE and Pool approaches are nearly identical. We expect this to

TABLE 1
Mean cophenetic correlation and Kullback–Leibler divergence with 95% confidence, for estimated vs true network for different values of v and n_i using the EM or Pool method

		Cophenetic correlation		Kullback–Leibler divergence	
n_i	v	EM	Pool	EM	Pool
20	50	0.19 (0.17; 0.21)	0.2 (0.18; 0.22)	240.37 (232.79; 247.94)	227.33 (220.13; 234.52)
30	50	0.26 (0.23; 0.28)	0.25 (0.23; 0.28)	126.61 (123.81; 129.41)	121.81 (119.1; 124.51)
50	50	0.6 (0.56; 0.64)	0.43 (0.39; 0.46)	75.62 (73.5; 77.74)	73.62 (71.54; 75.69)
100	50	0.88 (0.85; 0.9)	0.7 (0.67; 0.74)	33.04 (32.56; 33.52)	30.9 (30.44; 31.36)
500	50	0.99 (0.98; 0.99)	0.91 (0.89; 0.93)	23.64 (23.41; 23.88)	21.31 (21.1; 21.53)
1000	50	0.99 (0.99; 0.99)	0.9 (0.88; 0.92)	22.86 (22.59; 23.14)	20.53 (20.28; 20.78)
20	100	0.35 (0.32; 0.38)	0.35 (0.32; 0.37)	76.69 (74.05; 79.33)	72.36 (69.85; 74.86)
30	100	0.4 (0.37; 0.42)	0.39 (0.37; 0.42)	34.51 (33.76; 35.26)	33.14 (32.42; 33.87)
50	100	0.72 (0.68; 0.75)	0.69 (0.66; 0.72)	27.92 (27.2; 28.65)	27.26 (26.55; 27.97)
100	100	0.97 (0.96; 0.98)	0.96 (0.95; 0.97)	8.02 (7.88; 8.16)	7.85 (7.71; 7.98)
500	100	1 (0.99; 1)	1 (1; 1)	3.34 (3.31; 3.38)	3.18 (3.15; 3.21)
1000	100	1 (1; 1)	1 (1; 1)	2.95 (2.92; 2.98)	2.79 (2.77; 2.82)
20	1000	0.51 (0.48; 0.54)	0.51 (0.48; 0.54)	52.66 (51.04; 54.29)	49.61 (48.07; 51.16)
30	1000	0.61 (0.58; 0.64)	0.61 (0.58; 0.64)	22.5 (22.05; 22.95)	21.59 (21.16; 22.02)
50	1000	0.81 (0.78; 0.84)	0.81 (0.78; 0.84)	20.49 (19.91; 21.08)	20.02 (19.44; 20.59)
100	1000	0.99 (0.98; 0.99)	0.99 (0.99; 0.99)	4.47 (4.36; 4.58)	4.42 (4.31; 4.52)
500	1000	1 (1; 1)	1 (1; 1)	0.71 (0.7; 0.72)	0.71 (0.7; 0.72)
1000	1000	1 (1; 1)	1 (1; 1)	0.41 (0.41; 0.42)	0.41 (0.4; 0.42)
20	10,000	0.53 (0.5; 0.55)	0.52 (0.5; 0.55)	53.15 (51.26; 55.04)	50.07 (48.28; 51.86)
30	10,000	0.65 (0.61; 0.68)	0.64 (0.61; 0.68)	21.91 (21.46; 22.35)	21.01 (20.59; 21.44)
50	10,000	0.83 (0.8; 0.85)	0.82 (0.79; 0.85)	19.88 (19.29; 20.48)	19.42 (18.84; 20.01)
100	10,000	0.99 (0.99; 1)	0.99 (0.99; 1)	4.19 (4.11; 4.27)	4.14 (4.06; 4.22)
500	10,000	1 (1; 1)	1 (1; 1)	0.59 (0.58; 0.6)	0.59 (0.58; 0.6)
1000	10,000	1 (1; 1)	1 (1; 1)	0.28 (0.27; 0.28)	0.28 (0.27; 0.28)

be caused by the fact that the MLE method is initilized with the Pool estimates and stops after few iterations; presumably a better estimate cannot be found in these simple scenarios.

3.2. *Computation time for the RCM model.* Next we tested the performance of the different methods in terms of computation time. Figure 1 shows computation times of the methods with varying values of the dimension of the data, and demonstrates that the increased performance of the EM method comes at an extra cost in computation time.

