Biostatistics (2018) **19**, 2, *pp.* 153–168
doi:10.1093/biostatistics/kxx032
Advance Access publication on July 1, 2017

Automatic module selection from several microarray gene expression studies

ALIX ZOLLINGER*

Swiss Institute of Bioinformatics, SIB-BCF, Genopode Building, 1015 Lausanne, Switzerland and Ecole Polytechnique Fédérale de Lausanne, EPFL-FSB-MATHAA-STAT, Station 8, 1015 Lausanne, Switzerland
alix.leboucq@alumni.epfl.ch

ANTHONY C. DAVISON, DARLENE R. GOLDSTEIN

Ecole Polytechnique Fédérale de Lausanne, EPFL-FSB-MATHAA-STAT, Station 8, 1015 Lausanne, Switzerland

Summary

Independence of genes is commonly but incorrectly assumed in microarray data analysis; rather, genes are activated in co-regulated sets referred to as modules. In this article, we develop an automatic method to define modules common to multiple independent studies. We use an empirical Bayes procedure to estimate a sparse correlation matrix for all studies, identify modules by clustering, and develop an extreme-value-based method to detect so-called scattered genes, which do not belong to any module. The resulting algorithm is very fast and produces accurate modules in simulation studies. Application to real data identifies modules with significant enrichment and results in a huge dimension reduction, which can alleviate the computational burden of further analyses.

Keywords: Clustering; Empirical Bayes estimation; Estimation of large covariance matrices; GAP statistic; GPD mixture; Meta-analysis; Module.

1. Introduction

Analysis of microarray gene expression data commonly assumes that expression values of different genes are independent. This simplifying assumption is false, however: genes function as groups that are referred to as modules, and analysis based on modules rather than genes may produce a large dimension reduction. This idea was applied in Wirapati *and others* (2008), who used coexpression modules in breast cancer meta-analysis. They selected a few prototype genes having specific roles in breast cancer, then grouped genes based on the correlation of their expression values with the prototypes, thus defining coexpression modules and module scores for use in a meta-analysis. This method is valuable in the context of microarray data for several reasons: it is biologically meaningful—genes do not act alone, but are usually activated in groups; it is statistically helpful—considering a small group of coexpression modules rather than tens of thousands of genes reduces the number of hypotheses to be tested, thereby increasing the power of

*To whom correspondence should be addressed.

the analysis; and detecting interesting groups of genes is more reliable than identifying single genes, since differentially expressed genes can vary greatly from study to study while their common modes of action may not (Ein-Dor *and others*, 2005, 2006). There is thus a better chance of identifying a common differentially expressed module among studies than any particular gene.

Up to now, identification of modules has often been based on prior knowledge of specific genes that can be considered as candidate prototypes. This pre-analysis phase of identifying prototypes entails case-by-case collaboration with biologists, which is not always possible. This motivates the development of an automatic module identification procedure that would not require expert prior knowledge and could be applied to any dataset, with the goal of dimension reduction.

In order to define modules, we base our approach on the correlation matrix, defining modules to be groups of genes with highly correlated expression. We assume that the covariance matrix is sparse, as probably many genes have no or only weak correlation with others, and only some of them are very correlated, defining the modules. Estimating a sparse covariance matrix will aid the detection of modules by highlighting groups of genes that are highly correlated.

When the number of individuals N is much smaller than the number of variables G, $N \ll G$, the usual sample covariance matrix has rank at most $N-1$, is not a consistent estimator of the population covariance matrix, and is often ill-conditioned. Several methods exist to estimate a population covariance matrix in the high-dimensional case, many of which require extra assumptions for consistent estimation. Bai and Shi (2011) identify four classes of methods for covariance estimation when $G \gg N$: shrinkage, factor models, Bayesian approaches and random matrix theory methods. Pourahmadi (2011) reviews methods to estimate covariance matrices for two contents: generalized linear models and regularization.

Thresholding methods are very popular (Ledoit and Wolf, 2004, 2012; Bickel and Levina, 2008); these fix a threshold under which all entries of the covariance matrix are set to zero. Cai and Liu (2011) extended this idea by defining an adaptive threshold for each entry of the sample covariance matrix, which attains the optimal rate of convergence in the spectral norm.

Graphical models allow estimation of the covariance matrix through its inverse, the precision matrix, whose entries give information on conditional independence. This may be estimated by the graphical lasso (Friedman *and others*, 2008), but several other estimators exist (Rothman *and others*, 2008; Rothman, 2012; Guo *and others*, 2011; Banerjee and Ghosal, 2014).

Other methods (e.g., Carvalho *and others*, 2008) carry out factor modeling, using known variables and latent factors to represent known modules, defined by biologists, and unknown modules, that one wishes to retrieve. Fan *and others* (2008, 2011) also use factor models to estimate large covariance matrices.

Bayesian methods often model the covariance matrix through either a non-informative prior or an inverse Wishart prior. Pourahmadi (2011) identifies several methods in Bayes and empirical Bayes covariance matrix estimation. One can model the logarithm of the covariance matrix using the log matrix prior, which introduces a multivariate normal prior with many hyperparameters. Although more flexible than an inverse Wishart prior, this prior is not easily interpreted. Carvalho *and others* (2007) describe an efficient method for direct sampling from the hyper-Wishart distribution. Conlon *and others* (2012) developed a Bayesian model to incorporate information on co-regulated genes that are close on the chromosome and co-transcribed. They construct a hierarchical Bayesian model to obtain the gene-specific posterior probability of differential expression, including co-regulation information, and show that this supplementary information improves their previous model, which assumes genes to be independent. Zhao *and others* (2012) introduced a Bayesian model to integrate microarray data with pathways obtained from KEGG. Corander *and others* (2013) developed a predictive supervised Bayesian classifier for several classes, assumed to be represented by multivariate Gaussian distributions and known in advance, while the covariance matrix is assumed to be block-diagonal. One can also model the correlation matrices, as done by Barnard *and others* (2000), who use log-normal priors on variances, independent of a prior on the matrix of correlations R.

Abadir *and others* (2012) noticed that orthogonal matrices from the orthogonal decomposition of the sample covariance matrix are never ill-conditioned, and therefore they focus on estimating the eigenvalues. Fan *and others* (2013) apply the adaptive thresholding estimator of Cai and Liu (2011) to principal orthogonal complements, in order to construct a sparse covariance matrix.

