Stable Estimation of a Covariance Matrix Guided by Nuclear Norm Penalties

Eric C. Chi[a], Kenneth Lange[b]

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

Abstract

Estimation of covariance matrices or their inverses 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. In this paper we present an intuitive prior that shrinks the classic sample covariance estimator towards a stable target. We prove that our estimator is consistent and asymptotically efficient. Thus, it gracefully transitions towards the sample covariance matrix as the number of samples grows relative to the number of covariates. We demonstrate the utility of our estimator in four standard situations – regression, canonical correlation analysis, discriminant analysis, and EM clustering – when the number of samples is dominated by or comparable to the number of covariates.

Keywords: Covariance estimation, Regularization, Condition number, Canonical correlation analysis, Discriminant analysis, Clustering

1. Introduction

Estimates of covariance matrices and their inverses play a central role in many core statistical methods, ranging from least squares regression to EM clustering. In these applications it is crucial to obtain estimates that are not just non-singular but also well-conditioned. It is well known that the sample covariance matrix

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

is the maximum likelihood estimates of the population variance $\boldsymbol{\Omega}$ of a random sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ from a multivariate normal distribution. When the number of components p of $\boldsymbol{y}$ exceeds the sample size n, the sample covariance $\boldsymbol{S}$ is no longer invertible. Even when p is close to n, $\boldsymbol{S}$ becomes ill-conditioned and small perturbations in measurements can lead to disproportionately large fluctuations in its entries. To deal with this dilemma and to stabilize estimation generally, one can add a penalty that steers covariance estimates towards well-conditioned values.

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 $\boldsymbol{y}_i \sim N(\boldsymbol{0}, \boldsymbol{I}_{10})$. Figure 1 presents boxplots of the sorted eigenvalues of the sample covariance matrix $\boldsymbol{S}$ 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 $\frac{p}{n}$ approaches $\tau \in (0, 1)$, then the eigenvalues of $\boldsymbol{S}$ tend to the Marĉenko-Pastur law (Marĉenko and Pastur, 1967), which is supported on the interval $([1-\sqrt{\tau}]^2, [1+\sqrt{\tau}]^2)$. Thus, the distortion worsens as τ 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 four fundamental statistical methods - linear regression, canonical correlation analysis, 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 robust estimation of structured and unstructured covariance matrices. Section 3 specifies our Bayesian prior and derives the maximum a posteriori estimator under the prior. Section 4 proves that the estimator is consistent and asymptotically efficient. Section 5 illustrates the estimator for some common tasks in statistics. Finally, Section 7 discusses limitations,

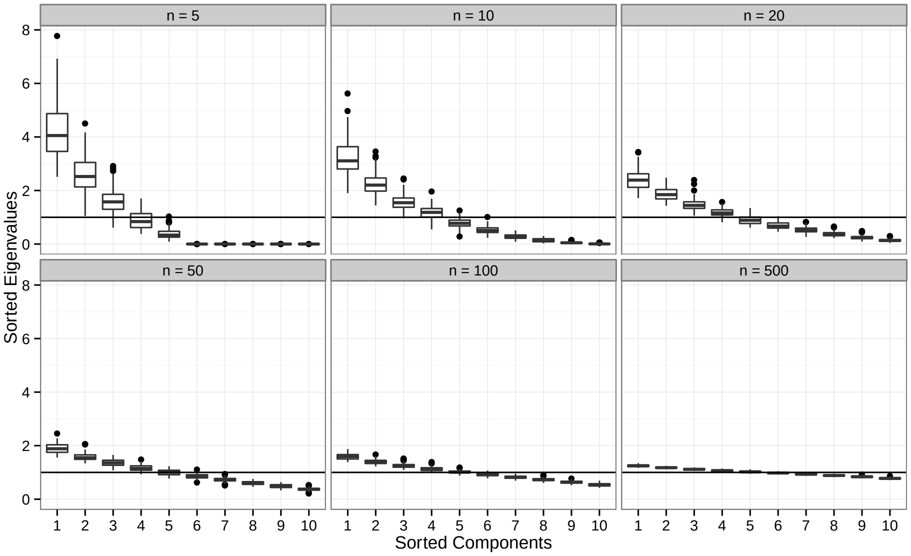


Figure 1: Boxplots of the sorted eigenvalues of the sample covariance matrix S over 100 random trials. Here the number of components $p = 10$, and the sample size n is drawn from the set $\{5, 10, 20, 50, 100, 500\}$.

generalizations, and further applications of the estimator.

2. Related Work

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 make sense 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 current paper, we do not assume any special structure. Our sole concern is to improve the condition number of the sample covariance matrix.

Thus, we work in the context of rotationally-invariant estimators first proposed by Stein (1975). If $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^t$ is the spectral decomposition of $\boldsymbol{S}$, then Stein suggests alternative estimators of the form

$$\hat{\boldsymbol{\Sigma}} \quad = \quad \boldsymbol{U}\,\mathrm{diag}(\hat{d}_1, \ldots, \hat{d}_p)\,\boldsymbol{U}^t$$

that change the eigenvalues but not the eigenvectors of $\boldsymbol{S}$. In particular, Stein (1975); Haff (1991); Ledoit and Wolf (2004) and Warton (2008) study the family

$$\hat{\boldsymbol{\Sigma}} \quad = \quad (1-\gamma)\boldsymbol{S} + \gamma\boldsymbol{T} \tag{1}$$

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

$$\hat{d}_i \quad = \quad (1-\gamma)d_i + \gamma\rho.$$

Ledoit and Wolf (2004, 2012) show that linear shrinkage works well when $\frac{p}{n}$ is large or the population eigenvalues are close to one another. On the other hand, if $\frac{p}{n}$ is small or the population eigenvalues are dispersed, linear shrinkage yields marginal improvements over the sample covariance. Nonlinear shrinkage estimators are also possible (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.

