Computational Statistics and Data Analysis 103 (2016) 284–303

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Ridge estimation of inverse covariance matrices from high-dimensional data

Wessel N. van Wieringen [a,b,*], Carel F.W. Peeters [a]

[a] Department of Epidemiology & Biostatistics, VU University Medical Center Amsterdam, Postbus 7057, 1007 MB Amsterdam, The Netherlands

[b] Department of Mathematics, VU University Amsterdam, 1081 HV Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 24 June 2015
Received in revised form 20 May 2016
Accepted 21 May 2016
Available online 31 May 2016

Keywords:
Graphical modeling
High-dimensional precision matrix estimation
Multivariate normal
ℓ_2-penalization
Precision matrix

ABSTRACT

The ridge estimation of the precision matrix is investigated in the setting where the number of variables is large relative to the sample size. First, two archetypal ridge estimators are reviewed and it is noted that their penalties do not coincide with common quadratic ridge penalties. Subsequently, starting from a proper ℓ_2-penalty, analytic expressions are derived for two alternative ridge estimators of the precision matrix. The alternative estimators are compared to the archetypes with regard to eigenvalue shrinkage and risk. The alternatives are also compared to the graphical lasso within the context of graphical modeling. The comparisons may give reason to prefer the proposed alternative estimators.

1. Introduction

Let $\mathbf{Y}_i,\ i = 1, \ldots, n$, be a p-dimensional random variate drawn from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. The maximum likelihood (ML) estimator of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ maximizes:

$$\mathcal{L}(\boldsymbol{\Omega}; \mathbf{S}) \propto \ln |\boldsymbol{\Omega}| - \mathrm{tr}(\mathbf{S}\boldsymbol{\Omega}), \tag{1}$$

where $\mathbf{S}$ is the sample covariance estimate. If $n > p$, the log-likelihood achieves its maximum for $\hat{\boldsymbol{\Omega}}^{\mathrm{ML}} = \mathbf{S}^{-1}$.

In the high-dimensional setting where $p > n$, the sample covariance matrix is singular and its inverse is undefined. Consequently, so is $\hat{\boldsymbol{\Omega}}^{\mathrm{ML}}$. A common workaround is the addition of a penalty to the log-likelihood (1). The ℓ_1-penalized estimation of the precision matrix was considered almost simultaneously by Yuan and Lin (2007), Banerjee et al. (2008), Friedman et al. (2008), and Yuan (2008). This (graphical) lasso estimate of $\boldsymbol{\Omega}$ has attracted much attention due to the resulting sparse solution and has grown into an active area of research (cf. Lu, 2010; Pourahmadi, 2011; Witten et al., 2011; Hsieh et al., 2012; Rothman and Forzani, 2014). Juxtaposed to situations in which sparsity is an asset are situations in which one is intrinsically interested in more accurate representations of the high-dimensional precision matrix. In addition, the true (graphical) model need not be (extremely) sparse in terms of containing many zero elements. In these cases we may prefer usage of a regularization method that shrinks the estimated elements of the precision matrix proportionally (Fu, 1998) in possible conjunction with some form of post-hoc element selection. It is such estimators we consider.

* Principal corresponding author at: Department of Mathematics, VU University Amsterdam, 1081 HV Amsterdam, The Netherlands.
E-mail addresses: w.vanwieringen@vumc.nl (W.N. van Wieringen), cf.peeters@vumc.nl (C.F.W. Peeters).

http://dx.doi.org/10.1016/j.csda.2016.05.012

We thus study ridge estimation of the precision matrix (and, mirrored, ridge estimation of the covariance matrix). We first review two archetypal ridge estimators and note that their penalties do not coincide with what is perceived to be the common ridge penalty (Section 2). Subsequently, starting from a common ridge penalty, analytic expressions are derived for alternative ridge estimators of the precision matrix in Section 3. This section, in addition, studies properties of the alternative estimators and proposes a method for choosing the penalty parameter. In Section 4 the alternative estimators are compared to their corresponding archetypes w.r.t. eigenvalue shrinkage. In addition, the risks of the various estimators are assessed under multiple loss functions, revealing the superiority of the proposed alternatives. Section 5 compares the alternative estimators to the graphical lasso in a graphical modeling setting using oncogenomics data. This comparison points to certain favorable behaviors of the proposed alternatives with respect to loss, sensitivity, and specificity. In addition, Section 5 demonstrates that the alternative ridge estimators yield more stable networks vis-à-vis the graphical lasso, in particular for more extreme p/n ratios. This section thus provides empirical evidence in the graphical modeling setting of what is tacitly known from regression (subset selection) problems: ridge penalties coupled with post-hoc selection may outperform the lasso. We conclude with a discussion (Section 6).