3.3. *Evaluation of the hypothesis testing.* Finally we investigate the performance of the p-value for the hypothesis test suggested in (2.8). To do this, we simulate from the hierarchical model with $k = 3$ and a range of different values for p, v, and n_i. For these simulations we used a Ψ matrix with a diagonal of

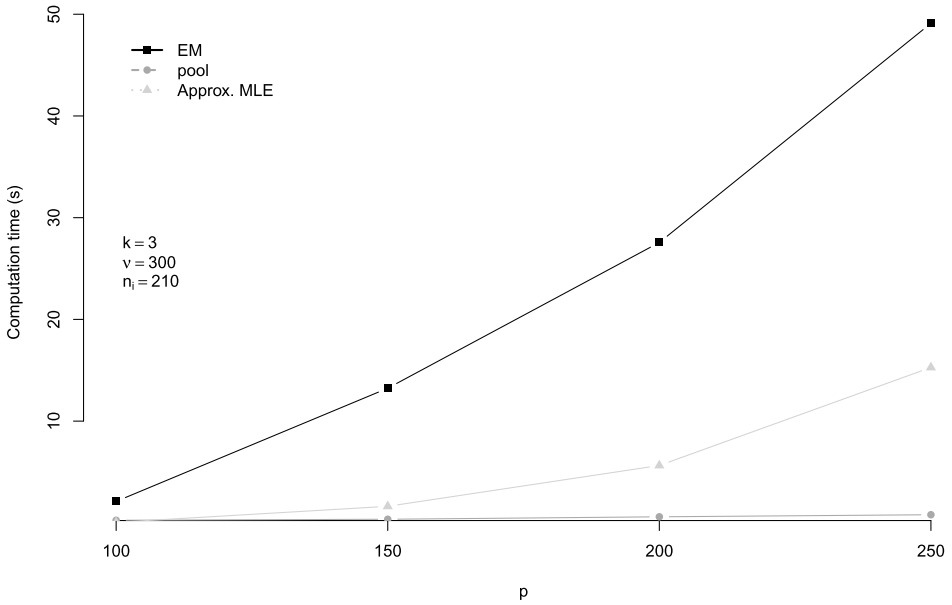

FIG. 1. *The mean computation time of* 10 *fits with varying dimension* p.

ones and 0.5 for off-diagonal values. Simulations were done 100 times for each scenario, and 500 permutations were done for each simulation. Results summarized as boxplots of the p-values obtained in the 100 simulations for each scenario are shown in Supplementary Figure A.3 [Bilgrau et al. (2018)]. We find that for heterogenous data (e.g., $p = 20$, $\nu = 30$) the null-hypothesis is clearly rejected if $n_i > p$. When increasing ν thus making the groups more similar, more observations are needed to reject the null hypothesis, while for identical groups, that is, $\nu = \infty$, the null-hypothesis is generally not rejected. The p-values obtained from the permuation test thus performs as intended.

4. DLBCL meta-analysis. Diffuse large B-cell lymphoma (DLBCL) is an aggressive cancer subtype accounting for 30%–58% of non-Hodgkin's lymphomas (NHL) which constitutes about 90% of all lymphomas [International Lymphoma Study Group (1997)].

4.1. *Data and preprocessing.* A large amount of DLBCL gene expression datasets are now available online at the NCBI (National Center for Biotechnology Information) Gene Expression Omnibus (GEO) website. 10 large-scale DLBCL gene expression studies were downloaded and preprocessed using custom brainarray chip definition files (CDF) [Dai et al. (2005)] and RMA-normalized using the R-package `affy` [Gautier et al. (2004)]. The corresponding GEO-accession numbers and microarray platforms used are seen in Table 2. The downloaded data yield

TABLE 2
Overview of studies used with GEO accession number from the NCBI Gene expression omnibus website, the relevant reference, array types used in the study, and number of samples and features on the used array

	GEO no.	Name	Reference	Used arrays	*n*
1	GSE56315	CHEPRETRO	Dybkær et al. (2015)	hgu133plus2	89
2	GSE19246	BCCA	Williams et al. (2010)	hgu133plus2	177
3	GSE12195	CUICG	Compagno et al. (2009)	hgu133plus2	136
4	GSE22895	HMRC	Jima et al. (2010)	hugene10st	101
5	GSE31312	IDRC	Visco et al. (2012)	hgu133plus2	469
6	GSE10846	LLMPP R-CHOP	Lenz et al. (2008)	hgu133plus2	181
7	GSE10846	LLMPP CHOP	Lenz et al. (2008)	hgu133plus2	233
8	GSE34171	MDFCI	Monti et al. (2012)	hgu133plus2, snp6	90
9	GSE34171	MDFCI	Monti et al. (2012)	hgu133a, hgu133b	78
10	GSE22470	MMML	Salaverria et al. (2011)	hgu133a	271
11	GSE4475	UBCBF	Hummel et al. (2006)	hgu133a	221

a total of 2046 samples with study sizes in the range 78–469. The summarization using brainarray CDFs to Ensembl gene identifiers facilitates cross-platform integration.

