Electronic Journal of Statistics
Vol. 7 (2013) 3059–3088
ISSN: 1935-7524
DOI: 10.1214/13-EJS872

Prediction in abundant high-dimensional linear regression

R. Dennis Cook

School of Statistics
University of Minnesota
e-mail: dennis@stat.umn.edu

Liliana Forzani

Instituto de Matemática Aplicada del Litoral and Facultad de Ingeniería Química
CONICET and UNL
e-mail: liliana.forzani@gmail.com

and

Adam J. Rothman

School of Statistics
University of Minnesota
e-mail: arothman@umn.edu

Abstract: An abundant regression is one in which most of the predictors contribute information about the response, which is contrary to the common notion of a sparse regression where few of the predictors are relevant. We discuss asymptotic characteristics of methodology for prediction in abundant linear regressions as the sample size and number of predictors increase in various alignments. We show that some of the estimators can perform well for the purpose of prediction in abundant high-dimensional regressions.

AMS 2000 subject classifications: Primary 62J05; secondary 62H12.
Keywords and phrases: Inverse regression, least squares, Moore-Penrose inverse, sparse covariance estimation.

Received January 2013.

Contents

1. Introduction

The classical linear model for the regression of a univariate response Y on a vector X of p predictors $X^{(j)}$ $(j = 1, \ldots, p)$ can be written as

$$Y = \mu_Y + \beta^T(X - \mu_X) + \epsilon, \tag{1}$$

where μ_Y and μ_X are the marginal means of Y and X, and $\beta \in \mathbb{R}^p$ is a vector of unknown regression coefficients. We assume that the error $\epsilon \sim N_1(0, \sigma_\epsilon^2)$, that $X \sim N_p(\mu_X, \Sigma)$ with $\Sigma > 0$ and that ϵ is independent of X. The assumption of multivariate normality is used primarily to facilitate our theoretical development. Simulation results indicate that modest deviations from normality do not affect our results qualitatively, provided that the linear model (1) still holds. Let $\sigma_Y^2 = \operatorname{var}(Y)$, so that $\sigma_\epsilon^2 = \sigma_Y^2 - \beta^T\Sigma\beta$, and let $\Delta = \operatorname{var}(X|Y) = \Sigma - \sigma_{XY}\sigma_{XY}^T/\sigma_Y^2$, where $\sigma_{XY} = \operatorname{cov}(X, Y) \in \mathbb{R}^p$ and $\beta = \Sigma^{-1}\sigma_{XY}$. We assume also that the data (Y_i, X_i), $(i = 1, \ldots, n)$, consist of n independent copies of (Y, X).

Least squares is surely the most common method of estimating β when $n \gg p$, but there seems to be no corresponding widely used estimator when $n = O(p)$. In particular, when $n \leq p$, β is unidentified in the cases model $Y_i = \mu_Y + \beta^T(X_i - \mu_X) + \epsilon_i$ $(i = 1, \ldots, n)$, and additional structure or regularization is required. Several regularizing methods exist and their appropriateness depends on application-specific requirements. The introduction of shrinkage or sparsity in estimators of β via penalized least squares is now a widely accepted constraint to facilitate progress. Frank and Friedman (1993) introduced the

bridge estimators for which ridge regression (Hoerl and Kennard, 1970) and the lasso (Tibshirani, 1996) are special cases. Alternative penalty functions have been proposed, including the smoothly clipped absolute deviation penalty (Fan and Li, 2001), the elastic net penalty (Zou, 2005), and the adaptive lasso penalty (Zou, 2006). Methods to estimate a sparse β that incorporate regularized estimators of $\operatorname{var}(X)$ have also been proposed (Witten and Tibshirani, 2009; Jeng and Daye, 2011). Working in the context of high-dimensional linear models with nonstochastic predictors, Shao and Deng (2012) showed that the ridge estimator of β typically does not give rise to an L_2-consistent estimator of the population fitted values, even with a sparsity condition imposed on the projection of β onto the row space of the design matrix. While these various methods have been shown to perform well in certain settings, there is a need to develop new methods for high-dimensional regressions when shrinkage or sparse estimators of β perform poorly.

Cook, Forzani and Rothman (2012) recently studied dimension reduction in high-dimensional regressions by modeling the inverse regression of X on Y as $\mathrm{E}(X|Y) = \mathrm{E}(X) + \Gamma[g(Y) - \mathrm{E}\{g(Y)\}]$, where $\Gamma \in \mathbb{R}^{p\times d}$ is unknown with rank $d \le p$, $g : \mathbb{R} \to \mathbb{R}^d$ is an unknown vector-valued function and $\operatorname{var}(X|Y) = \Delta \in \mathbb{R}^{p\times p}$ is positive definite. It follows from Cook (2007) and Cook and Forzani (2008) that $R(X) = (\Gamma^T\Delta^{-1}\Gamma)^{-1}\Gamma\Delta^{-1}\{X - \mathrm{E}(X)\}$ is the minimal sufficient reduction for the regression of Y on X. Cook, Forzani and Rothman (2012) studied the asymptotic properties of several estimators of $R(X)$: given a user-specified weight matrix $\widehat{W}$, they used weighted least squares to construct an estimator $\widehat{\Gamma}$ of Γ, leading to the estimator $\widehat{R}_{\widehat{W}}(X) = (\widehat{\Gamma}^T\widehat{W}\widehat{\Gamma})^{-1}\widehat{\Gamma}^T\widehat{W}(X-\bar{X})$ of $R(X)$. They studied several cases, including when $\widehat{W}$ was the inverse of the residual sample covariance matrix from the regression of X on Y, and when $\widehat{W}$ was a sparse estimator of Δ^{-1}. Conditioning on the observed values of the responses, they showed that $\widehat{R}_{\widehat{W}}(X_N) - R(X_N)$ depends on four terms that converge to 0, where X_N is an independent copy of X. This convergence rate depends on the rate at which $\widehat{W}$ converges to its population value W, the agreement between W and Δ^{-1}, and the signal rate in the regression. They showed also that root-n consistent estimation of $R(X_N)$ is possible when $p/n \to [0, 1)$, the signal rate is *abundant* and $X|Y$ is normal (Cook, Forzani and Rothman, 2012, prop. 6.2). However, they did not consider convergence rates of actual predictions of Y, reasoning instead that the regression of Y on $\widehat{R}$ could be studied using graphical methods or addressed using non-parametric methods when the dimension of $\widehat{R}$ is small, as often seems to be the case.

In this article we use various estimators of β in model (1) as essential ingredients for predicting Y at a new independent observation X_N of X from an abundant regression, in which the addition of predictors accumulates information on the response. Since the regressions we consider are allowed to be abundant, we do not impose sparsity on β or constrain it otherwise. Instead, we study predictions based on the least squares estimator of β when $n > p + 2$, a natural estimator of β when Σ is known, estimators of β based on the Moore-Penrose inverse of sample versions of Σ and Δ, and ultimately an estimator of β based on a sparse estimator of Δ^{-1}.

Our study links with the approach of Cook, Forzani and Rothman (2012) in the following ways. Our assumption of multivariate normality of (Y_i, X_i) means that if we condition on Y_i, then we cover a special case of the their model with $E(X|Y) = \Gamma(Y - \mu_Y)$; however, unlike Cook, Forzani and Rothman (2012), our technical results do not condition on Y. This also allowed us to consider estimators not covered by Cook, Forzani and Rothman (2012) and to study directly the convergence rates of prediction of Y. The closest point of commonality between the two studies is described in the preamble to Section 4. Taken together, the two studies indicate strongly that when appropriate it is better to deal with abundant regressions in which $n < p$ through restrictions on the conditional variance of $X|Y$ rather than the marginal variance of X.

2. Preliminaries

Let $\bar{Y}$ and $\bar{X}$ denote the sample means of Y and X, and let $\widehat{\beta}$ denote a generic estimator of β. Specific instances of $\widehat{\beta}$ will be studied in subsequent sections. The predicted value $\widehat{Y}_N$ of Y at a new observation X_N on X is then $\widehat{Y}_N = \bar{Y} + \widehat{\beta}^T(X_N - \bar{X})$. Since the term $\bar{Y}$ will be common to all estimators considered, we judge the relative merits of different estimators of β by studying the order of $D_N = \widehat{\beta}^T(X_N - \bar{X}) - \beta^T(X_N - \mu_X)$ as n and p approach infinity in various alignments. If $D_N = O_p\{r(n,p)\}$ and $r(n,p) \to 0$ as $n, p \to \infty$ then the sample predictions converge to the population prediction at rate at least r^{-1}. When details permit, we will consider the order of the variance $V = \mathrm{var}(D_N)$. If $V = O\{r^2(n,p)\}$, then again the sample predictions converge to the population prediction at rate at least r^{-1}. Additionally, all estimators $\widehat{\beta}$ are independent of $\bar{X}$ and consequently there is no predictive bias since then $\mathrm{E}(D_N) = 0$.

Information about the response accumulates as new predictors are added to an abundant regression. Several of the estimators we consider depend explicitly on a measure of the rate at which this accumulation occurs, which we refer to as the signal rate $h(p)$. We assume throughout this article that $h(p) = O(p)$, since this seems appropriate for most applications. Let R^2_{YX} denote the usual squared population multiple correlation coefficient for the regression of Y on X. Then the signal rate can be expressed as

$$h(p) = \frac{\sigma_{XY}^T \Delta^{-1} \sigma_{XY}}{\sigma_Y^2} = \frac{R^2_{YX}}{1 - R^2_{YX}}. \tag{2}$$

We see from the first expression for $h(p)$ that if σ_{XY} falls in a reducing subspace of Δ with eigenvalues that are bounded away from 0 and ∞ then $h(p) \asymp \|\sigma_{XY}\|^2$ as $p \to \infty$, where the notation $a_n \asymp b_n$ means that $a_n = O(b_n)$ and $b_n = O(a_n)$. If, in addition, sufficiently many elements of σ_{XY} are non zero then we can have $h(p) \asymp p$. We refer to regressions as abundant if $h(p) \to \infty$ and as sparse if $h(p) \asymp 1$. Clearly, a regression is abundant if and only if $R^2_{YX} \to 1$. When considered in the context of model (1), the signal rate (2) is the same as the signal rate defined by Cook, Forzani and Rothman (2012, §4.2).

Let $\varphi_{\max}(A)$ and $\varphi_{\min}(A)$ denote the largest and smallest eigenvalue of the matrix A, and let $\|v\|$ denote the length of a vector $\|v\|$. The signal rate is minimized over the directions $\sigma_{XY}/\|\sigma_{XY}\|$ when σ_{XY} lies in the span of the eigenvector corresponding to $\varphi_{\max}(\Delta)$. In that worst case scenario, $h(p) \asymp \|\sigma_{XY}\|^2/\varphi_{\max}(\Delta)$, and $\|\sigma_{XY}\|^2$ needs to increase faster that $\varphi_{\max}(\Delta)$ to have an abundant regression.

The forms for $h(p)$ stated in the next two lemmas may provide additional intuition. In preparation, let $\rho(u, v|w)$ denote the matrix of conditional correlations between the elements of the vectors u and v given w. Unconditional correlation matrices are written without the conditioning argument, and ordinary pairwise correlations result when u and v are scalar variables. Let $\omega_p = \sum_{j=1}^{p} \rho^2(X^{(j)}, Y)/\{1 - \rho^2(X^{(j)}, Y)\}$.

Lemma 2.1. *Assume that the eigenvalues of $\rho(X, X|Y)$ are bounded away from 0 and ∞ as $p \to \infty$. Then $h(p) \asymp \omega_p$.*

This lemma says essentially that if the conditional correlation matrix $\rho(X, X|Y)$ is well behaved as $p \to \infty$ and a sufficient number of predictors are marginally correlated with Y then the regression is abundant. For instance, if the marginal correlations $\rho^2(X^{(j)}, Y)$ are bounded away from 0 and 1 then $h(p) \asymp p$.

The next lemma describes the change in h when adding a single new predictor $X^{(p+1)}$ to a regression with p predictors X.

Lemma 2.2. *Let α denote the coefficient of Y in the population regression of $X^{(p+1)}$ on X and Y, and let R^2 denote the squared population multiple correlation coefficient for the regression of $X^{(p+1)}$ on X given Y. Then*

$$h(p+1) = h(p) + \frac{\sigma_Y^2}{\operatorname{var}(X^{(p+1)})} \frac{\alpha^2}{(1 - R^2)}.$$

The result in this lemma indicates that conditional collinearity, as measured by R^2, between the predictors X in the regression and the new predictor $X^{(p+1)}$ may result in a substantial increase in the signal rate, provided that there is a sufficient relationship α^2 between the new predictor and the response adjusting for the predictors already in the regression.

3. Forward regression estimators

3.1. Σ *known*

Let $\widehat{\sigma}_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})/(n-1)$, which is an unbiassed estimator of σ_{XY} relative to the joint distribution of X and Y. In this section we consider the natural estimator $\widehat{\beta} = \Sigma^{-1}\widehat{\sigma}_{XY}$, assuming that Σ is know. Although this setting is used primarily as a reference point, Σ could be known in some computer experiments where the predictors are generated as inputs to a computer code.

Proposition 3.1. *Assume that model (1) holds and that Σ is known. Then with $\widehat{\beta} = \Sigma^{-1}\widehat{\sigma}_{XY}$, we have $V \asymp p/n$.*

As a consequence of this result, we see that knowledge of Σ does not really suggest advances in methodology since we may still need $n \gg p$ for useful results. The lack of progress here may be a reflection of the fact that model (1) is conditional on X while the estimator is based on marginal moments. If this is so, then better rates might be obtained by using an estimator of Σ, even if Σ is known.

3.2. Σ estimated

We divide the discussion of the case when Σ is estimated by the relationship between n and p. Consider first regressions in which $n > p + 2$. Let $\widehat{\Sigma} = \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T/(n-1)$. Then $\widehat{\Sigma}^{-1}$ exists with probability 1 and we can use the usual ordinary least squares estimator $\widehat{\beta} = \widehat{\Sigma}^{-1}\widehat{\sigma}_{XY}$. Let $\kappa^2 = p/\{nh(p)\}$.

