Penalized Normal Likelihood and Ridge Regularization of Correlation and Covariance Matrices

David I. Warton

High dimensionality causes problems in various areas of statistics. A particular situation that rarely has been considered is the testing of hypotheses about multivariate regression models in which the dimension of the multivariate response is large. In this article a ridge regularization approach is proposed in which either the covariance or the correlation matrix is regularized to ensure nonsingularity irrespective of the dimensionality of the data. It is shown that the proposed approach can be derived through a penalized likelihood approach, which suggests cross-validation of the likelihood function as a natural approach for estimation of the ridge parameter. Useful properties of this likelihood estimator are derived, discussed, and demonstrated by simulation. For a class of test statistics commonly used in multivariate analysis, the proposed regularization approach is compared with some obvious alternative regularization approaches using generalized inverse and data reduction through principal components analysis. Essentially, the approaches considered differ in how they shrink eigenvalues of sample covariance and correlation matrices. This leads to predictable differences in power properties when comparing the use of different regularization approaches, as demonstrated by simulation. The proposed ridge approach has relatively good power compared with the alternatives considered. In particular, a generalized inverse is shown to perform poorly and cannot be recommended in practice. Finally, the proposed approach is used in analysis of data on macroinvertebrate biomasses that have been classified to species.

KEY WORDS: Cross-validation; Generalized inverse; High-dimensional data; Multivariate analysis; Multivariate regression.

1. INTRODUCTION

A challenging problem in statistical modeling is the estimation of and inference about parameters in multivariate regression, when the number of response variables is not small compared with the number of observations. Consider, therefore, a typical such situation, where the rows of an \( N \times p \) matrix \( \mathbf{Y} \) are assumed to be independent observation vectors from \( p \)-variate distributions with variance matrices that are constant and means that are to be modeled as a function of a set of \( N \)-vectors of explanatory variables \( \{x_1, x_2, \ldots, x_p\} \). Denote the \( N \times p \) mean matrix of \( \mathbf{Y} \) by \( \mu \), denote the constant variance matrix by \( \Sigma \), denote and let its correlation matrix be \( \mathbf{R} \). In this article we focus on the problem of testing some general linear hypothesis about parameters in the model for \( \mu \) when \( N \) is not large compared with \( p \).

High dimensionality poses two specific problems in this situation: First, large-\( N \)-fixed-\( p \) asymptotics fail (e.g., Anderson 2003). Second, the usual sample estimate of the variance matrix \( \hat{\Sigma} \) (denoted by \( \hat{\Sigma} \)) is unstable for large \( p \), becoming progressively closer to singularity as \( p \) approaches \( N \) (e.g., Ledoit and Wolf 2004). This becomes a problem in hypothesis testing, particularly when testing a hypothesis concerning the effect of some explanatory variable(s) on all response variables simultaneously. The standard multivariate test statistics are all functions of \( \hat{\Sigma}^{-1} \), so the test statistic will become unstable as \( p \) approaches \( N \) and will not be calculable for \( p > N \). Data sets for which \( p > N \) often arise in various subject areas; for example, in community ecology it is common to measure the abundance of many different species at relatively few sites and to study the relationship between abundance and the site environments at the community level, as discussed in Section 5.

A common approach to deal with singularity is regularization. Ridge regression (Hoerl and Kennard 1970) was one of the earliest methods to make use of this approach. Previous authors have proposed a range of regularization approaches for improving the behavior of \( \hat{\Sigma} \) when the number of variables is large. Bayesian (e.g., Brown, Fearn, and Haque 1999; Daniels and Kass 2001) and minimax approaches (Dey and Srinivasan 1985) have been proposed to regularize \( \hat{\Sigma} \) by shrinking sample eigenvalues. Ridge regularization also has been proposed, with the ridge parameter either estimated to minimize loss as measured using the Frobenius norm (Ledoit and Wolf 2004) or estimated using cross-validation to minimize some measure of prediction error (Friedman 1989; Krzanowski, Jonathan, McCarthy, and Thomas 1995; Hoffbeck and Longbrege 1996; Zhong, Zeng, Ma, Liu, and Zhu 2005).

In this article we pursue a ridge approach, specifically in the context of multivariate regression. Our approach is novel in that ridge regularization of \( \hat{\Sigma} \) or \( \hat{\mathbf{R}} \) is derived through penalized likelihood, which motivates the use of the likelihood as an objective function to maximize when estimating the ridge parameter. This naturally leads to the definition of stable likelihood-based test statistics.

Our methodology is related to that of Huang, Liu, Pourahmadi, and Liu (2006), who used a different penalty term to derive a penalized normal likelihood approach for covariance selection. Huang et al. (2006) suggested covariance selection by shrinking the off-diagonal elements of the Cholesky decomposition of \( \hat{\Sigma}^{-1} \) toward zero. In contrast, our choice of penalty leads to linear shrinkage of the eigenvalues of \( \hat{\Sigma} \) or \( \hat{\mathbf{R}} \). Note that our goal of regularization differs from that of Huang et al. (2006); here we want to find a stable, positive definite estimator of \( \Sigma \) to use for inference in multivariate regression.

In Section 2 we describe ridge regularization of the sample estimator of the correlation matrix and identify some important properties of this approach. In particular, we show that it can be derived using a penalized maximum likelihood approach. This result leads us to propose a new method of estimating...
the ridge parameter in Section 3, where we also describe some useful properties of this estimation approach. This method of estimation leads to the definition of suitable test statistics in high-dimensional situations. In Section 4 we explore the power properties of these statistics and compare them with those of some related alternative approaches for dealing with high dimensionality. Finally, in Section 5 we apply the new approach to abundance data from community ecology.