After RMA normalization and summarization, the data were brought to a common scale by quantile normalizing all data to the common cumulative distribution function of all arrays. Lastly, the datasets were reduced to 11,573 common genes represented in all studies and array platforms. Supplementary Figure A.4 [Bilgrau et al. (2018)] shows a plot of the first and second principal components of the combined dataset. We see a clear split on the first principal component, indicitating a possible batch effect and heterogeneous data, and thus a situation where the EM estimator might offer an advantage compared to the simpler Pool approach.

4.2. *Analysis*. For each dataset the scatter matrix $\boldsymbol{S}_i$ of the top 300 most variable genes (as measured by the pooled variance across all studies) was computed as the sufficient statistics along with the number of samples.

The parameters of the RCM were estimated using the EM algorithm and yielded the 300×300 matrix $\hat{\boldsymbol{\Psi}}$, $\hat{\nu} = 773.16$, and ICC = 0.0021. The RCM was fitted using three different initial sets of parameters which all converged to the same parameter estimates. Log-likelihood traces, iterations used, and computation times are seen in Figure 2. From the parameter estimate, the common expected covariance $\hat{\boldsymbol{\Sigma}} = (\hat{\nu} - p - 1)^{-1}\hat{\boldsymbol{\Psi}}$ was computed and subsequently scaled to the corresponding correlation matrix $\hat{\boldsymbol{R}}$.

Despite the low ICC value the permutation test yielded a p-value for the null hypothesis of study homogeneity of 0.002, clearly rejecting it. This means a significant difference has been detected between the estimated covariance structures

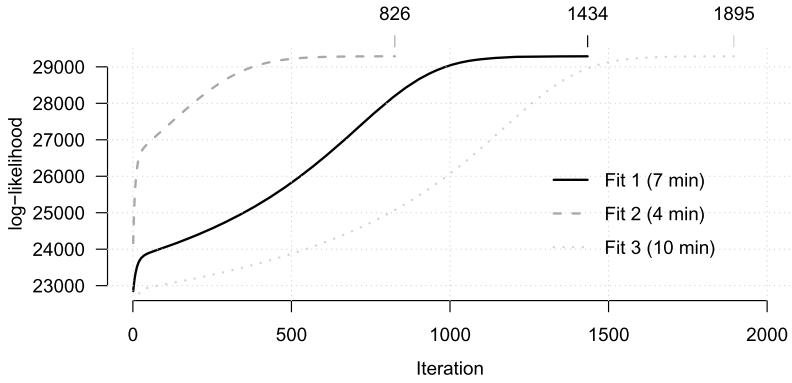

FIG. 2. *The trace of the log-likelihood for three different starting values of* $\boldsymbol{\Psi}$ *and* ν *using the EM algorithm and computational times in minutes. The number of iterations used for each fit is shown above.*

across studies. This low ICC might suggest selecting the most variable genes bias the ICC towards inter-study homogeneity of covariances. To further investigate the low ICC value we randomly sampled 300 genes and estimated the ν parameter 100 times. This gave a value of ν ranging from 383.5 to 394.76 with a mean of 388.58, corresponding to an ICC ranging from 0.0106 to 0.012 with a mean of 0.0113; histograms are shown in Supplementary Figure A.7 [Bilgrau et al. (2018)]. This indicates a bias towards more homogeneity for the high variance selected genes.

For simplicity we employed a standard network analysis to the estimated common correlation matrix $\hat{\boldsymbol{R}}$ across all studies. To identify clusters with high internal correlation, we used agglomerative hierarchical clustering with Ward-linkage and distance measure defined as 1 minus the absolute value of the correlation. The dendrogram was arbitrarily pruned at a height which produced five modules. The Modules are given different colors. Figure 3 shows the heatmap, associated network modules and suggested function.

We checked if the identified modules were prognostic for overall survival (OS) in the CHOP and R-CHOP-treated cohort datasets of GSE10846. To do this, the eigengene [Horvath (2011)] for each module was computed. The module eigengene is the first principal component of the expression matrix of the module which thus can be represented by a linear combination of the module genes. We also report the amount of variation the eigengene represents by calculating the explained variation of the first pricipal component. Multiple Cox proportional hazards model for OS was fitted with the module eigengenes as covariates. For the prognostically interesting and tightly clustered olivegreen module, the Kaplan–Meier estimates were computed for groups arising when dichotomizing the values of the corresponding eigengene as above or below the median value. These results are shown in Figure 4. The proportion of variance explained by the eigengene in the CHOP