Proposition 3.2. *Assume that model (1) holds and that $n > p + 2$. Let $\widehat{\beta} = \widehat{\Sigma}^{-1}\widehat{\sigma}_{XY}$. Then $V = O\{\kappa^2(n+1)/(n-p-2)\}$.*

This proposition implies that if $p/n \to r \in [0, 1)$ then $V = O(\kappa^2)$. From the definition of κ we see that there is a synergy between the sample size and the signal rate, the signal rate serving to multiply the sample size to produce an effective sample size of $nh(p)$. For instance, if $h(p) \asymp p$ then $V = O(n^{-1})$ and we obtain the usual root-n convergence rate, although we need not have $n \gg p$. These results suggest that we might reasonably expect useful predictions in an abundant regression with, say, $n = 1000$ and $p = 750$. On the other hand, if the regression is sparse then $V = O(p/n)$ and we are back to the usual requirement that $n \gg p$. Consider next regressions in which $h(p) \asymp p$ and $n - p = c > 2$, where c is a constant. In such regressions, $p/n \to 1$ and $V = O(1)$, which suggest that we might not obtain useful prediction in abundant regressions when, say, $n = 1000$ and $p = 997$, depending on the size of V.

Comparing Propositions 3.1 and 3.2, we conclude that if Σ is known it can still be better to form $\widehat{\beta}$ using $\widehat{\Sigma}$ instead of its known population value. While this result might seem counterintuitive, it appears to be an instance of the general paradox described by Henmi and Eguchi (2004).

We turn next to regressions in which $p > n$. It follows from classical results in linear model theory that a best linear unbiassed predictor can be estimated by taking $\widehat{\beta}$ to be any solution of the linear equations $\widehat{\Sigma}\widehat{\beta} = \widehat{\sigma}_{XY}$ (Christensen, 1987, §VI.3). Here we use the specific estimator $\widehat{\beta} = \widehat{\Sigma}^{-}\widehat{\sigma}_{XY}$, where A^{-} denotes the Moore-Penrose inverse of A.

Proposition 3.3. *Assume that model (1) holds, that $n < p$, and that the eigenvalues of Σ are bounded away from 0 and ∞ as $p \to \infty$. Let $\widehat{\beta} = \widehat{\Sigma}^{-}\widehat{\sigma}_{XY}$. Then $V \asymp 1$ if either (a) $n/p \to r \in [0, 1)$ or (b) $p - n$ is constant, so $n/p \to 1$, and $h(p) \asymp p$.*

This proposition requires that the eigenvalues of Σ be bounded, which is required also by current methods for estimating a sparse covariance matrix. It

indicates that we should not necessarily expect useful predictions when $n < p$ and the Moore-Penrose inverse is used in $\widehat{\beta}$, since then V converges to a positive constant. Of course, we could obtain good predictions if that constant is sufficiently small, but generally the result is not promising for the use of Moore-Penrose inverses. The primary issue is apparently that predictions outside of $\text{span}(X^{(1)}, \ldots, X^{(p)})$ can be relatively variable, as discussed in the next section.

The results of this section exclude regressions in which $n = p+j$ for $j = 0, 1, 2$. If $n = p+1$ or $n = p+2$ then $\widehat{\Sigma}^{-1}$ still exists with probability 1, but the variance of $\widehat{\Sigma}^{-1}$ does not exist (von Rosen, 1988), which can lead to erratic results in practice. Similar comments apply when $n = p$ (Cook and Forzani, 2011). The methodology discussed here should be avoided when $n = p + j$ for $j = 0, 1, 2$.

3.3. Prediction at estimable functions with Σ estimated

It seems reasonable to expect that we may get useful results when predicting at estimable functions $\beta^T X_N$; that is, at points X_N that are linear combinations of $X_1, \ldots, X_n$. This ensures that the predictions are unbiassed. If $n > p$ then $\text{span}(X_1 - \mu_X, \ldots, X_n - \mu_X) = \mathbb{R}^p$ and the restriction places no constraint on X_N. If $n < p$ then we must have $X_N - \mu_X \in \text{span}(X_1 - \mu_X, \ldots, X_n - \mu_X)$ so the restriction to estimable functions in some sense keeps X_N close to the observed predictors.

Assuming that $X_N|(\bar{X}, \widehat{\Sigma}) \sim N_p(\bar{X}, \widehat{\Sigma})$ ensures that $\beta^T X_N$ is an estimable function and that the distribution of X_N is similar to the observed data. Under this assumption, if $n < p$ then $X_N - \bar{X} \in \text{span}(\widehat{\Sigma})$, so X_N can always be represented as a linear combination of the observed predictors. The reasoning in the asymptotic analysis of the predictions follows the same general steps as given previously in Section 3.2, but the details are different and the cases $n > p$ and $n < p$ can be addressed at the same time.

Proposition 3.4. *Assume model (1) and that, given $(\bar{X}, \widehat{\Sigma})$, the new predictors $X_N|(\bar{X}, \widehat{\Sigma}) \sim N_p(\bar{X}, \widehat{\Sigma})$. Let $\widehat{\beta} = \widehat{\Sigma}^{-}\widehat{\sigma}_{XY}$. Then $V \asymp \min(n, p)\{nh(p)\}^{-1} + n^{-1}$.*

According to this proposition, if $n > p$ then $V \asymp \kappa^2$, which is the same rate obtained in Proposition 3.2 when $p/n \to r \in [0, 1)$. If $n < p$ then $V \asymp h^{-1}(p) + n^{-1}$ and the convergence rate depends on the relationship between the signal rate and the sample size. In particular, $V \asymp n^{-1}$ if $h(p) \asymp p$. Overall, Proposition 3.4 indicates that we can get favorable convergence rates for predictions at points that are close to the observed data.

4. Alternative estimators when $n < p$

The ordinary least squares estimator performs well asymptotically when $n > p + 2$, as indicated in Proposition 3.2, and the estimator $\widehat{\beta} = \widehat{\Sigma}^{-}\widehat{\sigma}_{XY}$ performs well when $n < p$ and $X_N|(\bar{X}, \widehat{\Sigma}) \sim N_p(\bar{X}, \widehat{\Sigma})$, as described in Proposition 3.4. In this section consider the class of predictors that may be suitable for settings where $n < p$ and $X_N \sim N_p(\mu_X, \Sigma)$.

Since $\Sigma^{-1} = \Delta^{-1} - \Delta^{-1}\sigma_{XY}\sigma_{XY}^T\Delta^{-1}/[\sigma_Y^2\{1+h(p)\}]$, the coefficient vector can be expressed as $\beta = \Delta^{-1}\sigma_{XY}/\{1+h(p)\}$. Let $\hat{\sigma}_Y^2$ denote the marginal sample variance of Y and let $\widehat{\Delta} = \widehat{\Sigma} - \hat{\sigma}_{XY}\hat{\sigma}_{XY}^T/\hat{\sigma}_Y^2$, so that $(n-1)\widehat{\Delta}$ follows a Wishart $W_p(\Delta, n-2)$ distribution. Let $\widehat{\Omega}$ be an estimator of Δ^{-1}. Then we consider estimators of β of the form

$$\widehat{\beta} = \widehat{\Omega}\hat{\sigma}_{XY}/\{1+\hat{h}(p)\}, \tag{3}$$

where $\hat{h}(p) = \hat{\sigma}_{XY}^T\widehat{\Omega}\hat{\sigma}_{XY}/\hat{\sigma}_Y^2$ and $\hat{\sigma}_Y^2$ is the marginal sample variance of Y. If $n > p+2$ and $\widehat{\Omega} = \widehat{\Delta}^{-1}$ then the estimator in (3) is equal to ordinary least squares estimator $\widehat{\Sigma}^{-1}\hat{\sigma}_{XY}$ whose behaviour was characterized in Proposition 3.2. Otherwise, (3) represents a new class of estimators that may have advantages over the previously discussed estimators when $n < p$. We discuss the estimator with $\widehat{\Omega} \propto \widehat{\Delta}^-$ in §4.1. The general behaviour of the estimator is characterized in §4.2.

The reduction $\widehat{R}_{\widehat{W}}(X)$, studied by (Cook, Forzani and Rothman, 2012) and described in the Introduction, simplifies when $f = Y - \bar{Y}$, $\widehat{\Delta}$ is non-singular and $\widehat{W} = \widehat{\Delta}^{-1}$. Under those conditions

$$\widehat{R}_{\widehat{W}}(X) = \frac{\hat{\sigma}_{XY}^T\widehat{\Delta}^{-1}}{\hat{\sigma}_Y\hat{h}(p)}(X - \bar{X}).$$

The coefficient vector $\widehat{B} = \widehat{\Delta}^{-1}\hat{\sigma}_{XY}^T/\{\hat{\sigma}_Y\hat{h}(p)\}$ for this reduction is proportional to $\widehat{\beta}$ in (3) when $\widehat{\Omega} = \widehat{\Delta}^{-1}$, but otherwise they differ. The reduction coefficient vector $\widehat{B}$ is invariant to scale changes in Y, which is appropriate for a reduction but not for estimation of β; $\widehat{B}$ is not and was not intended to be an estimator of β since the linear model (1) played no direct role in the study by Cook, Forzani and Rothman (2012). If $p/n \to r \in [0,1)$ and model (1) holds then, from Proposition 3.2, $D_N = O_p(\kappa)$ and, from Proposition 6.2 of Cook, Forzani and Rothman (2012), $\widehat{R}_{\widehat{W}}(X_N) - R(X_N) = O_p(\kappa)$. This indicates that $\widehat{\beta}^T(X_N - \bar{X})$ and $\widehat{R}_{\widehat{W}}(X)$ have the same convergence rate, although this result is not implied directly by Cook, Forzani and Rothman (2012).

4.1. $\widehat{\Omega} \propto \widehat{\Delta}^-$ and $\Delta = \delta^2 I_p$

Recall from Proposition 3.3 that we obtained the weak result that $V \asymp 1$ when $n < p$, the eigenvalues of Σ are bounded and the Moore-Penrose inverse of $\widehat{\Sigma}$ is used in $\widehat{\beta} = \widehat{\Sigma}^-\hat{\sigma}_{XY}$. We were surprised to find that stronger results can be obtained when the unbiassed version the Moore-Penrose inverse of $\widehat{\Delta}$ is used in (3) instead.

Proposition 4.1. *Assume that model (1) holds, that $n+1 < p$ with $n/p \to r \in (0,1)$ and that $\Delta = \delta^2 I_p$. Let $\widehat{\Omega} = [p(p-n+1)/\{(n-1)(n-2)\}]\widehat{\Delta}^-$. Then with $\widehat{\beta}$ as defined in (3), we have $D_N = O_p\{h^{-1/2}(p)\}$.*

The scaled version of the Moore-Penrose inverse used in this proposition is unbiassed, $\mathrm{E}(\widehat{\Omega}) = \Delta^{-1}$. The proposition requires that $\Delta = \delta^2 I_p$, since this is the only case for which we have the technical equipment to compute the required moments of $\widehat{\Omega}$. We anticipate that a similar result holds for a general Δ with bounded eigenvalues. Nevertheless, comparing the results of Proposition 4.1 with the corresponding result $D_N = O_p(1)$ from Proposition 3.1 for Σ known suggests that knowledge of Δ may be more useful than knowledge of Σ. We elaborate on this point following Corollary 4.1.

4.2. General $\widehat{\Omega}$

There are many regularized covariance estimators we could use in place of $\widehat{\Omega}$; Pourahmadi (2011) gives a review of several methods. In some applications, it may be reasonable to assume that Δ^{-1} is sparse or approximately sparse. A natural estimator that exploits this condition is that obtained by L_1-penalized likelihood (Yuan and Lin, 2007; Friedman, Hastie and Tibshirani, 2008; Rothman et al., 2008). Let $\mathrm{diag}(A)$ denote the diagonal matrix with diagonal elements the same as those of the square matrix A. We define this estimator by

$$\widehat{\Theta}_\lambda = \arg\min_{\Theta \succ 0} [\mathrm{tr}\{\Theta \mathrm{diag}^{-1/2}(\widehat{\Delta}) \widehat{\Delta} \mathrm{diag}^{-1/2}(\widehat{\Delta})\} - \log|\Theta| + \lambda \sum_{i \neq j} |\theta_{ij}|], \quad (4)$$

$$\widehat{\Delta}_\lambda^{-1} = \mathrm{diag}^{-1/2}(\widehat{\Delta}) \widehat{\Theta}_\lambda \mathrm{diag}^{-1/2}(\widehat{\Delta}),$$

where $\lambda \geq 0$ is a tuning parameter. The penalization is done on the inverse correlation scale, as suggested by Rothman et al. (2008). This ensures that our estimator is invariant to scaling of the variables. We used the graphical lasso algorithm (Friedman, Hastie and Tibshirani, 2008) to compute the inverse correlation matrix estimator $\widehat{\Theta}_\lambda$, selecting λ by k-fold cross validation to minimizing prediction error. The QUIC algorithm of Hsieh et al. (2011) could also be used and performs similarly.

We need to gauge the rate at which $\widehat{\Omega}$ converges Δ^{-1} to characterize generally the asymptotic behaviour of predictions based on (3). Let $S = \Delta^{1/2}(\widehat{\Omega} - \Delta^{-1})\Delta^{1/2}$ and let $\|S\|$ denote the spectral norm of S. The rates given in the next proposition require $\widehat{\Omega}$ to be chosen so that $\|S\|^2 = O_p(\omega^2)$ and $\|\mathrm{E}(S^2)\| = O(\omega^2)$ as $\omega \to 0$. The rate ω^{-1} depends on the particular estimator, but the conditions are not harsh and hold for many estimators, including when $\widehat{\Omega} = \widehat{\Delta}_\lambda^{-1}$.

Proposition 4.2. *Assume that model (1) holds and that β is estimated as given in (3). Assume also that $||S||^2 = O_p(\omega^2)$ and $\|\mathrm{E}(S^2)\| = O(\omega^2)$ as $\omega \to 0$. Then*

$$D_N = O_p(\kappa^2) + O_p\{\kappa h^{-1/2}(p)\} + O_p(\omega) + O_p(n^{-1/2}).$$

We know from Proposition 3.2 that $D_N = O_p(\kappa)$ when $n > p + 2$. Although Proposition 4.2 holds regardless of the relationship between n and p, it does not reduce to Proposition 3.2 when $n > p + 2$ and $\widehat{\Omega}^{-1} = \widehat{\Delta}^{-1}$ because of additional bounding necessary to incorporate a general $\widehat{\Omega}$. The main purpose of Proposition 4.2 is to address the case where $p > n$.