3. Maximum a Posteriori Estimation with a Novel Prior

Adding a penalty can be accomplished by imposing a prior $\pi(\boldsymbol{\Omega})$ on $\boldsymbol{\Omega}$. The prior we advocate is designed to steer the eigenvalues of $\boldsymbol{\Omega}$ away from the extremes of 0 and ∞. The reasonable choice

$$\pi(\boldsymbol{\Omega}) \quad \propto \quad e^{-\frac{\lambda}{2}\left[\alpha\|\boldsymbol{\Omega}\|_* + (1-\alpha)\|\boldsymbol{\Omega}^{-1}\|_*\right]},$$

relies on the nuclear norms of $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}^{-1}$, a positive strength constant λ, and an admixture constant $\alpha \in (0, 1)$. This is a proper prior on the set of invertible matrices because

$$e^{-\frac{\lambda}{2}\left[\alpha\|\boldsymbol{\Omega}\|_* + (1-\alpha)\|\boldsymbol{\Omega}^{-1}\|_*\right]} \quad \leq \quad e^{-\eta\lambda\|\boldsymbol{\Omega}\|_F}$$

for some positive constant η by virtue of the equivalence of vector norms on $\mathbb{R}^{p^2}$. The normalizing constant of $\pi(\boldsymbol{\Omega})$ is irrelevant in the ensuing discussion. Consider therefore minimization of the objective function

$$f(\boldsymbol{\Omega}) \;=\; \frac{n}{2}\ln\det\boldsymbol{\Omega} + \frac{n}{2}\operatorname{tr}(\boldsymbol{S}\boldsymbol{\Omega}^{-1}) + \frac{\lambda}{2}\left[\alpha\|\boldsymbol{\Omega}\|_* + (1-\alpha)\|\boldsymbol{\Omega}^{-1}\|_*\right].$$

The maximum of $-f(\boldsymbol{\Omega})$ occurs at the posterior mode. In the limit as λ tends to 0, $-f(\boldsymbol{\Omega})$ reduces to the loglikelihood. In the sequel we will refer to our MAP covariance estimate by the acronym NECM (nuclear estimate of a covariance matrix).

Fortunately, three of the four terms of $f(\boldsymbol{\Omega})$ can be expressed as functions of the eigenvalues e_i of $\boldsymbol{\Omega}$. The trace contribution presents a greater challenge. 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 $\boldsymbol{\Omega} = \boldsymbol{V}\boldsymbol{E}\boldsymbol{V}^t$ denote the spectral decomposition of $\boldsymbol{\Omega}$ with positive diagonal entries e_i ordered from largest to smallest. In view of von Neumann-Fan inequality (Mirsky, 1975), we can assert that

$$-\operatorname{tr}(\boldsymbol{S}\boldsymbol{\Omega}^{-1}) \;\leq\; -\sum_{i=1}^{p}\frac{d_i}{e_i},$$

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

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

We reject the negative root as inconsistent with $\boldsymbol{\Omega}$ being positive definite. For the special case $n = 0$ of no data, all $e_i = \sqrt{(1-\alpha)/\alpha}$, and the prior mode occurs at a multiple of the identity matrix. A simple rearrangement of the right-hand side of equation (2) shows that e_i depends on λ and n only through the ratio $\frac{\lambda}{n}$.

Holding all but one variable fixed in formula (2), one can demonstrate after a fair amount of algebra that

$$\begin{aligned} e_i &= d_i + \frac{\lambda(1-\alpha-\alpha d_i^2)}{n} + O\left(\frac{1}{n^2}\right), \quad n \to \infty \\ e_i &= \sqrt{\frac{1-\alpha}{\alpha}} + \left[\sqrt{\frac{1-\alpha}{\alpha}}\frac{nd_i}{2(1-\alpha)} - \frac{n}{2\alpha}\right]\frac{1}{\lambda} + O\left(\frac{1}{\lambda^2}\right), \quad \lambda \to \infty. \end{aligned} \tag{3}$$

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 $\boldsymbol{\Omega}$ toward the prior mode. Choice of the constants λ and α is an issue. To match the prior to the scale of the data, we recommend determining α as the solution to the equation

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

Cross validation leads to a reasonable choice of λ. For the sake of brevity, we omit further details. For a summary other approaches to this subject, consult Ledoit and Wolf (2004).

Figure 2 shows the nonlinear shrinkage function (2) applied at four different values of $\frac{\lambda}{n}$ with $\alpha = 0.5$. As $\frac{\lambda}{n}$ increases, the eigenvalue d is shrunk towards the target eigenvalue 1. The rate of shrinkage, however, is nonlinear. Eigenvalues greater than the target are pulled more aggressively towards the target than eigenvalues less than the target.

4. Consistency and Asymptotic Efficiency

In proving consistency, we will need 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

$$\sum_{i=1}^p (a_i - b_i)^2 \leq \|\boldsymbol{A} - \boldsymbol{B}\|_F^2. \tag{4}$$

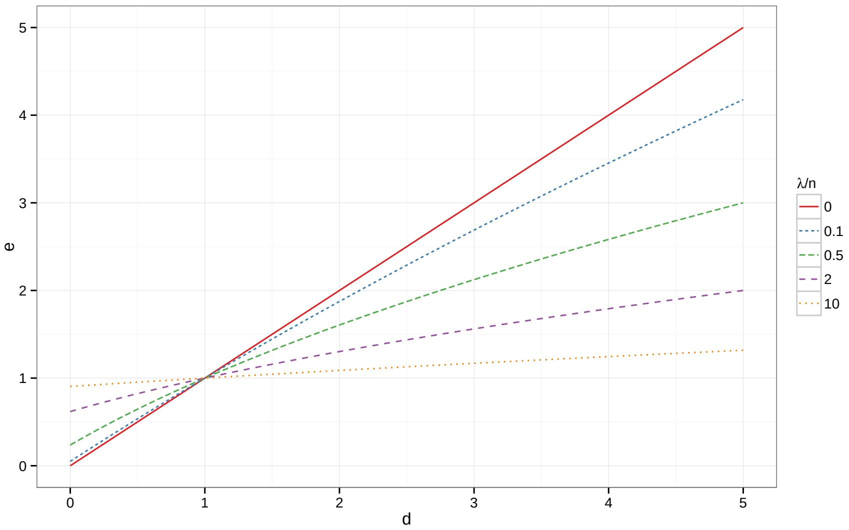


Figure 2: The nonlinear shrinkage function for four different values of λ/n and $\alpha = 0.5$.