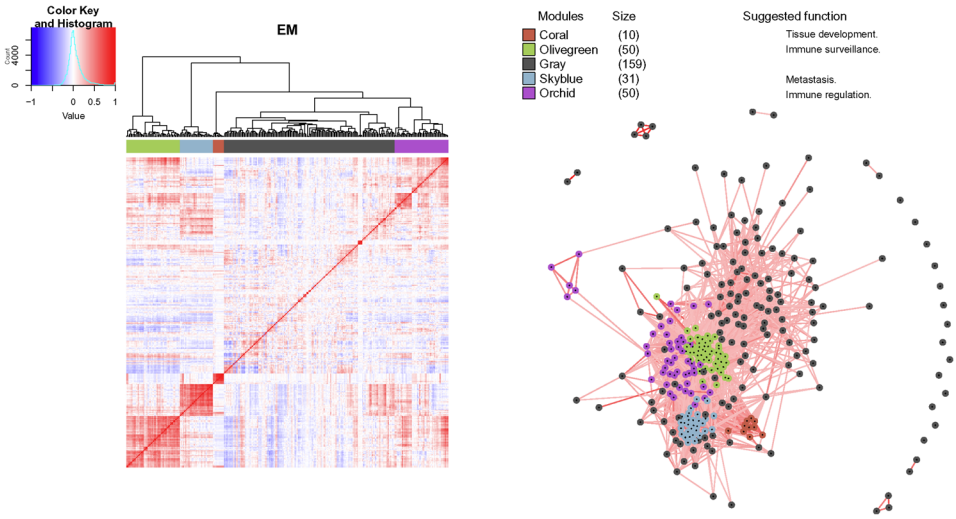

FIG. 3. *Heatmap and correlation network for the estimated correlation matrices of the top* 300 *genes for the DLBCL data using the EM method. The network is cut at a height producing five clusters.*

and R-CHOP datasets for respectively the coral, olivegreen, gray, skyblue, and orchid modules were 0.72, 0.6, 0.11, 0.7, 0.31, and 0.77, 0.55, 0.11, 0.7, 0.31.

Next, the modules were screened for biological relevance using GO (Gene Ontology) Biological Process, Molecular Function, and Cellular Component as well as REACTOME and KEGG pathway enrichment analysis. This was done using the g:profiler web server [Reimand et al. (2016)] via the accompanying R-package `gProfileR` [Reimand, Kolde and Arak (2016)]. Since we pre-selected the top 300 genes by variance, the enrichment analysis was done using only these as the background genes. Top genes for each module, ranked by connectivity, are shown in Table 3, while results of the enrichment analysis for each of the modules are shown in Supplementary Table A.3 [Bilgrau et al. (2018)]. Inspection of the enrichment analysis and most connected genes allowed us to hypothesize that the coral module is involved in "Tissue development", the skyblue module is involved in "metastasis" (strong association to extracellular processes), the orchid module involved in "immune regulation" (large overlap with GO:0002376-immune system process), and the olivegreen module involved in "immune surveillance" (strong association with GO:0006952-defense response and GO:0045087-innate immune response).

From the gene enrichment and survival analysis the olivegreen module appeared particularly interesting, as we notice a strong involvement of immune response and an association between high value of the eigengene expression and poor survival, which eventually could make these patients candidates for experimental immunotherapies. Several of the genes, for example, S100A8, S100A9,

TABLE 3

The identified modules, their sizes, and member genes. The genes are sorted decreasingly by their intra-module connectivity (sum of the incident edge weights). Only the top 40 genes are shown

Gray	**Olivegreen**	**Orchid**	**Skyblue**	**Coral**
$n = 159$	$n = 50$	$n = 50$	$n = 31$	$n = 10$
MYBL1	FCER1G	CD2	COL5A2	KRT6A
BATF	C1QB	CD3D	COL1A2	SPRR1A
STAP1	C1QA	GIMAP4	COL3A1	SPRR1B
CYB5R2	GBP1	PTGDS	THBS2	KRT13
TNFRSF13B	RARRES3	TRAT1	COL6A3	SPRR3
CD44	IDO1	CCL19	COL1A1	S100A2
MARCKSL1	CD14	CLU	COL5A1	KRT14
LRMP	LILRB2	ADAMDEC1	VCAN	DSP
HCK	SERPING1	TRBC2	FAP	KRT5
MME	PSTPIP2	ITM2A	PLOD2	SFN
LMO2	GZMA	LGALS2	MMP2	
VPREB3	CCL8	ITK	SULF1	
BCL2A1	VSIG4	PLA2G2D	MXRA5	
BLNK	NKG7	IL7R	DCN	
HLA-DOB	IFNG	PLA2G7	LUM	
RRAS2	GBP2	ENPP2	SPARC	
JADE3	CXCL10	IL18	POSTN	
STAG3	SLAMF7	CHI3L1	COL15A1	
BACH2	FGL2	TFEC	TMEM45A	
CCND2	CD163	CXCL13	COL11A1	
PDGFD	CXCL11	CCL21	CTSK	
NCF2	GZMH	CSTA	EMP1	
SPINK2	ALDH1A1	MMP9	AEBP1	
MNDA	CXCL9	LYZ	TGFBI	
MS4A1	GZMK	HSD11B1	GJA1	
CD22	GZMB	APOC1	EGFL6	
EBI3	KCNJ2	CXCL14	PLS3	
OSBPL10	LAG3	C3	TIMP1	
GPR137B	CPVL	MAL	ANXA1	
GRHPR	IGSF6	CYP27B1	TNFAIP6	
CHST2	LGMN	LAMP3	SPP1	
SORL1	MT2A	CHIT1		
IGF2BP3	MT1G	PLAC8		
SYBU	IFI27	SELL		
TCL1A	CD8A	KLRB1		
ZNF804A	MS4A4A	CD69		
SLC12A8	CRTAM	ROBO1		
CTGF	S100A9	ORM1		
FCRL2	MARCO	S1PR1		
DUSP5	S100A8	CCR7		