Corollary 4.1. *Under the conditions of Proposition 4.2, if $p > n$ then*

$$D_N = O_p(\kappa^2) + O_p(\omega) + O_p(n^{-1/2}). \tag{5}$$

If, in addition, Δ is known then

$$D_N = O_p(\kappa^2) + O_p(n^{-1/2}). \tag{6}$$

In reference to the rate in (5), $\kappa^2 > n^{-1/2}$ if and only if $p/h > \sqrt{n}$, and then the rate reduces to $O_p(\omega) + O_p(\kappa^2)$. On the other hand, if $p/h < \sqrt{n}$ the rate reduces to $O_p(\omega) + O_p(n^{-1/2})$. If $h(p) \asymp p$ and Δ is known then the rate in (6) reduces to $O_p(n^{-1/2})$. This rate stands in contrast to the rate $O_p(1)$ when Σ is known in Proposition 3.1. In effect, knowledge of Δ is much more important for prediction than knowledge of Σ.

The next corollary addresses the estimator $\widehat{\Delta}_\lambda^{-1}$ in the context of Proposition 4.2.

Corollary 4.2. *Under the conditions of Proposition 4.2, let $\widehat{\Omega} = \widehat{\Delta}_\lambda^{-1}$ with $\lambda \asymp (\log p/n)^{1/2}$. Assume that, as $p \to \infty$, the eigenvalues of Δ are bounded away from 0 and ∞ and that the number of non zero off-diagonal elements of Δ^{-1} is bounded. Then, when $h(p) \asymp p$, $D_N = O_p(n^{-1/2} \log^{1/2} p)$.*

Since $\widehat{\Delta}_\lambda^{-1}$ requires that the eigenvalues of Δ be bounded, the condition $h(p) \asymp p$ will hold when $\|\sigma_{XY}\| \asymp p$ so that many predictors are marginally correlated with the response. The condition that the number of non zero off-diagonal elements of Δ^{-1} be bounded is perhaps the most stringent theoretical condition required for $\widehat{\Delta}_\lambda^{-1}$. This condition could be relaxed to allow the number of non-zero off-diagonal elements in Δ^{-1} to grow slowly (Rothman et al., 2008). We expect that with additional assumptions, a convergence rate bound depending on the row sparsity of Δ^{-1} could be obtained (Ravikumar et al., 2011). The alternative sparse estimator of Δ^{-1} proposed by Cai, Liu and Luo (2011) could also be explored within this context to achieve rates of convergence depending on the approximate row sparsity of Δ^{-1}. Our conclusions from a variety of simulations is that $\widehat{\Delta}_\lambda^{-1}$ also works well when there are many non zero off-diagonal elements.

5. Simulation

5.1. Overview

Let $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda})}$ be the proposed estimator of β obtained via (3) using $\widehat{\Omega} = \widehat{\Delta}_{\hat{\lambda}}^{-1}$, where $\hat{\lambda}$ is selected with 5 fold cross validation, minimizing prediction error. When $p > n + 2$, let $\widehat{\beta}_{\widehat{\Delta}}$ be the proposed estimator obtained via (3) using $\widehat{\Omega} = [p(p - n + 1)/\{(n - 1)(n - 2)\}]\widehat{\Delta}^{-}$.

To illustrate regressions with $p > n$, we set $n = p/2$ and evaluated the performance of $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda})}$, $\widehat{\beta}_{\widehat{\Delta}}$, and $\widehat{\beta}_{\widehat{\Sigma}^-} = \widehat{\Sigma}^{-}\widehat{\sigma}_{XY}$. For $p < n$, we set $n = 2p$ and compared $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda})}$ to $\widehat{\beta}_{\widehat{\Sigma}^{-1}} = \widehat{\Sigma}^{-1}\widehat{\sigma}_{XY}$.

For each of 100 replications, we generated a realization of n independent copies of the random vector (Y, X). Multiple joint distribution specifications were considered. Performance was measured with the prediction error, defined as

$$\frac{1}{1000}\sum_{k=1}^{1000}\{\widehat{\beta}^T(X_{N,k}-\bar{X})-\beta^T(X_{N,k}-\mu_X)\}^2, \tag{7}$$

where $X_{N,k}, k = 1, \ldots, 1000$ are independent copies of X.

We selected the tuning parameter for $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda})}$ from $\{10^{-5+0.5j} : j = 0, \ldots, 12\}$ when $p < 128$ and from $\{10^{-3+0.5j} : j = 0, \ldots, 8\}$ when $p \geq 128$. In the simulations with $n = 2p$, there were no selected tuning parameters on the boundaries of these sets. When $n = p/2$, there was a small fraction of selections on the lower boundaries, especially at sample sizes $n = 8$ and $n = 16$ for which there was limited information for 5-fold cross-validation.

5.2. Inverse regression simulation

5.2.1. Model description

In this simulation, Y is standard normal and

$$X = \sigma_{XY}Y + \varepsilon, \tag{8}$$

where $\varepsilon \sim N_p(0, \Delta)$, $\varepsilon \perp\!\!\!\perp Y$ and $\beta = \Delta^{-1}\sigma_{XY}/(1+\sigma_{XY}^T\Delta^{-1}\sigma_{XY})$. We generated σ_{XY} to have round(p^α) nonzero entries, where $\alpha = 1/2$ and 1, with values independently drawn from the standard normal distribution. Two covariance structures for Δ were used, $\Delta_1 = I_p$ and Δ_2 with entries $\delta_{2ij} = a\ 0.9^{|i-j|}$ where we set $a = (1 + 0.9^2)/(1 - 0.9^2)$ to make the expected signal rates, over simulation replications, similar for Δ_1 and Δ_2, both being proportional to p^α. We used $p = 16, 32, 64, 128$, and 256.

5.2.2. Results when $n = p/2$

For Δ_1 with $n = p/2$, we plotted the average prediction error curves for $\widehat{\beta}_{\widehat{\Delta}}$, $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda})}$, and $\widehat{\beta}_{\widehat{\Sigma}^-}$ in Fig. 1a for $h \asymp p^{1/2}$ and in Fig. 1b. for $h \asymp p$. All three estimators appear to give consistent predictions as n and p grow and $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda})}$ performs best, which is expected since Δ_1^{-1} is diagonal. As our theory suggests, faster convergence occurs when $h \asymp p$. Propositions 4.1 and 4.2 guarantee prediction consistency as $p \to \infty$ in this setting for $\widehat{\beta}_{\widehat{\Delta}}$ and $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda})}$. Since Σ has unbounded eigenvalues, Proposition 3.3 does not apply, and we do not have theory to guarantee consistency for $\widehat{\beta}_{\widehat{\Sigma}^-}$.

For Δ_2 with $n = p/2$, we plotted the average prediction error curves for $\widehat{\beta}_{\widehat{\Delta}}$, $\widehat{\beta}_{\widehat{\Delta}(\lambda)}$, and $\widehat{\beta}_{\widehat{\Sigma}^-}$ in Fig. 1c for $h \asymp p^{1/2}$ and in Fig. 1d. for $h \asymp p$. We see that the three estimators give consistent predictions as n and p grow, where again $\widehat{\beta}_{\widehat{\Delta}(\lambda)}$ performs best, which is expected since Δ_2^{-1} is tri-diagonal. We do

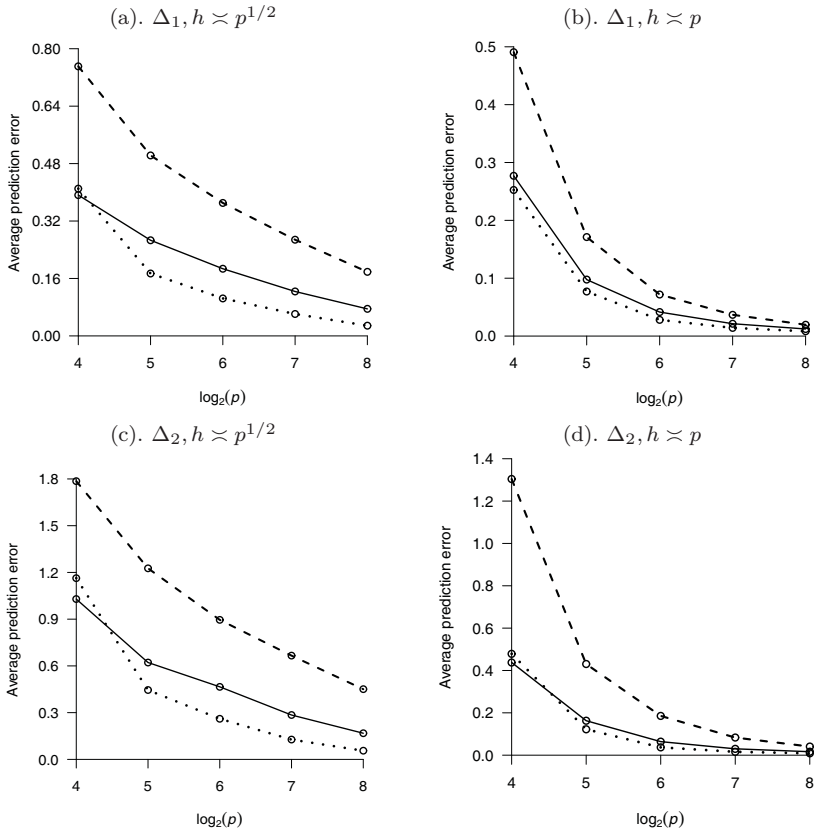

FIG 1. *Average prediction error for the inverse regression simulation based on 100 replications for* $\widehat{\beta}_{\widehat{\Sigma}^-}$ *(solid),* $\widehat{\beta}_{\widehat{\Delta}}$ *(dashes), and* $\widehat{\beta}_{\widehat{\Delta}(\lambda)}$ *(dots), with* $n = p/2$.

not have theory to guarantee consistency for $\widehat{\beta}_{\widehat{\Delta}}$ in this case because Δ_2 is not proportional to I_p.

5.2.3. Results when $n = 2p$

When $n = 2p$, we plotted the average prediction error curves for $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda})}$ and $\widehat{\beta}_{\widehat{\Sigma}^{-1}}$ in Fig. 2a and 2b for Δ_1; and in Fig. 2c and 2d for Δ_2. A pattern similar to the $n = p/2$ case is illustrated: $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda})}$ performs best and both estimators appear to give consistent predictions as n increases. Proposition 3.2 guarantees consistency for $\widehat{\beta}_{\widehat{\Sigma}^{-1}}$ under these settings as $p \to \infty$.

5.2.4. Results when Σ *and* Δ *are known*

We also investigated the prediction performance of the estimator $\Sigma^{-1}\widehat{\sigma}_{XY}$ of β that uses the known Σ, and of the estimator of β that uses the known Δ,

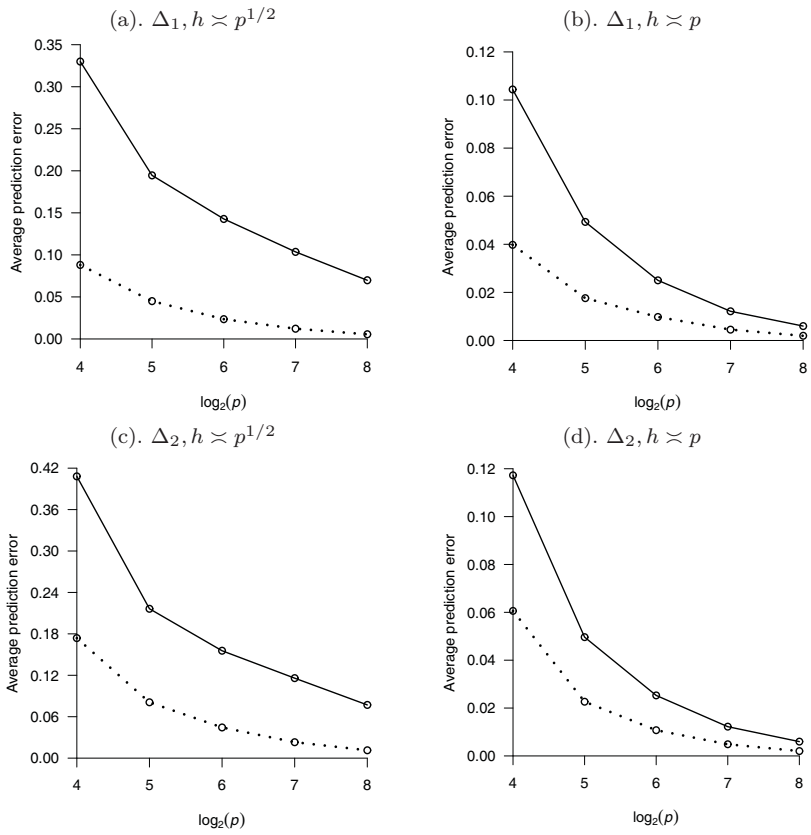

FIG 2. *Average prediction error for the inverse regression simulation based on 100 replications for* $\widehat{\beta}_{\widehat{\Sigma}^{-1}}$ *(solid) and* $\widehat{\beta}_{\widehat{\Delta}(\lambda)}$ *(dots), with* $n = 2p$.

obtained via (3) with $\widehat{\Omega} = \Delta^{-1}$. For the same simulation as described in Section 5.2.1, we plotted the average prediction error curves for Δ_1 in Fig. 3a and 3b for $n = p/2$; and in Fig. 3c and 3d for $n = 2p$. Proposition 3.1 guarantees the inconsistency of using the known Σ under these settings, which is clearly illustrated, and Proposition 4.2 guarantees the consistency of using the known Δ, which is also clearly illustrated. The curves for Δ_2 were essentially the same as those for Δ_1 and consequently they were omitted.