This is a consequence of Fan's inequality because $\sum_{i=1}^p a_i^2 = \|\boldsymbol{A}\|_F^2$ and $\sum_{i=1}^p b_i^2 = \|\boldsymbol{B}\|_F^2$. If the two matrices $\boldsymbol{A} = \boldsymbol{U}\operatorname{diag}(\boldsymbol{a})\boldsymbol{U}^t$ and $\boldsymbol{B} = \boldsymbol{U}\operatorname{diag}(\boldsymbol{b})\boldsymbol{U}^t$ are simultaneously diagonalizable, then equality holds in inequality (4). We will also need the inequalities

$$\sqrt{1+x} \quad \leq \quad 1 + \frac{x}{2} \quad \text{and} \quad \sqrt{1+x} \quad \geq \quad 1 + \frac{x}{2} - \frac{x^2}{8} \tag{5}$$

for nonnegative x. Verification 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 $\boldsymbol{\Omega}$ with eigenvalues ω_1 through ω_p. Inequality (4) 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 $\boldsymbol{S}_n$. To make this reasoning rigorous, we must show that the asymptotic expansion (3) is uniform as the eigenvalues d_{ni} converge to the eigenvalues ω_i. This is where the inequalities (5) come into play. Indeed, we have

$$\begin{aligned}\frac{\lambda(1-\alpha)}{n}-\frac{n}{2\lambda\alpha}\frac{x^2}{8} &\leq \; e_{ni}-d_{ni} \;\leq \; \frac{\lambda(1-\alpha)}{n} \qquad (6)\\ x &= \frac{4\lambda\alpha d_{ni}}{n}+\frac{4\lambda^2\alpha(1-\alpha)}{n^2}.\end{aligned}$$

The identity

$$\|\boldsymbol{S}_n-\boldsymbol{\Omega}_n\|_F^2 = \sum_{i=1}^{p}(d_{ni}-e_{ni})^2$$

finishes the proof that $\boldsymbol{\Omega}_n$ tends to $\boldsymbol{\Omega}$.

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

$$\sqrt{n}(\boldsymbol{\Omega}_n-\boldsymbol{\Omega}) = \sqrt{n}(\boldsymbol{S}_n-\boldsymbol{\Omega})+\sqrt{n}(\boldsymbol{\Omega}_n-\boldsymbol{S}_n)$$

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

$$\|\sqrt{n}(\boldsymbol{\Omega}_n-\boldsymbol{S}_n)\|_F^2 = n\sum_{i=1}^{p}(d_{ni}-e_{ni})^2$$

tends almost surely to 0 owing to the bounds (6) and the convergence of d_{ni} to ω_i.

5. Applications

Several common statistical procedures are potential beneficiaries of shrinkage estimation of sample covariance matrices. Here we illustrate how NECM applies to regression, canonical correlation analysis, discriminant analysis, and clustering. In each setting we compare NECM to a few alternative covariance estimators. A comprehensive comparison of regularized estimators is beyond the scope of this paper, and we simply document the fact that NECM is competitive with representative existing methods.

5.1. Covariance Regularized Regression

Consider the linear regression problem

$$\hat{\mathbf{b}} \quad = \quad \arg\min_{\boldsymbol{b}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2.$$

When the $n \times p$ design $\boldsymbol{X}$ has full column rank, the problem admits the classical solution $\hat{\mathbf{b}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$. When the design $\boldsymbol{X}$ is singular or nearly so, ridge regression is a reasonable remedy (Hoerl and Kennard, 1970). Ridge regression minimizes the regularized criterion

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|_2^2 + \frac{1}{2}\lambda\|\boldsymbol{b}\|_2^2$$

and delivers the explicit solution $\hat{\mathbf{b}} = (\boldsymbol{X}^t\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^t\boldsymbol{y}$. This is hardly the only way to attack regression in problematic situations. Recall that when $\boldsymbol{X}$ has full column rank, the sample precision matrix $\boldsymbol{\Theta}$ exists, and the standard regression estimator $\hat{\mathbf{b}}$ can be expressed in terms of the block structure

$$\boldsymbol{\Theta} \quad = \quad \begin{pmatrix} \boldsymbol{\Theta}_{xx} & \boldsymbol{\Theta}_{xy} \\ \boldsymbol{\Theta}_{xy}^t & \boldsymbol{\Theta}_{yy} \end{pmatrix}.$$

as $\hat{\mathbf{b}} = -\frac{\boldsymbol{\Theta}_{\boldsymbol{xy}}}{\Theta_{yy}}$. This fact suggests that we substitute a regularized estimate of the precision matrix in $\hat{\mathbf{b}} = -\frac{\boldsymbol{\Theta}_{\boldsymbol{xy}}}{\Theta_{yy}}$ when $\boldsymbol{X}$ is either rank deficient or ill-conditioned. The Scout method (Witten and Tibshirani, 2009) pursues this strategy. Let $\boldsymbol{S}$ denote the sample covariance of the augmented data matrix $\tilde{\boldsymbol{X}} = \begin{pmatrix}\boldsymbol{X} & \boldsymbol{y}\end{pmatrix}$, namely

$$\boldsymbol{S} \quad = \quad \begin{pmatrix} \boldsymbol{S}_{xx} & \boldsymbol{S}_{xy} \\ \boldsymbol{S}_{xy}^t & \boldsymbol{S}_{yy} \end{pmatrix}.$$