Estimation of large covariance matrices is thus well studied. Some methods focus on the estimation of a covariance matrix common to several studies, either by estimating a single covariance matrix for all groups, or by independently estimating the covariance matrices of each group and identifying their similarities (e.g., Gaskins and Daniels, 2013; Guo *and others*, 2011). In Section 2, we present an empirical Bayes procedure of the former type that estimates a sparse correlation matrix common to several studies.

2. Empirical Bayes estimation of a large sparse correlation matrix

2.1. *Model*

Based on correlations from several studies of the same phenomenon, we estimate a common sparse correlation matrix and use it to discover modules. The key idea is to treat the results from applying the Fisher transformation to the correlation coefficients as normally distributed. This transformation was also used by Hafdahl (2007) to combine correlation matrices in a fixed effects meta-analysis, but inference there was instead based on a spike-and-slab model (Mitchell and Beauchamp, 1988) whose parameters can be estimated by maximum likelihood.

We start from the sample correlation matrices from the gene expression matrices of S independent studies, denoted by $R^{(1)}, \ldots, R^{(S)}$, of size $G \times G$, from which we obtain the vectors $r^{(1)}, \ldots, r^{(S)}$, containing the vectorized upper diagonal parts of the correlation matrices. Applying Fisher's transformation to these vectors leads to a set of vectors $\{Z^{(1)}, \ldots, Z^{(S)}\}$, with elements

$$Z_i^{(s)} = \frac{1}{2} \log \left(\frac{1 + r_i^{(s)}}{1 - r_i^{(s)}} \right), \quad i = 1, \ldots, G(G-1)/2, \quad s = 1, \ldots, S.$$

We now assume that each of the correlations satisfies the model

$$Z_i^{(s)} \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma_s^2), \quad \theta_i \sim (1-p)\delta(\theta) + p\mathcal{N}(0, \tau^2), \; i = 1, \ldots, G(G-1)/2, \; s = 1, \ldots, S, \quad (2.1)$$

where the Dirac function $\delta(\theta)$ puts unit mass at $\theta = 0$, and all the variables are taken to be conditionally independent. The prior for θ allows some of the correlations to be exactly zero, thus imposing sparsity on the estimated common correlation matrix, but allows the variances of the transformed correlations to depend upon the study, owing for example to differences in study sizes. Calculations in Section 2.1 of the supplementary material available at *Biostatistics* online establish that the posterior density for θ_i is

$$\pi\left(\theta_i \mid Z_i^{(1)}, \ldots, Z_i^{(S)}\right) = p_z \sqrt{b}\, \varphi \left(\frac{\theta_i - b^{-1} \sum_{s=1}^{S} Z_i^{(s)}/\sigma_s^2}{b^{-1/2}} \right) + (1 - p_z)\delta(\theta_i), \quad (2.2)$$

where φ denotes the standard normal density function, $p_z = w_1/(w_1 + w_2)$, and

$$b = \frac{1}{\tau^2} + \sum_{s=1}^{S} \frac{1}{\sigma_s^2}, \quad w_1 = \frac{p}{\sqrt{b}\tau} \exp\left\{ \frac{1}{2b} \left(\sum_{s=1}^{S} \frac{Z^{(s)}}{\sigma_s^2} \right)^2 \right\} \prod_{s=1}^{S} \frac{1}{\sigma_s} \varphi\left(\frac{Z^{(s)}}{\sigma_s} \right), \quad w_2 = (1-p) \prod_{s=1}^{S} \frac{1}{\sigma_s} \varphi\left(\frac{Z^{(s)}}{\sigma_s} \right).$$

As we have an explicit expression for the joint marginal density of the $Z_i^{(s)}$, we can estimate the parameters $p, \sigma_1, \ldots, \sigma_S$ and τ by maximizing the independence log likelihood given by

$$l(\sigma_1, \ldots, \sigma_S, p, \tau) = \sum_{i=1}^{G(G-1)/2} \log \left[(1-p)\tilde{w}_2 + \frac{p\tilde{w}_2}{\tau\sqrt{b}} \exp \left\{ \frac{1}{2b} \left(\sum_{s=1}^{S} \frac{Z_i^{(s)}}{\sigma_s^2} \right)^2 \right\} \right], \quad \tilde{w_2} = \prod_{s=1}^{S} \frac{1}{\sigma_s} \varphi \left(\frac{Z^{(s)}}{\sigma_s} \right).$$

In practice we base these estimates on a randomly selected subset of the possible correlations for 10^5 pairs of genes; this is much faster and still gives precise estimates of the parameters. If convergence fails for 10^5 pairs, we try again with $10^6, 10^7, \ldots$, pairs. Using all $G(G-1)/2$ possible pairs would give spuriously precise estimates, because $l(\sigma_1, \ldots, \sigma_S, p, \tau)$ does not account for the dependence between the correlations.

The posterior mean is a natural summary for θ, but, as we seek a sparse estimate, we prefer the posterior median $\tilde{\theta}$, which performs a form of soft shrinkage (Johnstone and Silverman, 2005; Davison, 2008); see Section 2.2 of the supplementary material available at *Biostatistics* online. The common correlation matrix estimate $\tilde{R}$ is then obtained by applying the inverse of Fisher's transformation to the matrix $\tilde{\Theta}$ having entries $\tilde{\theta}$ in the upper diagonal part and filling the lower diagonal part by symmetry. Our estimate is not in general a correlation matrix, as it is not constrained to be positive definite. However, we shall use it for clustering genes, for which positive definiteness is not essential. Moreover, the method does not require all studies to have the same set of genes, because we only use the correlation from studies that have a particular pair of genes.

We extract modules, groups of highly correlated genes, from $\tilde{R}$ by using $1 - \tilde{R}$ as a dissimilarity matrix in a hierarchical clustering algorithm. The method returns a hierarchical tree, which we cut at the desired height to obtain the modules.

Figure 1 summarizes the entire procedure.

