Application of a New Ridge Estimator of the Inverse Covariance Matrix to the Reconstruction of Gene-Gene Interaction Networks

Wessel N. van Wieringen[1,2,⋆] and Carel F.W. Peeters[1]

[1] Department of Epidemiology and Biostatistics, VU University medical center, P.O. Box 7057, 1007 MB Amsterdam, The Netherlands
{w.vanwieringen,cf.peeters}@vumc.nl
[2] Deptartment of Mathematics, VU University Amsterdam, 1081 HV Amsterdam, The Netherlands

Abstract. A proper ridge estimator of the inverse covariance matrix is presented. We study the properties of this estimator in relation to other ridge-type estimators. In the context of Gaussian graphical modeling, we compare the proposed estimator to the graphical lasso. This work is a brief exposé of the technical developments in [1], focussing on applications in gene-gene interaction network reconstruction.

Keywords: Gaussian graphical model, Gene-gene interaction networks, Multivariate normal, Penalized estimation, Precision matrix.

1 Introduction

1.1 Scientific Background

Molecular biology aims to understand the molecular processes that occur in the cell. That is, which molecules present in the cell interact, and how are the interactions coordinated? For many cellular process, it is unknown which genes play what role.

A valuable source of information to uncover gene-gene interactions are (onco)genomics studies. Such studies comprise samples from n individuals with, e.g., cancer of the same tissue. Each sample is interrogated molecularly and the expression levels of many (p) genes are measured simultaneously. The resulting p-dimensional data vector is denoted $\mathbf{Y}_{i,*}$ for individual $i = 1, \ldots, n$.

From these data the gene-gene interaction network may be unraveled when the presence (absence) of a gene-gene interaction is operationalized as a conditional (in)dependency between the corresponding gene pair. Then, under the assumption of multivariate normality, $\mathbf{Y}_{i,*} \sim \mathcal{N}(\mathbf{0}_{p\times 1}, \boldsymbol{\Sigma})$, the absence of direct gene-gene interactions corresponds to zeros in the inverse covariance matrix $\boldsymbol{\Omega} \equiv \boldsymbol{\Sigma}^{-1}$ (also known as the precision matrix, whose elements are proportional to partial correlations). For instance, $(\boldsymbol{\Omega})_{1,2} = 0 \Leftrightarrow Y_1 \perp\!\!\!\perp Y_2 \,|\, Y_3, \ldots, Y_p$.

⋆ Corresponding author.

C. di Serio et al. (Eds.): CIBB 2014, LNCS 8623, pp. 170–179, 2015.
DOI: 10.1007/978-3-319-24462-4_15

Hence, the gene-gene interaction network is found by inversion of the covariance matrix and (subsequent) determination of its support. When dealing with data, $\mathbf{\Sigma}$ is estimated by its sample counterpart: $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_{i,*}\mathbf{Y}_{i,*}^{\mathrm{T}}$.

In genomics the data are often high-dimensional, in the sense of $p > n$. In such situations the sample covariance matrix $\mathbf{S}$ is singular and the sample precision matrix is not defined. But even if $p < n$ and p approaches n, the sample precision matrix yields inflated partial correlations. Both situations require some form of regularization to obtain a well-behaved estimate of the precision matrix, and consequently of the gene-gene interaction network.

1.2 Ridge-Type Covariance Estimators

A penalized covariance estimator traditionally referred to as the 'ridge estimator' is:

$$\hat{\mathbf{\Sigma}}_{r_I}(\lambda_{r_I}) = \mathbf{S} + \lambda_{r_I}\mathbf{I}_{p\times p} \qquad \text{for } \lambda_{r_I} > 0.$$

It could be considered a ridge estimator in the sense that it is an ad-hoc fix of the singularity of $\mathbf{S}$, much like how ridge regression was originally introduced [2]. The inverse of $\hat{\mathbf{\Sigma}}_{r_I}(\lambda_{r_I})$ would then form the basis for inference on the gene-gene interaction network.

