Stable Estimation of a Covariance Matrix Guided by Nuclear Norm Penalties

Eric C. Chi[a] and **Kenneth Lange**[b]

[a]Department of Electrical and Computer Engineering, Rice University, Texas, USA

[b]Departments of Human Genetics, Biomathematics, and Statistics, University of California, Los Angeles, California, USA

Abstract

Estimation of a covariance matrix or its inverse plays a central role in many statistical methods. For these methods to work reliably, estimated matrices must not only be invertible but also well-conditioned. The current paper introduces a novel prior to ensure a well-conditioned maximum a posteriori (MAP) covariance estimate. The prior shrinks the sample covariance estimator towards a stable target and leads to a MAP estimator that is consistent and asymptotically efficient. Thus, the MAP estimator gracefully transitions towards the sample covariance matrix as the number of samples grows relative to the number of covariates. The utility of the MAP estimator is demonstrated in two standard applications – discriminant analysis and EM clustering – in this sampling regime.

Keywords

Covariance estimation; Regularization; Condition number; Discriminant analysis; EM Clustering

1. Introduction

Estimation of a covariance matrix or its inverse plays a central role in many statistical methods, ranging from least squares regression to EM clustering. In these applications it is crucial to obtain estimates that are not only non-singular but also stable under small perturbations in sample values. It is well known that the sample covariance matrix

$$\boldsymbol{S} = \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{y}_j - \overline{\boldsymbol{y}})\,(\boldsymbol{y}_j - \overline{\boldsymbol{y}})^t$$

is the maximum likelihood estimate of the population covariance $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ of a random sample $\mathbf{y}_1, \ldots, \mathbf{y}_n \in \mathbb{R}^p$ from a multivariate normal distribution. When the number of components p of of each sample point exceeds the sample size n, the sample covariance $\mathbf{S}$ is no longer invertible. Even when n slightly exceeds p, the estimate $\mathbf{S}$ can be unstable. Introducing a penalty in the maximum likelihood framework offers a reliable means of stabilizing covariance estimation.

To motivate our choice of penalization, consider the eigenvalues of the sample covariance matrix in a simple simulation experiment. We drew n independent samples from a 10-dimensional multivariate normal distribution $\mathbf{y}_i \sim N(\mathbf{0}, \mathbf{I}_{10})$. Figure 1 presents boxplots of the sorted eigenvalues of the sample covariance matrix $\mathbf{S}$ over 100 trials for sample sizes n drawn from the set {5, 10, 20, 50, 100, 500}. The boxplots descend from the largest eigenvalue on the left to the smallest eigenvalue on the right. The figure vividly illustrates the previous observation that the highest eigenvalues tend to be inflated upwards above 1, while the lowest eigenvalues are deflated downwards below 1 (Ledoit and Wolf, 2004, 2012). In general, if the sample size n and the number of components p approach ∞ in such a way that the ratio p/n approaches $\zeta \in (0, 1)$, then the eigenvalues of $\mathbf{S}$ tend to the Marêenko-Pastur law (Marêenko and Pastur, 1967), which is supported on the interval $\left(\left[1 - \sqrt{\zeta} \right]^2, \left[1 + \sqrt{\zeta} \right]^2 \right)$. Thus, the distortion worsens as ζ approaches 1. The obvious remedy is to pull the highest eigenvalues down and push the lowest eigenvalues up.

In this paper, we introduce a novel prior which effects the desired adjustment on the sample eigenvalues. Maximum a posteriori (MAP) estimation under the prior boils down to a simple nonlinear transformation of the sample eigenvalues. In addition to proving that our estimator has desirable theoretical properties, we also demonstrate its utility in extending two fundamental statistical methods – discriminant analysis and EM clustering - to contexts where the number of samples n is either on the order of or dominated by the number of parameters p.

The rest of our paper is organized as follows. Section 2 discusses the history of stable estimation of structured and unstructured covariance matrices. Section 3 specifies our Bayesian prior and derives the MAP estimator under the prior. Section 4 proves that the estimator is consistent and asymptotically efficient. Section 5 reports finite sample studies comparing our MAP estimator to relevant existing estimators. Section 6 illustrates the estimator for some common tasks in statistics. Finally, Section 7 discusses limitations, generalizations, and further applications of the estimator.

2. Related Work

Structured estimation of covariance matrices can be attacked from two complementary perspectives: generalized linear models and regularization (Pourahmadi, 2011, 2013). In this work we consider the problem from the latter perspective. Regularized estimation of covariance matrices and their inverses has been a topic of intense scrutiny (Wu and Pourahmadi, 2003; Bickel and Levina, 2008), and the current literature reflects a wide spectrum of structural assumptions. For instance, banded covariance matrices are appropriate for time series and spatial data, where the order of the components is important.

It is also helpful to impose sparsity on a covariance matrix, its inverse, or its factors in a Cholesky decomposition or other factorization (Huang et al., 2006; Rohde and Tsybakov, 2011; Cai and Zhou, 2012; Ravikumar et al., 2011; Rajaratnam et al., 2008; Khare and Rajaratnam, 2011; Fan et al., 2011; Banerjee et al., 2008; Friedman et al., 2008; Hero and Rajaratnam, 2011, 2012; Peng et al., 2009).

In this work, we do not assume any special prior structure. Our sole concern is to directly address the distortion in the eigenvalues of the sample covariance matrix. Thus, we work in the context of rotationally-invariant estimators first proposed by Stein (1975). If $S = U \, D U^t$ is the spectral decomposition of S, then Stein suggests alternative estimators of the form

$$\hat{\sum} = U \operatorname{diag}(e_1, \ldots, e_p) \, U^t$$

that modify the eigenvalues but not the eigenvectors of S. In particular, Stein (1975); Haff (1991); Ledoit and Wolf (2004) and Warton (2008) study the family

$$\hat{\sum} = (1-\gamma)S + \gamma T \quad (1)$$

of linear shrinkage estimators, where $\gamma \in [0, 1]$ and $T = \rho I$ for some $\rho > 0$. The estimator (1) obviously entails

$$e_i = (1-\gamma)d_i + \gamma \rho.$$

A natural choice of ρ, and one taken by the popular estimator of Ledoit and Wolf (2004), is the mean of the sample eigenvalues, namely $\hat{\sigma} = (1/p) \operatorname{tr}(S)$. Under the assumption that $y_i \sim N(0, \sigma I)$, $\hat{\sigma}$ is the maximum likelihood estimate of σ. With this choice, the linear estimator becomes a mixture of the covariance model with the greatest number of degrees of freedom and the simplest non-trivial model. For the rest of this paper, we will assume that $\rho = \hat{\sigma}$. We will refer to the Ledoit and Wolf linear estimator by the acronym LW. In the class of linear estimators, the LW estimator specifies an asymptotically optimal value for γ based on the data.

Ledoit and Wolf (2004, 2012) show that linear shrinkage works well when p/n is large or the population eigenvalues are close to one another. On the other hand, if p/n is small or the population eigenvalues are dispersed, linear shrinkage yields marginal improvements over the sample covariance. Nonlinear shrinkage estimators may present avenues for further improvement (Dey and Srinivasan, 1985; Daniels and Kass, 2001; Sheena and Gupta, 2003; Pourahmadi et al., 2007; Ledoit and Wolf, 2012; Won et al., 2012). Our shrinkage estimator is closest in spirit to the estimator of Won et al. (2012), who put a prior on the condition number of the covariance matrix.