The Scout method proceeds in two stages. Stage one computes a regularized covariance estimate $\hat{\boldsymbol{\Theta}}_{xx}$ of $\boldsymbol{\Theta}_{xx}$,

$$\hat{\boldsymbol{\Theta}}_{xx} \quad = \quad \arg\max_{\boldsymbol{\Theta}_{xx}} \; \log(\det \boldsymbol{\Theta}_{xx}) - \operatorname{tr}(\boldsymbol{S}_{xx}\boldsymbol{\Theta}_{xx}) - J_1(\boldsymbol{\Theta}_{xx}), \tag{7}$$

subject to a penalty $J_1(\boldsymbol{\Theta}_{xx})$ that steers solutions toward sparsity or some other desired structure. Stage two solves the optimization problem

$$\hat{\boldsymbol{\Theta}} \quad = \quad \arg\max_{\boldsymbol{\Theta}} \; \log(\det \boldsymbol{\Theta}) - \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Theta}) - J_2(\boldsymbol{\Theta}) \tag{8}$$

subject to the constraint $\boldsymbol{\Theta}_{xx} = \hat{\boldsymbol{\Theta}}_{xx}$, where $J_2(\boldsymbol{\Theta})$ is a second penalty. Witten and Tibshirani (2009) show that the Scout method generalizes a variety of popular penalized regression methods such as ridge regression, the elastic net (Zou and Hastie, 2005), and the lasso (Tibshirani, 1996), for appropriate choices of J_1 and J_2. When $J_2(\boldsymbol{\Theta}) = \sum_{ij} \|\theta_{ij}\|^{p_2}$ for $p_2 = 1$ or 2, then the solution of the second stage is equivalent to the simpler optimization problem (Witten and Tibshirani, 2009, Claim 1)

$$\hat{\mathbf{b}} \;=\; \arg\min_{\boldsymbol{b}} \boldsymbol{b}^t \hat{\boldsymbol{\Theta}}_{xx}^{-1} \boldsymbol{b} - 2\boldsymbol{S}_{xy}^t \boldsymbol{b} + \lambda_2 \|\boldsymbol{b}\|_{p_2}^{p_2}. \qquad (9)$$

Although the two stage procedure is general, it is primarily motivated by the choice $J_1(\boldsymbol{\Theta}_{xx}) = \sum_{ij} |\theta_{ij}|$. This ℓ_1 (lasso) penalty shrinks elements in $\boldsymbol{\Theta}_{xx}$ toward zero. Recall that under a multivariate normal assumption, $\theta_{ij} = 0$ if and only if the ith and jth variables are conditionally independent. The Witten and Tibshirani method "scouts" for variables that are truly correlated, conditional on all other variables. This motivation is reasonable, for example in microarray data, where a large fraction of the big pool of covariates are conditionally independent. Precision matrix estimation is broken down into two stages because it is undesirable to shrink the partial correlations between the covariate and response variables.

For the ridge penalty $J_1(\boldsymbol{\Theta}_{xx}) = \sum_{ij} \theta_{ij}^2$, the precision matrix estimator is a Stein shrinkage estimator with the ith eigenvalue of the precision matrix determined by

$$e_i^{-1} \;=\; \frac{-d_i + \sqrt{d_i^2 + 8\lambda_1/n}}{4\lambda_1/n}.$$

Straightforward algebraic manipulations lead to the asymptotic expansions

$$\begin{aligned} e_i^{-1} &= d_i^{-1} - \frac{16\lambda_1}{d_i^3}\frac{1}{n} + O\left(\frac{1}{n^2}\right), \quad n \to \infty \\ e_i^{-1} &= O\left(\frac{1}{\sqrt{\lambda_1}}\right), \quad \lambda_1 \to \infty. \end{aligned}$$

Thus, the regularized precision estimates behave as desired as $n \to \infty$. As expected, the precision estimate is shrunk towards zero as $\lambda_1 \to \infty$.

We now compare covariance regularized regression under the six scenarios described in Witten and Tibshirani (2009), substituting NECM in the second estimation stage (9) along with a lasso penalty. We specifically consider

NECM, the elastic net, and Scout(1,1), and Scout(2,1), where Scout(p_1, p_2) means that $J_1(\boldsymbol{\Theta}_{xx}) = \sum_{ij} |\theta_{ij}|^{p_1}$ and $J_2(\boldsymbol{\Theta}_{xx}) = \sum_{ij} |\theta_{ij}|^{p_2}$. In all simulations, data were generated via the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{I})$. We split our simulated data set into three partitions: training, validation, and testing denoted by $\cdot/\cdot/\cdot$. For a grid of λ_1 and λ_2 values, we estimated $\hat{\mathbf{b}}$ and chose the $\hat{\mathbf{b}}$ corresponding to the pair (λ_1, λ_2) with the smallest prediction error. We standardized both the responses and covariates. The six scenarios were:

1. 20/20/200 observations, 8 predictors: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^t, \sigma = 3$, and $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = 2^{-|i-j|}$.

2. Same as scenario 1 except $\beta_i = 0.85$ for $i = 1, \ldots, 8$.

3. 100/100/400 observations, 40 predictors: $\beta_i = 0$ for $i = 1, \ldots, 10$ and $i = 21, \ldots, 30$ and 2 otherwise, $\sigma = 15$, and $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = 0.5$ for $i \neq j$ and $\Sigma_{ii} = 1$.

4. 50/50/400 observations, 40 predictors: $\beta_i = 3$ for $i = 1, \ldots, 15$ and 0 otherwise, and $\sigma = 15$. The first 15 predictors satisfy