2. Archetypal ridge estimators

Ridge estimators of the precision matrix currently in use can be roughly divided into two archetypes (cf. Ledoit and Wolf, 2004; Schäfer and Strimmer, 2005a). The first archetypal form of ridge estimator commonly is a convex combination of $\mathbf{S}$ and a positive definite (p.d.) target matrix $\boldsymbol{\Gamma}$: $\hat{\boldsymbol{\Omega}}^{\mathrm{I}}(\lambda_{\mathrm{I}}) = [(1-\lambda_{\mathrm{I}})\mathbf{S} + \lambda_{\mathrm{I}}\boldsymbol{\Gamma}]^{-1}$, with $\lambda_{\mathrm{I}} \in (0, 1]$. A common (low-dimensional) target choice is $\boldsymbol{\Gamma}$ diagonal with $(\boldsymbol{\Gamma})_{jj} = (\mathbf{S})_{jj}$ for $j = 1, \ldots, p$. This estimator has the desirable property of shrinking to $\boldsymbol{\Gamma}^{-1}$ when $\lambda_{\mathrm{I}} = 1$ (maximum penalization). The estimator can be motivated from the bias–variance tradeoff as it seeks to balance the high-variance, low-bias matrix $\mathbf{S}$ with the lower-variance, higher-bias matrix $\boldsymbol{\Gamma}$. It can also be viewed as resulting from the maximization of the following penalized log-likelihood:

$$\ln|\boldsymbol{\Omega}| - (1-\lambda_{\mathrm{I}})\mathrm{tr}(\mathbf{S}\boldsymbol{\Omega}) - \lambda_{\mathrm{I}}\mathrm{tr}(\boldsymbol{\Omega}\boldsymbol{\Gamma}). \tag{2}$$

The penalized log-likelihood (2) is obtained from the original log-likelihood (1) by the replacement of $\mathbf{S}$ by $(1-\lambda_{\mathrm{I}})\mathbf{S}$ and the addition of a penalty. The estimate $\hat{\boldsymbol{\Omega}}^{\mathrm{I}}(\lambda_{\mathrm{I}})$ can thus be viewed as a penalized ML estimate.

The second archetype finds its historical base in ridge regression, a technique that started as an ad-hoc modification for dealing with singularity in the least squares normal equations. The archetypal second form of the ridge precision matrix estimate would be $\hat{\boldsymbol{\Omega}}^{\mathrm{II}}(\lambda_{\mathrm{II}}) = (\mathbf{S} + \lambda_{\mathrm{II}}\mathbf{I}_p)^{-1}$ with $\lambda_{\mathrm{II}} \in (0, \infty)$. It can be motivated as an ad-hoc fix of the singularity of $\mathbf{S}$ in the high-dimensional setting, much like how ridge regression was originally introduced by Hoerl and Kennard (1970). Alternatively, this archetype too can be viewed as a penalized estimate, as it maximizes (see also Warton, 2008):

$$\ln|\boldsymbol{\Omega}| - \mathrm{tr}(\mathbf{S}\boldsymbol{\Omega}) - \lambda_{\mathrm{II}}\mathrm{tr}(\boldsymbol{\Omega}\mathbf{I}_p). \tag{3}$$

The penalties in (2) and (3) are non-concave (their second order derivatives equal the null-matrix $\mathbf{0}$). This, however, poses no problem under the restriction of a p.d. solution $\boldsymbol{\Omega}$ as the Hessian of both (2) and (3) equals $-\boldsymbol{\Omega}^{-2}$. More surprising is that neither penalty of the two current archetypes resembles the precision-analogy of what is commonly perceived as the ridge ℓ_2-penalty: $\frac{1}{2}\lambda\|\boldsymbol{\Omega}\|_2^2 = \frac{1}{2}\lambda\sum_{j_1=1}^{p}\sum_{j_2=1}^{p}[(\boldsymbol{\Omega})_{j_1,j_2}]^2$.

The graphical lasso uses a penalty that is in line with the ℓ_1-penalty of lasso regression. It is a similar objective we have in the remainder. We embark on the derivation of alternative Type I and Type II (graphical) ridge estimators using a proper ℓ_2-penalty. Consider Fig. 1 to get a flavor of the behavior of both the archetypal ridge precision matrix estimators and our alternatives (receiving analytic justification in Section 3). It is seen that ridge estimation based on a proper ridge penalty induces (slight) differences in behavior. Differences that will be shown to point to the preferability of the alternative estimators in Section 4.

3. Alternative ridge estimators of the precision matrix

In this section we derive analytic expressions for alternative Type I and Type II ridge precision estimators. In addition, we explore their moments (Section 3.3) and consistency (Section 3.4) as well as methods for choosing the penalty parameter (Section 3.5). Proofs (as indeed all proofs in the remainder) are deferred to Appendix A.

3.1. Type I

In this section an analytic expression for an alternative Type I ridge precision estimator is given. Before arriving at a proposition containing some properties of this estimator, we employ the following lemma:

Lemma 1 (*Alternative Type I Ridge Precision Estimator*)**.** *Amend the log-likelihood (1) with the ℓ_2-penalty*

$$\frac{\lambda_a}{2}\mathrm{tr}\left[(\boldsymbol{\Omega}-\mathbf{T})^{\mathrm{T}}(\boldsymbol{\Omega}-\mathbf{T})\right], \tag{4}$$