Recall that the condition number κ of a matrix is the ratio of its largest singular value to its smallest singular value. For a symmetric matrix, the singular values are the absolute values

of the eigenvalues, and for a covariance matrix they are the eigenvalues themselves. The best conditioned matrices are multiples of the identity matrix and have $\kappa = 1$. A well-conditioned covariance estimate is one where κ is not too large, say in excess of 1000.

When n does not greatly exceed p, Figure 1 shows that the sample covariance matrix is often ill-conditioned. To address this defect, Won et al. perform maximum likelihood estimation constrained to the set of positive definite matrices whose condition number does not exceed a threshold κ_{max}. Let $\ell(\Sigma)$ denote the negative loglikelihood, namely

$$\ell(\textstyle\sum) = \frac{n}{2}\ln\det\textstyle\sum + \frac{n}{2}\mathrm{tr}(\boldsymbol{S}\textstyle\sum^{-1}).$$

Won et al. seek an Σ that solves

$$\begin{aligned}\text{minimize} \quad & \ell(\textstyle\sum) \\ \text{subject to} \quad & \lambda_{\max}(\textstyle\sum)/\lambda_{\min}(\textstyle\sum) \le \kappa_{\max},\end{aligned}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ denote the largest and smallest eigenvalues of Σ respectively and $\kappa_{\max}$ 1 is a tuning parameter. When $\kappa_{\max} = 1$, the unique solution is $\sigma\hat{\boldsymbol{I}}$. As $\kappa_{\max}$ increases, the solution converges to the sample covariance matrix. In practice, $\kappa_{\max}$ is determined by cross-validation. We will use the acronym CNR (condition number regularized) in referring to Won et al.'s estimator for this problem.

Won et al. show that CNR has improved finite sample performance compared to linear estimators in simulations, but the greatest gains arise when only a few eigenvalues of the population covariance are much larger than the rest. The gains diminish when this does not hold. The main contribution of the estimator that we describe next is that it provides superior performance in scenarios where CNR loses its competitive edge.

3. Maximum a Posteriori Estimation with a Novel Prior

Adding a penalty is equivalent to imposing a prior $\pi(\Sigma)$ on the population covariance Σ. To ensure well-conditioned posterior estimates, let us suggest a prior that steers the eigenvalues of Σ away from the extremes of 0 and ∞. Recall that the nuclear norm of a matrix Σ, which we denote by $\|\Sigma\|_*$, is the sum of the singular values of Σ. Intuitively, we want to make sure that neither $\|\Sigma\|_*$ nor $\|\Sigma^{-1}\|_*$ is too large. Consequently, we propose the prior

$$\pi(\textstyle\sum) \propto e^{-\frac{\lambda}{2}[\alpha\|\sum\|_* + (1-\alpha)\|\sum^{-1}\|_*]},$$

where $\lambda > 0$ determines the strength of the prior and $\alpha \in (0, 1)$ determines the tradeoffs between the two nuclear norms.

This is a proper prior on the set of invertible matrices. One can demonstrate this fact by comparing the nuclear norm $\|\Sigma\|_*$ to the Frobenius norm $\|\Sigma\|_F$, which coincides with the Euclidean norm of the vector of singular values of Σ. In view of the equivalence of vector norms on $\mathbb{R}^{p(p+1)/2}$,

$$\frac{\alpha}{2}\Big\|\sum\Big\|_* \geq \eta\Big\|\sum\Big\|_F$$

for some positive constant η. Integrating the resulting inequality

$$e^{-\frac{\lambda}{2}[\alpha\|\sum\|_* + (1-\alpha)\|\sum^{-1}\|_*]} \leq e^{-\eta\lambda\|\sum\|_F}$$

over Σ demonstrates that the prior is proper. The normalizing constant of $\pi(\Sigma)$ is irrelevant in the ensuing discussion. Consider therefore minimization of the objective function

$$f\Big(\sum\Big)=\frac{n}{2}\ln\det\sum+\frac{n}{2}\mathrm{tr}\Big(\boldsymbol{S}\sum{}^{-1}\Big)+\frac{\lambda}{2}[\alpha\Big\|\sum\Big\|_* + (1-\alpha)\Big\|\sum{}^{-1}\Big\|_*].$$

The maximum of $-f(\Sigma)$ occurs at the posterior mode. In the limit as λ tends to 0, $-f(\Sigma)$ reduces to the loglikelihood $-\ell(\Sigma)$. In the sequel, we will refer to our MAP covariance estimate by the acronym CERNN (Covariance Estimate Regularized by Nuclear Norms).

The minimizer of the objective $f(\Sigma)$ can be found by extracting the spectral decomposition of Σ. Three of the four terms of $f(\Sigma)$ can be expressed as functions of the eigenvalues e_i of Σ. The trace term presents a greater challenge. As before, let $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^t$ denote the spectral decomposition of $\boldsymbol{S}$ with nonnegative diagonal entries d_i ordered from largest to smallest. Likewise, let $\Sigma = \boldsymbol{V}\boldsymbol{E}\boldsymbol{V}^t$ denote the spectral decomposition of Σ with positive diagonal entries e_i ordered from largest to smallest. In view of the von Neumann-Fan inequality (Mirsky, 1975), we can assert that

$$-\mathrm{tr}\Big(\boldsymbol{S}\sum{}^{-1}\Big) \leq -\sum_{i=1}^{p}\frac{d_i}{e_i},$$

with equality if and only if $\boldsymbol{V} = \boldsymbol{U}$. Consequently, we make the latter assumption and replace $f(\Sigma)$ by

$$g(\boldsymbol{E})=\frac{n}{2}\sum_{i=1}^{p}\ln e_i+\frac{n}{2}\sum_{i=1}^{p}\frac{d_i}{e_i}+\frac{\lambda}{2}\left[\alpha\sum_{i=1}^{p}e_i+(1-\alpha)\sum_{i=1}^{p}\frac{1}{e_i}\right]$$

using the cyclic permutation property of the trace function. At a stationary point of $g(\boldsymbol{E})$, we have

$$0 = \frac{n}{e_i} - \frac{nd_i + \lambda(1-\alpha)}{e_i^2} + \lambda\alpha.$$

The solution to this essentially quadratic equation is

$$e_i = \frac{-n + \sqrt{n^2 + 4\lambda\alpha[nd_i + \lambda(1-\alpha)]}}{2\lambda\alpha}. \quad (2)$$

We reject the negative root as inconsistent with Σ being positive definite. For the special case of no data when $n = 0$, all $e_i = \sqrt{(1-\alpha)/\alpha}$, and the prior mode occurs at a multiple of the identity matrix.

In contrast to the solution (2), CNR shrinks the sample eigenvalues d_i according to the formulas

$$e_i = \begin{cases} \kappa_{\max}\tau^* & d_i \geq \kappa_{\max}\tau^* \\ d_i & \tau^* < d_i < \kappa_{\max}\tau^* \\ \tau^* & d_i \leq \tau^* \end{cases} \quad (3)$$

for a $\tau^* > 0$ that is determined from the data. Thus, CNR truncates extreme sample eigenvalues while leaving moderate ones untouched.