2. RIDGE ESTIMATION OF Σ OR R

Consider a multivariate regression model for \( Y \) with estimated means stored in \( \hat{\mu} \), where \( g \) parameters are used in estimating each column of \( \hat{\mu} \). The maximum likelihood estimator of the error variance matrix is

\[
\hat{\Sigma} = \frac{1}{N}(Y - \hat{\mu})^T(Y - \hat{\mu}).
\]

If \( p > N \), then this matrix will be singular. This problem can be resolved using ridge regularization,

\[
\hat{\Sigma}_c = \hat{\Sigma} + \kappa I,
\]

where \( I \) is the \( p \times p \) identity matrix, \( \kappa > 0 \) is the ridge parameter, and the subscript of \( \hat{\Sigma}_c \) indicates that it is a ridge or regularized estimator with ridge parameter \( \kappa \). Regularization of \( \hat{\Sigma} \) was discussed by Friedman (1989) in, for example, the context of discriminant analysis.

When the response variables have been measured on different scales, or when they have substantially different variability, it may be more appropriate to regularize on the standardized scale. In such a situation, the sample estimator \( \hat{R} \) of the correlation matrix could be regularized, rather than \( \hat{\Sigma} \). We write

\[
\hat{R} = \hat{\Sigma}^{-1/2}\hat{\Sigma}_d\hat{\Sigma}_d^{-1/2},
\]

where \( \hat{\Sigma}_d \) is the diagonal matrix formed from \( \hat{\Sigma} \) using the corresponding diagonal elements. Rather than using a regularized estimate of the form of (1) (i.e., \( \hat{R}_c = \hat{R} + \kappa I \)), we propose rescaling this matrix as

\[
\hat{R}_c = \hat{R} + (1 - \lambda)\lambda I,
\]

where \( \lambda = 1/(1+\kappa) \in (0, 1] \). We choose \( \hat{R}_c \) as our ridge-regularized estimator because it is a correlation matrix—that is, it has diagonal elements equal to unity, unlike \( \hat{R}_c \). Being proportional to \( \hat{R}_c \) and \( \hat{R}_c \) inherits the property of being nonsingular for \( \lambda \in (0, 1) \).

Unlike Hastie, Buja, and Tibshirani (1995) and Breiman and Friedman (1997), we use regularization not to improve the parameter estimates fitted in the multivariate regression model (\( \hat{\beta} \)), but rather as a way to ensure nonsingularity of the variance matrix (\( \hat{\Sigma} \)). Under the method of this article, estimates of \( \beta \) are not a function of the regularization parameter \( \lambda \), and the role of regularization is in inference about (not estimation of) regression parameters.

2.1 Properties of the Estimator

The following properties follow directly either from the definition or from previously established results:

- \( \hat{R}_c \) can be considered a shrinkage estimator in which \( \hat{R} \) has been shrunk toward \( I \). Thus \( \hat{R}_c \) has the same eigenvectors as \( \hat{R} \), and the eigenvalues of \( \hat{R} \) have been shrunk toward 1; that is, the diagonal matrix of eigenvalues \( \lambda \) has been replaced by \( \lambda\lambda + (1 - \lambda)I \). These properties have been noticed by others who used \( \Sigma_c \) in place of \( \Sigma \) (Friedman 1989; Brown et al. 1999), and they can be exploited to reduce the computational time for regularized discriminant analysis in high-dimensional settings (Hastie and Tibshirani 2004).
- \( \hat{R}_c \) can be calculated by multiplying all off-diagonal elements of \( \hat{R} \) by \( \lambda \).
- The proposed regularization will reduce the bias inherent in eigenvalues of sample correlation matrices (Friedman 1989), provided that a reasonable choice of regularization parameter \( \lambda \) is used. Bias generally will not be completely removed.
- \( \hat{R}_c \) manages a bias–variance trade-off; \( \hat{R} \) has potentially high variance but relatively little bias (being asymptotically unbiased for large \( N \), fixed \( p \)), and \( I \) has no variance, but is biased for \( \hat{R} \) in general. Thus, loosely speaking, \( \hat{R}_c \) can be thought of as a weighted average of a variance term and a bias term, and we aim to choose \( \lambda \) to optimize this bias–variance trade-off. Ledoit and Wolf (2004) previously made this point for \( \hat{\Sigma}_c \).

An additional property provides our central new result, namely that \( \hat{R}_c \) can be derived as the penalized likelihood estimate of \( R \) for multivariate normal data, with a penalty term proportional to \( \text{tr}(R^{-1}) \). A penalty term of this form penalizes estimates of \( R \) that are ill-conditioned and approach singularity, because in such cases the diagonal elements of \( R^{-1} \) approach \( \infty \).

**Theorem 1.** Consider an \( N \times p \) matrix \( Y \) of multivariate normal data with mean \( \mu \) and covariance matrix \( \Sigma = \Sigma_d^{-1/2} \Sigma_d^{-1/2} \), where \( \Sigma_d \) is the diagonal matrix formed from \( \Sigma \). Let \( \hat{R} \) be the maximum likelihood estimator of the correlation matrix \( \Sigma \).

The ridge estimator of the correlation matrix, \( \hat{R}_c = \hat{R} + (1 - \lambda)\lambda I \), is the maximum penalized likelihood estimator with penalty term proportional to \( -\text{tr}(R^{-1}) \).

The proof (indeed all proofs) can be found in the Appendix. Previous authors have applied ridge regularization to \( \hat{\Sigma} \) rather than to \( \hat{R} \) (Friedman 1989; Hoffbeck and Landgrebe 1996), and so we note that the link to penalized likelihood applies in this setting as well, with the proof following a similar method.

**Corollary 1.** The ridge estimator of the variance matrix, \( \hat{\Sigma}_c = \hat{\Sigma} + \kappa I \), is the maximum penalized likelihood estimator with penalty term proportional to \( -\text{tr}(\Sigma^{-1}) \).