Alternatively, a 'ridge estimator' popularized by [3] in the field of genomics, is (cf. [4,5]):

$$\hat{\mathbf{\Sigma}}_{r_{II}}(\lambda_{r_{II}}) = (1-\lambda_{r_{II}})\mathbf{S} + \lambda_{r_{II}}\mathbf{\Gamma} \qquad \text{for } \lambda_{r_{II}} \in (0,1].$$

In this latter expression $\mathbf{\Gamma}$ is a $(p \times p)$-dimensional, symmetric positive definite (p.d.) target matrix. The target matrix is chosen prior to estimation. Its role is to serve as a 'null estimate' towards which the covariance estimate is shrunken as $\lambda_{r_{II}}$ tends to one. In the remainder we will mainly consider the following choice: $\mathbf{\Gamma}$ diagonal with $\mathrm{diag}(\mathbf{\Gamma}) = \mathrm{diag}(\mathbf{S})$. This represents a reasonable choice in the absence of any prior knowledge on the Gaussian process. Again, when determining the support of the precision matrix the inverse of this second 'ridge estimator' could be used.

Neither of the two ridge estimators above is a proper ridge estimator, in the sense that neither can be formulated as the result from the maximization of a loss function augmented with what is commonly perceived as the ridge penalty: the sum of the square of its elements.

1.3 Overview

In Section 2 an alternative ridge estimator for the inverse covariance matrix is presented. In Section 3 the proposed estimator is compared with the traditional ridge-type estimators and the graphical lasso. Section 4 illustrates, using oncogenomics data, practical usage of the proposed estimator in a graphical modeling setting. Section 5 carries some concluding remarks, while Section 6 closes with a small description of the accompanying software.

2 Materials and Methods

2.1 An Alternative Ridge Inverse Covariance Estimator

We consider estimation of the inverse covariance matrix with conventional ridge regularization. The alternative ridge estimator of the inverse covariance matrix maximizes the following penalized log-likelihood:

$$\mathcal{L}^{\text{pen}}(\boldsymbol{\Omega}; \mathbf{S}, \mathbf{T}, \lambda_a) = \ln|\boldsymbol{\Omega}| - \text{tr}(\mathbf{S}\boldsymbol{\Omega}) - f^{\text{pen}}(\boldsymbol{\Omega}, \mathbf{T}, \lambda_a), \tag{1}$$

where λ_a is the penalty parameter, $\mathbf{T}$ denotes a symmetric p.d. target matrix, and $f^{\text{pen}}(\cdot,\cdot,\cdot)$ indicates the penalty function. The ridge penalty function amounts to:

$$f^{\text{pen}}(\boldsymbol{\Omega}, \mathbf{T}, \lambda_a) = \frac{1}{2}\lambda_a \text{tr}[(\boldsymbol{\Omega} - \mathbf{T})^{\text{T}}(\boldsymbol{\Omega} - \mathbf{T})]. \tag{2}$$

In case $\mathbf{T} = \mathbf{0}_{p\times p}$, the penalty function reduces to $f^{\text{pen}}(\boldsymbol{\Omega}, \mathbf{T}, \lambda_a) = f^{\text{pen}}(\boldsymbol{\Omega}, \lambda_a) = \frac{1}{2}\lambda_a \sum_{j_1,j_2=1}^{p}[(\boldsymbol{\Omega})_{j_1,j_2}]^2$, which corresponds to the common perception of the ridge penalty. The penalty function (2) is thus a generalized ridge penalty.

We show (cf. [1]) that there is an explicit solution that maximizes the penalized log-likelihood (1) with the general ridge penalty (2):

$$\hat{\boldsymbol{\Omega}}^{\text{ridge}}(\lambda_a) = \left\{ \left[\lambda_a \mathbf{I}_{p\times p} + \frac{1}{4}(\mathbf{S} - \lambda_a \mathbf{T})^2 \right]^{1/2} + \frac{1}{2}(\mathbf{S} - \lambda_a \mathbf{T}) \right\}^{-1}. \tag{3}$$

This ridge precision estimator is p.d. when $\lambda_a \in (0, \infty)$ and can be viewed as a penalized maximum likelihood (ML) estimator. Moreover, in the low-dimensional case the ridge estimator (2) reduces to $\mathbf{S}^{-1}$ as $\lambda_a \downarrow 0$. When λ_a tends to infinity, $\hat{\boldsymbol{\Omega}}^{\text{ridge}}(\lambda_a)$ shrinks to $\mathbf{T}$, much like the covariance estimator of [3] shrinks to a user-specified target. Thus, when $\mathbf{T}$ is diagonal and $\text{diag}(\mathbf{T}) = 1/\text{diag}(\mathbf{S})$ the inverse of estimator (3) mimics the behaviour of the latter. Similarly, choosing $\mathbf{T} = \mathbf{0}_{p\times p}$ yields a ridge estimator of the precision matrix that shrinks to the null matrix as does the inverse of $\hat{\boldsymbol{\Sigma}}_{r_I}(\lambda_{r_I})$. The explicit form of our ridge estimator (3) allows us to calculate the moments of the estimator and prove its consistency [1].