Figure 2 compares the eigenvalue solution paths obtained by CERNN and CNR to the solution paths of the linear estimator on a set of five sample eigenvalues (13.29, 5.73, 1.51, 0.55, 0.44). At each condition number κ, the regularization parameters (λ, $\kappa_{\max}$, γ) of the three methods are adjusted to give a condition number matching κ. Each path starts at a sample eigenvalue and ends at the mean $\hat{\sigma}$ of the sample eigenvalues. As desired, all three methods pull the large eigenvalues down towards $\hat{\sigma}$ and the small eigenvalues up towards $\hat{\sigma}$. There are important differences, however. Compared to the linear estimator, both of the non-linear estimators pull the larger eigenvalues down more aggressively and pull the smaller eigenvalues up less aggressively. The discrepancy in the treatment of high and low eigenvalues is more pronounced in CNR than in CERNN. We will see later in finite sample experiments that this more moderate approach usually leads to better performance.

Some insight to the limiting behavior of our estimator can be gained through asymptotic expansions. Holding all but one variable fixed in formula (2), one can demonstrate after a fair amount of algebra that

$$e_i = d_i + \frac{\lambda(1-\alpha-\alpha d_i^2)}{n} + O\left(\frac{1}{n^2}\right), \quad n \to \infty$$

$$e_i = \sqrt{\frac{1-\alpha}{\alpha}} + \left[\sqrt{\frac{1-\alpha}{\alpha}}\frac{nd_i}{2(1-\alpha)} - \frac{n}{2\alpha}\right]\frac{1}{\lambda} + O\left(\frac{1}{\lambda^2}\right), \quad \lambda \to \infty. \quad (4)$$

These asymptotic expansions accord with common sense. Namely, the data eventually overwhelms a fixed prior, and increasing the penalty strength for a fixed amount of data pulls the estimate of Σ toward the prior mode. These expansions are helpful in proving CERNN's asymptotic properties in Section 4. As we discuss next, the second asymptotic expansion is instrumental in selecting λ in a practical data-dependent way.

3.1. Choosing λ and α

To compute a CERNN estimate, we need to specify the parameters λ and α. Let us first consider how to choose α. A natural choice for it would match the prior to the scale of the data. Thus, we would determine α as the solution to the equation

$$p\sqrt{\frac{1-\alpha}{\alpha}} = \mathrm{tr}\left(\sqrt{\frac{1-\alpha}{\alpha}}\boldsymbol{I}\right) = \mathrm{tr}(\boldsymbol{S}),$$

namely

$$\hat{\alpha} = \left[1 + \left(\frac{1}{p}\mathrm{tr}(\boldsymbol{S})\right)^2\right]^{-1}. \quad (5)$$

Of course, we recognize that this choice of α results in the prior mode being $\sigma\hat{\boldsymbol{I}}$. For the remainder of the paper we will set α to $\hat{\alpha}$.

We recommend different strategies for choosing λ depending on whether the covariance estimation is being performed for a supervised or unsupervised task. Both strategies employ cross-validation. We defer describing the former strategy until we discuss an application to discriminant analysis. To choose λ in the unsupervised context, we implement cross-validation as follows. Let $\boldsymbol{Y} \in \mathbb{R}^{n \times p}$ denote the observed data. In K-fold cross validation, we partition the observations, or rows of $\boldsymbol{Y}$, into K disjoint sets. Let $\boldsymbol{Y}_k$ denote the kth subset, n_k denote the number of its rows, and $\hat{\sum}_\lambda^{(-k)}$ denote the estimate we obtain when we fit the data using all but the kth partition $\boldsymbol{Y}_k$. We have indicated with our notation that our estimate depends on λ but not α since we have fixed α at the value displayed in equation (5). For a grid of increasing regularization parameters, $\lambda = 0, \ldots, \lambda_{\max}$, we evaluate the predictive negative loglikelihood of $\hat{\sum}_\lambda^{(-k)}$ on the held out kth subgroup $\boldsymbol{Y}_k$

$$\ell_k(\hat{\sum}_\lambda^{(-k)}, \boldsymbol{Y}_k) = \frac{n_k}{2} \ln\det \hat{\sum}_\lambda^{(-k)} + \frac{n_k}{2} \mathrm{tr}\left(\frac{1}{n_k} \boldsymbol{Y}_k^t \boldsymbol{Y}_k \left[\hat{\sum}_\lambda^{(-k)}\right]^{-1}\right).$$

We select the λ that minimizes the average ℓ_k over the K folds, namely

$$\hat{\lambda} = \underset{\lambda \in \{0,\ldots,\lambda_{\max}\}}{\arg\min} \frac{1}{n} \sum_{k=1}^{K} \ell_k(\hat{\sum}_\lambda^{(-k)}, \boldsymbol{Y}_k).$$

We want to choose $\lambda_{\max}$ sufficiently large to adequately explore the dynamic range of possible solutions. The second expansion in (4) gives us a rough bound on how much each of the shrunken eigenvalues deviates from the prior mode. We choose $\lambda_{\max}$ so that those deviations are no more than some small fraction ε of the prior mode, namely we choose $\lambda_{\max}$ so that

$$\max_{i=1,\ldots,p} \left| \sqrt{\frac{1-\hat{\alpha}}{\hat{\alpha}}} \frac{nd_i}{2(1-\hat{\alpha})} - \frac{n}{2\alpha} \right| \frac{1}{\lambda_{\max}} \leq \varepsilon \sqrt{\frac{1-\hat{\alpha}}{\hat{\alpha}}},$$

where $\varepsilon \ll 1$, say 10^{-2}.

4. Consistency and Asymptotic Efficiency

To prove consistency, we will need to assemble various facts. First, suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ are two $p \times p$ symmetric matrices with ordered eigenvalues $\{a_i\}_{i=1}^p$ and $\{b_i\}_{i=1}^p$. Then one has

$$\sum_{i=1}^{p} (a_i - b_i)^2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_{\mathrm{F}}^2. \quad (6)$$

This is a consequence of Fan's inequality because $\sum_{i=1}^{p} a_i^2 = \|\boldsymbol{A}\|_{\mathrm{F}}^2$ and $\sum_{i=1}^{p} b_i^2 = \|\boldsymbol{B}\|_{\mathrm{F}}^2$. If the two matrices $\boldsymbol{A} = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{a})\boldsymbol{U}^t$ and $\boldsymbol{B} = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{b})\boldsymbol{U}^t$ are simultaneously diagonalizable, then equality holds in inequality (6). For nonnegative x the inequalities

$$\sqrt{1+x} \leq 1 + \frac{x}{2} \quad \text{and} \quad \sqrt{1+x} \geq 1 + \frac{x}{2} - \frac{x^2}{8} \quad (7)$$

are also relevant to our task. Verification of these inequalities will be left to the reader based on the fact that the derivatives of $\sqrt{1+x}$ alternate in sign. Functions having this property are said to be completely monotonic.