3. ESTIMATION OF \( \lambda \)

We propose as an estimator of the regularization parameter the maximizer of the normal likelihood, as estimated by \( K \)-fold cross-validation (Hastie, Tibshirani, and Friedman 2001). We first split the data into \( K \) roughly equal parts \( Y_k^T = [Y_k^T, \ldots, Y_k^T] \), with \( N_k \) observations in the \( k \)th group. \( Y_k^T \) is reserved as the validation data for the \( k \)th group, and all other observations from \( Y \) are used as the training data to
estimate \( \hat{\mu}_k \), \( \hat{R}_k \), and \( \hat{\Sigma}_k \), where \( \hat{\mu}_k \) is the \( N_k \times p \) matrix of means for \( Y_k \), estimated using the multivariate regression model. The observed likelihood for \( Y_k \), \( L(\hat{\mu}_k, \hat{\Sigma}_k; Y_k) \), then can be calculated for each \( k \). We then estimate \( \lambda \) to maximize the cross-validated likelihood function,

\[
\hat{\lambda} = \arg \max_{\lambda} \sum_{k=1}^{K} \log L(\hat{\mu}_k, \hat{\Sigma}_k; Y_k),
\]

where

\[
-2 \log L(\hat{\mu}_k, \hat{\Sigma}_k; Y_k) = \sum_{k=1}^{K} \log \left| \hat{\Sigma}_k \right| + \text{tr} \left[ (Y_k - \hat{\mu}_k) (\hat{\Sigma}_k)^{-1} (Y_k - \hat{\mu}_k)^T \right].
\]

This is a natural approach to use for estimating \( \lambda \), because cross-validation is commonly used to estimate regularization parameters, and the likelihood function is an obvious choice of objective function given the link between normal likelihood and ridge regularization offered by Theorem 1. This approach is closely related to that of Friedman (1989) and Hoffbeck and Landgrebe (1996), who regularized \( \Sigma \) using ridge regularization for discriminant analysis. Both used cross-validation to estimate the ridge parameter; Friedman (1989) proposed using the discriminant scores as the objective function, and Hoffbeck and Landgrebe (1996) proposed using the log-likelihood function, although without noting the relationship to penalized likelihood. For discriminant analysis with equal sample sizes in all groups, the objective function of Friedman (1989) is equivalent to that in (4).

Estimation of \( \hat{\lambda} \) is relatively straightforward through maximization of \( \sum_{k=1}^{K} \log L(\hat{\mu}_k, \hat{\Sigma}_k; Y_k) \) over the interval \( \lambda \in [0, 1] \) using numerical optimization. Experience with this function to date has always yielded either a single stationary point and global maximum or a monotonic increasing function maximized at the endpoint \( \lambda = 1 \). Nevertheless, the possibility of multiple maxima has not been disproved, so one should guard against this possibility by, for example, using multiple starting points for optimization algorithms or a diagnostic plot of \( \sum_{k=1}^{K} \log L(\hat{\mu}_k, \hat{\Sigma}_k; Y_k) \) against \( \lambda \in [0, 1] \).

3.1 Properties of \( \hat{\lambda} \)

The estimator \( \hat{\lambda} \) has some useful properties. In particular, we can show the following:

- For fixed \( p \), \( \hat{\lambda} \) as \( N \to \infty \); that is, \( \hat{\lambda} \) is consistent for \( \lambda \) (see Thm. 2).
- \( \hat{\lambda} \) has full rank, but \( \hat{\Sigma} \) does not, then there is probability zero that \( \hat{\lambda} = 1 \) (see Thm. 3). Thus we can guarantee that \( \hat{\lambda} \) is positive definite if \( \hat{\Sigma} \) is.
- As bias in sample eigenvalues of \( \hat{\Sigma} \) increases, \( \hat{\lambda} \) becomes progressively more unstable, and a greater degree of regularization is required.

Although we have demonstrated that \( \hat{\lambda} \) is consistent in the large-\( N \)-small-\( p \) setting, we do not expect consistency in high-dimensional settings where \( p \to \infty \). This is because sample eigenvalues are known to be inconsistent when \( p \to \infty \) (Bai and Saranadasa 1996); thus from Theorem 4, we expect \( \hat{\lambda} < 1 \).

**Theorem 2.** Consider increasing sample sizes such that as \( N \to \infty, N_k \to \infty \) \( \forall k \). If \( \hat{\mu}_k \to \mu \) and \( \hat{\Sigma}_k \to \Sigma \), then

\[
\hat{\lambda} \to 1.
\]

Similarly, strong consistency of \( \hat{\mu} \) and \( \hat{\Sigma} \) implies strong consistency of \( \hat{\lambda} \). The condition that \( N_k \to \infty \) \( \forall k \) is satisfied for \( K \)-fold cross-validation.

**Theorem 3.** If \( Y = (y_1^T, y_2^T, \ldots, y_N^T)^T \), then \( y_i \) is absolutely continuous (with respect to Lebesgue measure) and \( \text{rank}(\text{var}(y_i)) = p \).

**Theorem 4.** Stationary points of the observed likelihood estimated by \( K \)-fold cross-validation, \( \sum_{k=1}^{K} \log L(\hat{\mu}_k, \hat{\Sigma}_k; Y_k) \), have the form \( \hat{\lambda} = (\hat{\lambda} + 1)^{-1} \) where \( \hat{\lambda} \) satisfies

\[
\hat{\lambda} = \left[ \sum_{k=1}^{K} \sum_{j=1}^{p} w_{jk} \right]^{-1} \sum_{k=1}^{K} \sum_{j=1}^{p} w_{jk} (\hat{Z}_k\hat{Z}_k^T/N_k - \hat{\Lambda}^\dagger)_{jj}.
\]

(6)

\[
w_{jk} = N_k (\hat{\Lambda}^\dagger_{jj} + \hat{\lambda})^{-1},
\]

\( \hat{\Lambda}^\dagger \) has spectral decomposition \( \hat{\Lambda}^\dagger = \hat{P}^\dagger \hat{A}^\dagger (\hat{P}^\dagger)^T \), and

\[
\hat{Z}_k = (Y_k - \hat{\mu}_k^\dagger) (\hat{\Sigma}_k^\dagger)^{-1/2}.
\]

(7)
The proof has been omitted. It involves expressing \( L(\hat{\mu}_k, \hat{\Sigma}^{\hat{\lambda}}; Y_k) \) in terms of \((\hat{Z}_k^T \hat{Z}_k)_{jj}, \hat{\lambda}_{jj} \) and \( \kappa \), then solving for the stationary point.