2.2. *Selecting the optimal number of modules*

In clustering, where the goal is to discover previously unknown groups of objects, the number of clusters is unknown in advance and needs to be determined, either as an input of the algorithm, as in k-means, or as the cutting height of the tree in hierarchical clustering. Dudoit and Fridlyand (2002) present several methods for estimating the number of clusters. The optimal number of clusters is usually not assessed directly; rather, results from a clustering algorithm for different numbers of clusters k are compared, and the optimal number of clusters is defined as the configuration optimizing some quantity. Methods to determine the number of clusters are based on either internal or external indices. Internal indices typically use functions of between- or within-cluster sums of squares and exploit the observations used to produce the clustering; examples are the silhouette plot, the GAP statistic (Tibshirani *and others*, 2001), consensus clustering (Monti *and others*, 2003), or Clest (Dudoit and Fridlyand, 2002). External indices are based on measures of agreement between two clustering results, such as the Rand (1971) index. Some of these methods are based on resampling and are therefore very computationally intensive for the high-dimensional datasets arising in genomics.

We use the estimated common correlation matrix $\tilde{R}$ as input to a clustering algorithm, defining the dissimilarity matrix as $1 - \tilde{R}$. We estimate the optimal number of modules using a modified GAP statistic, which can be applied on large datasets even when only the distance matrix is available. We let $\mathcal{C}_k = \{C_1, \ldots, C_k\}$ denote a partition of the genes of size k, for a predefined number of modules $k = 1, \ldots, K_{\max}$,which attributes each gene to a module $C_l \in \mathcal{C}_k$, where $|C_l| = n_l$. We summarize each module by the sum of the pairwise distances between its genes, $D_l = \sum_{i,j \in C_l} d_{ij}$, and each partition

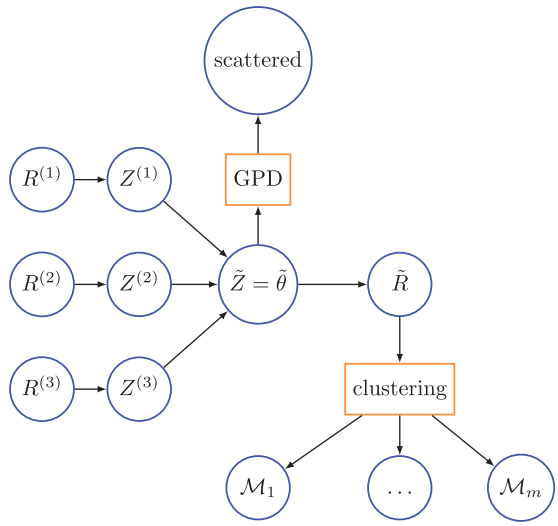

Fig. 1. Algorithm to estimate modules using correlation matrices $R^{(1)}, R^{(2)}, R^{(3)}$ from three studies, with corresponding matrices $Z^{(1)}, Z^{(2)}, Z^{(3)}$ on the Fisher scale. The empirical Bayes procedure yields the combined thresholded matrix $\tilde{Z}$, from which scattered genes are identified using the GPD mixture. The other elements of $\tilde{Z}$ are back-transformed using the inverse Fisher transformation, leading to $\tilde{R}$. Hierarchical clustering is applied to find modules $\mathcal{M}_1, \ldots, \mathcal{M}_m$, where m is estimated by the modified GAP procedure.

of k modules by the average pairwise distance between its genes,

$$W_k = \sum_{l=1}^{k} \frac{1}{2n_l} D_l.$$

The idea of the GAP statistic is to compare $\log W_k$ with its expected value computed under a reference distribution obtained by resampling. The optimal number of clusters $\widehat{k}$ is obtained as the value maximizing the GAP statistic,

$$\mathrm{GAP}_B(k) = \mathbb{E}_B(\log W_k) - \log W_k, \tag{2.3}$$

where B indicates the number of resamples used to estimate the reference distribution. Tibshirani *and others* (2001) generate a uniform reference distribution over the range of each observed feature (gene). In the original article, the authors therefore use the data matrix to obtain the reference distribution. Since we only have the common correlation matrix estimate, this resampling procedure is not directly applicable. Instead we break the structure of the correlation matrix, so that it has no module under the null hypothesis, i.e., the number of clusters $k = 1$. To obtain a reference distribution, we insert the elements of the upper triangular part of the common correlation matrix $\tilde{R}$ into a vector, randomly permute it, and reinject it as the upper triangular part of the resampled matrix $\tilde{R}^{(b)}$, which we fill in by imposing 1s on the diagonal and completing the rest by symmetry. The process is repeated B times, $W_k^{(b)}$ is calculated for each resample ($b = 1, \ldots, B$; $k = 1, \ldots, K_{\max}$), and estimates of $\mathbb{E}_B(\log W_k)$ and its standard deviation are obtained, i.e.,

$$\widehat{\mu} = \frac{1}{B} \sum_{b=1}^{B} \log W_k^{(b)}, \quad \mathrm{sd}_k = \left\{ \frac{1}{B} \sum_{b=1}^{B} \left(\log W_k^{(b)} - \widehat{\mu} \right)^2 \right\}^{1/2}.$$

The optimal number of clusters is chosen to minimize the GAP statistic (2.3), but to allow for its variability we should choose the smallest k satisfying (Tibshirani *and others*, 2001)

$$\mathrm{GAP}(k) \geq \mathrm{GAP}(k+1) - (1+1/B)^{1/2}\mathrm{sd}_{k+1}.$$

In the original method, resampling is performed for each k, which may be computationally intensive, so we used a dichotomized version. This tries several values of k in a predefined set, chooses the optimum for this set and then repeats this procedure with a more constrained set with values around the optimum, until the distance between the values in the set equals unity. The resulting algorithm to find the optimal number of clusters, $\widehat{k}_{\mathrm{best}}$, in a set of integers $\{K_{\min}, \ldots, K_{\max}\}$ is:

1. Construct a test set $\mathcal{K}_{\mathrm{set1}} = \{K_{\min} = k_1, \ldots, k_n = K_{\max}\}$, where $k_{i+1} - k_i = \Delta = (K_{\max} - K_{\min})/n$, with $n = 20$, say;
2. Use the GAP statistic to find the optimal $\widehat{k}_{\mathrm{set1}} \in \mathcal{K}_{\mathrm{set1}}$, and create a new test set $\mathcal{K}_{\mathrm{set2}}$ of size n around $\widehat{k}_{\mathrm{set1}}$, with $k_{i+1} - k_i = \Delta/2$;
3. Repeat step 2 until $\Delta = 1$, and then return the optimal $\widehat{k}_{\mathrm{best}}$.