Let $\boldsymbol{S}_n$ be the sample covariance matrix with eigenvalues d_{n1} through d_{np} for the first n sample points. The sequence $\boldsymbol{S}_n$ converges almost surely to the true covariance matrix $\boldsymbol{\Sigma}$

with eigenvalues ω_1 through ω_p. Inequality (6) therefore implies $\lim_{n\to\infty}\sum_{i=1}^{p}(d_{ni}-\omega_i)^2=0$. On this basis we will argue that $\lim_{n\to\infty}\sum_{i=1}^{p}(e_{ni}-\omega_i)^2=0$ as well, where the e_{ni} are the transformed eigenvalues of S_n. To make this reasoning rigorous, we must show that the asymptotic expansion (4) is uniform as the eigenvalues d_{ni} converge to the eigenvalues ω_i. This is where the inequalities (7) come into play. Indeed, we have

$$
\frac{\lambda(1-\alpha)}{n}-\frac{n}{2\lambda\alpha}\frac{x^2}{8} \leq e_{ni}-d_{ni} \leq \frac{\lambda(1-\alpha)}{n}
$$
$$
x=\frac{4\lambda\alpha d_{ni}}{n}+\frac{4\lambda^2\alpha(1-\alpha)}{n^2}. \tag{8}
$$

The identity

$$
\left\|S_n-\sum\nolimits_n\right\|_{\mathrm{F}}^2=\sum_{i=1}^{p}(d_{ni}-e_{ni})^2
$$

finishes the proof that Σ_n tends to Σ.

Now consider the question of asymptotic efficiency. The scaled difference $\sqrt{n}(S_n-\sum)$ tends in distribution to a multivariate normal distribution with mean $\mathbf{0}$ because the sequence of estimators S_n is asymptotically efficient (Ferguson, 1996). The representation

$$
\sqrt{n}(\sum\nolimits_n-\sum)= \sqrt{n}(S_n-\sum)+ \sqrt{n}(\sum\nolimits_n-S_n)
$$

and Slutsky's theorem (Ferguson, 1996) imply that $\sqrt{n}(\sum_n-\sum)$ tends in distribution to the same limit. In this regard note that

$$
\left\| \sqrt{n}(\sum\nolimits_n-S_n)\right\|_{\mathrm{F}}^2=n\sum_{i=1}^{p}(d_{ni}-e_{ni})^2
$$

tends almost surely to 0 owing to the bounds (8) and the convergence of d_{ni} to ω_i.

CNR and linear estimators are also asymptotically very well behaved. The asymptotic properties of CNR are stated in terms of the entropy risk, which is the expectation of the entropy loss given below

$$
\mathscr{L}_{\mathrm{E}}(\sum\nolimits_n,\sum)=\mathrm{tr}(\sum\nolimits^{-1}\sum\nolimits_n)-\mathrm{logdet}(\sum\nolimits^{-1}\sum\nolimits_n)-p. \tag{9}
$$

CNR has asymptotically lower entropy risk than the sample covariance (Won et al., 2012). But since the sample covariance is a consistent estimator of the covariance, it follows that CNR is also a consistent estimator.

Among all linear estimators, the LW estimator of Ledoit and Wolf (2004) is asymptotically optimal with respect to a quadratic risk. To make this claim more precise, consider the optimization problem,

$$\min_{\gamma_1, \gamma_2} \quad \|\tilde{\Sigma} - \Sigma\|_F^2$$
$$\text{subject to} \quad \tilde{\Sigma} = \gamma_1 \boldsymbol{I} + \gamma_2 \boldsymbol{S}_n, \quad (10)$$

where the weights γ_1 and γ_2 are allowed to be random and can therefore depend on the data. One can think of the solution $\Sigma^\star$ to (10) as being the best possible linear combination of $\boldsymbol{I}$ and the sample covariance $\boldsymbol{S}_n$. Even though $\Sigma^\star$ may not be an estimator, since it is allowed to depend on the unseen population covariance Σ, Ledoit and Wolf (2004) prove that the quadratic risk of LW has the same quadratic risk as $\Sigma^\star$ asymptotically. Their results are actually even stronger than stated here because p is allowed to increase along with n, under suitable, but somewhat complicated regularity conditions that we omit. Since the sample covariance is consistent, and its quadratic risk by definition exceeds the quadratic risk of $\Sigma^\star$, it follows that LW is also a consistent estimator.

5. Finite Sample Performance

Given their similar asymptotic behavior, to better understand the relative strengths of CERNN, CNR, and the optimal linear estimator, LW, we conducted a finite sample study almost identical to the one carried out by Won et al. (2012). We assessed the estimators based on two commonly used criteria, namely the entropy loss (9) and the quadratic loss

$$\mathcal{L}_Q(\hat{\Sigma}, \Sigma) = \|\hat{\Sigma}\Sigma^{-1} - \boldsymbol{I}\|_F^2.$$

In our experiments, we simulated data in which we varied the ratio p/n and the spread of the eigenvalues of the population covariance. As noted earlier, linear shrinkage improves on the sample covariance when p/n is large or the population eigenvalues are close to one another. Won et al. report that when the eigenvalues of the population covariance are bimodal, CNR dramatically outperforms linear estimators if very few eigenvalues take on the high values. As the proportion of large eigenvalues grows, however, the discrepancy diminishes. CERNN shrinks extreme eigenvalues in a manner similar to CRN but less drastically and shrinks intermediate eigenvalues similarly to linear estimators. In contrast, CNR leaves intermediate eigenvalues unchanged. Consequently, we anticipate that CERNN has the potential to improve on CNR in the latter case, where there is a need to shrink all eigenvalues, like linear estimators, but with extra emphasis on the extreme ones, unlike linear estimators.

We simulated p-dimensional zero mean multivariate normal samples with diagonal covariances with $p = 120$, 250, and 500. The eigenvalues took on either high values of $1 - \upsilon + \upsilon p$ or low values of $1 - \upsilon$ where $\upsilon = 0.1$. For each p, the number of high value eigenvalues ranged from a single high value to 10%, 20%, 30%, and 40% of p. For each p we drew one

of three sample sizes n such that the ratio $r = p/n$ took on values that were roughly 1.25, 2, or 4. For each choice of p, n, and r, we simulated 100 data sets and computed CNR, CERNN, and LW. For each data set, we chose an optimal $\kappa_{\max}$ and λ via 10-fold cross-validation. We used the R package **CondReg** to obtain the CNR estimate (Oh et al., 2013). LW specifies a penalization parameter γ based on the data (Ledoit and Wolf, 2004).

To expedite comparisons, we report the average ratio of the loss of other estimators to the loss of CERNN. When this ratio is less than one, the other estimator is performing better than CERNN, and when the ratio is greater than one, CERNN is performing better. Tables A.2 and A.3, in Appendix A, summarize comparisons under the quadratic loss and entropy loss respectively.

We first note that these experiments confirm what was previously observed by Won et al. (2012). Regardless of loss criterion, CNR typically outperforms linear estimators over a wide range of scenarios, but especially when few eigenvalues dominate and the ratio p/n is larger. Compared to CERNN, however, CNR soundly outperforms CERNN only in the extreme case of a single large eigenvalue. In all other scenarios, under either loss criterion, CERNN performs better. It is especially notable that CERNN performs very well in comparison to CNR in scenarios where CNR provides marginal improvement over linear estimators, namely when the fraction of high eigenvalues is highest at 40%. According to equation (3), recall that CNR only shrinks the most extreme sample eigenvalues and leaves the intermediate eigenvalues unchanged. It is not surprising then that it works best when there are very few large population eigenvalues and loses its competitive edge in the opposite circumstance. CERNN's more moderate approach of shrinking all eigenvalues, with extra emphasis on larger ones, appears to win out when there is a more balanced mix of high and low eigenvalues.