Theorem 4 suggests that the amount of shrinkage that can be expected is closely related to the amount of bias in sample eigenvalues. To see this, note that the \( j \)th sample eigenvalue \( \hat{\lambda}_{jj} \) can be interpreted as the sample variance of training data scores along an axis defined by the \( j \)th eigenvector of \( \hat{R}^{\hat{\lambda}} \). Now \((\hat{Z}_k^T \hat{Z}_k)_{jj}/N_k \) is an out-of-sample estimate of this same quantity, and \( \hat{\kappa} \) is a weighted-average difference of these training and validation data estimates—a measure of the overall bias in sample eigenvalues as estimated through cross-validation. If the eigenvalue estimates from training and validation data are similar, then \( \hat{\kappa} \) is near 0 (and thus \( \hat{\lambda} \) is near unity); that is, if cross-validation suggests little bias in sample eigenvalues, then there will be little shrinkage. Conversely, when there are substantial discrepancies between training and validation estimates of sample eigenvalues, we would expect \( \hat{\kappa} \) to be large and positive. We do not expect \( \hat{\kappa} \) to be negative in such instances, because the \( w_{jk} \)'s give greater weighting to the smaller, downward-biased sample eigenvalues.

3.2 Simulation: Sample Properties of \( \hat{\lambda} \)

We conduct a small simulation study to explore the sample properties of \( \hat{\lambda} \). Each simulated data set is an iid multivariate normal sample. An orthogonal \( 3 \times 3 \times 2 \times 3 \) design is used: \( N \in \{20, 80, 320\}, p \in \{5, 20, 80\}, R \) with either a first-order autoregressive correlation structure [\( \text{AR}(1), R_{ij} = \rho^{\left| i-j \right|} \)] or an exchangeable correlation structure (i.e., \( R_{ij} = \rho \)), and \( \rho \in \{.3, .5, .7\} \). Without loss of generality, \( \mu = 0 \) and \( \Sigma_d = I \).

For each simulation, we generate 100 sample data sets, and for each set we estimate \( \hat{\lambda} \) by cross-validation using the likelihood function.

Results for \( N \) and \( p \) are consistent across different choices of \( R \), so we show the results only when \( R \) has an AR(1) structure with \( \rho = .5 \) (Fig. 1), and note the following:

- As \( N \) increases, \( \hat{\lambda} \) increases toward 1 (as in Thm. 2).
- As \( p \) increases, \( \hat{\lambda} \) decreases (as suggested by Thm. 4).
- \( \text{var}(\hat{\lambda}) \) tends to decrease as \( N \) and \( p \) increase.

These are all desirable properties for an estimator of \( \lambda \). As more information becomes available about sample eigenvalues (i.e., as \( N \) increases and \( p \) decreases), \( \hat{\lambda} \) should indeed increase toward 1. Moreover, as more data become available for estimation of \( \hat{\lambda} \) (i.e., as \( N \) and \( p \) increase), \( \text{var}(\hat{\lambda}) \) indeed should decrease.

Although results are not presented for simulations that vary correlation structure and strength of correlation, we note that \( \hat{\lambda} \) tends to increase as the true eigenvalues \( A_{jj} \) become more variable (i.e., as \( p \) increases), as observed by Ledoit and Wolf (2004) for their estimator.

4. POWER OF HIGH–DIMENSIONAL TEST STATISTICS

Consider now a test statistic \( T(Y) \), based on the \( N \times p \) matrix of observations \( Y \). Many such test statistics involve \( R \) in some way, so adjustments will be needed when the dimensionality is high. We identify three competing types of adjustment:

- \( T_2(Y) \), the “ridge regularized test statistic,” where \( \hat{R}_k \) is used in calculating the test statistic in place of \( R_k \), as advocated earlier.
- \( T(Z_k) \), the “principal components test statistic,” constructed from \( Z_k \), the scores along the first \( s \) principal components of \( \hat{R}_k \).
- \( T(Y) \), the “generalized inverse test statistic,” in which a Moore–Penrose generalized inverse is used in place of \( R_k^{-1} \) in the test statistic, as suggested by various authors (e.g., Breiman and Friedman 1997).

Here \( T(Z_k) \) and \( T(Y) \) are considered competitors to \( T_2(Y) \), because they can be considered alternative methods of regularizing \( R_k \) to construct a test statistic. Note that although principal components analysis is more commonly applied to the sample correlation matrix of \( Y \), we apply it to the sample correlation matrix of residuals \( \hat{R}_k \), given our focus on approaches that regularize \( \hat{R}_k \).

We specifically consider the situation in which we test a hypothesis about parameters in the mean model, and the test statistic is \( T(Y) = g(\Omega_2 \Sigma_d^{-1/2} \hat{R}_k^{-1/2}) \) for some function \( g(\cdot) \) and some matrix of effects \( \Omega_2 \). \( T(Y) \) is a class of test statistics that includes classical multivariate test statistics.

In particular, if \( \Omega_2 = (B/N)^{1/2} \), where \( B \) is the “between groups matrix,” then Wilk’s statistic is \( |N \hat{\Sigma}/N \hat{\Sigma} + B| = (I + \hat{R}_k^{-1/2} \hat{\Sigma}_d^{-1/2} \hat{\Sigma}_d^{-1/2} \hat{R}_k^{-1/2})^{-1} \).