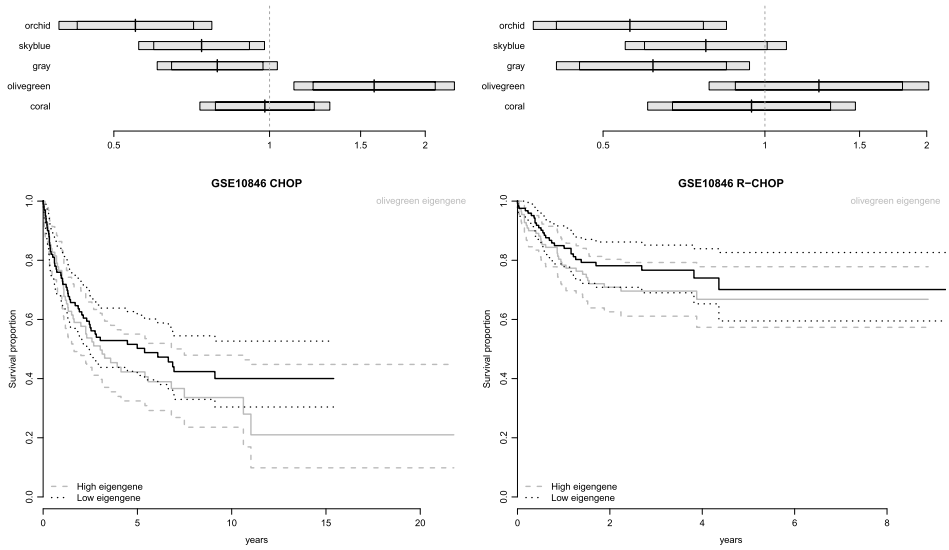

FIG. 4. *The top row shows* 95% *and* 99% *CI for the hazard ratio for each eigengene in the multiple Cox proportional hazards model containing all eigengenes in the CHOP and R-CHOP dataset. The bottom row shows Kaplan–Meier estimates* (*and* 95% *CI*) *for the overall survival for patients stratified by the dichotomized olivegreen eigengene.*

CD14, and CD163 with the highest connectivity in this module have been associated to immunotherapy [Cheng et al. (2008), Fulmer (2008), Stroncek et al. (2017)]. As prominent examples S100A8 (MRP8; calgranulin A) and the gene S100A9 (MRP14; calgranulin B) appear in the list. This is interesting as compelling research has shown that the S100 family of calciumbinding proteins maintain immunosuppressive myeloid-derived suppressor (MDS) cells at the tumor site [Fulmer (2008)]. Notably, in mice injected with lymphoma cells, knockout of S100A9 resulted in greater tumor infiltration of T-cells and less accumulation of MDS cells than that seen in wild-type mice [Cheng et al. (2008)]. The knockout mice had higher rates of tumor rejection and lower tumor size than their wild-type littermates. This result indicates that knockdown of these proteins may improve the outcome of immunotherapy strategies in patients with values of the eigengene of the olivegreen module.

Finally, we compared the network analyses based on the covariance matrix obtained by the EM to that obtained by the Pool methods, results are shown in the Supplementary Material [Bilgrau et al. (2018)]. The upper row of Supplementary Figure A.5 shows the heatmap and associated network modules for the Pool method, when the dendogram is cut at five modules, Supplementary Figure A.6 shows plots for the survival analysis, and top genes and gene enrichments are given in Supplementary Tables A.4 and A.5. For the Pool method, we chose for each module the same color as the module of the EM based clustering with most

overlapping genes. In the lower row of Figure A.5 a tangleram was constructed and the cophenetic correlation was calculated. We noticed generally a great overlap between the modules, but a low cophenetic correlation. With background in the simulation we anticipate the Pool method has lower efficiency than the EM method.