5.3. Elliptical t regression simulation

5.3.1. Model description

Let $T = (Y, X^T)^T$ and let $t_\nu^k(\mu, \Xi)$ denote the k dimensional elliptical t distribution with ν degrees of freedom and parameters $\mu \in \mathbb{R}^k$ and $\Xi \in \mathbb{R}^{k \times k}$ (Muirhead, 1982). This implies that $E(T) = \mu$ and $\text{var}(T) = \nu/(\nu - 2)\Xi$, when $\nu > 2$. In

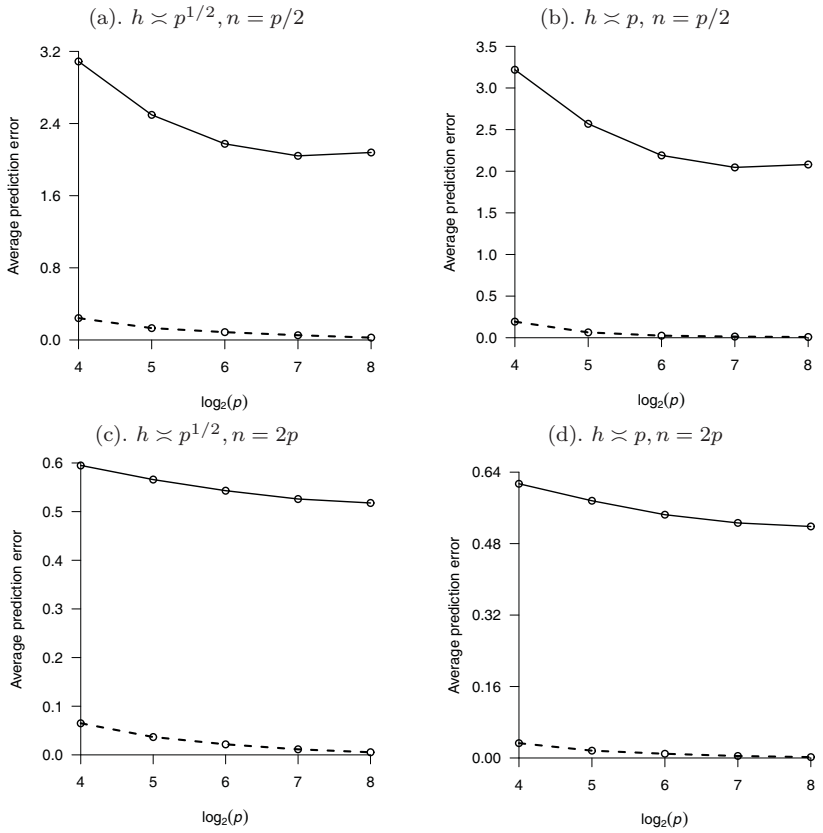

FIG 3. *Average prediction error for based on 100 replications using known* Σ *(solid) and known* Δ *(dashes) and* $\Delta = I_p$.

this simulation, $T \sim t_5^{p+1}(\mu, \Xi)$, where $\mu = 0$ and, representing Ξ according to Y and X, $\Xi_{YY} = 1$, $\Xi_{XY} = \sigma_{XY}$, $\Xi_{YX} = \Xi_{XY}^T$ and $\Xi_{XX} = \Sigma$. As a consequence, $Y|X$ follows an elliptical t with $\nu + p = 5 + p$ degrees of freedom. We generated σ_{XY} in the same way as in §5.2 and set $\Sigma = I_p + \sigma_{XY}\sigma_{XY}^T$, which has the same spirit as setting $\Delta = I_p$ in the multivariate normal simulations of §5.2. Since $\beta = \Sigma^{-1}\sigma_{XY}$, we have that $E(Y|X) = \beta^T X$ and $\text{var}(Y|X) = g(X)(1 - \beta^T\Sigma\beta)$, where g is some function. We used $p = 16, 32, 64, 128$, and 256.

5.3.2. Results

When $n = p/2$, we plotted the average prediction error curves for $\widehat{\beta}_{\widehat{\Delta}}$, $\widehat{\beta}_{\widehat{\Sigma}^-}$, and $\widehat{\beta}_{\widehat{\Delta}(\lambda)}$ in Fig. 4a for $h \asymp p^{1/2}$ and in Fig. 4b. for $h \asymp p$. The estimators appear to give consistent predictions as n and p grow with slightly worse absolute performance as compared to the multivariate normal simulations presented in Fig. 1.

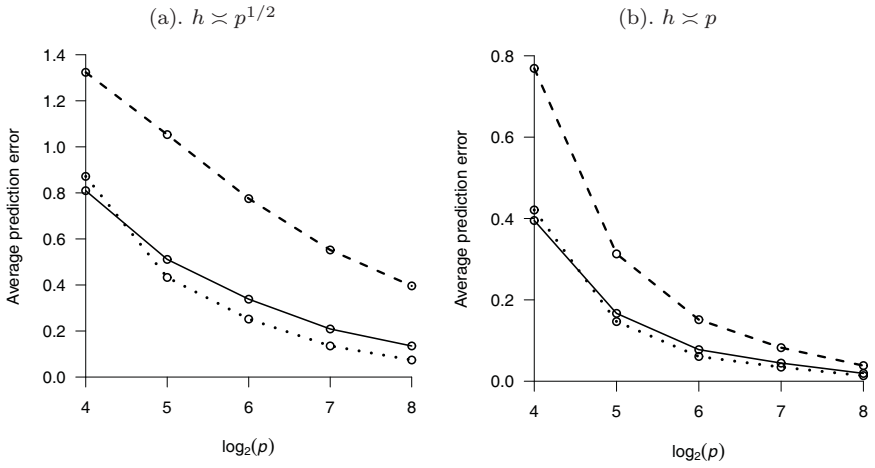

FIG 4. *Average prediction error for the elliptical t regression simulation based on 100 replications for* $\widehat{\beta}_{\widehat{\Sigma}^-}$ *(solid),* $\widehat{\beta}_{\widehat{\Delta}}$ *(dashes), and* $\widehat{\beta}_{\widehat{\Delta}(\lambda)}$ *(dots), with* $n = p/2$.

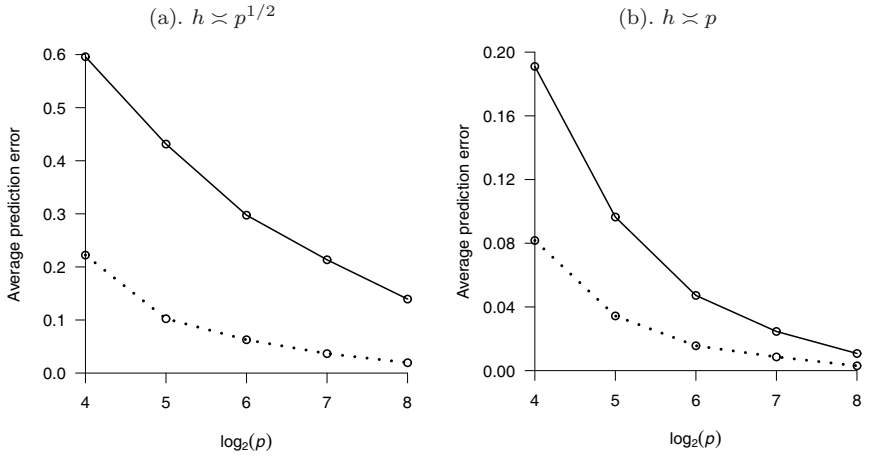

FIG 5. *Average prediction error for the elliptical t regression simulation based on 100 replications for* $\widehat{\beta}_{\widehat{\Sigma}^{-1}}$ *(solid) and* $\widehat{\beta}_{\widehat{\Delta}(\lambda)}$ *(dots), with* $n = 2p$.

When $n = 2p$, we plotted the average prediction error curves for $\widehat{\beta}_{\widehat{\Sigma}^{-1}}$ and $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda})}$ in Fig. 5a for $h \asymp p^{1/2}$ and in Fig. 5b. for $h \asymp p$. Both estimators appear consistent, but perform worse than in the multivariate normal simulations presented in Fig. 2.

We also ran this simulation using the $t_3^{p+1}(\mu, \Xi)$ distribution and noticed similar patterns to those illustrated above, but we recommend that our proposed methods only be applied to distributions that have fourth moments.

Following a referee's suggestion, we also ran this simulation using the $t_1^{p+1}(\mu, \Xi)$ distribution. We were unable to compute $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda})}$ because of numerical instability.

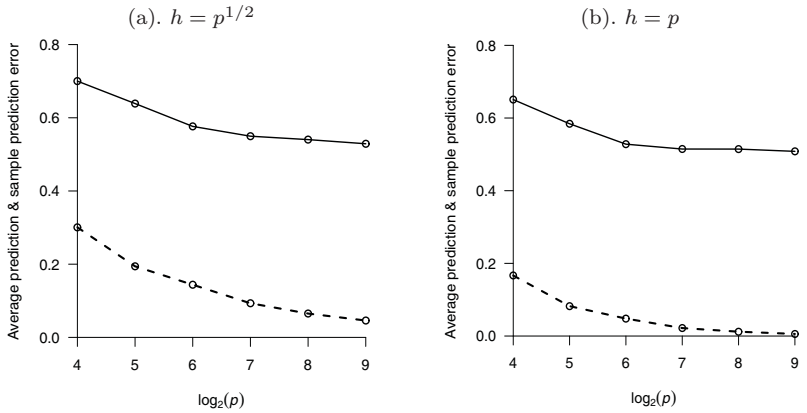

FIG 6. *Average prediction error for* $\widehat{\beta}_{\widehat{\Sigma}_-}$ *(solid) and average sample prediction error for* $\widehat{\beta}_{\widehat{\Sigma}_-}$ *(dashes), using the forward regression simulation based on 100 replications with* $n = p/2$.

The estimators $\widehat{\beta}_{\widehat{\Sigma}^{-1}}$, $\widehat{\beta}_{\widehat{\Delta}}$, and $\widehat{\beta}_{\widehat{\Sigma}_-}$ performed very poorly: most of their average prediction errors were between 10^4 and 10^7. This may not be surprising since in this case $\mathrm{E}(Y|X)$ and $\mathrm{var}(Y|X)$ exist, but the expectation of our performance criterion (7) does not exist since neither $E(X)$ nor $\mathrm{var}(X)$ exists.

5.4. Forward regression simulation

To illustrate Propositions 3.3 and 3.4, we constructed a simulation where $\mathrm{var}(X)$ has bounded eigenvalues as p grows. Specifically, $X \sim N_p(0, I_p)$ and $Y = \beta^T X + \epsilon$, where $\epsilon \sim N_1(0, 1 - \beta^T\beta)$ and $\epsilon \perp\!\!\!\perp X$. This implies $\sigma_Y^2 = 1$, $\beta = \sigma_{XY}$, $\Delta^{-1} = I_p + (1 - \beta^T\beta)^{-1}\beta\beta^T$ and $h = (1 - \beta^T\beta)^{-1}\beta^T\beta$. For $h = p^{1/2}$ and p, we set all elements of β equal to $[\{p(h+1)\}^{-1}h]^{1/2}$, and used $p = 16, 32, 64, 128, 256$, and 512 with $n = p/2$.

In addition to the prediction error, we measured performance with the sample prediction error, defined by (7), where the predictions were at 1000 independent copies of $X_N \sim N_p(\bar{X}, \widehat{\Sigma})$. We plotted the average prediction and sample prediction erros for $\widehat{\beta}_{\widehat{\Sigma}_-}$ in Fig. 6a for $h = p^{1/2}$ and in Fig. 6b. for $h = p$. The prediction error curve appears inconsistent and bounded as Proposition 3.3 guarantees and the sample prediction error curve appears consistent as Proposition 3.4 guarantees.

6. Data analysis

6.1. Overview

We illustrate an abundant regression with data introduced by Sæbø et al. (2007), where the percentage of fat in beef or pork samples is predicted with absorbance

spectral measurements. There are $p = 100$ wavelengths and $n = 103$ cases of which 54 are pork samples.

6.2. Pork samples

The pork samples provide an illustration of a regression with $p > n$. Many regularized regression procedures could be applied, including penalized least-squares, principal component regression, and partial least squares. Such alternative methods could perform well depending on characteristics of the regression. For instance, penalized least squares may be appropriate when β is sparse, but we do not anticipate sparsity for these data, as discussed in §6.3. Partial least squares might perform well when σ_{XY} lies in the span of the first few eigenvectors of Σ, although Chung and Keleş (2010) showed recently that its estimator of the coefficient vector in the linear regression of Y on X is inconsistent unless $p/n \to 0$.

We restrict comparisons of our proposed methods to the standard estimator $\widehat{\beta}_{\widehat{\Sigma}^-}$ since it is the default option in many software packages when $p \geq n$. We performed leave-one-out cross validation to compare the prediction performance of $\widehat{\beta}_{\widehat{\Delta}}$ and $\widehat{\beta}_{\widehat{\Sigma}^-}$. Due to numerical instability, $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda})}$ could not be computed for values of λ that cross-validation recommended. To circumvent this instability, we considered a similar estimator $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda},R)}$ obtained by replacing $\lambda \sum_{i \neq j} |\theta_{ij}|$ in (4) with $\lambda \sum_{i \neq j} \theta_{ij}^2$ (Rothman et al., 2008). For each excluded case, we selected the tuning parameter for $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda},R)}$ using 5-fold cross-validation on the remaining 53 cases, where validation prediction error was minimized and the optimal tuning parameter value was selected from $\{10^{-10+0.5j} : j = 0, 1, \ldots 24\}$. There were no selections on the boundary of this set.

The average squared prediction error, computed from the 54 left out cases, was 5.60 for $\widehat{\beta}_{\widehat{\Sigma}^-}$, 3.46 for $\widehat{\beta}_{\widehat{\Delta}}$ and 2.79 for $\widehat{\beta}_{\widehat{\Delta}(\hat{\lambda},R)}$. These prediction errors are represented with boxplots in Fig. 7a.

6.3. Pork and beef samples

Analysis of both the pork and beef samples illustrates a regression where $p \approx n$. We exclude the shortest and longest three wavelengths from the analysis to avoid the moment issues mentioned at the end of §3.2, leaving $p = 94$ and $n = 103$. The ordinary least squares fit is excellent in this case; the fit is represented with a response versus fitted values plot in Fig. 7b.