Theorem 5. Consider a test statistic based on \( Y \), which can be expressed as \( T(Y) = g(\Omega_2 \Sigma_d^{-1/2} \hat{R}_k^{-1/2}) \), for some function \( g(\cdot) \) and some matrix of effects \( \Omega_2 \). Let \( \hat{R} \) have spectral decomposition \( \hat{R} = P \hat{\Sigma} \hat{P}^T \) with eigenvalues in decreasing order. Then the following results hold:

a. The test statistic \( T(Y) \) can be expressed as

\[
T(Y) = g(\Omega_2 \text{diag}[\hat{\lambda}_{jj}^{-1/2}, 1 \leq j \leq p] \hat{P}^T),
\]

where \( \Omega_2 = \Omega_2 \Sigma_d^{-1/2} \hat{P} \) and diag(\( v \)) is a diagonal matrix formed from \( v \).

b. The ridge-regularized test statistic can be expressed as

\[
T_2(Y) = g(\Omega_2 \text{diag}[(\hat{\lambda}_{jj} + 1 - \lambda)^{-1/2}, 1 \leq j \leq p] \hat{P}^T).
\]

c. If the test statistic is rotation-invariant and scale-invariant, then

\[
T(Z_k) = g(\Omega_2 \text{diag}[(\hat{\lambda}_{jj} - 1)^{-1/2}, 1 \leq j \leq s] \hat{P}^T),
\]

where \( \hat{P} \) is an \( a \times b \) matrix of 0’s.

d. If rank(\( \hat{R}_k \)) = \( q \), then the generalized inverse test statistic can be expressed as

\[
T_g(Y) = g(\Omega_2 \text{diag}[(\hat{\lambda}_{jj} - 1)^{-1/2}, 1 \leq j \leq \min\{q, s\}] \hat{P}^T).
\]

This theorem shows the ramifications of regularizing \( \hat{R} \) for hypothesis testing in high-dimensional multivariate regression. All of the test statistics are functions of \( \Omega_2 \), the effects as expressed along the eigenvectors of \( \hat{R} \). The only difference is in the relative weightings given to the different eigenvectors (i.e., to the different columns of \( \Omega_2 \)), through modification of the \( \hat{\lambda}_{jj} \), which has several implications for power, as discussed next.
$T_\lambda(Y)$. In ridge regularization, the $\hat{A}_{jj}$ are shrunk toward 1. This increases the weighting given to scores along eigenvectors for which $\hat{A}_{jj} > 1$, and decreases the weighting given to eigenvector scores for which $\hat{A}_{jj} < 1$. If the true effect is expressed along eigenvectors for which $\hat{A}_{jj} > 1$ [i.e., if $\hat{A}_{jj} < 1 \implies (\Omega_2)_{jj} \approx 0$], then power is increased by ridge regularization. Conversely, power is decreased if the true effect is expressed along eigenvectors for which $\hat{A}_{jj} < 1$.

$T(Z_s)$. A principal components approach involves replacing $\hat{A}_{jj}^{-1}$ with 0 for $j > s$. Thus scores along the last $p-s$ eigenvectors are ignored, and $T(Z_s)$ has no ability to detect effects expressed along these eigenvectors. But $T(Z_s)$ will have relatively good power if effects are expressed along the first $s$ eigenvectors, because in this scenario data reduction has not lost information about the nature of effects.

$T_\lambda(Y)$. In taking a generalized inverse, $\hat{A}_{jj}^{-1}$ is replaced by 0 for $j > \text{rank}(R) = q$. $T_\lambda(Y)$ will behave like $T(Z_s)$, where $s = q$; that is, $T_\lambda(Y)$ will have good power when effects are expressed along the first $q$ eigenvectors but poor power otherwise.

For all of the foregoing regularizations of $T(Y)$, to various extents, power can be expected to increase when effects are expressed along dominant eigenvectors and to decrease in other cases. $T(Z_s)$ and $T_\lambda(Y)$ reduce the dimensionality of the problem, and in doing so gamble that the effects of interest are not lost in dimension reduction. How much of a gamble this is depends on the extent of dimension reduction, and the extent is necessarily considerable if $p \gg N$. In contrast, $T_\lambda(Y)$ bet-hedges; eigenvalues are shrunk toward unity, so dominant eigenvectors receive greater attention, but some attention is also given to other dimensions.

Figure 2 demonstrates schematically the implications of Theorem 5 for the power of different test statistics under two important scenarios: when effects are expressed along a dominant eigenvector and when they are not.

### 4.1 Power Simulation: Design

The implications of Theorem 5 for power properties of the different tests is demonstrated by simulation. In each simulation, we generate two iid multivariate normal samples each of size 10, with mean vectors $\mu_1 = 0, \mu_2 \neq \mu_1$. We test the null hypothesis $H_0: \mu_1 = \mu_2$ using the log-likelihood ratio statistic $[-2 \log(\Lambda)$, where $\Lambda$ is Wilk’s statistic], with the following adjustments for high dimensionality:

$T_\lambda(Y)$, ridge-regularized statistic using $\hat{\lambda}$. The ridge parameter is estimated using cross-validation of the likelihood function under the alternative hypothesis.

$T_\lambda(Y)$, ridge-regularized statistic using fixed $\lambda$. The ridge parameter is fixed at a value chosen a priori, $\lambda \in \{0, .25, .5, .75\}$.

$T(Z_s)$, $T_\lambda(Y)$.

Figure 2. Schematic diagram illustrating in two dimensions the effects on a test statistic $T(Y)$ of high-dimensionality adjustments, for the comparison of the means of two samples. Two situations are considered: the means differ either (a) along the first eigenvector or (b) along the second eigenvector of $R$. Ridge regularization, $T_\lambda(Y)$, shrinks eigenvalues toward 1, reducing the eccentricity of ellipses. Thus in (a), the ellipses appear further apart for $T_\lambda(Y)$ than for $T(Y)$, because the difference between means is larger relative to error along the first eigenvector. Thus power is improved by regularization in (a). In (b), the ellipses are brought closer together by regularization, so power is reduced. A principal components $T(Z_s)$ or generalized inverse approach $T_\lambda(Y)$ involves dimensionality reduction, visualized in the third column as reducing the dimensionality from two to one! In this situation, the means are projected onto the first eigenvector, and thus the mean shift remains in (a) but is completely removed in (b).
The log-likelihood ratio statistic is computed from $Z_q$, where $\text{rank (R)} = q$. Note that if $\hat{R}$ has full rank, then $T_-(Y) = T(Y)$; that is, no correction is applied.