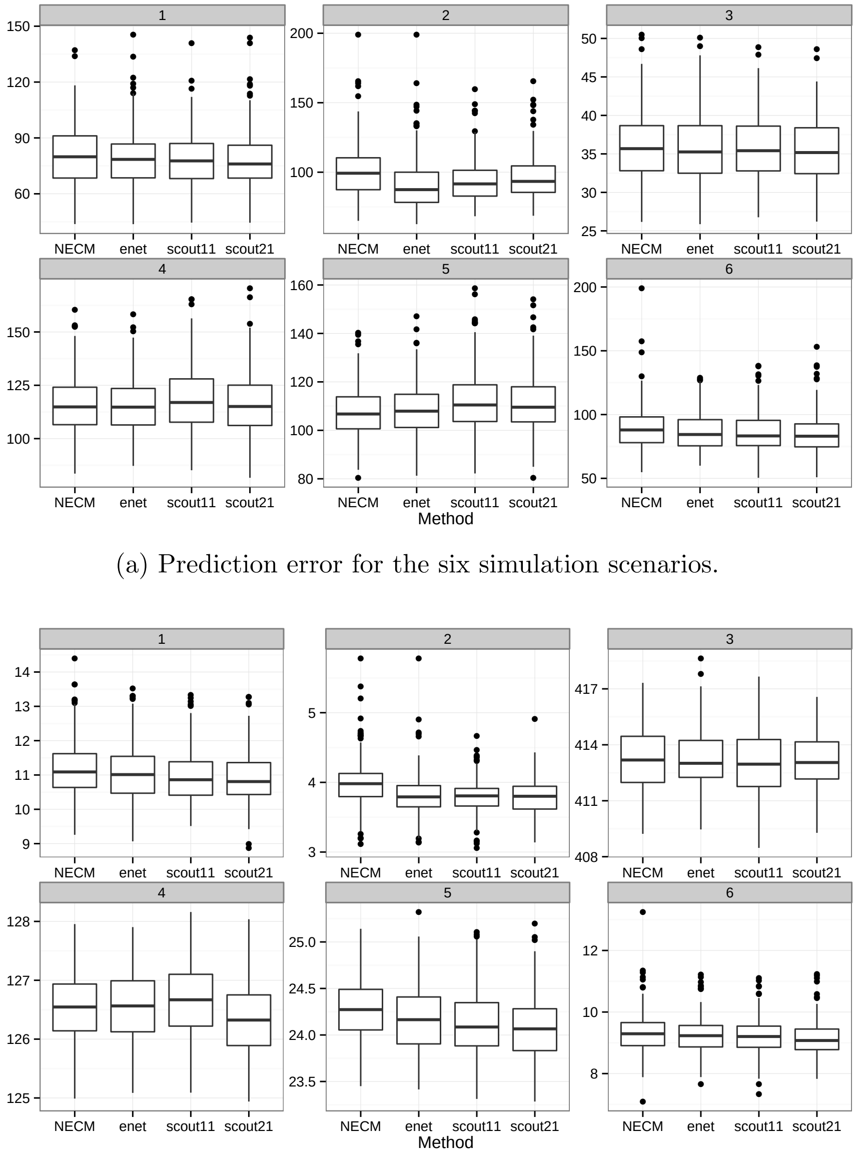
$$\boldsymbol{x}_i \quad = \quad \begin{cases} \boldsymbol{z}_1 + \boldsymbol{\epsilon}_i^x & i = 1, \ldots, 5 \\ \boldsymbol{z}_2 + \boldsymbol{\epsilon}_i^x & i = 6, \ldots, 10 \\ \boldsymbol{z}_3 + \boldsymbol{\epsilon}_i^x & i = 11, \ldots, 15 \end{cases},$$

   where $\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{z}_3 \sim N(0, \boldsymbol{I})$ and $\boldsymbol{\epsilon}_i^x$ are i.i.d. $N(0, 0.01\boldsymbol{I})$ for $i = 1, \ldots, 15$. The remaining predictors $\boldsymbol{x}_i$ are i.i.d. $N(0, \boldsymbol{I})$ for $i = 16, \ldots, 40$.

5. 50/50/400 observations, 50 predictors: $\beta_i = 2$ for $i \leq 8$ and 0 otherwise, $\sigma = 6$, and $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\Sigma_{ij} = 0.5$ for $i \neq j$ and $i, j \leq 9$, $\Sigma_{ij} = 0$ otherwise, and $\Sigma_{ii} = 1$.

6. Same as scenario 1, except $\boldsymbol{\beta} = (3, 1.5, 0, 0, 0, 0, -1, -1)^t$.

In scenarios 1, 3, 4, 5, and 6, $\boldsymbol{\beta}$ is sparse. Scenarios 1, 2, 4, 5, and 6 have a sparse precision matrix. In scenario 4 there are three blocks of highly correlated variables. Figure 3 shows box plots of the prediction error and ℓ_2 distance between the true and estimated regression coefficients over 200 replicates for each of the six simulation scenarios. Despite the fact that most of the simulation scenarios involve sparse precision matrices, all methods perform similarly.



(a) Prediction error for the six simulation scenarios.

(b) ℓ_2 distance between the estimated and true regression coefficients.

Figure 3: Comparison of NECM, the elastic net, and Scout(2,1) and Scout(1,1) on 200 simulations for six scenarios.

5.2. Canonical Correlation Analysis

Suppose we have two multidimensional random variables $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{y} \in \mathbb{R}^q$ representing two sets of measurements on a common set of subjects. The goal in canonical correlation analysis (CCA) is to determine a coordinate system that maximizes the cross-correlation between the two sets of measurements (Hotelling, 1936). In other words we seek the linear transformations $\boldsymbol{a}^t\boldsymbol{x}$ and $\boldsymbol{b}^t\boldsymbol{y}$ that are maximally correlated subject to the constraint that $\boldsymbol{a}^t\boldsymbol{x}$ and $\boldsymbol{b}^t\boldsymbol{y}$ have unit variance. If we denote the joint covariance matrix by

$$\boldsymbol{\Sigma} \;=\; \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{\boldsymbol{xy}} \\ \boldsymbol{\Sigma}_{\boldsymbol{yx}} & \boldsymbol{\Sigma}_{\boldsymbol{yy}}, \end{pmatrix}$$

then we can succinctly define the CCA problem as maximizing the criterion

$$\frac{\boldsymbol{a}^t\boldsymbol{\Sigma}_{\boldsymbol{xy}}\boldsymbol{b}}{[\boldsymbol{a}^t\boldsymbol{\Sigma}_{\boldsymbol{xx}}\boldsymbol{a}]^{\frac{1}{2}}[\boldsymbol{b}^t\boldsymbol{\Sigma}_{\boldsymbol{yy}}\boldsymbol{b}]^{\frac{1}{2}}}.$$

Assuming that $\boldsymbol{\Sigma}_{\boldsymbol{xx}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{yy}}$ are positive definite, the optimal $\boldsymbol{a}$ and $\boldsymbol{b}$ solve the generalized eigenvalue problems

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{xy}}\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{yx}}\boldsymbol{a} &= \upsilon\boldsymbol{\Sigma}_{\boldsymbol{xx}}\boldsymbol{a} \\ \boldsymbol{\Sigma}_{\boldsymbol{yx}}\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{xy}}\boldsymbol{b} &= \upsilon\boldsymbol{\Sigma}_{\boldsymbol{yy}}\boldsymbol{b}. \end{aligned}$$

Since CCA is invariant under affine transformations, without loss in generality, one can replace $\boldsymbol{\Sigma}$ by the sample correlation matrix $\boldsymbol{R}$ (Mardia et al., 1979, Theorem 10.2.4).