The olivegreen and coral modules seem to be so tightly regulated that they manifest themselves for both methods, which is also seen in the enrichment analyses. However, the size of the skyblue module is increased for the pool method by acquiring genes from the grey module identified by the EM method, but the overall enrichment is not changed. For the orchid module, we notice a number of genes ending up in the grey module for the pool method. This has the consequence that the immune regulation fingerprint disappers using the Pool method. Morever, if we look at the less correlated intramodular connections the noise plays a larger role leading to a less clear separation between the modules for the Pool method. This can have potential biological implications, when regulating hub genes resulting in intra-module cascades of reactions.

5. Discussion. The RCM for meta-analysis of covariance structures was shown to be superior to simple pooling as suggested previously in the literature. The estimated covariance matrix was also capable of providing a dissimilarity measure, which was able to pinpoint alternative biologically meaningful gene correlation networks in DLBCL, which can be used to formulate new hypothesis about the role of immune therapy in DLBCL.

However, the proposed testing is computationally demanding and only feasible when p is sufficiently small. This could for example, be overcome by improved and faster fitting procedures or by deriving the distribution of $\hat{\nu}$ under the null hypothesis. Yet the latter is seemingly intractable as $\hat{\nu}$ is a very complex function of the data. The fact that the null-hypothesis lies on the edge of the parameter space also seems to constrain the feasibility of deriving such a distribution. One might question whether the added utility of the ν parameter provides sufficient relaxation of the covariance homogeneity. Therefore, the present work should be considered a first step in the direction of explicitly modelling the inter-study variation of covariance matrices. It is also worth noticing, that although the suggested method proved to be superior to simple pooling, it only works for small or moderate numbers of features p. This can partly be alleviated by combining multiple studies to yield a sufficiently large total sample size $n_\bullet$ that allows for the estimation of large covariance matrices. Turning to using p-values seems tempting, but one should be aware, as with all hypothesis testing, that the exact threshold of ICC (or ν) needed to claim homogeneous studies is dependent on the sample size and the relevant effect size. In this respect the relevant effect size is unclear and will be problem dependent.

The moderate size of p is a severe drawback as many methods have been published concerning estimation of large covariance matrices by various regularization

methods [Friedman, Hastie and Tibshirani (2008), Meinshausen and Bühlmann (2006), van Wieringen and Peeters (2016)]. Therefore we believe this work could be further enriched by combining the method with regularized estimation. In the future such generalizations of the model to $p \gg n_\bullet$ is extremely interesting though out of scope for this article.

In conclusion the article demonstrates an advantageous model based way of conducting meta-analaysis of covariance matrices—especially in a setting with moderate number of features compared to the dimension. One should also notice the method seems to provide a generally applicable framework making it usable in other settings where multiple features are measured and believed to share a common covariance matrix obscured by group dependent noise.

Acknowledgements. We thank Martin Raussen, Jon Johnsen, as well as Niels Richard Hansen for their assistance on some of the mathematical proofs. The helpful comments from Steffen Falgreen, Andreas S. Pedersen, and reviewers were also much appreciated. The technical assistance from Alexander Schmitz, Julie S. Bødker, Ann-Maria Jensen, Louise H. Madsen, and Helle Høholt is also greatly appreciated.

SUPPLEMENTARY MATERIAL

Supplement A: Appendices (DOI: 10.1214/18-AOAS1136SUPPA; .pdf). Supplementary figures, tables and proofs available online.

Supplement B: Documents for reproducibility (http://github.com/AEBilgrau/RCM). The documents and other needed files to perform the analyses to reproduce this article. See the `README` file herein.

REFERENCES

AGNELLI, L., FORCATO, M., FERRARI, F., TUANA, G., TODOERTI, K., WALKER, B. A., MORGAN, G. J., LOMBARDI, L., BICCIATO, S. and NERI, A. (2011). The reconstruction of transcriptional networks reveals critical genes with implications for clinical outcome of multiple myeloma. *Clin. Cancer Res.* **17** 7402–7412.

BILGRAU, A. E. (2014). correlateR: Fast, efficient, and robust partial correlations. R package version 0.1. Available at http://github.com/AEBilgrau/correlateR.

BILGRAU, A. E., BRØNDUM, R. F., ERIKSEN, P. S., DYBKÆR, K. and BØGSTED, M. (2018). Supplement to "Estimating a common covariance matrix for network meta-analysis of gene expression datasets in diffuse large B-cell lymphoma." DOI:10.1214/18-AOAS1136SUPPA.

BORENSTEIN, M., HEDGES, L. V., HIGGINS, J. P. and ROTHSTEIN, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods* **1** 97–111.