To allow for difficulties in making inferences about Wilk’s statistic when $N$ is not large compared with $p$, we use permutation tests to estimate $p$ values, based on 999 permutations of group membership labels for the $N$ observations, after adjusting the two samples to have equal sample means. This provides a Monte Carlo approximation to the exact test of $H_0: \mu_1 = \mu_2$ when both samples are generated from the same $p$-variate distribution (Edgington 1995).

Simulations experiment with four quantities: $p \in [2, 30]; R$ is either AR(1) or exchangeable, as previously; $\rho \in [.3, .5, .7]$; and three types of alternative to $H_0$ are considered, in which $\mu_2$ is shifted:

1. Along the first eigenvector of $R$ by $a \sqrt{p} \Lambda_{jj}^{1/2}$ units
2. Along the last eigenvector of $R$ by $\frac{a}{\sqrt{p}} \Lambda_{jj}^{1/2}$ units
3. Along all eigenvectors: $a \Lambda_{jj}^{1/2}$ units along the $j$th eigenvector of $R$.

We use $a = .6$ to obtain intermediate power for most of the statistics and simulation conditions.

4.2 Power Simulation: Results

Patterns in power simulations are consistent across different choices of $R$, so we show results only for when $R$ has an AR(1) structure with $\rho = .5$ (Fig. 3). Results of power simulations demonstrate the various consequences of Theorem 5 for power properties of high-dimensional test statistics, particularly the following:

- Theorem 5 suggests that $T_2(Y)$ and $T(Z_1)$ would have higher power than $T(Y)$ for shift 1 but lower power for shift 2, especially for $T(Z_2)$. $T_-(Y) = T(Y)$ for $p < N$, and Figure 3 shows that the predicted patterns are observed for $p < 10$.
- $T(Z_1)$ and $T_-(Y)$ were expected to perform progressively worse with increasing $p$, and they perform especially poorly for $p > N$ in all simulations (Fig. 3). The effect of interest (the mean shift) is one-dimensional in simulations, but the proportion of dimensions excluded from $T(Z_1)$ and $T_-(Y)$ increases with $p$, so these “hit-or-miss” approaches have a progressively greater chance of “missing” as $p$ increases.
- For $T_2(Y)$, we would expect from Theorem 5 that as $\lambda$ decreases toward 0 (i.e., as the amount of regularization increases), power would increase for shift 1 and decrease for shift 2. This is precisely what we observe, with the rank order of $T_2(Y)$ being always as expected for $\lambda \in [0, .25, .5, .75]$.

We also note that $T_2(Y)$ behaves like $T_1(Y)$, where $\lambda \approx 1$ when $N$ is large and $p$ is small, but $\lambda$ decreases as $p$ increases relative

![Figure 3](image-url)
to $N$. This was expected given the properties of $\hat{\lambda}$ described in Section 3.1.

One conspicuous pattern in Figure 3 is the sharp decrease in power for $T_{-}(Y) [= T(Y)]$ when $10 < p < 20$, followed by a “kink” in the curve at $p \approx 20$. The decrease in power as $p \to N$ is well known and attributable to increasing instability in $\hat{\Sigma}$ (Bai and Saranadasa 1996). Note that there is no evidence of a similar trend for ridge-regularized statistics. The “kink” occurs because of the nature of the permutation testing approach used here, in which the rank of resampled data is $N - 1$, not $N$, for $T_{-}(Y)$.

5. PRACTICAL APPLICATION

Here the proposed method is applied to a data set from Pearson and Blackstock (1984), reanalyzed previously by Clarke and Ainsworth (1993) and Clarke, Somerfield, and Chapman (2006). This data set contains the biomasses of benthic macroinvertebrates, classified to species, from the Garroch Head dumping ground. Samples were collected at 12 different locations, individuals were sorted into species, and the abundance of each species was measured by calculating total biomass. Environmental variables also were collected at each location; for the purposes of this analysis, we consider two of them: log-transformed lead concentrations [log(Pb)] and depth.

The purpose of analysis is to identify whether there is a relationship between community macrobenthic abundance and environmental variables. The key feature of this data set that makes it interesting for the purposes of this article is that the number of variables is larger than the number of observations ($N = 12$, $p = 23$). This is a typical property of this type of data (e.g., Warton and Hudson 2004).

Table 1 presents a subset of the data, in which some important properties of the data can be seen. Note that there are many 0’s in the data, because not every species is observed in every site. A second feature apparent in Table 1 is that biomass appears to be strongly right-skewed (with the occasional very large biomass value).

We apply a log($y + 1$) transformation to biomass values and fit a homoscedastic least squares model. Although this approach has been shown to do a surprisingly good job of modeling the mean–variance relationship of this type of data set (Warton 2005), it nevertheless is a rather crude modeling approach and thus can be considered a first attempt that may be improved on in later work.

Therefore, we fit a linear model of the form

$$\mu = \beta_0 + \log(\text{Pb})\beta_1 + \text{depth}\beta_2$$

(8)

and use a likelihood statistic to test the hypothesis

$$H_0 : \beta_1 = \beta_2 = 0, \quad H_a : \text{otherwise}.$$ 

Because $N$ is small, $\hat{\lambda}$ is estimated through leave-one-out cross-validation using (4).

When (8) is fitted, the likelihood function is maximized at $\hat{\lambda} \approx .32$. Thus a considerable amount of shrinkage is applied, as might be expected because $p \approx 2N$. The likelihood expression from (4) appears to be a smooth function of $\lambda$ with a single maximum (Fig. 4), as appears to be the norm in our experience.

The ridge-regularized likelihood ratio statistic is $T_{32}(Y) \approx 124.9$. The $p$ value, evaluated using a permutation test with 999 permutations of the rows of the data, is $P \approx .017$. Thus we can conclude that there is some evidence of a relationship between log(Pb), depth, and macroinvertebrate abundance.