2.2 Extracting an Interaction Network

When turning to the application of ridge estimation in Gaussian graphical modeling of gene-gene interaction networks, the proposed estimator (3) yields (after standardization) an estimate of the partial correlation matrix. In doing so, an informed choice of the penalty parameter needs to be made. Hereto we utilize an approximate leave-one-out cross-validation (LOOCV) procedure [6]. Finally, one needs to decide which elements of the partial correlation matrix are indistinguishable from zero, for which we employ the local false discovery rate (FDR) procedure of [3].

3 Results

3.1 Comparison with the Traditional Ridge Estimators

We compare the proposed ridge estimator (3) with the two other 'ridge estimators', $\hat{\mathbf{\Sigma}}_{r_I}(\lambda_{r_I})$ and $\hat{\mathbf{\Sigma}}_{r_{II}}(\lambda_{r_{II}})$. Analytically, we study the rate of shrinkage of the estimators. The proposed ridge precision estimator (3) with $\mathbf{T} = \mathbf{0}_{p\times p}$ displays slower shrinkage (with increasing penalty parameter) to the null target than $[\hat{\mathbf{\Sigma}}_{r_I}(\lambda_{r_I})]^{-1}$. As the target is degenerate, this behaviour is to be preferred. The opposite is seen when studying the shrinkage rate of estimator (3) with $\text{diag}(\mathbf{T}) = 1/\text{diag}(\mathbf{S})$ in relation to $[\hat{\mathbf{\Sigma}}_{r_{II}}(\lambda_{r_{II}})]^{-1}$ with $\mathbf{\Gamma} = \mathbf{T}^{-1}$. That is, the former shrinks faster to $\mathbf{T}$ than the latter. Whenever $\mathbf{T}$ is close to $\mathbf{\Omega}$, faster shrinkage is desirable. In a simulation study we turn to the comparison of the risk of the proposed ridge estimator and its contenders. For the scenario's studied, the former performs favourably.

3.2 Comparison with the Graphical Lasso

For the application to Gaussian graphical modelling, the inverse covariance matrix is often estimated by means of the graphical lasso [7,8], as it performs automated edge selection. The lasso precision estimator maximizes (1) under the alternative penalty $f'^{\text{pen}}(\mathbf{\Omega}, \mathbf{T}, \lambda_l) = \lambda_l \|\mathbf{\Omega}\|_1 = \lambda_l \sum_{j_1,j_2=1}^{p} |(\mathbf{\Omega})_{j_1,j_2}|$. To accommodate the diagonal target matrix $\mathbf{T}$ (with $\text{diag}(\mathbf{T}) = 1/\text{diag}(\mathbf{S})$) this penalty function may be replaced by $f''^{\text{pen}}(\mathbf{\Omega}, \mathbf{T}, \lambda_l) = \|\mathbf{\Lambda} \circ \mathbf{\Omega}\|_1$ in which $\circ$ denotes the Hadamard product and $(\mathbf{\Lambda})_{j_1,j_2} = \lambda_l$ when $j_1 \neq j_2$ and zero otherwise (as is implemented in the `glasso`-package [9] by the option `penalize.diagonal=FALSE`).

We compare the proposed ridge and lasso estimators of the standardized precision matrix, as this forms the basis for inference on the conditional independence graph (the standardized precision matrix equals the partial correlation matrix up to the sign of off-diagonal elements). This is done in a data-driven manner, to avoid bias towards any of the estimators. Five curated breast cancer studies with gene expression data generated by the same (or comparable) Affymetrix platform [10] are used for this purpose. The full data set is limited to sets of genes that map to a pathway (as defined by the KEGG repository [11]). High-dimensionality is then realized by drawing subsets of the pathway data at sample sizes $n = 5, 10, 25$. For each draw the covariance matrix is estimated by means of lasso and ridge procedures. For both the LOOCV is used to choose their penalty parameters. The ridge estimate is then subjected to the local FDR procedure to decide on the presence/absence of gene-gene interactions.