Our interest is to investigate the presence of abundance. Let a subscript of $[j]$ indicate that an estimate is based on the first j predictors and let $\widehat{\beta} = \widehat{\Sigma}^{-1}\widehat{\sigma}_{XY}$. We will see how the estimate $\hat{h}(j) = \widehat{\beta}_{[j]}^T \widehat{\Sigma}_{[j]} \widehat{\beta}_{[j]} / (\hat{\sigma}_{Y[j]}^2 - \widehat{\beta}_{[j]}^T \widehat{\Sigma}_{[j]} \widehat{\beta}_{[j]})$ increases as j increases from 1 to p. Although these spectral predictors have a natural ordering, we will also consider random orderings in our investigation. To establish a benchmark comparison, within each of 500 replications we randomly selected $\alpha \times 100$ percent of the predictors and randomly permuted their case orderings. In this way the selected predictors with permuted case orderings had

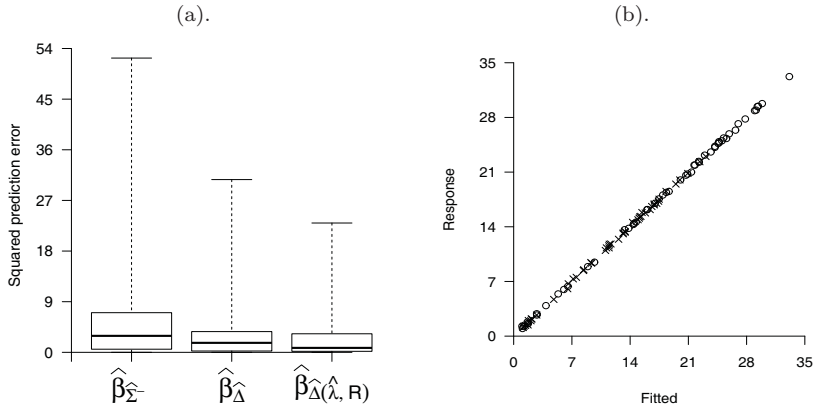

FIG 7. *Boxplots of the 54 squared prediction errors for the pork samples (a) and observed responses versus fitted values for both pork (circles) and beef (×'s) samples (b).*

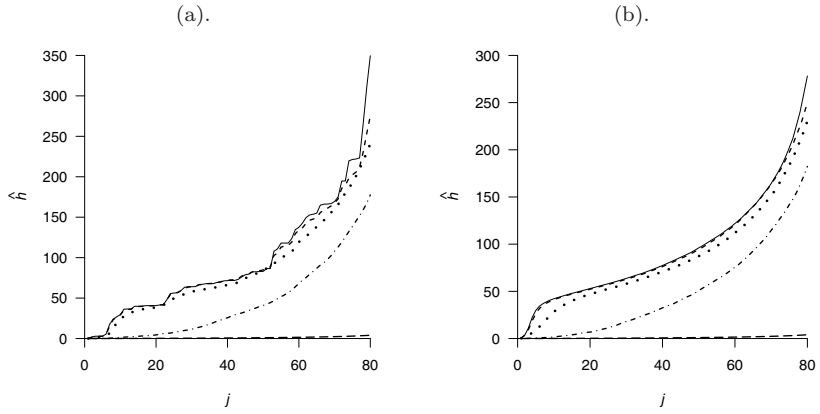

FIG 8. $\hat{h}(j)$ *(solid),* $\hat{h}_{0.1}(j)$ *(dashes),* $\hat{h}_{0.5}(j)$ *(dots),* $\hat{h}_{0.9}(j)$ *(dash–dot), and* $\hat{h}_1(j)$ *(long dashes) versus* j *when the predictors are ordered with increasing wavelength (a) and randomly ordered (b).*

no contribution to the regression and only the remaining $(1-\alpha)\times 100$ percent of the predictors could be relevant. The estimated signal rates $\hat{h}(j)$, $(j = 1, \ldots, p)$ were then computed within each replication. We report $\hat{h}_\alpha(j)$, defined as the average $\hat{h}(j)$ over the 500 replications with case orderings permuted for $\alpha \times 100$ percent of the predictors.

Using the natural predictor ordering, we plotted $\hat{h}(j)$, $\hat{h}_{0.1}(j), \hat{h}_{0.5}(j), \hat{h}_{0.9}(j)$, and $\hat{h}_1(j)$ versus j in Fig. 8a. We only show $j = 1, \ldots, 80$ for ease of illustration. It is clear that $\hat{h}(j)$ is growing rapidly as j increases indicating that an abundant signal is plausible in this regression. Permuting 10 percent of the predictors' case orderings has only a small effect on the signal rate estimate, while permuting 90 and 100 percent strongly attenuates the estimated signal rate.

We randomly permuted the predictors within each of 500 replications and plot the average value of $\hat{h}(j)$ versus j as well as $\hat{h}_{0.1}(j), \hat{h}_{0.5}(j), \hat{h}_{0.9}(j)$, and $\hat{h}_1(j)$ versus j in Fig. 8b. We again see the same pattern as with the natural ordering.

7. Discussion

A referee requested that we contrast our approach with that in an arXiv paper by Dicker (2012), who considered four methods of prediction in high dimensional linear models, including the ordinary least squares estimator $\widehat{\beta} = \widehat{\Sigma}^{-}\widehat{\sigma}_{XY}$ that we addressed in Propostions 3.2 and 3.3. Although stated a bit differently, our results in Proposition 3.2 are consistent with those in Dicker's Proposition 2 for the case when $n > p + 2$ (Dicker's $d = p + 1$). However, our results in Proposition 3.3 do not agree with those in Dicker's Proposition 2 for the case when $n < p$. In developing his result for $n < p$, Dicker claimed in effect that, for any full rank matrix $A \in \mathbb{R}^{p\times p}$,

$$A^{-1}\widehat{\Sigma}^{-}\widehat{\sigma}_{XY} = (A^T\widehat{\Sigma}A)^{-}A^T\widehat{\sigma}_{XY}. \tag{9}$$

Choosing $A = \Sigma$ enabled Dicker to restrict attention to the case $\Sigma = I_p$ without loss of generality. Equation (9) holds when $\widehat{\Sigma} > 0$, but not generally otherwise because the Moore-Penrose inverse is not equivariant; that is, $(A^T\widehat{\Sigma}A)^{-} \neq A^{-1}\widehat{\Sigma}^{-}A^{-T}$ (Cook and Forzani, 2011). Consequently, we find Dicker's result for $n < p$ to be of uncertain value. Dicker also considered James-Stein and Ridge estimators, which we did not evaluate since they are outside the scope of this article.

Acknowledgments

Research for this work was supported in part by National Science Foundation grants DMS-10-07547 (RDC) and DMS-11-05650 (AJR and RDC). The authors thank the Editors and referees for helpful comments on an earlier version of this article.

Appendix A: Preliminary results

In this appendix we give a series of preliminary results that will be used in the proofs of Appendix B. All notation is as provided in the body of the paper unless indicated otherwise. We also use additional notation: $a_n \asymp_p b_n$ means that $a_n = O_p(b_n)$ and $b_n = O_p(a_n)$, $W_p(\Sigma, d)$ stands for the p-dimensional Wishart distribution with covariance matrix Σ and d degrees of freedom, and χ^2_d stands for the chi-squared distribution with d degrees of freedom. We let F denote the $n \times p$ matrix with rows $(X_i - \bar{X})^T$, $(i = 1, \ldots, n)$, let $K_{p,p}$ denote the $p \times p$ commutation matrix, let ε denote the $n \times 1$ vector with elements consisting of the errors ϵ_i from model (1), so that $\varepsilon \sim N_n(0, \sigma^2_\epsilon I_p)$, and let vec : $\mathbb{R}^{p\times q} \mapsto \mathbb{R}^{pq}$

denote the operator that maps a matrix to a vector by stacking its columns. We used Z as local notation whose definition varies depending on the proof being developed.

The following algebraic relationships follow from the definition of $h(p) = \sigma_{XY}^T\Delta^{-1}\sigma_{XY}/\sigma_Y^2$ and from the identities $\beta = \Sigma^{-1}\sigma_{XY}$, $\sigma_Y^2 = \sigma_\epsilon^2 + \beta^T\Sigma\beta$ and $\Sigma = \Delta + \sigma_{XY}\sigma_{XY}^T/\sigma_Y^2$, which implies that

$$\begin{aligned}
\Sigma^{-1} &= \Delta^{-1} - \Delta^{-1}\sigma_{XY}\sigma_{XY}^T\Delta^{-1}/(\sigma_Y^2 + \sigma_{XY}^T\Delta^{-1}\sigma_{XY}).\\
\operatorname{tr}(\Sigma\Delta^{-1}) &= p + h(p) \asymp p, \qquad (10)\\
\operatorname{tr}(\Delta^{-1/2}\Sigma\Delta^{-1/2})^2 &\asymp p + h^2(p), \qquad (11)\\
\beta^T\Sigma\Delta^{-1}\Sigma\beta &= h(p)\sigma_Y^2 \asymp h(p), \qquad (12)\\
\beta^T\Sigma\Delta^{-1}\Sigma\Delta^{-1}\Sigma\beta &= \sigma_{XY}^T\Delta^{-1}\Sigma\Delta^{-1}\sigma_{XY} \asymp h^2(p) \qquad (13)\\
\sigma_{XY}^T\Sigma^{-1}\sigma_{XY} &= \sigma_{XY}^T\Delta^{-1}\sigma_{XY} - (\sigma_{XY}^T\Delta^{-1}\sigma_{XY})^2/(\sigma_Y^2 + \sigma_{XY}^T\Delta^{-1}\sigma_{XY})\\
&= \sigma_Y^2 h(p)/(1 + h(p))\\
\sigma_\epsilon^2 &= \sigma_Y^2 - \beta^T\Sigma\beta\\
&= \sigma_Y^2 - \sigma_{XY}^T\Sigma^{-1}\sigma_{XY}\\
&= \sigma_Y^2/(1 + h(p)).
\end{aligned}$$

As a consequence,

$$\sigma_\epsilon^2 = \sigma_{XY}^T\Sigma^{-1}\sigma_{XY}/h(p) = \beta^T\Sigma\beta/h(p) \asymp h^{-1}(p). \qquad (14)$$

The next series of results gives various moments involving $\widehat{\sigma}_{XY}$. All operators are with respect to the joint distribution of X and Y.

$$\begin{aligned}
\mathrm{E}\,(\widehat{\sigma}_{XY}) &= \sigma_{XY} \qquad (15)\\
\operatorname{var}(\widehat{\sigma}_{XY}) &= (\sigma_Y^2\Sigma + \sigma_{XY}\sigma_{XY}^T)/(n-1) \qquad (16)\\
\mathrm{E}\,(\widehat{\sigma}_{XY}\widehat{\sigma}_{XY}^T) &= (\sigma_Y^2\Sigma + n\sigma_{XY}\sigma_{XY}^T)/(n-1) \qquad (17)\\
\mathrm{E}\,(\widehat{\sigma}_{XY}^T\Delta^{-1}\widehat{\sigma}_{XY}) &= \sigma_Y^2\{h(p)(n+1) + p\}/(n-1) \qquad (18)\\
\operatorname{var}(\widehat{\sigma}_{XY}^T\Delta^{-1}\widehat{\sigma}_{XY}) &\asymp h^2(p)/n. \qquad (19)\\
h^{-1}(p)\mathrm{E}\,(\widehat{\sigma}_{XY}^T\Delta^{-1}\widehat{\sigma}_{XY} - \sigma_{XY}^T\Delta^{-1}\sigma_{XY}) &\asymp \kappa^2, \qquad (20)
\end{aligned}$$

with $\kappa^2 = p/\{nh(p)\}$.

Proofs of relationships (15)–(20) involve the moments of the Wishart matrix $W = (n-1)\widehat{\Sigma} = F^TF \sim W_p(\Sigma, n-1)$: $\mathrm{E}(W) = (n-1)\Sigma$ and $\operatorname{var}(W) = (n-1)(I_{p^2} + K_{p,p})(\Sigma\otimes\Sigma)$ (Magnus and Neudecker, 1979, Corollary 4.2). Results (15) and (16) follow from the identity $\widehat{\sigma}_{XY} = (n-1)^{-1}(W\beta + F\varepsilon)$, the independence of F and ε, $\beta = \Sigma^{-1}\sigma_{XY}$ and $\sigma_Y^2 = \sigma_\epsilon^2 + \beta^T\Sigma\beta$. These imply that $\mathrm{E}(\widehat{\sigma}_{XY}) = \mathrm{E}(W\beta)/(n-1) = \sigma_{XY}$, and

$$\begin{aligned}
\operatorname{var}(\widehat{\sigma}_{XY}) &= \{\operatorname{var}(W\beta) + \sigma_\epsilon^2\mathrm{E}\,(W)\}/(n-1)^2\\
&= \{(\beta^T\otimes I_p)\operatorname{var}(\operatorname{vec}(W))(\beta\otimes I_p) + \sigma_\epsilon^2\mathrm{E}\,(W)\}/(n-1)^2
\end{aligned}$$

$$
\begin{aligned}
&= \{(\beta^T \otimes I_p)(I_{p^2} + K_{p,p})(\Sigma \otimes \Sigma)(\beta \otimes I_p) + \sigma_\epsilon^2 \Sigma\}/(n-1) \\
&= \{(\beta^T \Sigma \beta + \sigma_\epsilon^2)\Sigma + \Sigma \beta \beta^T \Sigma\}/(n-1),
\end{aligned}
$$

where the final step, which implies (16), makes use of properties of $K_{p,p}$ from Magnus and Neudecker (1979, Theorem 3.1).

Result (17) is a direct consequence of (15) and (16), and (18) follows from (17) using that the trace is a cyclic operation, the definition of $h(p)$ and (10). Result (20) follows from (18) and the definition of $h(p)$.