To apply CCA to real data, we collect the $\boldsymbol{x}$ measurements into an $n \times p$ matrix $\boldsymbol{X}$ and the $\boldsymbol{y}$ measurements into an $n \times q$ matrix $\boldsymbol{Y}$. Provided n is sufficiently large compared to p and q, the sample correlation matrix $\boldsymbol{R}$ is a well conditioned estimate of $\boldsymbol{\Sigma}$. When $n < \max\{p, q\}$, as is typical with modern high-throughput experiments, one or both of the sample correlation matrices $\boldsymbol{R}_{xx}$ and $\boldsymbol{R}_{\boldsymbol{yy}}$ will be singular. Early attempts to combat singularity include the canonical ridge method (Vinod, 1976; Leurgans et al., 1993; González et al., 2008), which replaces the sample correlation matrix $\boldsymbol{R}$ with

$$\begin{pmatrix} \boldsymbol{R}_{xx} + \lambda_1\boldsymbol{I} & \boldsymbol{R}_{\boldsymbol{xy}} \\ \boldsymbol{R}_{\boldsymbol{yx}} & \boldsymbol{R}_{\boldsymbol{yy}} + \lambda_2\boldsymbol{I} \end{pmatrix}.$$

Alternatively, one could estimate the correlation matrix using NECM.

We now compare the CCA performance of NECM to ridge penalized estimates and the Ledoit-Wolf estimates on a study on the effects of nutrition in mice (Martin et al., 2007). The goal of the study was to determine genetic and dietary effects on the expression of liver genes involved in fatty acid catabolism. Forty mice were studied. On each mouse, a panel of 120 gene expression levels and a panel of 21 liver fatty acid concentrations were measured. Half of the mice were wild-type, while the other half were deficient in the PPARα receptor, which is an important modulator of lipid metabolism. Each half was divided into groups of four mice fed one of five diets differentiated by oils with varying fatty acid profiles. In the context of CCA we sought linear summaries of gene expression profiles and fatty acid concentration profiles that are most correlated. Since there are more gene expression measurements than study subjects, we had to resort to regularized covariance estimates. We applied 5-fold cross validation to select λ_1 and λ_2 for NECM and ridge penalization. The Ledoit-Wolf estimator does not employ regularization parameters and required no tuning. Figure 4 shows that the three methods identify very similar projections. In all three, the first component, in both expression and fatty acid space, captures much of the variation in genotype (wild-type versus PPARα deficient). In all three, the first two components together capture quite a bit of the variation due to diet.

5.3. Discriminant Analysis

Linear discriminant analysis is yet another area that stands to benefit from shrinkage estimation of the sample covariance matrix. The classical discriminant function

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

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 $\mathcal{C}_1, \ldots, \mathcal{C}_c$, then the standard estimator of $\boldsymbol{\Sigma}$ is