CHENG, P., CORZO, C. A., LUETTEKE, N., YU, B., NAGARAJ, S., BUI, M. M., ORTIZ, M., NACKEN, W., SORG, C., VOGL, T. et al. (2008). Inhibition of dendritic cell differentiation and accumulation of myeloid-derived suppressor cells in cancer is regulated by S100A9 protein. *J. Exp. Med.* **205** 2235–2249.

CHOI, J. K., YU, U., KIM, S. and YOO, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19** i84–i90.

CLARKE, C., MADDEN, S. F., DOOLAN, P., AHERNE, S. T., JOYCE, H., O'DRISCOLL, L., GALLAGHER, W. M., HENNESSY, B. T., MORIARTY, M., CROWN, J., KENNEDY, S. and CLYNES, M. (2013). Correlating transcriptional networks to breast cancer survival: A large-scale coexpression analysis. *Carcinogenesis* **34** 2300–2308.

COMPAGNO, M., LIM, W. K., GRUNN, A., NANDULA, S. V., BRAHMACHARY, M., SHEN, Q., BERTONI, F., PONZONI, M., SCANDURRA, M., CALIFANO, A. et al. (2009). Mutations of multiple genes cause deregulation of NF-κB in diffuse large B-cell lymphoma. *Nature* **459** 717–721.

DAI, M., WANG, P., BOYD, A. D., KOSTOV, G., ATHEY, B., JONES, E. G., BUNNEY, W. E., MYERS, R. M., SPEED, T. P., AKIL, H., WATSON, S. J. and MENG, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*. **33** e175.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

DERSIMONIAN, R. and LAIRD, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* **7** 177–188.

DYBKÆR, K., BØGSTED, M., FALGREEN, S., BØDKER, J. S., KJELDSEN, M. K., SCHMITZ, A., BILGRAU, A. E., XU-MONETTE, Z. Y., LI, L., BERGKVIST, K. S., LAURSEN, M. B., RODRIGO-DOMINGO, M., MARQUES, S. C., RASMUSSEN, S. B., NYEGAARD, M., GAIHEDE, M., MØLLER, M. B., SAMWORTH, R. J., SHAH, R. D., JOHANSEN, P., EL-GALALY, T. C., YOUNG, K. H. and JOHNSEN, H. E. (2015). A diffuse large B-cell lymphoma classification system that associates normal B-cell subset phenotypes with prognosis. *J. Clin. Oncol*. **33** 1379–1388.

EDDELBUETTEL, D. and FRANÇOIS, R. (2011). Rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **40** 1–18.

FRANÇOIS, R., EDDELBUETTEL, D. and BATES, D. (2012). RcppArmadillo: Rcpp integration for Armadillo templated linear algebra library. R package version 0.3.6.1.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

FULMER, T. (2008). Suppressing the suppressors. *SciBX* **1**(38). DOI:10.1038/scibx.2008.914.

GALILI, T. (2015). dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31** 3718–3720.

GAUTIER, L., COPE, L., BOLSTAD, B. M. and IRIZARRY, R. A. (2004). affy—Analysis of affymetrix GeneChip data at the probe level. *Bioinformatics* **20** 307–315.

HORVATH, S. (2011). *Weighted Network Analysis*: *Applications in Genomics and Systems Biology*. Springer, Berlin.

HUMMEL, M., BENTINK, S., BERGER, H., KLAPPER, W., WESSENDORF, S., BARTH, T. F., BERND, H.-W., COGLIATTI, S. B., DIERLAMM, J., FELLER, A. C. et al. (2006). A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N. Engl. J. Med.* **354** 2419–2430.

INTERNATIONAL LYMPHOMA STUDY GROUP (1997). A clinical evaluation of the international lymphoma study group classification of non-Hodgkin's lymphoma. *Blood* **89** 3909–3918.

IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264.

JIMA, D. D., ZHANG, J., JACOBS, C., RICHARDS, K. L., DUNPHY, C. H., CHOI, W. W., AU, W. Y., SRIVASTAVA, G., CZADER, M. B., RIZZIERI, D. A. et al. (2010). Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. *Blood* **116** e118–e127.

JOHNSON, W. E., LI, C. and RABINOVIC, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** 118–127.

LEE, J. A., DOBBIN, K. K. and AHN, J. (2014). Covariance adjustment for batch effect in gene expression data. *Stat. Med.* **33** 2681–2695. MR3256670

LENZ, G., WRIGHT, G. W., EMRE, N. T., KOHLHAMMER, H., DAVE, S. S., DAVIS, R. E., CARTY, S., LAM, L. T., SHAFFER, A., XIAO, W. et al. (2008). Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc. Natl. Acad. Sci. USA* **105** 13520–13525.