6. Applications

Several common statistical procedures can potentially benefit from shrinkage estimation of sample covariance matrices. Here we illustrate how CERNN applies to discriminant analysis and clustering.

6.1. Discriminant Analysis

The classical discriminant function

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}^k \sum\nolimits^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^t \sum\nolimits^{-1} \boldsymbol{\mu}_k + \ln \pi_k,$$

incorporates the mean $\boldsymbol{\mu}_k$ and prior probability π_k of each class k. A new observation $\boldsymbol{x}$ is assigned to the class k maximizing $\delta_k(\boldsymbol{x})$. If there are c classes $c_1, \ldots, c_c$, then the standard estimator of Σ is

$$\hat{\Sigma} = \frac{1}{n-c} \sum_{k=1}^{c} \sum_{i \in \mathscr{C}_k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^t,$$

where

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{|\mathscr{C}_k|} \sum_{i \in \mathscr{C}_k} \boldsymbol{x}_i.$$

One can obviously shrink $\hat{\Sigma}$ to moderate its eigenvalues. In quadratic discriminant analysis, a separate covariance matrix Σ_k is assigned to each class k. These are estimated in the usual way, and eigenvalue shrinkage is likely even more beneficial than in linear discriminant analysis. Friedman (1989) advocates regularized discriminant analysis (RDA), a compromise between linear and quadratic discriminant analysis that shrinks Σ_k toward a common Σ via a convex combination $\gamma\Sigma_k + (1 - \gamma)\Sigma$. Although Friedman also suggests shrinking toward class specific multiples of the identity matrix, we do not consider his more complicated version here. Guo et al. (2007) shrink covariance estimates towards the identity matrix and also apply lasso shrinkage on the centroids to obtain improved classification performance in microarray studies. The main difference between CERNN and these methods is that CERNN performs nonlinear shrinkage of the sample eigenvalues.

Since we are primarily interested in the case where all or most of the predictors are instrumental in grouping, we consider only Friedman's method in a comparison on three data sets from the UCI machine learning repository (Bache and Lichman, 2013). In the case of the E. Coli data set, we restricted analysis to the five most abundant classes. We split each data set into training and testing sets. In each experiment we used 1/5 of the data for training and 4/5 for testing. Table 1 records the number of samples per group in each set. In these data poor examples, linear discriminant analysis is not even viable since a common sample covariance estimate will be ill-conditioned if not singular. Nonetheless, our results show that the combination of separate covariances with regularization works well. We modeled a separate covariance for each class and used 5-fold cross validation to select c regularization parameters for CERNN. We used essentially the same procedure described in Section 3.1 except that we used the misclassification rate instead of the predictive negative loglikelihood. We also used 5-fold cross validation to select the single γ parameter for RDA. The testing errors in Table 1 demonstrate that CERNN performs well in comparison with RDA. Even when it does not perform as accurately, its drop off is small.

6.2. Covariance Regularized EM Clustering

We now show how CERNN stabilizes estimation in the standard EM clustering algorithm (McLachlan and Peel, 2000). Let $\varphi(\boldsymbol{y} \mid \boldsymbol{\mu}, \Sigma)$ denote a multivariate Gaussian density with mean $\boldsymbol{\mu}$ and covariance Σ. EM clustering assumes that the data arise from the mixture density

$$h(\boldsymbol{y}|\Xi) = \sum_{k=1}^{c} \pi_k \phi(\boldsymbol{y}|\boldsymbol{\mu}_k, \textstyle\sum_k)$$

with parameters $\Xi = \{\boldsymbol{\mu}_k, \sum_k, \pi_k\}_{k=1}^{c}$. The π_k are nonnegative mixture weights summing to 1. We are given n independent observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ from the mixture density and wish to estimate Ξ by maximizing the loglikelihood. It is well known that the loglikelihood is unbounded from above and plagued by many suboptimal local extrema (McLachlan and Peel, 2000). Hathaway (1985) proposed constraints leading to a maximum likelihood problem free of singularities and beset by fewer local extrema. Later it was shown that these constraints could be met by imposing bounds on the largest and smallest eigenvalues of the Σ_k (Ingrassia, 2004; Ingrassia and Rocci, 2007). These findings suggest that employing our Bayesian prior to estimate Σ_k can also improve the search for good local optima, since it shrinks the extreme sample eigenvalues to the sample mean eigenvalue.

If z_{ik} is the indicator function of the event that observation i comes from cluster k, then the complete data loglikelihood plus log-prior amounts to

$$\ell(\Xi) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} [\ln \pi_k + \ln \phi(\boldsymbol{y}_i|\boldsymbol{\mu}_k, \textstyle\sum_k)] - \frac{\lambda}{2} \left[\alpha_k \| \textstyle\sum_k \|_* + (1-\alpha_k) \| \textstyle\sum_k^{-1} \|_* \right].$$

Straightforward application of Bayes rule yields the conditional expectation

$$w_{ik} = E[z_{ik}|\boldsymbol{Y}, \Xi] = \frac{\pi_k \phi(\boldsymbol{y}_i|\boldsymbol{\mu}_k, \sum_k)}{\sum_{l=1}^{c} \pi_l \phi(\boldsymbol{y}_i|\boldsymbol{\mu}_l, \sum_l)}.$$

These weights should be subscripted by the current iteration number m, but to avoid clutter we omit the subscripts. If we set

$$w_k = \sum_{i=1}^{n} w_{ik} \quad \text{and} \quad \boldsymbol{S}_k = \frac{1}{w_k} \sum_{i=1}^{n} w_{ik} (\boldsymbol{y}_i - \boldsymbol{\mu}_k)(\boldsymbol{y}_i - \boldsymbol{\mu}_k)^t,$$

then the EM updates are $\pi_k = \frac{w_k}{n}$, $\boldsymbol{\mu}_k = \frac{1}{w_k} \sum_{i=1}^{n} w_{ik} \boldsymbol{y}_i$, and

$$\sum_k = \arg\min_{\sum} \frac{w_k}{2} \mathrm{logdet} \sum + \frac{w_k}{2} \mathrm{tr}(\textstyle\sum^{-1} \boldsymbol{S}_k) + \frac{\lambda}{2} \left[\alpha_k \| \textstyle\sum \|_* + (1-\alpha_k) \| \textstyle\sum^{-1} \|_* \right].$$

We now address two practical issues. First, there is the question of how to choose α_k. In the previous examples we sought a stable estimate of a single covariance matrix. Here we seek c covariance matrices whose imputed data change from iteration to iteration. In accord with

our previous choice for a, we set a_k to be $(1/p)\,\text{tr}(S_k)$. This a_k changes dynamically as S_k changes. Second, it is possible for $w_{ik} \approx 0$ for all i for a given k at some point as the algorithm iterates. Indeed, finite machine precision may assign w_k the value 0 at some point, making the updates for Σ_k and $\boldsymbol{\mu}_k$ undefined. Consequently, we only update Σ_k and $\boldsymbol{\mu}_k$ when $w_k > 0$. This is not a major issue, however, since this scenario only arises when no data points should be assigned to the kth cluster.

Besides the work of Ingrassia and Rocci (2007), similar approaches have been employed previously. Fraley and Raftery (2007) suggest a restricted parameterization of covariance matrices. While they offer a menu of parameterizations that cover a range of degrees of freedom, each model has a fixed number of degrees of freedom. One advantage of our model is that the degrees of freedom may be tuned continuously with a single parameter λ.