The sensitivity and specificity of the resulting ridge and lasso inferred conditional independencies are compared. Hereto we define a 'consensus truth' based on overlapping edges. The resulting sensitivity and specificity of edge retrieval is comparable between the proposed ridge and the lasso estimators. An alternative comparison focusses on the loss of the estimates of the standardized precision matrix. Then, the proposed ridge estimator clearly yields a lower loss. These observations are consistent over the sample sizes, pathways, and data sets considered. Figure 1 visualizes these observations for a particular pathway.

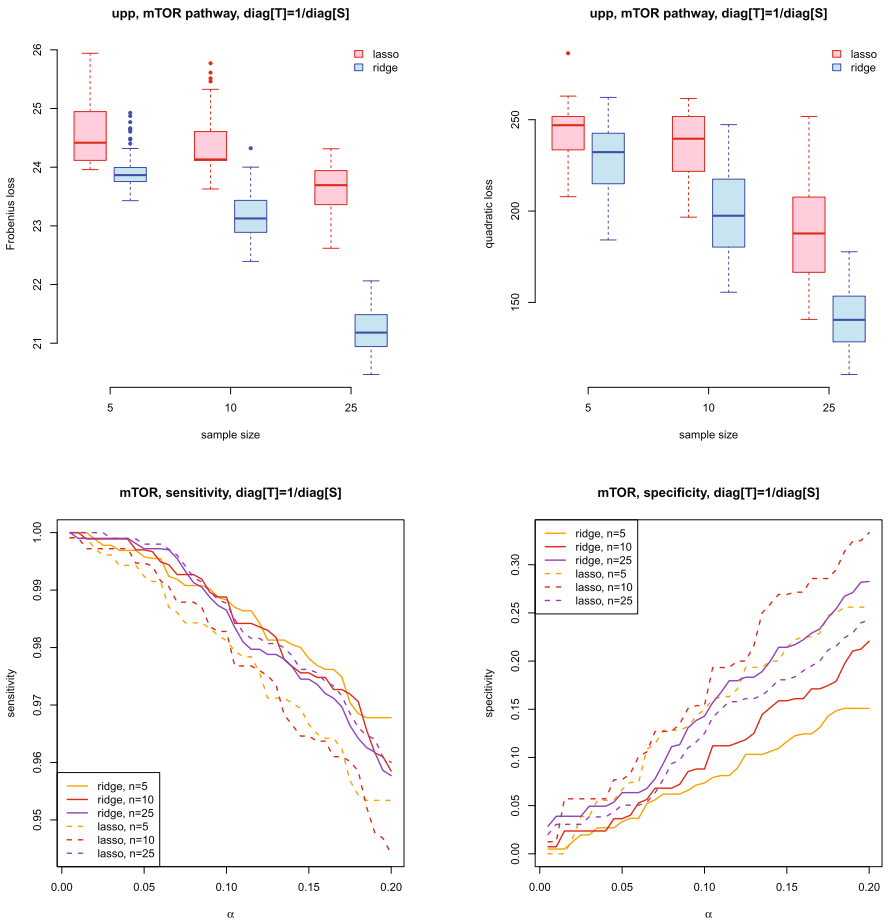

Fig. 1. The upper panels depict a loss comparison between the alternative ridge and the corresponding graphical lasso estimators for the mTOR-pathway on the UPP breast cancer data [10]. The loss is determined with a proxy of the standardized population precision matrix for the mTOR-pathway. The upper left-hand panel depicts Frobenius loss while the upper right-hand panel depicts quadratic loss. The lower panels depict a sensitivity and specificity comparison between the alternative ridge and graphical lasso estimators, again on the mTOR-pathway data. The evaluation of edge retrieval sensitivity and specificity requires knowledge of the true conditional dependencies. As such knowledge is absent we resort to defining a 'consensus truth', comprised of those conditional dependencies that appear in the top $100\alpha\%$ of at least 4 out of the 5 breast cancer data sets by both methods (graphical lasso and alternative ridge paired with local FDR edge selection). The parameter α ranges from .005 to .20, corresponding to what is believed to be biologically plausible (in terms of network density). Sensitivity (specificity) for a particular combination of n and α is then estimated as the median sensitivity (specificity) over the generated subsamples over all data sets. The lower left-hand panel gives sensitivity results while the lower right-hand panel gives specificity results.