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

where

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

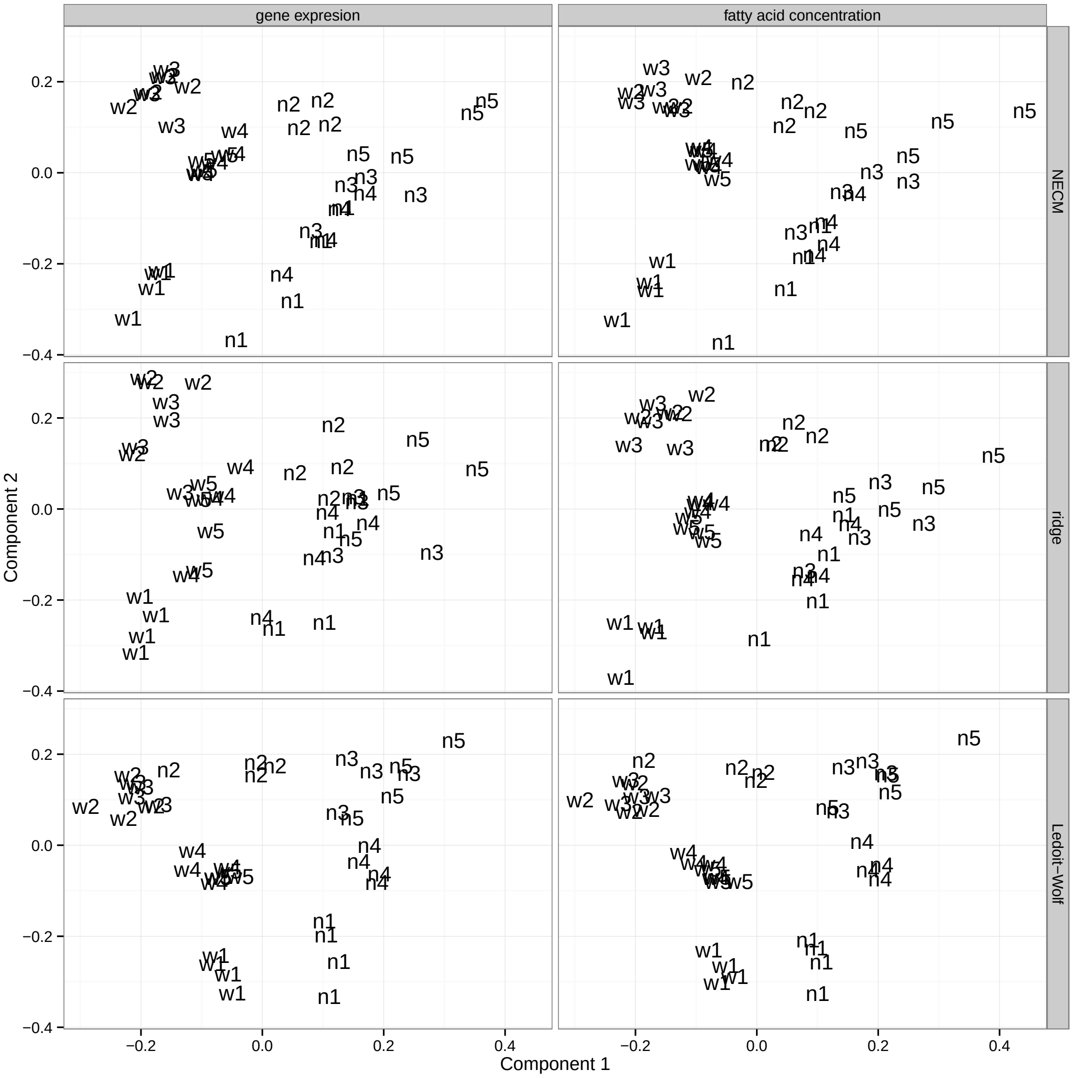


Figure 4: Regularized CCA on mouse nutrition data: Each point is the projection of a sample onto the first two CCA components. The left panel shows projections of the gene expression space, and the right panel shows the projections of the fatty acid space. The rows correspond to projections under the three regularized covariances estimates (NECM, ridge, Ledoit-Wolf). Samples are denoted by the concatenation of genotype (w: wild-type, n: PPARα deficient) and diet (1: COC, 2: FISH, 3: LIN, 4: REF, 5:SUN).

One can obviously shrink $\hat{\mathbf{\Sigma}}$ to moderate its eigenvalues. In quadratic discriminant analysis, a separate covariance matrix $\mathbf{\Sigma}_k$ 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 $\mathbf{\Sigma}_k$ toward a common $\mathbf{\Sigma}$ via a convex combination $\alpha\mathbf{\Sigma}_k + (1-\alpha)\mathbf{\Sigma}$. 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 NECM and these methods is that NECM 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, even linear discriminant analysis is not viable since a common sample covariance estimate will be ill-conditioned if not singular. Nonetheless, out 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 k regularization parameters for NECM and a single α parameter for (Friedman, 1989). The testing errors in Table 1 demonstrate that NECM performs well in comparison with RDA. Even when it does not perform as accurately, its drop off is small.

6. Covariance Regularized EM Clustering

We now show how NECM stabilizes estimation in the standard EM clustering algorithm (McLachlan and Peel, 2000). Let $\phi(\boldsymbol{y} \mid \boldsymbol{\mu}, \mathbf{\Sigma})$ denote a multivariate Gaussian density with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$. EM cluster-

Table 1: Comparison of NECM 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)	NECM	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

ing revolves around the admixture density

$$h(\boldsymbol{y} \mid \Xi) \;=\; \sum_{k=1}^{c} \pi_k \, \phi(\boldsymbol{y} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

with parameters $\Xi = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k\}_{k=1}^{c}$. The π_k are nonnegative admixture weights summing to 1. We are given n independent observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ and wish to estimate Ξ. If z_{ik} is the indicator function of the event that observation i comes from cluster k, then the complete data loglikelihood plus logprior amounts to

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

Straightforward application of Bayes rule yields the conditional expectation

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

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

$$\boldsymbol{\Sigma}_k \;=\; \arg\min_{\boldsymbol{\Sigma}} \frac{w_k}{2} \log\det \boldsymbol{\Sigma} + \frac{w_k}{2} \operatorname{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{S}_k) + \frac{\lambda}{2} \left[\alpha \|\boldsymbol{\Sigma}\|_* + (1-\alpha)\|\boldsymbol{\Sigma}^{-1}\|_*\right].$$

To further stabilize the estimation procedure, one can put a Dirichlet prior with rates γ_k on the cluster probabilities π_k. This leads to modified updates

$$\pi_k \quad = \quad \frac{w_k + \gamma_k - 1}{n + \sum_{k=1}^{c} \gamma_k - c}.$$

Finally, we address two practical issues. First, there is the question of how to choose α. 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. We could estimate a separate α_k for each cluster, but doing so leads to unstable estimates. Instead we simply fix α at $\frac{1}{2}$ for all clusters. This action shrinks all covariance matrices nonlinearly towards the identity matrix and provides good adaptivity to the data without sacrificing numerical stability. Second, while placing an appropriate Dirichlet prior on π_k ensures that it will be strictly positive for all iterations, it is still possible for $w_{ik} \approx 0$ for all i for a given k. If this happens, the updates for $\mathbf{\Sigma}_k$ and $\boldsymbol{\mu}_k$ may behave poorly. As a precaution, we refuse to update $\mathbf{\Sigma}_k$ and $\boldsymbol{\mu}_k$ when $w_{ik} \approx 0$ for all i.

Similar approaches have been employed previously. Fraley and Raftery (2007) suggest a restricted parameterization of the 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 adapted to the data by choosing the regularization parameter λ by cross-validation. Indeed this approach was proposed by Ruan et al. (2011), who used an ℓ_1 penalty to enforce sparsity in the estimated precision matrices. Sparse precision matrices corresponds to the prior belief that many of the covariates are conditionally independent. While this may be true for some data sets, our regularized covariance estimates does not make this assumption.

Figure 5 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 used $\gamma_k = 2$ for all k, fixed $\lambda = 10$, and set $c = 10$. We ran our algorithm 100 times using random initializations with the k-means++ algorithm (Arthur and Vassilvitskii, 2007) and kept the clustering that gave the greatest penalized likelihood. The resulting clustering is quite sensible. The only missteps are splitting cluster 1 into two clusters and merging clusters 2 and 10. The latter decision is reasonable given how much clusters 2 and 10 overlap.