Figure 3 shows the results of clustering with our algorithm on a simulated data set. A total of 60 data points were generated from a mixture of 10 bivariate normals corresponding to 59 parameters in the most general case. The number of observations per cluster ranged from 3 to 11. We set the number of clusters c to be 10 and considered several different λ values over a wide range (0.1, 10, 100, 10000). We could choose λ in a completely data dependent way through cross-validation, but our main concern is to stabilize the procedure so fine-tuning λ is not of paramount importance, especially when doing so complicates the procedure. We ran our algorithm 500 times using random initializations with the k-means++ algorithm (Arthur and Vassilvitskii, 2007) and kept the clustering that gave the greatest value of the expected log likelihood

$$\sum_{i=1}^{n}\sum_{k=1}^{c} w_{ik}\,[\ln \pi_k + \ln \phi(\boldsymbol{y}_i | \boldsymbol{\mu}_k, \textstyle\sum_k)].$$

The resulting clusterings are quite good over our broad range of λ values. It is notable that for three out of the four values of λ, even clusters 2 and 10, which overlap, are somewhat distinguished. The only missteps occur when $\lambda = 10$, where cluster 1 is split into two clusters, and clusters 2 and 10 have been merged. The latter decision is reasonable given how much clusters 2 and 10 overlap.

7. Discussion

The initial insight of Stein (1975) has led to several methods for shrinkage estimation of a sample covariance matrix S. These methods preserve the eigenvectors of S while pushing S towards a multiple of the identity matrix. Our Bayesian prior does precisely this in a nonlinear fashion. Finite sample experiments comparing CERNN and CNR show that CERNN and CNR complement each other. CNR performs better when only a few eigenvalues of the population covariance are very large. CERNN performs better when there is a more uniform distribution of high and low eigenvalues. Both estimators perform at least as well and often better than the simple and asymptotically optimal LW estimator.

CERNN does require a singular value decomposition (SVD), as does CNR. Although highly optimized routines for accurately computing the SVD are readily available, such calculations are not cheap. Recently introduced algorithms from randomized numerical linear algebra may provide computational relief (Halko et al., 2011; Mahoney, 2011). If one can tolerate a small loss in accuracy, the SVD of a randomly sampled subset of the data or a random projection of the data can give an acceptable surrogate SVD.

Applications extend well beyond the classical statistical methods illustrated here. For example, in gene mapping with pedigree data, a covariance matrix is typically parameterized as a mixture of three components, one of which is the global kinship coefficient matrix capturing the relatedness between individuals in the study (Lange, 2002). The kinship matrix can be estimated from a high density SNP (single nucleotide polymorphism) panel rather than calculated from possibly faulty genealogical records. Because a typical study contains thousands of individuals typed at hundreds of thousands of genetic markers, this application occurs in the regime $n \ll p$. The construction of networks from gene co-expression data is another obvious genetic example (Horvath, 2011). Readers working in other application areas can doubtless think of other pertinent examples.

Acknowledgments

References

Arthur, D.; Vassilvitskii, S. *k*-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. SODA '07. Society for Industrial and Applied Mathematics; Philadelphia, PA, USA. 2007. p. 1027-1035.

Bache, K.; Lichman, M. UCI machine learning repository. 2013. URL http://archive.ics.uci.edu/ml

Banerjee O, El Ghaoui L, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. Journal of Machine Learning Research. Jun. 2008 9:485–516.

Bickel PJ, Levina E. Regularized estimation of large covariance matrices. Annals of Statistics. 2008; 36 (1):199–227.

Cai T, Zhou H. Minimax estimation of large covariance matrices under ℓ_1 norm. Statistica Sinica. 2012; 22:1319–1378.

Daniels MJ, Kass RE. Shrinkage estimators for covariance matrices. Biometrics. 2001; 57 (4):1173–1184. [PubMed: 11764258]

Dey DK, Srinivasan C. Estimation of a covariance matrix under Stein's loss. Annals of Statistics. 1985; 13 (4):1581–1591.

Fan J, Liao Y, Mincheva M. High-dimensional covariance matrix estimation in approximate factor models. Annals of Statistics. 2011; 39 (6):3320–3356. [PubMed: 22661790]

Ferguson, TS. CRC Texts in Statistical Science Series. Chapman and Hall; 1996. A course in large sample theory.

Fraley C, Raftery AE. Bayesian regularization for normal mixture estimation and model-based clustering. Journal of Classification. Sep; 2007 24 (2):155–181.

Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9 (3):432–441. [PubMed: 18079126]

Friedman JH. Regularized discriminant analysis. Journal of the American Statistical Association. 1989; 84 (405):165–175.

Guo Y, Hastie T, Tibshirani R. Regularized linear discriminant analysis and its application in microarrays. Biostatistics. 2007; 8 (1):86–100. [PubMed: 16603682]

Haff LR. The variational form of certain Bayes estimators. The Annals of Statistics. 1991; 19 (3): 1163–1190.

Halko N, Martinsson PG, Tropp JA. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review. 2011; 53 (2):217–288.

Hathaway RJ. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. Annals of Statistics. 1985; 13 (2):795–800.

Hero A, Rajaratnam B. Large-scale correlation screening. Journal of the American Statistical Association. 2011; 106 (496):1540–1552.

Hero A, Rajaratnam B. Hub discovery in partial correlation graphs. Information Theory, IEEE Transactions on. 2012; 58 (9):6064–6078.

Horvath, S. Weighted Network Analysis. Applications in Genomics and Systems Biology. Springer; New York: 2011.

Huang JZ, Liu N, Pourahmadi M, Liu L. Covariance matrix selection and estimation via penalised normal likelihood. Biometrika. 2006; 93 (1):85–98.

Ingrassia S. A likelihood-based constrained algorithm for multivariate normal mixture models. Statistical Methods and Applications. 2004; 13 (2):151–166.

Ingrassia S, Rocci R. Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. Computational Statistics & Data Analysis. 2007; 51 (11):5339–5351.

Khare K, Rajaratnam B. Wishart distributions for decomposable covariance graph models. Annals of Statistics. 2011; 39:514–555.

Lange, K. Statistics for Biology and Health. 2. Springer-Verlag; New York: 2002. Mathematical and Statistical Methods for Genetic Analysis.

Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis. 2004; 88 (2):365–411.

Ledoit O, Wolf M. Nonlinear shrinkage estimation of large-dimensional covariance matrices. Annals of Statistics. 2012; 40 (2):1024–1060.

Mahoney MW. Randomized algorithms for matrices and data. Foundations and Trends in Machine Learning. Feb; 2011 3 (2):123–224.

Marêenko VA, Pastur LA. Distribution of eigenvalues for some sets of random matrices. Mathematics of the USSR-Sbornik. 1967; 1 (4):507–536.

McLachlan, G.; Peel, D. Finite Mixture Models. Wiley; New York: 2000.

Mirsky L. A trace inequality of John von Neumann. Monatshefte für Mathematik. 1975; 79:303–306.

Oh S-Y, Rajaratnam B, Won J-H. CondReg: Condition Number Regularized Covariance Estimation. R package version 0.16. 2013

Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. Journal of the American Statistical Association. 2009; 104 (486):735–746. [PubMed: 19881892]

Pourahmadi M. Covariance estimation: The GLM and regularization perspectives. Statistical Science. 2011; 26 (3):369–387.