When the data are high dimensional, this algorithm avoids getting stuck at local minima and tests fewer values of k than the original method.

2.3. *Scattered genes*

Scattered genes do not belong to any module and should have low correlations with other genes. Tseng and Wong (2005) and Langfelder and Horvath (2007) developed methods to identify them.

Thalamuthu *and others* (2006) compare six clustering methods based on their ability to detect groups of genes in simulated microarray gene expression data. They conclude that in the presence of scattered genes, tight clustering (Tseng and Wong, 2005) and model-based clustering are best overall, while five of the methods recover the groups accurately when scattered genes are absent. We therefore identify scattered genes and remove them from the estimated correlation matrix, $\tilde{R}$, prior to clustering, in order to ensure accurate module detection.

In order to detect scattered genes, we use the variance of the Fisher transformed matrix, $\tilde{\Theta}$, for each gene:

$$T_g = \mathrm{var}\left(|\tilde{\Theta}_{gj}|\right), \quad j \neq g, \quad g,j = 1, \ldots, G, \tag{2.4}$$

where $\tilde{\Theta}$ is the matrix having entries $\tilde{\theta}$, the posterior median of θ, in the upper diagonal part, and with its lower diagonal filled by symmetry, as defined in Section 2.1. We identify scattered genes as those with the largest values of $-\log T_g$, i.e., the smallest variances. Simulations presented in Figure 2 of the supplementary material available at *Biostatistics* online show that this procedure seems to detect scattered genes quite well. Setting a threshold above which $-\log T_g$ corresponds to a scattered gene is not easy, mainly because we do not know its distribution. We tried resampling to obtain a null distribution with which to compare the observed value, but this is computationally intensive, especially since the data are high dimensional. As we are working with the tail of a distribution, extreme value theory can be useful in helping to determine a threshold.

Threshold selection remains an open question in statistics of extreme values, although several methods have been developed, e.g., by Wadsworth and Tawn (2012), who also provide a good review of the topic. In our context, we seek a method to identify scattered genes. The tail of the distribution of $-\log T_g$ mostly

contains scattered genes, with other genes mixed in, but we believe that the most extreme values are likely to stem from the former. We therefore decided to model the tail of the distribution as a mixture of two generalized Pareto distributions (GPD), one for the scattered genes, with parameters σ_1 and ξ, and the other with parameters σ_2 and ξ. The shape parameters are taken equal to obtain stable fits; see Rootzén and Zholud (2015). An observed value of our statistic t above a threshold u is thus assumed to have the density

$$f(t) = pf(t-u;\sigma_1,\xi) + (1-p)f(t-u;\sigma_2,\xi), \quad t > u, \tag{2.5}$$

where $f(x;\sigma,\xi) = \sigma^{-1}(1+\xi x/\sigma)_+^{-1/\xi-1}$, with $a_+ = \max(a,0)$ for real a. In order to ensure that the observed values are a mixture of scattered and non-scattered genes, we usually take u to be the 0.9 quantile of the observed values of T_g. The likelihood can be easily obtained, and parameters are estimated by maximum likelihood. Identification of scattered genes is done by computing, for each gene g, the estimated posterior probability of belonging to the first mixture component, based on the value of the corresponding statistic t_g defined in equation (2.4), i.e.,

$$\frac{\widehat{p}f(t_g;\widehat{\sigma}_1,\widehat{\xi})}{\widehat{p}f(t_g;\widehat{\sigma}_1,\widehat{\xi}) + (1-\widehat{p})f(t_g;\widehat{\sigma}_2,\widehat{\xi})}.$$

Genes are declared scattered if their posterior probability is larger than 0.5.

3. Numerical examples

3.1. *Simulation design*

We present results for two simulation designs. The first was introduced by Langfelder and Horvath (2007), who provide in their supplementary material available at *Biostatistics* online a simulation design intended to mimic real gene expression data, coded in their package, WGCNA. Briefly, they generate a module seed (the true module eigengene) for each module, and all genes belonging to that module are correlated with the seed. Modules consist of genes distributed symmetrically around the eigengene. The second design, described in Zollinger *and others* (2015), generates a gene expression matrix with a block diagonal covariance matrix from which we get a correlation matrix R_t of the form

$$R_t = \begin{pmatrix} A_1 & & 0 \\ & \ddots & \\ 0 & & A_B \end{pmatrix}, \quad A_b = \begin{pmatrix} 1 & \rho_b & \cdots & \rho_b \\ \rho_b & & & \vdots \\ \vdots & & \ddots & \rho_b \\ \rho_b & \cdots & \rho_b & 1 \end{pmatrix}, \quad b = 1,\ldots,B, \tag{3.6}$$

where the correlation ρ_b inside each block varies as a parameter in the simulations and the block sizes may vary. Scattered genes, if present, are generated as genes uncorrelated with all other genes. We then add Gaussian noise to the gene expression matrix, producing three noisy gene expression matrices. For each we compute the matrix of pairwise correlations, which are input to our model. The procedure is illustrated in Figure 3 of the supplementary material available at *Biostatistics* online, and will be particularly useful for comparing our method with WGCNA (Langfelder and Horvath, 2007) and tight clustering (Tseng and Wong, 2005), which require gene expression matrices as input.

In the simulations, and unless otherwise specified, the number of genes is $G = 100$, the number of studies is $S = 3$, the number of modules is $B = 10$, and the number of simulation replicates is

100. The dissimilarity matrix used for the clustering algorithm is $1 - \tilde{R}$, where $\tilde{R}$ is obtained from the procedure described in Section 2. We compare k-means or hierarchical clustering algorithms with a variety of linkage functions, associated with either the GAP statistic (Tibshirani *and others*, 2001) or consensus clustering (Monti *and others*, 2003). Results of these simulations, presented in Figure 4 of the supplementary material available at *Biostatistics* online, suggest using hierarchical clustering with Ward's link and the modified GAP statistic of Section 2.2 to select the optimal number of clusters.