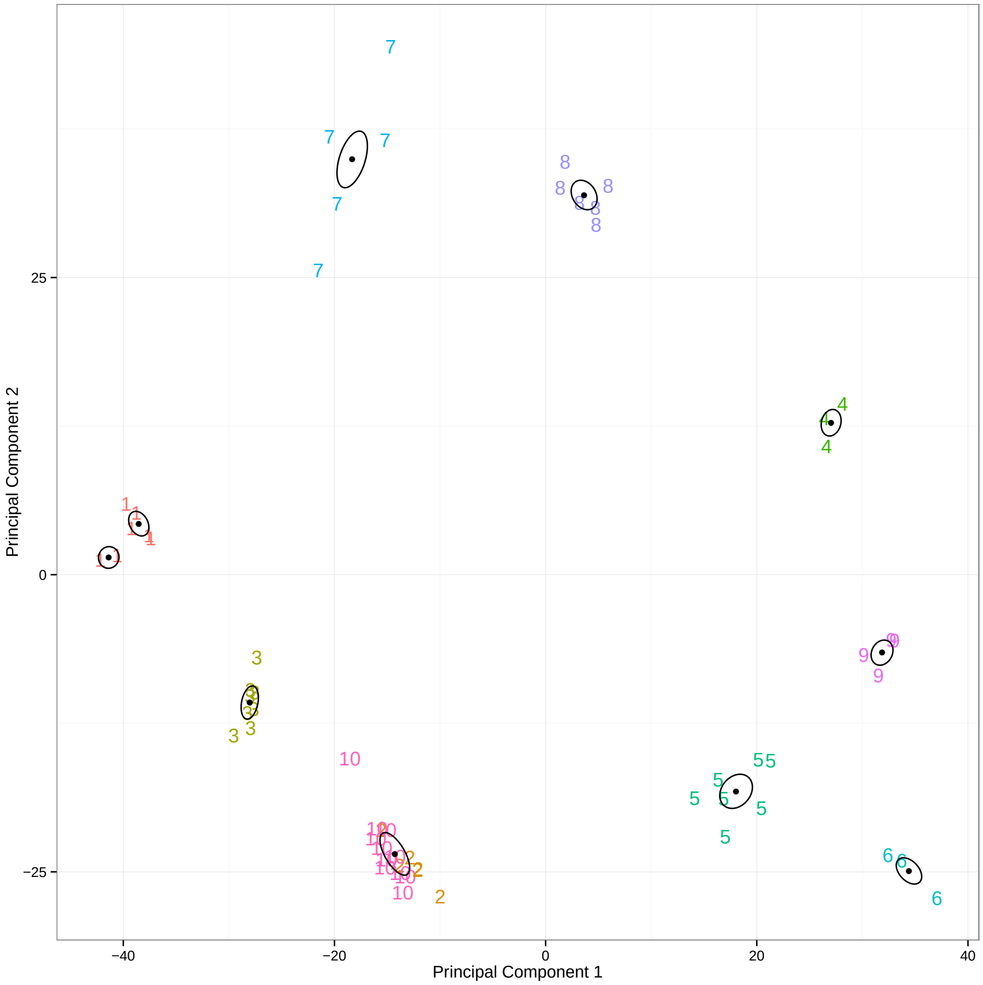


Figure 5: NECM clustering projected onto the first two principal components of the data. Ellipses depict the first two eigenvectors (and their corresponding eigenvalues) of the estimated covariances of each cluster.

7. Discussion

The initial insight of Stein (1975) has led to several methods for shrinkage estimation of a sample covariance matrix $\boldsymbol{S}$. These methods preserve the eigenvectors of $\boldsymbol{S}$ while pushing $\boldsymbol{S}$ towards a multiple of the identity matrix. Our Bayesian prior does precisely this in a nonlinear fashion. In our four examples it appears that exerting shrinkage is desirable, but the particular kind of shrinkage is immaterial. In its favor NECM has the advantage of making very few assumptions and requiring only simple calculations. Even in the covariance-regularization examples, where the data were generated using sparse precision matrices, there was relatively little loss in accuracy using NECM.

NECM does require a singular value decomposition (SVD). Although highly optimized routines for accurately computing the SVD are readily available, such calculations are not cheap. Randomized 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.

References

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

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

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

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

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

Daniels, M. J., Kass, R. E., 2001. Shrinkage estimators for covariance matrices. Biometrics 57 (4), 1173–1184.

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

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

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

Fraley, C., Raftery, A. E., Sep. 2007. Bayesian regularization for normal mixture estimation and model-based clustering. J. Classif. 24 (2), 155–181.

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

Friedman, J. H., 1989. Regularized discriminant analysis. Journal of the American Statistical Association 84 (405), 165–175.

González, I., Déjean, S., Martin, P. G. P., Baccini, A., 1 2008. CCA: An R package to extend canonical correlation analysis. Journal of Statistical Software 23 (12), 1–14.

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

Haff, L. R., 1991. The variational form of certain Bayes estimators. The Annals of Statistics 19 (3), 1163–1190.

Halko, N., Martinsson, P. G., Tropp, J. A., 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. 53 (2), 217–288.

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

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

Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

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

Hotelling, H., 1936. Relations between two sets of variants. Biometrika 28, 321–377.

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

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

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

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

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

Leurgans, S. E., Moyeed, R. A., Silverman, B. W., 1993. Canonical correlation analysis when the data are curves. Journal of the Royal Statistical Society. Series B (Methodological) 55 (3), 725–740.

Mahoney, M. W., Feb. 2011. Randomized algorithms for matrices and data. Found. Trends Mach. Learn. 3 (2), 123–224.

Marĉenko, V. A., Pastur, L. A., 1967. Distribution of eigenvalues for some sets of random matrices. Mathematics of the USSR-Sbornik 1 (4), 507–536.

Mardia, K. V., Kent, J. T., Bibby, J. M., 1979. Multivariate Analysis. Academic Press.

Martin, P. G. P., Guillou, H., Lasserre, F., Déjean, S., Lan, A., Pascussi, J.-M., SanCristobal, M., Legrand, P., Besse, P., Pineau, T., 2007. Novel aspects of PPARα-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. Hepatology 45 (3), 767–777.

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

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

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

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

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

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

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

Ruan, L., Yuan, M., Zou, H., Jun. 2011. Regularized parameter estimation in high-dimensional gaussian mixture models. Neural Comput. 23 (6), 1605–1622.

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

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

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58 (1), 267–288.

Vinod, H. D., May 1976. Canonical ridge and econometrics of joint production. Journal of Econometrics 4 (2), 147–166.

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

Witten, D. M., Tibshirani, R., 2009. Covariance-regularized regression and classification for high dimensional problems. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (3), 615–636.

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

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

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Ser. B 67 (2), 301–320.