Pourahmadi, M. High-Dimensional Covariance Estimation: With High-Dimensional Data. Wiley; 2013.

Pourahmadi M, Daniels MJ, Park T. Simultaneous modelling of the Cholesky decomposition of several covariance matrices. Journal of Multivariate Analysis. 2007; 98 (3):568–587.

Rajaratnam B, Massam H, Carvalho CM. Flexible covariance estimation in graphical Gaussian models. Annals of Statistics. 2008; 36 (6):2818–2849.

Ravikumar P, Wainwright MJ, Raskutti G, Yu B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Electronic Journal of Statistics. 2011; 5:935–980.

Rohde A, Tsybakov AB. Estimation of high-dimensional low-rank matrices. Annals of Statistics. 2011; 39 (2):887–930.

Sheena Y, Gupta AK. Estimation of the multivariate normal covariance matrix under some restrictions. Statistics & Decisions. 2003; 21 (4):327–342.

Stein, C. Estimation of a covariance matrix, 39th A. Meet. Institute of Mathematical Statistics; Atlanta.. 1975.

Warton DI. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. Journal of the American Statistical Association. 2008; 103 (481):340–349.

Won J-H, Lim J, Kim S-J, Rajaratnam B. Condition-number-regularized covariance estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2012

Wu WB, Pourahmadi M. Nonparametric estimation of large covariance matrices of longitudinal data. Biometrika. 2003; 90 (4):831–844.

Appendix A. Finite Sample Comparisons

Table A.2

Average ratio of quadratic loss of CNR and LW to that of CERNN. Values less than 1 indicate better performance than CERNN (bold). Standard deviations are in parentheses.

	singleton					
	CNR/CERNN			**LW/CERNN**		
$p =$	125	250	500	125	250	500
$p/n = 4$	**0.42** (0.13)	**0.19** (0.03)	**0.12** (0.01)	12.4 (6.05)	17.9 (3.97)	33.2 (4.23)
2	**0.36** (0.07)	**0.23** (0.03)	**0.20** (0.02)	9.28 (2.76)	18.1 (2.66)	31.7 (2.09)
1.25	**0.37** (0.04)	**0.29** (0.03)	**0.27** (0.02)	8.87 (1.59)	16.9 (1.48)	26.7 (1.18)

	10% high					
	CNR/CERNN			**LW/CERNN**		
$p =$	125	250	500	125	250	500
$p/n = 4$	1.15 (0.20)	1.80 (0.12)	2.77 (0.13)	4.91 (0.53)	8.01 (0.42)	17.0 (0.66)
2	1.18 (0.11)	1.66 (0.08)	1.84 (0.06)	4.42 (0.30)	6.85 (0.28)	12.3 (0.41)
1.25	1.46 (0.08)	1.39 (0.05)	1.81 (0.05)	4.14 (0.21)	6.08 (0.20)	10.3 (0.29)

	20% high					
	CNR/CERNN			**LW/CERNN**		
$p =$	125	250	500	125	250	500
$p/n = 4$	4.28 (0.70)	8.52 (0.86)	18.5 (1.06)	8.72 (0.88)	23.5 (1.62)	78.5 (3.32)
2	1.97 (0.16)	2.51 (0.13)	5.25 (0.16)	6.10 (0.37)	12.9 (0.53)	31.1 (1.00)
1.25	1.45 (0.07)	1.95 (0.07)	4.90 (0.15)	4.98 (0.20)	9.42 (0.31)	20.9 (0.61)

	30% high					
	CNR/CERNN			**LW/CERNN**		
$p =$	125	250	500	125	250	500
$p/n = 4$	11.9 (1.90)	37.8 (3.17)	128 (6.03)	16.9 (1.74)	62.6 (3.49)	244 (7.32)
2	3.97 (0.36)	6.65 (0.34)	13.6 (0.31)	10.7 (0.67)	31.4 (1.24)	95.0 (2.22)
1.25	2.15 (0.10)	3.76 (0.12)	10.6 (0.30)	7.71 (0.36)	18.6 (0.63)	51.8 (1.56)

	40% high					
	CNR/CERNN			**LW/CERNN**		
$p =$	125	250	500	125	250	500
$p/n = 4$	25.6 (3.18)	83.2 (5.06)	304 (10.7)	28.3 (2.54)	105 (4.60)	407 (10.9)

40% high					
		CNR/CERNN			**LW/CERNN**
p =	125	250	500	125	250	500
2	11.0 (1.10)	28.2 (1.61)	68.9 (3.48)	20.1 (1.23)	66.9 (2.28)	234 (4.74)
1.25	3.66 (0.22)	6.39 (0.14)	18.9 (0.27)	13.5 (0.68)	38.6 (1.05)	119 (1.62)

Table A.3

Average ratio of entropy loss of CNR and LW to that of CERNN. Values less than 1 indicate better performance than CERNN (bold). Standard deviations are in parentheses.

singleton					
		CNR/CERNN			**LW/CERNN**
p =	125	250	500	125	250	500
p/n = 4	**0.83** (0.11)	**0.56** (0.05)	**0.40** (0.03)	5.38 (2.24)	9.20 (2.10)	18.7 (2.61)
2	**0.78** (0.07)	**0.59** (0.03)	**0.49** (0.02)	5.24 (1.63)	11.9 (2.08)	24.6 (2.17)
1.25	**0.79** (0.04)	**0.66** (0.03)	**0.56** (0.02)	5.75 (1.15)	12.8 (1.37)	24.4 (1.30)

10% high					
		CNR/CERNN			**LW/CERNN**
p =	125	250	500	125	250	500
p/n = 4	1.11 (0.12)	1.34 (0.07)	1.62 (0.06)	1.43 (0.11)	1.83 (0.09)	3.19 (0.11)
2	1.04 (0.07)	1.08 (0.04)	1.16 (0.03)	1.29 (0.07)	1.34 (0.05)	1.81 (0.05)
1.25	1.13 (0.05)	0.96 (0.03)	1.26 (0.03)	1.29 (0.05)	1.20 (0.03)	1.38 (0.03)

20% high					
		CNR/CERNN			**LW/CERNN**
p =	125	250	500	125	250	500
p/n = 4	1.90 (0.19)	2.71 (0.15)	3.74 (0.10)	2.16 (0.19)	4.16 (0.20)	8.27 (0.17)
2	1.19 (0.06)	1.20 (0.05)	1.97 (0.03)	1.73 (0.10)	2.93 (0.10)	5.57 (0.11)
1.25	**0.92** (0.03)	1.03 (0.02)	2.14 (0.04)	1.42 (0.05)	2.18 (0.06)	4.05 (0.07)

30% high					
		CNR/CERNN			**LW/CERNN**
p =	125	250	500	125	250	500
p/n = 4	2.50 (0.20)	4.04 (0.17)	6.49 (0.15)	2.69 (0.17)	5.06 (0.17)	9.06 (0.15)
2	1.50 (0.08)	1.72 (0.06)	2.39 (0.02)	2.44 (0.12)	4.58 (0.12)	8.53 (0.10)
1.25	**0.96** (0.03)	1.24 (0.02)	2.64 (0.02)	2.11 (0.10)	3.78 (0.10)	7.13 (0.07)