3.2. *Simulations without scattered genes*

We first compare the performances of our approach and of WGCNA, a method for automatic module identification developed by Langfelder and Horvath (2007), in terms of module detection in the absence of scattered genes. Using the simulation design presented in Figure 3 of the supplementary material available at *Biostatistics* online, we let all blocks have the same correlation $\rho \in \{0.1, \ldots, 0.9\}$ and used Gaussian noise with fixed variance $\sigma_\epsilon = 0.5$. Results presented in Figure 5 of the supplementary material available at *Biostatistics* online correspond to a hierarchical clustering algorithm with Ward's link and the number of clusters, $B = 10$, provided to both methods. Our method detects the modules almost perfectly if $\rho \geq 0.4$, and its performance remains reasonable for $\rho \approx 0.3$ but deteriorates for smaller ρ. We also used the simulation design from Langfelder and Horvath (2007), with three studies, 100 simulations, Nmodules $= 10$, Npatients $= 50$, nGenes $= 100$, eigengenes generated according to a standard normal distribution, with 20% of the genes in each module, and no gray genes (i.e., scattered genes, see Section 3.3), minimum correlation 0.1, and maximum correlation 1, background noise between 0.3 and 0.9 with six levels, and no negative correlation. Comparisons using the Rand index, presented in Figure 5 of the supplementary material available at *Biostatistics* online, show that our approach detects modules better here than does WGCNA. Different levels of noise added in the design of Langfelder and Horvath (2007) affect the results only slightly. In these simulations, clusters are defined using the fixed-cut tree procedure, with the true number of clusters given to both methods. Langfelder and Horvath (2007) suggest using their dynamic approach, which may give different results; we use this below when introducing scattered genes.

In these first simulations, the number of modules was known to both methods. We also performed simulations where the number of modules must be estimated, using our simulation design, but with a matrix with $G = 500$ genes divided into $B = 50$ modules and no scattered genes. Results in Figure 6 of the supplementary material available at *Biostatistics* online show that here our method has a better Rand index than WGCNA for every value of the intra-module correlation ρ, and can retrieve the number of modules perfectly, whereas WGCNA tends to select too many modules.

3.3. *Simulations with scattered genes*

We now compare our empirical Bayes approach, using the GPD mixture to detect scattered genes, with WGCNA (Langfelder and Horvath, 2007) and tight clustering (Tseng and Wong, 2005), in terms of detection of scattered genes and accuracy of the gene modules. Tight clustering was not used in the comparison of Section 3.2, as it finds scattered genes automatically. For each method, we record the numbers of correctly- and incorrectly identified scattered genes, from which we compute the sensitivity and specificity of scattered-gene detection, and the Rand index of the final partition compared to the correct partition. Detection of scattered genes is completely automatic for each method, but the number of modules is provided to the algorithms. For our method, scattered genes are detected by fitting the GPD mixture described in Section 3.3. They are then removed, and the resulting correlation matrix is used as input to a hierarchical clustering algorithm with Ward's linkage. Tight clustering and WGCNA have their own procedures for the detection of scattered genes. We compute the Rand index on the entire set of genes, treating the scattered genes as a single module. In these simulations, we used $G = 500$

genes of which 50 are scattered, $B = 50$ modules, and used 100 replicates for each value of the intra-module correlation $\rho_b = 0.1, \ldots, 0.9$ ($b = 1, \ldots, 50$) in (3.6). Results in Figure 7 of the supplementary material available at *Biostatistics* online show that the Rand index for our method is at least as good as for the other methods, especially for smaller correlations. Scattered gene detection using the GPD mixture tends to be conservative, with very few false positives (high specificity) coupled with somewhat low sensitivity.

We now turn to comparisons in which the number of modules must be estimated, with the simulation design and values of the simulation parameters as described above. For the modified GAP statistics, we used 10 permutations, step size $\Delta = 10$, and k from 2 to 200. We compare our procedure with WGCNA and tight clustering using the Rand index, sensitivity and specificity regarding the detection of scattered genes, and the number of modules detected; see Figure 2. Our method again has the best Rand index, and accurately estimates the true number of modules. Tight clustering estimates the number of modules perfectly, but its performance heavily depends on one parameter setting of the algorithm which must be set to around $B + 5$, where B is the targeted number of clusters. Setting different values for this parameter leads to much worse performance. Despite setting the parameter to $B + 5$, tight clustering has a lower Rand index than our method, perhaps owing to poor detection of scattered genes. WGCNA has a low Rand index when the intra-module correlation is low, rising to almost unity for $\rho > 0.8$. The same phenomenon appears for the number of modules; WGCNA tends to select far too many modules for low correlations, finding the correct number only when $\rho > 0.8$. Concerning the detection of scattered genes, our method again tends to be very conservative; it identifies very few scattered genes but has a high specificity. The sensitivity and specificity plots of Figure 2 cannot be looked at separately, as we hope for high values of both. Although WGCNA has high sensitivity for the detection of scattered genes, its specificity is very low: it has many false positives. Our approach therefore seems to be a good compromise: it has high Rand index, and tends to select scattered genes conservatively. Indeed, we prefer high specificity to high sensitivity, if a choice must be made. Since scattered genes are excluded from modules, we prefer that some of them are incorrectly included in modules, rather than that relevant genes are excluded from them.

We performed sensitivity analyses for all settings presented above by generating three matrices, one having gene names permuted at random to produce modules of the same size and strength but containing genes different from the other two matrices. We then ran all three methods on these studies, and compared the modules found with the truth. Results in Figures 8 and 9 of the supplementary material available at *Biostatistics* online show that our method is more robust to such errors than are the other methods.

4. Real data

4.1. *Our procedure*

Zollinger *and others* (2015) consider four studies (Mok *and others*, 2009; Yoshihara *and others*, 2009; Lili *and others*, 2013; Cancer Genome Atlas Research Network, 2011) that compare serous ovarian cancer and normal control samples; see Table 1. For each study, the sample correlation matrix based on the 9803 genes common to these datasets was obtained and our empirical Bayes procedure was applied. It took about 15 min to obtain the common correlation matrix, the estimated modules and their optimal number selected by the dichotomized GAP algorithm with $\Delta = 10$, 10 permutations and k in the range 100–1000. Our method identified 200 modules from 7 to 446 genes, with 14 scattered genes found by the GPD method of Section 3.3. The estimated number of scattered genes seems low, and others may have been erroneously assigned to modules.