We now justify the final relationship (19). Let

$$
U = \Delta^{-1/2} W \Delta^{-1/2} \sim W_p(\Delta^{-1/2} \Sigma \Delta^{-1/2}, n-1).
$$

The mean and variance of a quadratic form in ε are given by (Magnus and Neudecker, 1979, Corollary 4.1)

$$
\begin{aligned}
\mathrm{E}(\varepsilon^T \Lambda \varepsilon) &= \sigma_\epsilon^2 \mathrm{tr}\{\Lambda\} \\
\mathrm{var}(\varepsilon^T \Lambda \varepsilon) &= 2\sigma_\epsilon^4 \mathrm{tr}\{\Lambda^2\},
\end{aligned}
$$

where Λ is an n-dimensional symmetric matrix, and $\mathrm{var}(\lambda^T \varepsilon) = \sigma_\epsilon^2 \lambda^T \lambda$ for $\lambda \in \mathbb{R}^n$. Using these moments, the independence of F and ε, writting $\widehat{\sigma}_{XY} = (n-1)^{-1}(W\beta + F\varepsilon)$, and letting $V = (n-1)^4 \mathrm{var}(\widehat{\sigma}_{XY}^T \Delta^{-1} \widehat{\sigma}_{XY})$ and $G = F^T \Delta^{-1} F$, we have

$$
\begin{aligned}
V &= (n-1)^4 \{\mathrm{var}_X \mathrm{E}_{\varepsilon|X}\left(\widehat{\sigma}_{XY}^T \Delta^{-1} \widehat{\sigma}_{XY}\right) + \mathrm{E}_X \mathrm{var}_{\varepsilon|X}\left(\widehat{\sigma}_{XY}^T \Delta^{-1} \widehat{\sigma}_{XY}\right)\} \\
&= \mathrm{var}_X\left\{\sigma_\epsilon^2 \mathrm{tr}\{G\} + \beta^T W^T \Delta^{-1} W \beta\right\} + \mathrm{E}_X\left\{2\sigma_\epsilon^4 \mathrm{tr}\{G^2\}\right. \\
&\quad \left. + 4\,\sigma_\epsilon^2 \beta^T W^T \Delta^{-1} F F^T \Delta^{-1} W \beta\right\} \\
&= \mathrm{var}\{\mathrm{tr}(\sigma_\epsilon^2 U + \beta^T \Delta^{1/2} U^2 \Delta^{1/2} \beta)\} + 2\sigma_\epsilon^4 \mathrm{E}\{\mathrm{tr}(U^2)\} \\
&\quad + 4\sigma_\epsilon^2 \mathrm{E}(\beta^T \Delta^{1/2} U^3 \Delta^{1/2} \beta) \\
&= I + II + III, \qquad (21)
\end{aligned}
$$

where the term labels – I, II and III – are defined implicitly. Using the moments of a Wishart matrix, it can next be shown that $I/(n-1)^4 \asymp h^2(p)/n$, $II/(n-1)^4 \asymp 1/n$ and $III/(n-1)^4 \asymp h(p)/n$. The conclusion follows because $h(p)$ is a monotonically increasing function of p. We conclude our discussion of (19) by giving the details on the orders of the terms in (21).

Using an expression for $\mathrm{E}(U^3)$ from Letac and Massan (2004, page 295), we have

$$
\begin{aligned}
\frac{III}{(n-1)^4} &\asymp \sigma_\epsilon^2 n^{-3} [\{\mathrm{tr}^2(\Delta^{-1/2} \Sigma \Delta^{-1/2}) + n\mathrm{tr}(\Delta^{-1/2} \Sigma \Delta^{-1/2})^2\} \beta^T \Sigma \beta \\
&\quad + n\beta^T \Sigma \Delta^{-1} \Sigma \beta \mathrm{tr}(\Delta^{-1/2} \Sigma \Delta^{-1/2}) + n^2 \beta^T \Sigma \Delta^{-1} \Sigma \Delta^{-1} \Sigma \beta] \\
&\asymp n^{-1} h(p),
\end{aligned}
$$

where the final relationship follows from the definition of $h(p)$, (10)–(14) and $p/\{h(p)n\} = O(1)$. The calculations for the term $II/(n-1)^4$ follow similarly.

Using an expression for $\mathrm{E}(U^2)$ from Letac and Massan (2004, page 308), we have

$$\begin{aligned}\frac{II}{(n-1)^4} &\asymp \sigma_\epsilon^4 n^{-3}\left[\mathrm{tr}^2(\Delta^{-1/2}\Sigma\Delta^{-1/2}) + n\mathrm{tr}(\Delta^{-1/2}\Sigma\Delta^{-1/2})^2\right] \\ &\asymp 1/n,\end{aligned}$$

where the final relationship follows from (10), (11), (14) and $p/\{h(p)n\} = O(1)$.

The evaluation of the term $I/(n-1)^4$ is more involved. For this term it is sufficient to consider the orders of $I_1 = (n-1)^{-4}\mathrm{var}\{\mathrm{tr}(\sigma_\epsilon^2 U)\}$ and $I_2 = (n-1)^{-4}\mathrm{var}(\beta^T\Delta^{1/2}U^2\Delta^{1/2}\beta)$. Now, $I_1 = \sigma_\epsilon^4(n-1)^{-4}\mathrm{vec}^T(I_p)\mathrm{var}\{\mathrm{vec}(U)\}\mathrm{vec}(I_p)$. Using an expression for $\mathrm{var}\{\mathrm{vec}(U)\}$ from Magnus and Neudecker (1979, Theorem 4.4), it follows that $I_1 = O(1/n)$, again using (11), (14) and the fact that $p/\{h(p)n\} = O(1)$. Let $Z = \Sigma^{-1/2}X$, $\sigma_{ZY} = \mathrm{cov}(Z, Y)$, $W_I \sim W_p(I_p, n-1)$ and σ_0 be an orthonormal basis for $\mathrm{span}^\perp(\sigma_{XY})$. For the term I_2, from $\Delta = \Sigma - \sigma_{XY}\sigma_{XY}^T/\sigma_Y^2$ and $\sigma_Y^2 = \sigma_\epsilon^2 + \beta^T\Sigma\beta = \sigma_\epsilon^2 + \sigma_{XY}^T\Sigma^{-1}\sigma_{XY}$, we have

$$\begin{aligned}\Delta^{-1} &= \Sigma^{-1} - \Sigma^{-1}\sigma_{XY}\sigma_{XY}^T\Sigma^{-1}/(-\sigma_Y^2 + \sigma_{XY}^T\Sigma^{-1}\sigma_{XY}) \\ &= \Sigma^{-1} + \Sigma^{-1/2}\sigma_{ZY}\sigma_{ZY}^T\Sigma^{-1/2}/\sigma_\epsilon^2.\end{aligned}$$

This allows us to express

$$\begin{aligned}\beta^T\Delta^{1/2}U^2\Delta^{1/2}\beta &= \beta^T\Delta^{1/2}U\Delta^{1/2}\Delta^{-1}\Delta^{1/2}U\Delta^{1/2}\beta \\ &= \beta^T\Sigma^{1/2}W_I\Sigma^{1/2}\Delta^{-1}\Sigma^{1/2}W_I\Sigma^{1/2}\beta \\ &= \sigma_{ZY}^T W_I\Sigma^{1/2}\Delta^{-1}\Sigma^{1/2}W_I\sigma_{ZY} \\ &= \sigma_{ZY}^T W_I^2\sigma_{ZY} + (\sigma_{ZY}^T W_I\sigma_{ZY})^2/\sigma_\epsilon^2.\end{aligned}$$

Now, let $P = \sigma_{ZY}\sigma_{ZY}^T/||\sigma_{ZY}||^2$ be the projection onto σ_{ZY} and let $Q = I_p - P = \sigma_0\sigma_0^T$ denote the orthogonal projection. Then,

$$\begin{aligned}\sigma_{ZY}^T W_I^2\sigma_{ZY} &= \sigma_{ZY}^T W_I(P+Q)W_I\sigma_{ZY} \\ &= (\sigma_{ZY}^T W_I\sigma_{ZY})^2/||\sigma_{ZY}||^2 + \sigma_{ZY}^T W_I\sigma_0\sigma_0^T W_I\sigma_{ZY}\end{aligned}$$

and then since $||\sigma_{ZY}||^2 = \beta^T\Sigma\beta$,

$$\beta^T\Delta^{1/2}U^2\Delta^{1/2}\beta = (\sigma_{ZY}^T W_I\sigma_{ZY})^2\{(\beta^T\Sigma\beta)^{-1} + \sigma_\epsilon^{-2}\} + \sigma_{ZY}^T W_I\sigma_0\sigma_0^T W_I\sigma_{ZY}.$$

We next make use of certain orthogonality relations. Write $W_I = NN^T$ where $N \in \mathbb{R}^{p\times(n-1)}$ is a matrix of independent standard normal variates. Define $N_1 = N^T\sigma_{ZY}/(\beta^T\Sigma\beta)^{1/2} \in \mathbb{R}^{(n-1)}$ and $N_2 = N^T\sigma_0 \in \mathbb{R}^{(n-1)\times(p-1)}$ so that N_1 and N_2 are independent and each comprised of independent standard normal variates. Then we have

$$\begin{aligned}\beta^T\Delta^{1/2}U^2\Delta^{1/2}\beta &= (\beta^T\Sigma\beta)^2(N_1^T N_1)^2\{(\beta^T\Sigma\beta)^{-1} + \sigma_\epsilon^{-2}\} \\ &\quad + (\beta^T\Sigma\beta)N_1^T N_2 N_2^T N_1 \\ &= T_1 + T_2.\end{aligned}$$

To study the order of $\mathrm{var}(I_2)$ is enough to study the order of the variances of T_1 and T_2. Since $N_1^T N_1 \sim \chi^2_{n-1}$, $\mathrm{E}(N_1^T N_1)^r \asymp n^r$, $\mathrm{var}(N_1^T N_1) = 2(n-1)$ and $\mathrm{var}(N_1^T N_1)^2 \asymp n^3$. As a consequence, using (14) and $\sigma_Y^2 = \sigma_\epsilon^2 + \beta^T \Sigma \beta$,

$$\mathrm{var}(T_1) \asymp \{(\beta^T \Sigma \beta)^{-1} + \sigma_\epsilon^{-2}\}^2 n^3 \asymp h(p)^2 n^3.$$

Since $N_1^T N_2 N_2^T N_1 | N_1 \sim N_1^T N_1 \chi^2_{p-1}$ and $p/\{h(p)n\} \asymp O(1)$, we have

$$\begin{aligned} \mathrm{var}(T_2) &= \mathrm{var}_{N_1}(\mathrm{E}(T_2|N_1)) + \mathrm{E}_{N_1}\mathrm{var}(T_2|N_1) \\ &= (\beta^T\Sigma\beta)^2(p-1)^2\mathrm{var}(N_1^T N_1) + 2(\beta^T\Sigma\beta)^2(p-1)\mathrm{E}(N_1^T N_1)^2 \\ &\asymp p^2 n + pn^2 \\ &\asymp h^2(p)n^3. \end{aligned}$$

This implies that the order of I_2 is $h^2(p)/n$ and, together with the order of I_1, the order of $I/(n-1)^4$ is $h^2(p)/n$.

Appendix B: Proofs

Proof of Lemma 2.1. Let ρ_{XY} denote the $p \times 1$ vector of correlations $\rho(X^{(j)}, Y)$, let r_{XY} denote the $p \times 1$ vector with elements $\rho(X^{(j)}, Y)/\{1 - \rho^2(X^{(j)}, Y)\}^{1/2}$ $(j = 1, \ldots, p)$, and recall that $\rho_{XX|Y}$ is the $p \times p$ matrix of conditional predictor correlations given Y. Using the relationship

$$\rho^2(X^{(j)}, Y) = 1 - \mathrm{var}(X^{(j)}|Y)/\mathrm{var}(X^{(j)}),$$

which follows from the joint normality of $(X^{(j)}, Y)$, the signal rate can be expressed as

$$\begin{aligned} h(p) &= \sigma_{XY}^T \Delta^{-1} \sigma_{XY}/\sigma_Y^2 = \rho_{XY}^T \{\mathrm{diag}^{-1/2}(\Sigma)\Delta\mathrm{diag}^{-1/2}(\Sigma)\}^{-1}\rho_{XY} \\ &= \rho_{XY}^T \{\mathrm{diag}^{-1/2}(\Sigma)\mathrm{diag}^{1/2}(\Delta)\rho_{XX|Y}\mathrm{diag}^{1/2}(\Delta)\mathrm{diag}^{-1/2}(\Sigma)\}^{-1}\rho_{XY} \\ &= r_{XY}^T \rho_{XX|Y}^{-1} r_{XY}. \end{aligned}$$

Consequently, $\varphi_{\max}^{-1}(\rho_{XX|Y}) r_{XY}^T r_{XY} \le h(p) \le \varphi_{\min}^{-1}(\rho_{XX|Y}) r_{XY}^T r_{XY}$. □

Proof of Lemma 2.2. For notational convenience we use the subscript 1 to denote X and the subscript 2 to denote the added predictor X^{p+1}. Now, it can be checked that

$$\begin{aligned} \Delta^{-1} &= \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \Delta_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + C^{-1} \begin{pmatrix} \Delta_{11}^{-1}\Delta_{12}\Delta_{12}^T\Delta_{11}^{-1} & -\Delta_{11}^{-1}\Delta_{12} \\ -\Delta_{12}^T\Delta_{11}^{-1} & 1 \end{pmatrix}, \end{aligned} \quad (22)$$

where $\Delta_{22} = \mathrm{var}(X^{p+1}|Y)$ is a scalar, $\Delta_{11} = \mathrm{var}(X|Y) \in \mathbb{R}^{p\times p}$, and

$$C = \Delta_{22} - \Delta_{21}\Delta_{11}^{-1}\Delta_{12} = \Delta_{22}(1 - R^2_{21|Y})$$

with $R_{21|Y}$ being the multiple correlation coefficient from the regression of X^{p+1} on X given Y. To get $h(p+1)$ we need to pre and post multiply Δ^{-1} by $(\sigma_{1Y}^T, \sigma_{2Y})/\sigma_Y$ and its transpose. Using (22) after some simplifications we get

$$\begin{aligned} h(p+1) &= h(p) + \frac{(\sigma_{1Y}^T \Delta_{11}^{-1} \Delta_{12} - \sigma_{2Y})^2}{\sigma_Y^2 \Delta_{22}(1 - R_{21|Y}^2)} \\ &= h(p) + \frac{\sigma_Y^2}{\Delta_{22}} \frac{\{(\sigma_{1Y}^T \Delta_{11}^{-1} \Delta_{12} - \sigma_{2Y})/\sigma_Y^2\}^2}{(1 - R_{21|Y}^2)}. \end{aligned}$$