4 Illustration

In this section we illustrate the reconstruction of a gene-gene interaction network from gene expression data using our R-implementation (see Section 6 below) of the proposed ML ridge estimator of the precision matrix. We employ breast cancer gene expression data by The Cancer Genome Atlas (TCGA) [12] of the mitogen-activated protein kinases (MAPK) pathway (as defined by KEGG).

For purposes of reproducibility we first provide the R-code that loads and 'processes' the data. It starts by activation of the necessary R-packages:

```
> library(biomaRt)
> library(cgdsr)
> library(KEGG.db)
> library(rags2ridges)
```

To get a list of all human genes and additional relevant information:

```
> ensembl = useMart("ensembl", dataset="hsapiens_gene_ensembl")
> geneList <- getBM(attributes=c("external_gene_name",
+                   "entrezgene"), mart = ensembl)
> geneList <- geneList[!is.na(geneList[,2]),]
```

Obtain the entrez IDs [13] of the genes that map to the MAPK pathway:

```
> kegg2entrez <- as.list(KEGGPATHID2EXTID)
> entrezIDs <- as.numeric(kegg2entrez[which(names(kegg2entrez)
+                          %in% "hsa04010")][[1]])
> entrez2name <- match(entrezIDs, geneList[,2])
> geneList <- geneList[entrez2name[!is.na(entrez2name)],]
```

Specify data set details (repository, TCGA study, samples, and profile):

```
> tcgaDB <- CGDS("http://www.cbioportal.org/public-portal/")
> cancerStudy <- "brca_tcga"
> caseList <- getCaseLists(tcgaDB, cancerStudy)[1,1]
> mrnaProf <- "brca_tcga_pub_mrna"
```

Extract the pathway expression data:

```
> Y <- getProfileData(tcgaDB, geneList[,1], mrnaProf, caseList)
> for (j in 1:ncol(Y)){
+      Y[,j] <- as.numeric(levels(Y[,j])[Y[,j]]) }
> Y <- data.matrix(Y)
```

Filter no-data samples and genes:

```
> sRemove <- which(rowSums(is.na(Y)) > ncol(Y)/10)
> Y <- Y[-sRemove,]
> gRemove <- which(colSums(is.na(Y)) > 0)
> Y <- Y[,-gRemove]
> Y <- sweep(Y, 2, colMeans(Y))
```