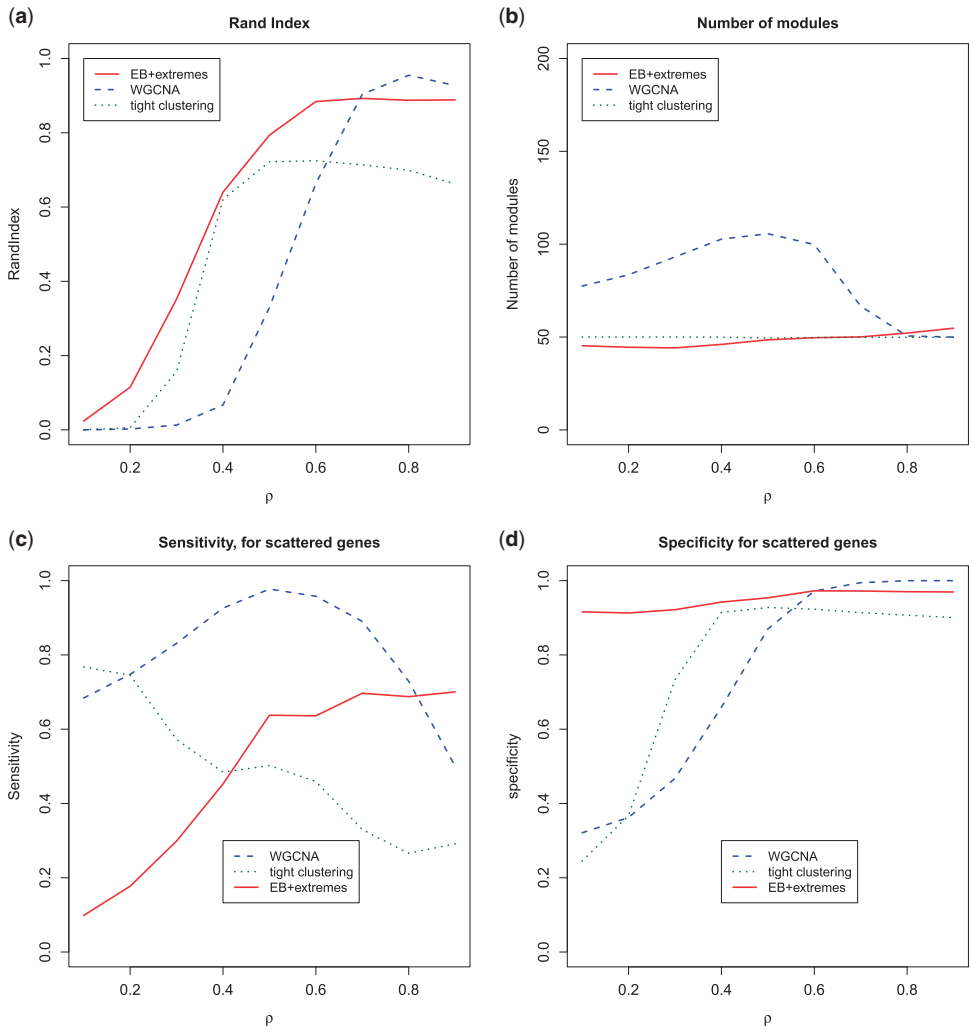

Fig. 2. Comparison of our method to WGCNA and tight clustering, using Rand index, number of modules compared to the true value ($B = 50$), and sensitivity and specificity for the detection of scattered genes.

4.2. *Comparison with WGCNA*

We also analyzed the same four datasets using WGCNA. Due to many missing values in the Yoshihara *and others* (2009) dataset, we could not fit WGCNA without removing most of the genes (only 832 genes met their selection criterion). Fitting to the other three studies took about 40 min, and yielded 286 modules and 4880 scattered genes; this corresponds to about half of the genes included in the analysis, and seems rather high. The Rand index between the clusterings returned by our method and WGCNA is only 0.011, indicating almost complete disagreement, but further direct comparison does not seem pertinent, because the truth is unknown.

Table 1. *Summary of the four studies included in the meta-analysis. SOC denotes serous ovarian cancer*

Study	MOK	YOS
Reference	Mok *and others* (2009)	Yoshihara *and others* (2009)
Data source	GEO GSE18520	GEO GSE12470
Platform	Affymetrix U133 Plus 2.0	Agilent Human 1A
Samples (SOC/controls)	53/10	43/10
Unique genes	20827	16546
Study	**LIL**	**TCGA**
Reference	Lili *and others* (2013)	Cancer Genome Atlas Research Network (2011)
Data source	GEO GSE38666	`curatedOvarianData`
Platform	Affymetrix U133 Plus 2.0	Affymetrix U133A
Samples (SOC/controls)	18/12	570/8
Unique genes	21049	12981

4.3. *Enrichment analysis*

To assess the enrichment of the modules defined by our method and by WGCNA, and in order to highlight their biological functions, if any, we used the Molecular Signature database (Liberzon *and others*, 2011, MSig DB), which provides pathways from several repositories. For our analysis, we retained the pathways from KEGG, Reactome, Biocarta, and GO, available under the c2 and c5 collections from the MSig database.

After downloading all the pathways, we kept only those containing between 5 and 200 genes, as suggested in Tseng *and others* (2012), and those having genes in common with our module set. The enrichment analysis included 6894 genes and 2336 pathways for our method, and 3573 genes and 1876 pathways for WGCNA. The scattered genes were removed from this enrichment analysis, which partially explains the lower number of genes for WGCNA. Enrichment of each module in each pathway for each method was tested using Fisher's exact test, and the test with the smallest p-value, or the most enriched term, was recorded for each module. Results are presented in Table 2, where p-values are adjusted using the Benjamini and Hochberg (1995) procedure. Almost all modules are enriched in one of the pathways included, and some are strongly enriched, with very small p-values, suggesting that modules found by our procedure are biologically relevant. Figure 3 suggests that these modules tend to be more enriched than those defined by WGCNA, based on the enriched pathways common to the two methods. We also compared the distribution of the sizes of the gene clusters detected by our method and WGCNA with the distribution of the sizes of KEGG pathways. Our method tends to detect clusters of comparable size to KEGG pathways, whereas WGCNA tends to detect much smaller modules (cf. Figure 10 from the supplementary material available at *Biostatistics* online).

5. Conclusion and discussion

We have outlined an empirical Bayes procedure for the estimation of a large sparse correlation matrix common to multiple, independent studies. The resulting estimate is not in general a correlation matrix, as it is not constrained to be positive definite, but it can anyway be used to estimate modules. The method is simple and fast, can be applied to large sets of genes, and simulations suggest that it works quite well and is insensitive to outliers. The detection of gene modules is facilitated by the sparsity of the estimated matrix. Moreover, we need not deal with one correlation matrix per study in order to identify the common modules, but need just one matrix $\tilde{R}$, which simplifies module detection.