For the purposes of comparison, consider the alternative high-dimensional approaches described in Section 4.1. Using a generalized inverse, $T_{-}(Y) \approx 15.0$, and $P \approx .22$, as estimated by permutation. Using a principal components approach, $T(Z) \approx 14.6$ and $P \approx .95!$ Neither test statistic detects the patterns in macroinvertebrate abundance deemed to be significant at the .05 level by the ridge-regularized statistic.

The obvious next step in the analysis of this data set is to explore the nature of the species–environment relationship. This might involve, for example, the use of graphical techniques such as small multiples (Tufte 1983) and multiple testing to determine the species in which responses to environmental variables are expressed. Such analyses, although interesting in their own right, are beyond the scope of this article.
6. DISCUSSION

In the context of chemometrics, where the interest lies in predicting a single response variable from many explanatory variables, Frank and Friedman (1993) have compared principal components regression and partial least squares with ridge regression. Our results described in Theorem 5 and Figure 2 are analogous to those of Frank and Friedman (1993), although in our case there are many response variables rather than many explanatory variables. This means that we apply shrinkage to the correlation or covariance matrix of \( Y \) rather than to the covariance matrix of \( X \). Frank and Friedman (1993) noted that the advantage of shrinkage regularization of \( \beta \) diminished quickly as \( p \) increased, due to uncertainties in estimation of \( P \) and \( A \). Thus for the type of application considered in Section 5, shrinkage might offer little assistance, although this is certainly of interest in further investigations.

APPENDIX: PROOFS

Proof of Theorem 1

The log-likelihood function for \( Y \) with penalty term \(-c \text{tr}(R^{-1})/2\) can be written as

\[
\log L(\mu, \Sigma; Y) = -\frac{Np}{2} \log(2\pi) - \frac{N}{2} \log |\Sigma| - \frac{1}{2} \text{tr}((Y - \mu)^T \Sigma^{-1} (Y - \mu)) - \frac{c}{2} \text{tr}(R^{-1}).
\]

Letting \( S = (Y - \hat{\mu})^T (Y - \hat{\mu})/N \), we can express \(-2 \log(L)\) as

\[
-2 \log L(\mu, \Sigma; Y) = Np \log(2\pi) + N \log |\Sigma| + N \text{tr}(\Sigma \Sigma^{-1}) + c \text{tr}(R^{-1})
\]

\[
= Np \log(2\pi) + N \log |\Sigma| + N \log |\Sigma_d| + N \log |R|
\]

\[
+ N \text{tr} \left[ \left( \Sigma_d^{-1/2} S \Sigma_d^{-1/2} + \frac{c}{N} I \right) R^{-1} \right]
\]

\[
= Np \log(2\pi) + N \log |\Sigma_d K_1| + N \log |K_1^{-1} R|
\]

\[
+ N \text{tr}[K_1 R^{-1}], \tag{A.2}
\]

where \( K_1 = \Sigma_d^{-1/2} S \Sigma_d^{-1/2} + I c / N \).

With respect to \( R \), (A.2) has a unique minimum when \( K_1 R^{-1} = I \), that is, when

\[
\hat{R}_d = \left( 1 - \frac{c}{N} \right)^{-1} S_d. \tag{A.3}
\]

The restriction \( c \in [0, N) \) in the theorem ensures that \( \hat{S}_d \) is defined.

Thus

\[
\hat{R}_k = \left( 1 - \frac{c}{N} \right) R + \frac{c}{N} I = \lambda R + (1 - \lambda) I,
\]

where \( \lambda = 1 - c / N \) and \( \lambda \in (0, 1] \).

Proof of Theorem 2

Differentiating the likelihood expression of (5) with respect to \( \kappa = 1/\lambda - 1 \), we obtain

\[
\frac{\partial}{\partial \kappa} (-2 \log L(\hat{\mu}_k, \hat{\Sigma}_k; Y_k)) = \sum_{j=1}^p (\hat{\lambda}_k)^{-1} \left[ \kappa N_k - (\hat{Z}_k^T \hat{Z}_k)_{jj} + N_k \hat{\lambda}_k \hat{\lambda}_k \right], \tag{A.4}
\]

where \( \hat{Z}_k \) is as defined in (7). As \( N \to \infty \), \( \hat{\mu}_k \xrightarrow{P} \mu_k \), \( \hat{R}_k \xrightarrow{P} R \), \( \hat{\Sigma}_k \xrightarrow{P} \Sigma_d \), and

\[
\frac{1}{N_k} (Y_k - \hat{\mu}_k)^T (Y_k - \hat{\mu}_k) \xrightarrow{P} \Sigma
\]

\[
\frac{1}{N_k} \hat{Z}_k^T \hat{Z}_k \xrightarrow{P} P^T \Sigma_d^{-1/2} \Sigma \Sigma_d^{-1/2} P = P^T R P = \Lambda.
\]

where \( R = P A P^T \) is the spectral decomposition of \( R \).
Applying these convergence results to (A.4), ∀k, we have
\[ \frac{1}{N_k} \frac{\partial}{\partial \tilde{\mu}_k} \left( -2 \log L(\tilde{\mu}_k^k, \tilde{\Sigma}_k^k; Y_k) \right) \to \frac{1}{p} \sum_{j=1}^{p} \left( A_{jj} + k \right)^2 \cdot \frac{\partial}{\partial \tilde{\mu}_k} \left( -2 \log L(\tilde{\mu}_k^k, \tilde{\Sigma}_k^k; Y_k) \right). \]

This is positive for \( k > 0 \). If \( R \) has full rank, then \( A_{jj} > 0 \) ∀j; thus \( \sum_{k=1}^{K} -2 \log L(\tilde{\mu}_k^k, \tilde{\Sigma}_k^k; Y_k) \) exists and has a unique minimum at \( k = 0 \). So \( \tilde{\lambda} \to \lambda \), and thus \( \lambda \to 1 \).