MATTIUSSI, V., TUMMINELLO, M., IORI, G. and MANTEGNA, R. N. (2011). Comparing correlation matrix estimators via Kullback–Leibler divergence. Preprint, DOI:10.2139/ssrn.1966714.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MONTI, S., CHAPUY, B., TAKEYAMA, K., RODIG, S. J., HAO, Y., YEDA, K. T., INGUILIZIAN, H., MERMEL, C., CURRIE, T., DOGAN, A. et al. (2012). Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer Cell* **22** 359–372.

PHIPSON, B. and SMYTH, G. K. (2010). Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.* **9** Art. 39, 14. MR2746025

R CORE TEAM (2012). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

REIMAND, J., KOLDE, R. and ARAK, T. (2016). gProfileR: Interface to the 'g:Profiler' toolkit. R package version 0.6.1.

REIMAND, J., ARAK, T., ADLER, P., KOLBERG, L., REISBERG, S., PETERSON, H. and VILO, J. (2016). g:Profiler—A web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44** W83–W89.

SALAVERRIA, I., PHILIPP, C., OSCHLIES, I., KOHLER, C. W., KREUZ, M., SZCZEPANOWSKI, M., BURKHARDT, B., TRAUTMANN, H., GESK, S., ANDRUSIEWICZ, M. et al. (2011). Translocations activating IRF4 identify a subtype of germinal center-derived B-cell lymphoma affecting predominantly children and young adults. *Blood* **118** 139–147.

SHROUT, P. E. and FLEISS, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **86** 420–428.

SOKAL, R. R. and ROHLF, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon* **11** 33–40.

STRONCEK, D. F., BUTTERFIELD, L. H., CANNARILE, M. A., DHODAPKAR, M. V., GRETEN, T. F., GRIVEL, J. C., KAUFMAN, D. R., KONG, H. H., KORANGY, F., LEE, P. P., MARINCOLA, F., RUTELLA, S., SIEBERT, J. C., TRINCHIERI, G. and SELIGER, B. (2017). Systematic evaluation of immune regulation and modulation. *J. Immunother. Cancer* **5** 21.

VAN WIERINGEN, W. N. and PEETERS, C. F. W. (2016). Ridge estimation of inverse covariance matrices from high-dimensional data. *Comput. Statist. Data Anal.* **103** 284–303. MR3522633

VISCO, C., LI, Y., XU-MONETTE, Z. Y., MIRANDA, R. N., GREEN, T. M., TZANKOV, A., WEN, W., LIU, W., KAHL, B., D'AMORE, E. et al. (2012). Comprehensive gene expression profiling and immunohistochemical studies support application of immunophenotypic algorithm for molecular subtype classification in diffuse large B-cell lymphoma: A report from the international DLBCL Rituximab-CHOP consortium program study. *Leukemia* **26** 2103–2113.

WILLIAMS, P. M., LI, R., JOHNSON, N. A., WRIGHT, G., HEATH, J.-D. and GASCOYNE, R. D. (2010). A novel method of amplification of FFPET-derived RNA enables accurate disease classification with microarrays. *J. Mol. Diagnostics* **12** 680–686.

XIE, Y. (2013). *Dynamic Documents with R and Knitr*. CRC Press, Boca Raton, FL.

A. E. BILGRAU
DEPARTMENT OF HAEMATOLOGY
AALBORG UNIVERSITY HOSPITAL
SDR. SKOVVEJ 15
DK-9000 AALBORG
DENMARK
AND
DEPARTMENT OF MATHEMATICAL SCIENCES
AALBORG UNIVERSITY
FREDRIK BAJERS VEJ 7G
DK-9220 AALBORG Ø
DENMARK
E-MAIL: anders.ellern.bilgrau@gmail.com

K. DYBKÆR
M. BØGSTED
DEPARTMENT OF HAEMATOLOGY
AALBORG UNIVERSITY HOSPITAL
SDR. SKOVVEJ 15
DK-9000 AALBORG
DENMARK
AND
DEPARTMENT OF CLINICAL MEDICINE
AALBORG UNIVERSITY HOSPITAL
SDR. SKOVVEJ 15
DK-9000 AALBORG Ø
DENMARK
E-MAIL: k.dybkaer@rn.dk
mboegsted@dcm.aau.dk

R. F. BRØNDUM
DEPARTMENT OF HAEMATOLOGY
AALBORG UNIVERSITY HOSPITAL
SDR. SKOVVEJ 15
DK-9000 AALBORG
DENMARK
E-MAIL: rfb@rn.dk

P. S. ERIKSEN
DEPARTMENT OF MATHEMATICAL SCIENCES
AALBORG UNIVERSITY
FREDRIK BAJERS VEJ 7G
DK-9220 AALBORG Ø
DENMARK
E-MAIL: svante@math.aau.dk