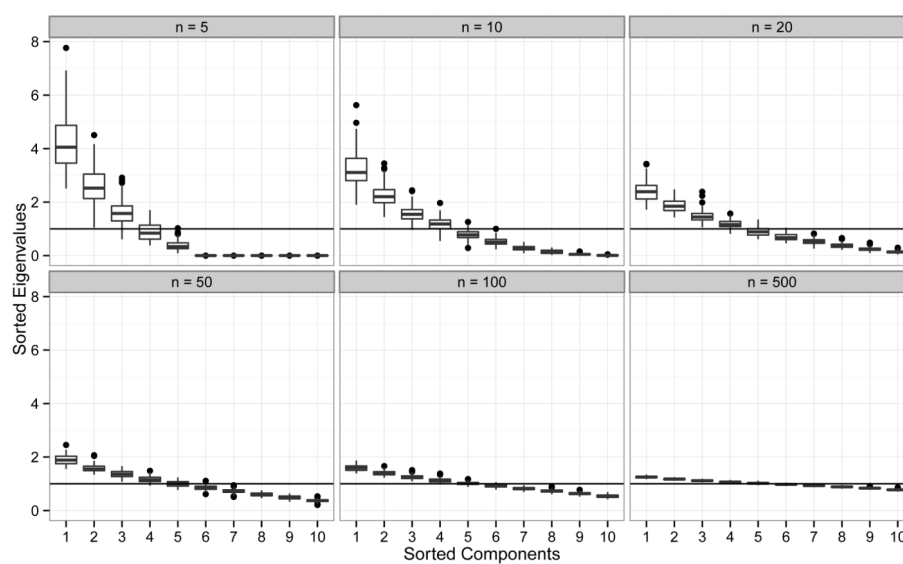| 40% high					
		CNR/CERNN			**LW/CERNN**
p =	125	250	500	125	250	500
p/n = 4	2.80 (0.17)	4.44 (0.13)	7.29 (0.13)	2.79 (0.15)	4.91 (0.12)	8.45 (0.12)
2	2.09 (0.11)	3.04 (0.11)	4.23 (0.12)	2.81 (0.13)	4.91 (0.12)	8.51 (0.09)
1.25	1.13 (0.05)	1.33 (0.02)	2.66 (0.02)	2.67 (0.11)	4.79 (0.09)	8.64 (0.08)

Figure 1.
Boxplots of the sorted eigenvalues of the sample covariance matrix S over 100 random trials. Here the number of components $p = 10$, and the sample size n is drawn from the set $\{5, 10, 20, 50, 100, 500\}$.
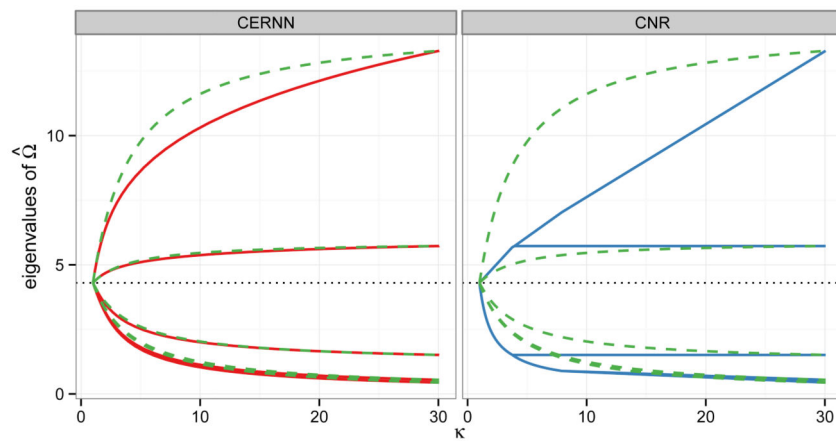
Figure 2.
A comparison of the solution paths for CERNN (left panel, solid line) and CNR (right panel, solid line) against the path for the linear estimator (both panels, dashed line) for sample eigenvalues (13.29, 5.73, 1.51, 0.55, 0.44). The dotted line indicates the mean of the sample eigenvalues.
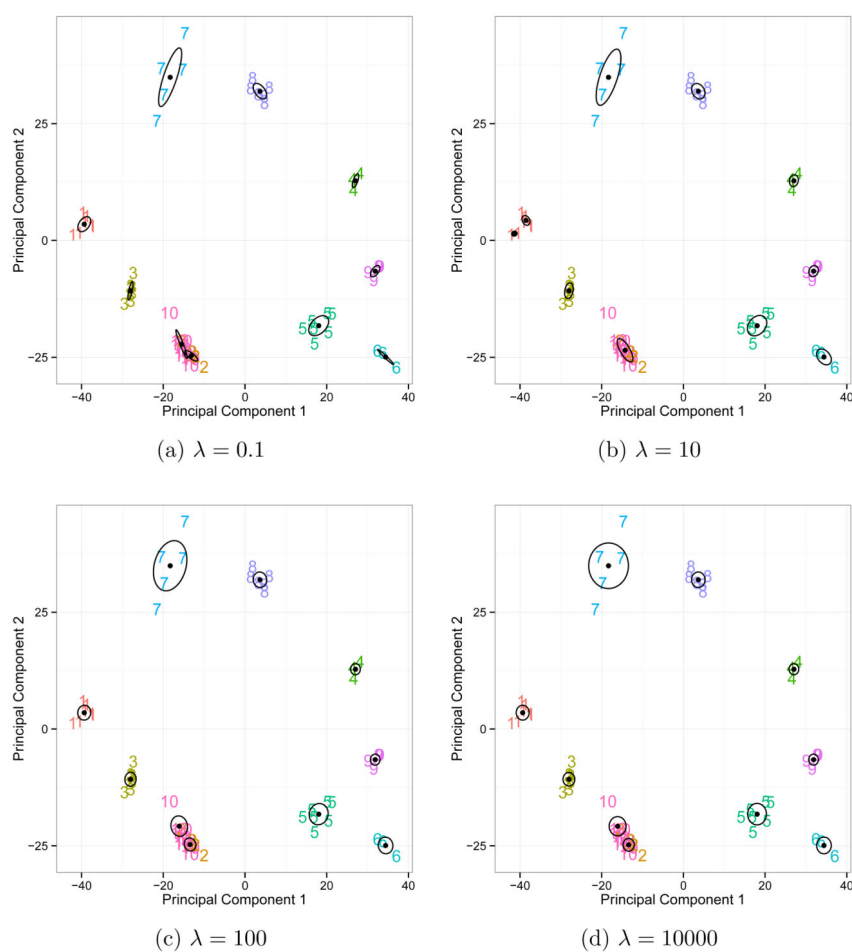
Figure 3.
CERNN clustering projected onto the first two principal components of the data. Ellipses depict the first two eigenvectors (and their corresponding eigenvalues) of the estimated covariances of each cluster.

Table 1

Comparison of CERNN and RDA on three data sets from the UCI machine learning repository. The fourth column indicates the number of parameters (mean and covariance) per group in the QDA model. The fifth and sixth columns breakdown the number of samples per group. The last two columns report the classification success rate in the test set.

data	p	c	$\frac{p(p+3)}{2}$	samples (train)	samples (test)	CERNN	RDA
wine	13	3	104	13/13/10	46/58/38	0.859	0.627
seeds	7	3	35	14/15/13	56/55/57	0.929	0.935
ecoli	7	5	35	30/17/7/3/9	113/60/28/17/43	0.670	0.705