Proof of Theorem 3

The following lemma is required to prove Theorem 3.

Lemma A.1. For any p-variate sample \( Y = (y_1^T, y_2^T, \ldots, y_n^T)^T \),
\[ \hat{\lambda} = 1 \implies \text{colspace}(Y_k) \subseteq \text{colspace}(Y_k^k - \hat{\mu}_k). \]

Proof. Let \( \tilde{R}^k = \tilde{P}^k \tilde{A}^k (\tilde{P}^k)^T \) be the spectral decomposition of \( \tilde{R}^k \), with eigenvalues in decreasing order. It can be shown that \( \tilde{R}^k = \tilde{P}^k (\hat{\lambda}_k \tilde{X}^k + (1 - \lambda_k) I) (\tilde{P}^k)^T \).

Now if we let \( Z_k = (Y_k - \hat{\mu}_k^k) (\tilde{X}^k)^T - (\tilde{P}^k)^T \), then we can write (5) as
\[ -2 \log L(\hat{\mu}_k, \hat{\Sigma}_k; Y_k) = C_2 + N_k \log(\lambda_k \lambda + (1 - \lambda_k) I)
+ tr[Z_k (\hat{\lambda}_k \tilde{X}^k + (1 - \lambda_k) I)^{-1} \tilde{Z}_k^T], \]

where \( C_2 \) is a constant with respect to \( \lambda_k \),
\[ = C_2 + \sum_{j=1}^p N_k \log(\lambda_k \lambda + (1 - \lambda_k) I)
+ \sum_{j=1}^p \frac{|(Z_k^j)^T Z_k^j|}{[(\lambda_k \lambda + (1 - \lambda_k) I)^{1/2}]} \cdot (Z_k^j)^T \tilde{Z}_k^j. \]

Because \( (\lambda_k \lambda + (1 - \lambda_k) I)^{1/2} \geq 0 \) and \( |(Z_k^j)^T Z_k^j| \), \( j \leq p \), \( -2 \log L(\hat{\mu}_k, \hat{\Sigma}_k; Y_k) \) is continuous for \( 0 \leq \lambda_k < 1 \).

If \( \forall k \) such that \( \text{rank}(\tilde{R}^k) = q < p \), then \( (\lambda_k \lambda + (1 - \lambda_k) I)^{1/2} \) is rank-deficient, and so
\[ \exists (j, k), q < j \leq p \left( (Z_k^j)^T Z_k^j > 0 \right) \implies \hat{\lambda} < 1. \]

Using the converse argument,
\[ \hat{\lambda} = 1 \implies (Y_k, q < j \leq p) \left( (Z_k^j)^T Z_k^j = 0 \right) \implies \forall k, \text{colspace}(Y_k - \hat{\mu}_k^k) (\tilde{X}^k)^{-1/2} \subseteq \text{colspace}(P_1^k, P_2^k, \ldots, P_q^k) \implies \forall k, \text{colspace}(Y_k) \subseteq \text{colspace}(Y_k^k - \hat{\mu}_k). \]

Now the proof of Theorem 3 follows straightforwardly. From Lemma A.1,
\[ P(\hat{\lambda} = 1) = \text{rank}(\tilde{R}) < p \]
\[ \leq P(\forall k, \text{colspace}(Y_k) \subseteq \text{colspace}(Y_k^k - \hat{\mu}_k^k) | \text{rank}(\tilde{R}) < p) \]
But \( Y_k \) is a p-dimensional object, and \( \text{rank}(Y_k^k - \hat{\mu}_k^k) \leq (\text{rank}(\tilde{R}) < p) \), so
\[ P(\forall k, \text{colspace}(Y_k) \subseteq \text{colspace}(Y_k^k - \hat{\mu}_k^k) | \text{rank}(\tilde{R}) < p) = 0. \]

Proof of Theorem 5

a. The result follows because
\[ \Omega Y \hat{\Sigma}^{-1/2} \tilde{R}^{-1/2} = \Omega Y \hat{\Sigma}^{-1/2} \tilde{P}^{-1/2} \tilde{P}^{-1/2} = \Omega Z \hat{\Lambda}^{-1/2} \tilde{P}^{-1/2}. \]

b. The result follows because when we replace \( \hat{R} \) with \( \hat{R} \), we have
\[ \Omega Y \hat{\Sigma}^{-1/2} |\hat{\Lambda} + (1 - \lambda) I|^{-1/2} \tilde{P}^{-1/2} \tilde{P}^{-1/2} = \Omega Z |\hat{\Lambda} + (1 - \lambda) I|^{-1/2}. \]

c. Let \( Z = Y \hat{\Sigma}^{-1/2} \tilde{P} \). Now scale and rotation invariance of \( T(Y) \) implies that
\[ T(Z) = T(Y) = g(\Omega Z \hat{\Lambda}^{-1/2} \tilde{P}^{-1/2}). \]

The jth column of \( \Omega Z \hat{\Lambda}^{-1/2} \tilde{P}^{-1/2} \) is the effect expressed in the jth variable of Z. Let
\[ I_{a \times b} = (I_{a \times b} a \times (a-b)) \text{ for } a < b. \]
Then \( Z_a = Z_{a \times a} \), the sample variance of \( Z_a \), is \( I_{a \times p} A_{a \times p} \), and the effects matrix for \( Z_a \) is \( \Omega Z_{a \times p} \), so that
\[ T(Z_a) = g(\Omega Z_{a \times p} A_{a \times p})^{-1/2} I_{a \times p} \tilde{P}^{-1/2} = g(\Omega Z \tilde{P}^{-1/2} I_{a \times p} \tilde{P}) \]
d. If \( \text{rank}(\tilde{R}) = q \), then the Moore–Penrose inverse of \( \hat{R} \) or calculated as \( \tilde{R}^{-1} = \tilde{P} \tilde{P}^{-1} \tilde{P}^{-1/2} \).

REFERENCES


