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.