Table 2. *Top 25 enriched modules. The p-values correspond to Fisher's exact tests and are corrected by the Benjamini and Hochberg (1995) procedure*

	EB			WGCNA		
Rank	ID	Pathway	*p*-value	ID	Pathway	*p*-value
1	175	Reactome peptide chain elongation	2.83e-76	3	Cell cycle process	1.25e-25
2	133	Cell cycle process	5.70e-29	38	Reactome influenza viral RNA transcription and replication	1.39e-18
3	159	KEGG antigen processing and presentation	6.29e-28	14	Reactome 3 utr mediated translational regulation	8.62e-16
4	167	Reactome interferon signaling	3.75e-27	169	Reactome interferon alpha beta signaling	8.87e-15
5	127	Reactome s phase	6.24e-25	170	KEGG asthma	2.25e-13
6	199	Reactome RNA pol i promoter opening	2.47e-23	50	Structural constituent of ribosome	5.98e-13
7	54	KEGG ecm receptor interaction	2.25e-16	59	KEGG ecm receptor interaction	2.03e-12
8	7	Biocarta no2il12 pathway	7.88e-13	55	Reactome antigen presentation folding assembly and peptide loading of class i mhc	2.24e-12
9	188	Chemokine activity	6.81e-11	2	Substrate specific channel activity	2.47e-10
10	30	Proteinaceous extracellular matrix	9.83e-11	1	Reactome tca cycle and respiratory electron transport	1.49e-08
11	118	Cellular defense response	3.34e-08	80	Reactome interferon alpha beta signaling	5.20e-08
12	37	Reactome unfolded protein response	9.81e-07	217	KEGG oxidative phosphorylation	1.21e-07
13	91	Synaptogenesis	7.63e-06	25	Reactome pol switching	1.95e-06
14	78	Response to bacterium	8.58e-06	63	Reactome immunoregulatory interactions between a lymphoid and a non lymphoid cell	3.39e-05
15	63	Cell projection part	1.32e-05	201	Protease inhibitor activity	3.39e-05
16	145	Biocarta spry pathway	1.54e-05	209	Reactome notch1 intracellular domain regulates transcription	3.39e-05
17	172	Reactome respiratory electron transport	2.84e-05	34	Collagen binding	3.86e-05
18	60	KEGG vascular smooth muscle contraction	3.95e-05	6	RNA export from nucleus	1.05e-04
19	25	Reactome cell cell junction organization	4.14e-05	56	Sensory perception	1.05e-04
20	67	Epidermis development	4.14e-05	151	Regulation of mitosis	1.13e-04
21	99	Biocarta atm pathway	4.14e-05	9	KEGG RNA polymerase	1.16e-04
22	31	Synapse part	5.42e-05	7	Response to wounding	1.59e-04
23	22	Rhodopsin like receptor activity	5.53e-05	17	Leukocyte activation	2.28e-04
24	41	Organic anion transmembrane transporter activity	5.53e-05	144	Small gtpase regulator activity	2.28e-04
25	52	Transmembrane receptor protein kinase activity	5.53e-05	246	Carbon carbon lyase activity	2.28e-04

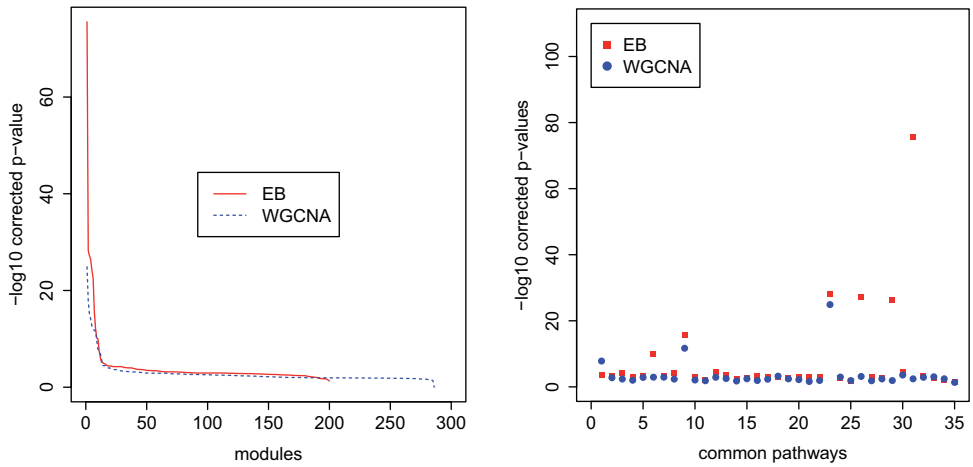

Fig. 3. Comparison of the enrichment of the modules defined by WGCNA and our empirical Bayes method. The *p*-values are corrected by the Benjamini–Hochberg procedure and are on the $\log_{10}$ scale. *Left*: comparison for all pathways (which may differ between methods). *Right*: Comparison of the enrichment of those pathways found by both methods.

We have also developed methods for the automatic selection of modules, first identifying scattered genes by fitting a mixture of GPD to the tail of a transformed variance; this procedure works well in our simulations but detects few scattered genes in the real data application. Using standard hierarchical clustering methods, we could readily identify modules, and simulations suggest that our method is equivalent to or better than its competitors. Our proposed modification of the GAP statistic allows it to be applied to larger datasets than could be handled by the original version. Application of our procedure to real data showed promising results: most of the modules it defines are significantly enriched in some biological pathways, and, in particular, we have been able to define more enriched modules than WGCNA. Our method is also faster, simpler and involves setting fewer parameters than competing methods, which makes it easier to apply in practice.

Since the method is based only on correlation matrices, and not on a model for raw data, it can be applied more broadly.

6. Software

An R package to perform the analyses is available from github (https://github.com/azolling/EBmodules).

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank M. Delorenzi for useful comments on the real-data application and the reviewers for their helpful remarks. *Conflict of Interest*: None declared.

REFERENCES

ABADIR, K. M., DISTASO, W. AND ZIKES, F. (2012). Design-free estimation of large variance matrices. *Technical Report*. Imperial College London.