The lemma follows since

$$\begin{aligned} \mathrm{E}(X_2|X_1, Y) = \mathrm{E}(X_2) &+ \Delta_{21}\Delta_{11}^{-1}\{X_1 - \mathrm{E}(X_1)\} \\ &- ((\Delta_{21}\Delta_{11}^{-1}\sigma_{1Y} - \sigma_{2Y})/\sigma_Y^2)\{Y - \mathrm{E}(Y)\} \end{aligned}$$

and thus the term $\{(\sigma_{1Y}^T \Delta_{11}^{-1} \Delta_{12} - \sigma_{2Y})/\sigma_Y^2\}^2$ is the squared coefficient of Y in the population regression of X_2 on X_1 and Y. □

In Propositions 3.1, 3.2, 3.3, 4.1 and 4.2 we studied the order of V under model (1) for a new independent $X_N \sim N(\mu_X, \Sigma)$. In order to do that we write

$$\begin{aligned} D_N &= \widehat{\beta}^T(X_N - \bar{X}) - \beta^T(X_N - \mu_X) \\ &= (\widehat{\beta} - \beta)^T(X_N - \mu_X) + \widehat{\beta}^T(\mu_X - \bar{X}) \\ &= I + II. \end{aligned} \tag{23}$$

Squaring D_N and expanding to get $V = \mathrm{E}(D_N^2)$, the cross product term have mean 0 because X_N is a new independent observation. Now, since $\bar{X}$ is independent of $\widehat{\beta}$,

$$\begin{aligned} V &= E\{(\widehat{\beta} - \beta)^T(X_N - \mu_X) - \widehat{\beta}^T(\bar{X} - \mu_X)\}^2 \\ &= E\{(\widehat{\beta} - \beta)^T\Sigma(\widehat{\beta} - \beta)\} + \frac{1}{n}E\{\widehat{\beta}^T\Sigma\widehat{\beta}\} \qquad (24) \\ &= I + II. \qquad (25) \end{aligned}$$

Proof of Proposition 3.1. In this case $\widehat{\beta} = \Sigma^{-1}\widehat{\sigma}_{XY}$ with Σ known. Re-expressing (24) by using (15) and (17),

$$\begin{aligned} V &= E\{(\widehat{\sigma}_{XY} - \sigma_{XY})^T\Sigma^{-1}(\widehat{\sigma}_{XY} - \sigma_{XY})\} + \frac{1}{n}E\{\widehat{\sigma}_{XY}^T\Sigma^{-1}\widehat{\sigma}_{XY}\} \\ &= \frac{n+1}{n(n-1)}\mathrm{tr}\{(\sigma_Y^2 I_p + n\Sigma^{-1}\sigma_{XY}\sigma_{XY}^T)\} - \sigma_{XY}^T\Sigma^{-1}\sigma_{XY} \\ &= \frac{n+1}{n(n-1)}(\sigma_Y^2 p + 2\sigma_{XY}^T\Sigma^{-1}\sigma_{XY}) \\ &= \frac{n+1}{n(n-1)}(\sigma_Y^2 p + 2\beta^T\Sigma\beta). \end{aligned}$$

Since $\beta^T\Sigma\beta$ is bounded, the conclusion follows. □

For Proposition 3.2 and 3.3, $\widehat{\beta} = W^- W\beta + W^- F^T\varepsilon$, with $W = F^T F \sim W_p(\Sigma, n-1)$. Substituting $\widehat{\beta}$ into the first term I of V from (25), we have

$$\begin{aligned} I &= \mathrm{E}\{\beta^T(I_p - W^- W)\Sigma(I_p - W^- W)\beta\} + \mathrm{E}\{\mathrm{tr}(FW^-\Sigma W^- F^T)\}\sigma_\epsilon^2 \\ &= \mathrm{E}\{\beta^T(I_p - W^- W)\Sigma(I_p - W^- W)\beta\} + \mathrm{tr}\{\Sigma\mathrm{E}(W^-)\}\sigma_\epsilon^2. \end{aligned}$$

In a similar way we have $n\,II = \mathrm{E}(\beta^T WW^-\Sigma W^- W\beta) + \mathrm{tr}\{\Sigma\mathrm{E}(W^-)\}\sigma_\epsilon^2$. Plugging I and II in (25),

$$\begin{aligned} V &= \mathrm{E}\{\beta^T(I_p - W^- W)\Sigma(I_p - W^- W)\beta\} + (1+n^{-1})\mathrm{tr}\{\Sigma\mathrm{E}(W^-)\}\sigma_\epsilon^2 \\ &\quad + n^{-1}\mathrm{E}(\beta^T WW^-\Sigma W^- W\beta) = T_1 + T_2 + T_3. \end{aligned} \tag{26}$$

where the three terms – T_1, T_2 and T_3 – are defined implicitly.

Proof of Proposition 3.2. Since $n > p+2$, $W^- = W^{-1}$. This implies that $T_1 = 0$ and $T_3 = n^{-1}\beta^T\Sigma\beta = O(n^{-1})$ since $\beta^T\Sigma\beta$ is bounded. Term T_2 involves $\mathrm{tr}\{\Sigma\mathrm{E}(W^-)\} = \mathrm{tr}\{\mathrm{E}(W_I^{-1})\}$, where $W_I \sim W_p(I_p, n-1)$. Using Theorem 3.1 from von Rosen (1988), $\mathrm{E}(W_I^{-1}) = (n-p-2)^{-1}I_p$ and thus $T_2 = \sigma_\epsilon^2 p(n+1)/n(n-p-2)$, which implies the desired conclusion from (14). □

Proof of Proposition 3.3. In the justification of this result we use bounding and the first moment $\mathrm{E}(W_I^-) = (n-1)/\{p(p-n)I_p\}$ (Cook and Forzani, 2011, Theorem 3.1). Let $\varphi_{\max}$ and $\varphi_{\min}$ be the largest and smallest eigenvalues of Σ. In the following all inequalities become equalities when $\Sigma = I_p$, so $\varphi_{\max} = \varphi_{\min} = 1$.

Write $W = \Sigma^{1/2}ZZ^T\Sigma^{1/2}$, where $Z \in \mathbb{R}^{p\times(n-1)}$ is a matrix of independent standard normal random variables. Then the matrices W^-, WW^- and $WW^-\Sigma WW^-$ can be bounded above as follows

$$\begin{aligned} W^- &= \Sigma^{1/2}Z(Z^T\Sigma Z)^{-2}Z^T\Sigma^{1/2} \\ &\le \varphi_{\min}^{-2}\Sigma^{1/2}Z(Z^TZ)^{-2}Z^T\Sigma^{1/2} \\ &= \varphi_{\min}^{-2}\Sigma^{1/2}W_I^-\Sigma^{1/2} \end{aligned}$$

and therefore

$$\mathrm{E}(W^-) \le \varphi_{\min}^{-2}\frac{n-1}{p(p-n)}\Sigma.$$

Now,

$$\begin{aligned} WW^- &= \Sigma^{1/2}ZZ^T\Sigma Z(Z^T\Sigma Z)^{-2}Z^T\Sigma^{1/2} \\ &= \Sigma^{1/2}Z(Z^T\Sigma Z)^{-1}Z^T\Sigma^{1/2} \qquad (27) \\ &\le \varphi_{\min}^{-1}\Sigma^{1/2}Z(Z^TZ)^{-1}Z^T\Sigma^{1/2}. \\ WW^-\Sigma WW^- &= \Sigma^{1/2}Z(Z^T\Sigma Z)^{-1}Z^T\Sigma^2 Z(Z^T\Sigma Z)^{-1}Z^T\Sigma^{1/2} \\ &\le \varphi_{\max}WW^-, \qquad (28) \end{aligned}$$

where we used (27) for the last matrix. The distribution of $Z(Z^TZ)^{-1}Z^T$ is invariant under orthogonal transformation of the columns of Z. Consequently,

its expectation is of the form cI_p, and it follows that $c = (n-1)/p$ since $\mathrm{E}(Z(Z^TZ)^{-1}Z^T) = cI_p$ implies $cp = \mathrm{tr}\{\mathrm{E}(Z(Z^TZ)^{-1}Z^T)\} = n-1$. Therefore $\mathrm{E}(WW^-) \le \varphi_{\min}^{-1}(n-1)/p\Sigma$. In the same way, letting $Z_0 \in \mathbb{R}^{p\times(p-n+1)}$ be orthogonal to Z, we get $\mathrm{E}(Z_0(Z_0^TZ_0)^{-1}Z_0^T) = (p-n+1)/pI_p$. Plugging Z_0 into (27),

$$(I_p - WW^-)\Sigma(I_p - WW^-) = \Sigma^{-1/2}Z_0(Z_0^T\Sigma^{-1}Z_0)^{-1}Z_0^TZ_0(Z_0^T\Sigma^{-1}Z_0)^{-1}Z_0^T\Sigma^{-1/2}$$

and therefore we get

$$\mathrm{E}((I_p - WW^-)\Sigma(I_p - WW^-)) \quad \le \quad \varphi_{\max}^2\varphi_{\min}^{-2}(p-n+1)/p\Sigma.$$

Lower bounds can be established similarly by replacing $\varphi_{\min}$ with $\varphi_{\max}$. We next use these bounds to obtain orders for the three term on the right hand side of (26), finding that $T_1 \asymp (p-n+1)/p$, $T_2 \asymp \sigma_\epsilon^2 n/(p-n) \asymp n/\{h(p)(p-n)\}$ by (14) and, using (28) and $\mathrm{tr}\{\Sigma\} \asymp p$, $T_3 \asymp p^{-1}$, which imply the desired conclusions. □

Proof of Proposition 3.4. In this case we write

$$\begin{aligned} D_N &= \widehat{\beta}^T(X_N - \bar{X}) - \beta^T(X_N - \mu_X) \\ &= (\widehat{\beta} - \beta)^T(X_N - \bar{X}) - \beta^T(\bar{X} - \mu_X). \end{aligned}$$

For a new observation $X_N|(\bar{X}, \widehat{\Sigma}) \sim N(\bar{X}, \widehat{\Sigma})$,

$$V = \mathrm{E}\{(\beta - \widehat{\beta})^T\widehat{\Sigma}(\beta - \widehat{\beta})\} + n^{-1}\beta^T\Sigma\beta.$$

Substituting $\widehat{\beta} = W^-W\beta + W^-F\varepsilon$ into the expression for V, leads to

$$\begin{aligned} V &\asymp \mathrm{E}\{\beta^T(I_p - WW^-)\widehat{\Sigma}(I_p - W^-W)\beta\} + \mathrm{E}(\varepsilon^TFW^-\widehat{\Sigma}W^-F^T\varepsilon) + n^{-1}\beta^T\Sigma\beta \\ &\asymp \{nh(p)\}^{-1}\mathrm{E}\{\mathrm{tr}(WW^-)\} + n^{-1}, \end{aligned}$$

where we use that $\widehat{\Sigma} = W/(n-1)$ implies $\widehat{\Sigma}(I_p - WW^-) = 0$, the independence of ε and F to compute the second expectation, $W = FF^T$ and (14). The result is now a consequence of the fact that $\mathrm{tr}\{WW^-\} = \min(p, n-1)$, which is trivial for $n > p$ and it follows from (27) for $n \le p$. □

Proof of Proposition 4.1 and Proposition 4.2. For both Propositions 4.1 and 4.2, we require the order, as $n, p \to \infty$, of D_N from (23) for $\widehat{\beta}$ given in (3) with $\widehat{\Omega}$ as defined in its corresponding propositions. Define

$$\begin{aligned} B &= h^{-1}(p)\{1 + h(p)\} = h^{-1}(p)(1 + \sigma_{XY}^T\Delta^{-1}\sigma_{XY}/\sigma_Y^2) \text{ and} \\ \widehat{B} &= h^{-1}(p)\{1 + \hat{h}(p)\} = h^{-1}(p)(1 + \widehat{\sigma}_{XY}^T\widehat{\Omega}\widehat{\sigma}_{XY}/\hat{\sigma}_Y^2). \end{aligned}$$

The analysis is facilitated by using the representation $\widehat{\beta} = \widehat{B}^{-1}\widehat{\Omega}\widehat{\sigma}_{XY}/h(p)$. $B \asymp 1$, and we will show later that

$$\widehat{B} - B = O_p(\kappa^2 + n^{-1/2}) \text{ under the conditions of Proposition 4.1 and} \quad (29)$$

$$\widehat{B} - B = O_p(\omega + \kappa^2 + n^{-1/2}) \text{ under the conditions of Proposition 4.2,} \quad (30)$$

and as a consequence in both cases $\widehat{B} \asymp_p 1$.