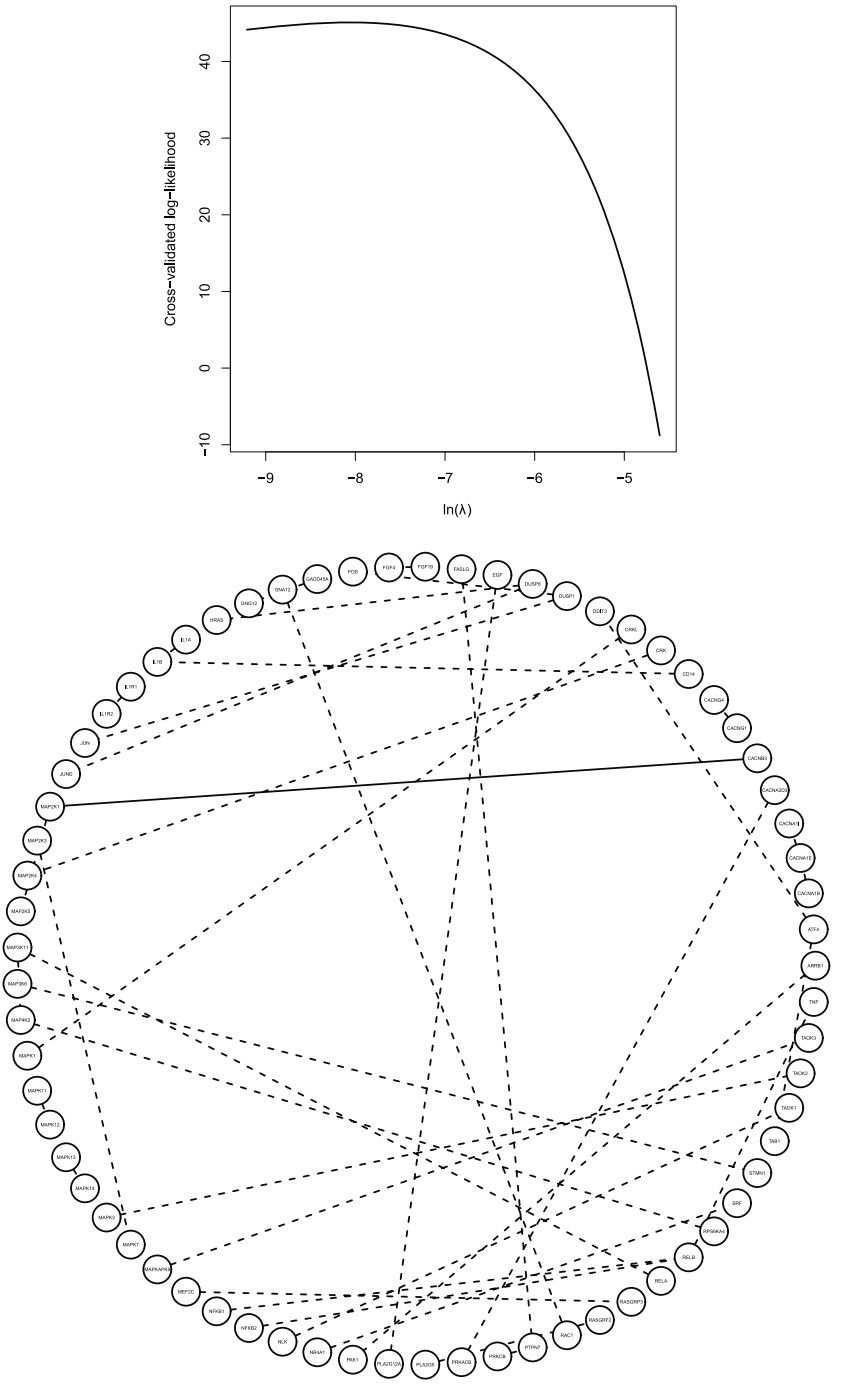

Fig. 2. Upper-panel: cross-validated log-likelihood. Bottom-panel: the inferred conditional independence graph of the MAPK pathway. Dashed lines indicate negative precision elements while solid lines indicate positive precision elements.

This concludes the executions in R required to obtain TCGA breast cancer data of the MAPK pathway as defined by KEGG. The gene expression data comprises $n = 496$ samples and $p = 259$ genes.

Finally, we turn to the reconstruction of the gene-gene interaction network of the MAPK pathway by means of the proposed ML ridge estimator of the precision matrix.[1] The target we use is $\mathbf{T} = \varphi \mathbf{I}_{p \times p}$, where φ denotes the average of the inverse (nonzero) eigenvalues of $\mathbf{S}$. Under this choice (3) is rotation equivariant, which is computationally advantageous (see Section 6). First, one needs to make an informed choice on the penalty parameter λ_a. This is done via the approximate LOOCV procedure (in which $\varphi \mathbf{I}_{p \times p}$ is the default target option):

```
> CVres <- optPenalty.aLOOCV(Y, 0.0001, 0.01, step=100)
```

The thus obtained cross-validated log-likelihood profile is plotted against the (logarithm of the) penalty parameter (see the upper-panel of Figure 2). The cross-validated log-likelihood achieves an optimum close to $\ln(\lambda_a) = -8.112$. This rather small value (little regularization) is due to the relative 'low-dimensionality' of the data.

With the optimal penalty parameter at hand the penalized ML ridge estimate of the precision matrix is obtained through:

```
> penPrec <- ridgeS(covML(Y), CVres$optLambda)
```

The object `penPrec` contains the desired estimate of the precision matrix that forms the basis for inferring the conditional independencies in the MAPK pathway data. Hereto the ML ridge estimate of the precision matrix is standardized to have a unit diagonal. The local FDR procedure of [3] is then applied to the off-diagonal elements of this standardized precision matrix. Edges corresponding to such elements with a posterior probability exceeding 0.95 are considered to be present in the gene-gene interaction network.

```
> P0 <- sparsify(penPrec, threshold="localFDR",
+                FDRcut=0.95)$sparsePrecision
```

The resulting sparsified precision matrix is visualized by its implied conditional independence graph:

```
> Ugraph(P0, type="fancy", prune=TRUE)
```

This gives an impression of the gene-gene interaction network underlying the MAPK pathway (see the bottom-panel of Figure 2).