BAI, J. AND SHI, S. (2011). Estimating high dimensional covariance matrices and its applications. *Annals of Economics and Finance* **12**, 199–215.

BANERJEE, S. AND GHOSAL, S. (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electronic Journal of Statistics*, **8**, 2111–2137.

BARNARD, J., MCCULLOCH, R. AND MENG, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10**, 1281–1312.

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289–300.

BICKEL, P. J. AND LEVINA, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics* **36**, 2577–2604.

CAI, T. AND LIU, W. (2011). Adaptive thresholding of sparse covariance matrix estimation. *Journal of the American Statistical Association* **106**, 672–684.

CANCER GENOME ATLAS RESEARCH NETWORK. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615.

CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. AND WEST, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association* **103**, 1438–1456.

CARVALHO, C. M., MASSAM, HELENE AND WEST, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika* **94**, 647–659.

CONLON, E. M., POSTIER, B. L., METHÉ, B. A., NEVIN, K. P. AND LOVLEY, D. R. (2012). A Bayesian model for pooling gene expression studies that incorporates co-regulation information. *PloS One* **7**, e52137.

CORANDER, J., KOSKI, T., PAVLENKO, T. AND TILLANDER, A. (2013). Bayesian block-diagonal predictive classifier for Gaussian data. In: Kruse, R., Berthold, M. R., Moewes, C., Gil, M. A., Grzegorzewski, P. and Hryniewicz, O. (editors), *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*. Berlin: Springer, pp. 543–551.

DAVISON, A. C. (2008). Some challenges for statistics. *Statistical Methods and Applications* **17**, 167–181.

DUDOIT, S. AND FRIDLYAND, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* **3**, 1–21.

EIN-DOR, L., KELA, I., GETZ, G., GIVOL, D. AND DOMANY, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **21**, 171–178.

EIN-DOR, L., ZUK, O. AND DOMANY, E. (2006). Thousand of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences* **103**, 5923–5928.

FAN, J., FAN, Y. AND LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147**, 186–197.

FAN, J., LIAO, Y. AND MINCHEVA, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics* **39**, 3320–3356.

FAN, J., LIAO, Y. AND MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B* **75**, 603–680.

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

GASKINS, J. T. AND DANIELS, M. J. (2013). A nonparametric prior for simultaneous covariance estimation. *Biometrika* **100**, 125–138.

GUO, J., LEVINA, E., MICHAILIDIS, G. AND ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98**, 1–15.

HAFDAHL, A. R. (2007). Combining correlation matrices: Simulation analysis of improved fixed-effects methods. *Journal of Educational and Behavioral Statistics* **32**, 180–205.

JOHNSTONE, I. M. AND SILVERMAN, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Annals of Statistics* **33**, 1700–1752.

LANGFELDER, P. AND HORVATH, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology* **1**, 1–54.

LEDOIT, O. AND WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* **88**, 365–411.

LEDOIT, O. AND WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Annals of Statistics* **40**, 1024–1060.

LIBERZON, A., SUBRAMANIAN, A., PINCHBACK, R., THORVALDSDÓTTIR, H., TAMAYO, P. AND MESIROV, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740.

LILI, L. N., MATYUNINA, L. V., WALKER, L., BENIGNO, B. B. AND MCDONALD, J. F. (2013). Molecular profiling predicts the existence of two functionally distinct classes of ovarian cancer stroma. *BioMed Research International* **2013**, 1–9.

MITCHELL, T. J. AND BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**, 1023–1032.

MOK, S. C., BONOME, T., VATHIPADIEKAL, V., BELL, A., JOHNSON, M. E., WONG, K.-K., PARK, D.-C, HAO, K., YIP, D. K. P., DONNINGER, H. *and others*. (2009). A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: Microfibril-associated glycoprotein 2. *Cancer Cell* **16**, 521–532.

MONTI, S., TAMAYO, P., MESIROV, J. AND GOLUB, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* **52**, 91–118.

POURAHMADI, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science* **26**, 369–387.

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* **66**, 846–850.

ROOTZÉN, H. AND ZHOLUD, D. (2015). Efficient estimation of the number of false positives in high-throughput screening. *Biometrika* **102**, 695–704.

ROTHMAN, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika* **99**, 733–740.

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. AND ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2**, 494–515.

THALAMUTHU, A., MUKHOPADHYAY, I., ZHENG, X. AND TSENG, G. C. (2006). Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics* **22**, 2405–2412.

TIBSHIRANI, R. J., WALTHER, G. AND HASTIE, T. J. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: series B* **63**, 411–423.

TSENG, G., GHOSH, D. AND FEINGOLD, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research* **40**, 3785–3799.

TSENG, G. C. AND WONG, W. H. (2005). Tight clustering: A resampling-based approach for identifying stable and tight patterns in data. *Biometrics* **61**, 10–16.

WADSWORTH, J. L. AND TAWN, J. A. (2012). Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling. *Journal of the Royal Statistical Society: Series B* **74**, 543–567.

WIRAPATI, P., SOTIRIOU, C., KUNKEL, S., FARMER, P., PRADERVAND, S., HAIBE-KAINS, B., DESMEDT, C., IGNATIADIS, M., SENGSTAG, T., SCHÜTZ, F. *and others*. (2008). Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Research* **10**, R65.

YOSHIHARA, K., TAJIMA, A., KOMATA, D., YAMAMOTO, T., KODAMA, S., FUJIWARA, H., SUZUKI, M., ONISHI, Y., HATAE, M., SUEYOSHI, K. *and others*. (2009). Gene expression profiling of advanced-stage serous ovarian cancers distinguishes novel subclasses and implicates ZEB2 in tumor progression and prognosis. *Cancer Science* **100**, 1421–1428.

ZHAO, Y., CHEN, M.-H., PEI, B., ROWE, D., SHIN, D.-G., XIE, W., YU, F. AND KUO, L. (2012). A Bayesian approach to pathway analysis by integrating gene–gene functional directions and microarray data. *Statistics in Biosciences* **4**, 105–131.

ZOLLINGER, A., DAVISON, A. C. AND GOLDSTEIN, D. R. (2015). Meta-analysis of incomplete microarray studies. *Biostatistics* **16**, 686–700.

[*Received May 17, 2016; revised May 9, 2017; accepted for publication May 23, 2017*]