Turning to the first term I of the representation of D_N from (23), we have letting $\varepsilon_N = X_N - \mu_X \sim N(0, \Sigma)$,

$$\begin{aligned} I &= h^{-1}\widehat{B}^{-1}(\widehat{\sigma}_{XY}^T\widehat{\Omega} - \sigma_{XY}^T\Delta^{-1})\varepsilon_N + h^{-1}\widehat{B}^{-1}(B - \widehat{B})B^{-1}\sigma_{XY}^T\Delta^{-1}\varepsilon_N \\ &= I_1 + I_2. \end{aligned} \quad (31)$$

Using $\text{var}(\varepsilon_N) = \Sigma$ and (13) it follows $\text{var}(h^{-1}(p)\sigma_{XY}^T\Delta^{-1}\varepsilon_N) \asymp 1$ and thus $I_2 = O_p(\kappa^2 + n^{-1/2})$ under Proposition 4.1 and $I_2 = O_p(\omega + \kappa^2 + n^{-1/2})$ under Proposition 4.2. To find the order of I_1 we represent it in three terms, $I_1 = I_{11} + I_{12} + I_{13}$, as $I_{11} = h^{-1}(p)(\widehat{\sigma}_{XY} - \sigma_{XY})^T\Delta^{-1}\varepsilon_N$, $I_{12} = h^{-1}(p)\sigma_{XY}^T(\widehat{\Omega} - \Delta^{-1})\varepsilon_N$ and $I_{13} = h^{-1}(p)(\widehat{\sigma}_{XY} - \sigma_{XY})^T(\widehat{\Omega} - \Delta^{-1})\varepsilon_N$. Each term has mean 0, we and compute their orders individually: Using (15), (17), (11) and (13),

$$\begin{aligned} \text{var}(I_{11}) &= h^{-2}(p)\text{E}\{(\widehat{\sigma}_{XY} - \sigma_{XY})^T\Delta^{-1}\Sigma\Delta^{-1}(\widehat{\sigma}_{XY} - \sigma_{XY})\} \\ &= h^{-2}(p)\{\sigma_Y^2\text{tr}(\Delta^{-1}\Sigma)^2 + \sigma_{XY}^T\Delta^{-1}\Sigma\Delta^{-1}\sigma_{XY}\}/(n-1) \\ &\asymp h^{-2}(p)\{p + 2h^2(p)\}/n \asymp \kappa^2/h(p) + n^{-1}, \end{aligned}$$

and therefore $I_{11} \asymp_p h^{-1/2}\kappa + n^{-1/2}$.

For the term I_{12} under Proposition 4.1 we use an expression for $\text{var}\{\text{vec}(\widehat{\Omega})\}$ from Cook and Forzani (2011, Theorem 3.1),

$$\text{var}(\text{vec}(\widehat{\Omega})) \asymp c(I_{p^2} + K_{p,p}) + d\text{vec}(I_p)\text{vec}^T(I_p), \quad (32)$$

where $c \asymp 1/n \asymp 1/p$ and $d \asymp 1/n^2 \asymp 1/p^2$. Then, using the independence of $\widehat{\Omega}$ and ε_N,

$$\begin{aligned} \text{var}(I_{12}) &= h^{-2}(p)\text{var}(\sigma_{XY}^T(\widehat{\Omega} - \Delta^{-1})\varepsilon_N) \\ &= h^{-2}(p)\text{E}\{(\varepsilon_N^T \otimes \sigma_{XY}^T)\text{var}\{\text{vec}(\widehat{\Omega})\}(\varepsilon_N \otimes \sigma_{XY})\} \\ &\asymp h^{-2}(p)p^{-1}\text{E}\{\varepsilon_N^T\varepsilon_N\sigma_{XY}^T\sigma_{XY})\} \\ &\quad + h^{-2}(p)p^{-2}\text{E}\{(\varepsilon_N^T \otimes \sigma_{XY}^T)\text{vec}(I_p)\text{vec}^T(I_p)(\varepsilon_N \otimes \sigma_{XY})\} \\ &\asymp h^{-2}(p)p^{-1}\text{tr}\{\Sigma\}\sigma_{XY}^T\sigma_{XY} \\ &\asymp \kappa^2, \end{aligned}$$

since for this case $\sigma_{XY}^T\sigma_{XY} \asymp h(p)$, $\text{tr}\{\Sigma\} = p + \sigma_{XY}^T\sigma_{XY}/\sigma_Y^2 \asymp p + h(p)$ and $\kappa^2 \asymp h^{-1}(p)$.

For I_{12} under the assumptions of Proposition 4.2, recalling that $S = \Delta^{1/2}(\widehat{\Omega} - \Delta^{-1})\Delta^{1/2}$ with $\|\text{E}(S^2)\| = O(\omega^2)$, we have

$$\begin{aligned} \text{var}(I_{12}) &= h^{-2}\text{E}\{\sigma_{XY}^T(\widehat{\Omega} - \Delta^{-1})\varepsilon_N)\}^2 \\ &= h^{-2}(p)\sigma_{XY}^T\Delta^{-1/2}\text{E}(S\Delta^{-1/2}\Sigma\Delta^{-1/2}S)\Delta^{-1/2}\sigma_{XY} \end{aligned}$$

$$\begin{aligned} &\leq h^{-2}(p)\{1+h(p)\}\sigma_{XY}^T\Delta^{-1/2}\mathrm{E}\left(S^2\right)\Delta^{-1/2}\sigma_{XY} \\ &\leq 2h^{-1}(p)\sigma_{XY}^T\Delta^{-1}\sigma_{XY}\|\mathrm{E}(S^2)\|, \end{aligned}$$

since $\varphi_{\max}(\Delta^{-1/2}\Sigma\Delta^{-1/2}) = 1+h(p)$. Using the definition of $h(p)$ and the order of $\|\mathrm{E}(S^2)\|$, we get $I_{12} = O_p(\omega)$.

It can be shown that the term I_{13} has smaller order than I_{11} and I_{12} and, as a consequence, we have $I = O_p(\kappa)$ under Proposition 4.1 and $I = O_p(\omega + \kappa^2 + h^{-1/2}\kappa + n^{-1/2})$ under Proposition 4.2.

Turning to term II, we have $II = \widehat{B}^{-1}\widehat{\sigma}_{XY}^T\widehat{\Omega}(\mu_X - \bar{X})/h(p)$. The first factor $\widehat{B}^{-1} \asymp_p 1$ and consequently it is sufficient to consider $II_1 = \widehat{\sigma}_{XY}^T\widehat{\Omega}(\mu_X - \bar{X})/h(p)$.

For Proposition 4.1 the conclusion follows since $II_1 = O_p(n^{-1/2})$, which is no greater than $I = O_p(\kappa)$. To see that $II_1 = O_p(n^{-1/2})$ we write $\mathrm{var}(II_1) = \mathrm{E}\{\mathrm{var}(II_1|\widehat{\sigma}_{XY}, \bar{X})\} + \mathrm{var}\{\mathrm{E}(II_1|\widehat{\sigma}_{XY}, \bar{X})\}$ and study the two terms in this decomposition separately. Let $Z = \mu_X - \bar{X}$. Using the independence of $\widehat{\Omega}$, $\widehat{\sigma}_{XY}$ and $\bar{X}$, (32) and (17), and considering that $\mathrm{E}(Z^TZ) = \mathrm{tr}\Sigma/n$, we have

$$\begin{aligned} \mathrm{E}\{\mathrm{var}(II_1|\widehat{\sigma}_{XY}, \bar{X})\} &= \mathrm{E}\{(Z^T \otimes \widehat{\sigma}_{XY}^T)\mathrm{var}(\mathrm{vec}(\widehat{\Omega}))(Z \otimes \widehat{\sigma}_{XY})\}/h^2(p) \\ &\asymp \mathrm{E}\{Z^TZ\widehat{\sigma}_{XY}^T\widehat{\sigma}_{XY})\}/\{ph^2(p)\} \\ &\asymp \mathrm{tr}(\Sigma)\ \mathrm{tr}\{\sigma_Y^2\Sigma + n\sigma_{XY}^T\sigma_{XY}\}/\{pn(n-1)h^2(p)\}, \end{aligned}$$

From the definition of $h(p)$, $\mathrm{tr}(\Sigma) \asymp p$ and $p/\{nh(p)\} \asymp 1$, $\mathrm{E}\{\mathrm{var}(II_1|\widehat{\sigma}_{XY}, \bar{X})\} \asymp (nh)^{-1}$. Now, using again the independence of Z and $\widehat{\sigma}_{XY}$ and (17),

$$\begin{aligned} \mathrm{var}\{\mathrm{E}(II_1|\widehat{\sigma}_{XY}, \bar{X})\} &= \mathrm{var}(\widehat{\sigma}_{XY}^T\Delta^{-1}Z)/h^2(p) \\ &= \mathrm{E}(\widehat{\sigma}_{XY}^T\Delta^{-1}\mathrm{var}(Z)\Delta^{-1}\widehat{\sigma}_{XY})/h^2(p) \\ &= \mathrm{E}(\widehat{\sigma}_{XY}^T\Delta^{-1}\Sigma\Delta^{-1}\widehat{\sigma}_{XY})/\{nh^2(p)\} \\ &= \mathrm{tr}\{\Delta^{-1}\Sigma\Delta^{-1}(\sigma_Y^2\Sigma + n\sigma_{XY}^T\sigma_{XY})\}/\{n(n-1)h^2(p)\}, \end{aligned} \tag{33}$$

which implies $\mathrm{var}\{\mathrm{E}(II_1|\widehat{\sigma}_{XY}, \bar{X})\} \asymp n^{-1}$, using (11), (12) and $p/\{nh(p)\} \asymp 1$.

Using Lemma F.3 of Cook, Forzani and Rothman (2012), the order of II_1 under Proposition 4.2 is $O_p(\omega)$ plus the order of $II_1 = \widehat{\sigma}_{XY}^T\Delta^{-1}(\mu_X - \bar{X})/h(p)$, which does not depend on $\widehat{\Omega}$. This is exactly the variance computed in (33) and it is of order $n^{-1/2}$. Thus $II = O_p(\omega + n^{-1/2})$. Combining this with the order of I establishes the claimed order for D_N.

It is left to prove(29) and (30). To prove (29) we use the independence of $\widehat{\Omega}$ and $\widehat{\sigma}_{XY}$, (20), the definition of $h(p)$ and $p/\{nh(p)\} = O(1)$ to get

$$\mathrm{E}(\widehat{\sigma}_{XY}^T\widehat{\Omega}\widehat{\sigma}_{XY}) \asymp h(p). \tag{34}$$

From the independence of $\widehat{\Omega}$, $\widehat{\sigma}_{XY}$, (19), (32) and (34),

$$\begin{aligned} \mathrm{var}(\widehat{\sigma}_{XY}^T\widehat{\Omega}\widehat{\sigma}_{XY}) &= \mathrm{var}(\widehat{\sigma}_{XY}^T\Delta^{-1}\widehat{\sigma}_{XY}) \\ &\quad + \mathrm{E}\left((\widehat{\sigma}_{XY}^T \otimes \widehat{\sigma}_{XY}^T)\mathrm{var}(\mathrm{vec}(\widehat{\Omega}))(\widehat{\sigma}_{XY} \otimes \widehat{\sigma}_{XY})\right) \end{aligned}$$

$$\begin{aligned} &\asymp h^2(p)/n + \mathrm{E}\left(\widehat{\sigma}_{XY}^T\widehat{\sigma}_{XY}\right)^2/n \\ &\asymp h^2(p)/n + \mathrm{var}\left(\widehat{\sigma}_{XY}^T\widehat{\sigma}_{XY}\right)/n + \left(\mathrm{E}(\widehat{\sigma}_{XY}^T\widehat{\sigma}_{XY})\right)^2/n \\ &\asymp h^2(p)/n. \end{aligned} \tag{35}$$

Now, $\widehat{\sigma}_Y^2 = \sigma_Y^2 + O_p(n^{-1/2})$ implying $\widehat{B} - B \asymp h^{-1}(p)(\widehat{\sigma}_{XY}^T\widehat{\Omega}\widehat{\sigma}_{XY} - \sigma_{XY}^T\widehat{\Omega}\sigma_{XY})$. Then, (29) follows from (20) and (35).

Using Lemma F.3 of Cook, Forzani and Rothman (2012), (30) for Proposition 4.2 is of order ω plus the order of the same terms where we replace $\widehat{\Omega}$ by Δ. The order of those terms is a direct consequence of (19), (20) and the fact that $\widehat{\sigma}_Y^2 = \sigma_Y^2 + O_p(n^{-1/2})$. □

References

Cai, T. T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

Christensen, R. (1987). *Plane Answers to Complex Questions*. Wiley, New York. MR0897102

Chung, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of The Royal Statistical Society Series B* **72** 3–25. MR2751241

Cook, R. D. (2007). Fisher lecture: dimension reduction in regression. *Statist. Sci.* **22** 1–26. MR2408655

Cook, R. D. and Forzani, L. (2008). Principal fitted components for dimension reduction in regression. *Statist. Sci.* **23** 485–501. MR2530547

Cook, R. D. and Forzani, L. (2011). On the mean and variance of the generalized inverse of a singular Wishart matrix. *Electronic Journal of Statistics* **5** 146–158. MR2786485

Cook, R. D., Forzani, L. and Rothman, A. J. (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *Ann. Statist.* **40** 353–384. MR3014310

Dicker, L. (2012). Dense signals, linear estimators, and out-of-sample predictions for high-dimensional linear models. arXiv:1102.2952v3 [math.ST].

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.

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

Henmi, M. and Eguchi, S. (2004). A paradox converning nuisance parameters and projected estimating functions. *Biometrika* **91** 929–941. MR2126042

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

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S. and Ravikumar, P. K. (2011). Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in Neural Information Processing Systems*, **24** 2330–2338. MIT Press, Cambridge, MA.

Jeng, X. J. and Daye, Z. J. (2011). Sparse covariance thresholding for high-dimensional variable selection. *Statistica Sinica* **21** 625–657. MR2829849

Letac, G. and Massan, H. (2004). All invariant moments of the Wishart distribution. *Scand. J. Statist.* **31** 295–318. MR2066255

Magnus, J. R. and Neudecker, H. (1979). The commutation matrix: some properties and applications. *Ann. Statist.* **7** 381–394. MR0520247

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York. MR0652932

Pourahmadi, M. (2011). Modeling covariance matrices: the GLM and regularization perspectives. *Statistical Science* **26** 369–387. MR2917961

Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. *Electronic Journal of Statistics* **5** 935–980. MR2836766

Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515. MR2417391

Sæbø, S., Almøy, T., Aarøe, J. and Aastveit, A. H. (2007). ST-PLS: a multi-directional nearest shrunken centroid type classifier via PLS. *Journal of Chemometrics* **20** 54–62.

Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *Ann. Statist.* **40** 812–831. MR2933667

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc., Ser. B* **58** 267–288. MR1379242

von Rosen, D. (1988). Moments of the inverted Wishart distribution. *Scand. J. Statist.* **15** 97–109. MR0968156

Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high-dimensional problems. *Journal of The Royal Statistical Society Series B* **71** 615–636. MR2749910

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

Zou, H. (2005). Regularization and variable selection via the elastic net. *Journal of The Royal Statistical Society Series B* **67** 301–320. MR2137327

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469