5 Conclusion

We have presented a proper ML ridge estimator of the precision matrix, for which analytical properties can be proven. In a vis-á-vis comparison with other

[1] In the remainder of the illustration we use calls related to version 1.3 of our own R-package [15]. Please note that this package is in continual development so that certain calls may be depreciated or enhanced in future versions.

penalized inverse covariance estimators it was shown to yield a lower risk. Moreover, its performance is on a par with the graphical lasso with respect to the sensitivity and specificity of selected conditional independencies. Hence, the presented ridge estimator is a strong contender for inverse covariance estimation from high-dimensional data.

Currently, we are exploring the use of the target matrix $\mathbf{T}$. In this exposé we have limited ourselves to obvious choices. More sophisticated choices may be conceived. For instance, it may incorporate prior knowledge on the gene-gene interaction network as obtained from a pilot experiment or from repositories such as KEGG.

6 Software

The `R` [14] package `rags2ridges` [15] implements the proposed ridge precision estimator along with functions supporting subsequent graphical modeling. These additional functions enable, among others, (automated) penalty parameter selection, the evaluation of entropy and fit, support determination, (network) visualization, and network topology evaluation. The proposed estimator is analytic, making its computation friendly for a given penalty. When the chosen target implies rotation equivariance (i.e., the estimator leaves the eigenvectors of $\mathbf{S}$ unchanged), the search for an optimal penalty value and subsequent network extraction also become computationally efficient as the (relatively) expensive matrix square root can then be circumvented. In this situation only a single spectral decomposition and a single matrix inversion are required to obtain the complete solution path over any λ_a in the feasible domain. See the package documentation [15] for more information. The package is freely available from the Comprehensive `R` Archive Network [16].

Acknowledgments. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7, 2007-2013), Research Infrastructures action, under the grant agreement No. FP7-269553 (EpiRadBio project).

References

1. van Wieringen, W.N., Peeters, C.F.W.: Ridge Estimation of Inverse Covariance Matrices From High-Dimensional Data. arXiv:1403.0904 [stat.ME] (2014)
2. Hoerl, A.E., Kennard, R.W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12, 55–67 (1970)
3. Schäfer, J., Strimmer, K.: A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics. Stat. Appl. Genet. Mo. B. 4, Article 32 (2005)
4. Ledoit, O., Wolf, M.: A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. J. Multivariate Anal. 88, 365–411 (2004)

5. Fisher, T.J., Sun, X.: Improved Stein-Type Shrinkage Estimators for the High-Dimensional Multivariate Normal Covariance Matrix. Comput. Stat. Data An. 55, 1909–1918 (2011)
6. Vujačić, I., Abbruzzo, A., Wit, E.C.: A Computationally Fast Alternative to Cross-Validation in Penalized Gaussian Graphical Models. arXiv: 1309.621v2 [stat.ME] (2014)
7. Banerjee, O., El Ghaoui, L., d'Aspremont, A.: Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. J. Mach. Learn. Res. 9, 485–516 (2008)
8. Friedman, J., Hastie, T., Tibshirani, R.: Sparse Inverse Covariance Estimation with the Graphical Lasso. Biostatistics 9, 432–441 (2008)
9. Friedman, J., Hastie, T., Tibshirani, R.: `glasso`: Graphical Lasso-Estimation of Gaussian Graphical Models. `R` package, version 1.8 (2014), `http://cran.r-project.org/web/packages/glasso/index.html`
10. Schröder, M., Haibe-Kains, B., Culhane, A., Sotiriou, C., Bontempi, G., Quackenbush, J.: breastCancerMAINZ; breastCancerTRANSBIG; breastCancerUNT; breastCancerUPP; breastCancerVDX. `R` packages, versions 1.0.6 (2011), `http://compbio.dfci.harvard.edu/`
11. Kanehisa, M., Goto, S.: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27–30 (2000)
12. The Cancer Genome Atlas, `http://cancergenome.nih.gov/`
13. The National Center for Biotechnology Information, `http://www.ncbi.nlm.nih.gov/gene`
14. `R` Development Core Team: `R`: A Language and Environment for Statistical Computing. `R` Foundation for Statistical Computing, Vienna, Austria (2011)
15. Peeters, C.F.W., van Wieringen, W.N.: `rags2ridges`: Ridge Estimation of Precision Matrices from High-Dimensional Data. `R` package, version 1.3 (2014), `http://cran.r-project.org/web/packages/rags2ridges/index.html`
16. Comprehensive `R` Archive Network, `http://www.R-project.org/`