



universität
wien

DISSERTATION

Titel der Dissertation

„Shrinkage methods for prediction out-of-sample:
Performance and selection of estimators“

Verfasserin

Mag. Nina Huber, Bakk.

angestrebter akademischer Grad

Doctor of Philosophy (PhD)

Wien, 2013

Studienkennzahl lt. Studienblatt:

A 094 136

Dissertationsgebiet lt. Studienblatt:

Statistik und Operations Research

Betreuerin / Betreuer:

Univ. Prof. Mag. Dr. Hannes Leeb

To my beloved family and friends

To my supervisor Hannes Leeb

Contents

	Page
1 Introduction	1
2 Explicit finite-sample results	3
3 Asymptotic approximations	8
4 Selection of estimators	16
5 Discussion	21
Appendix	23
A Some results for the central and the noncentral chi-square distribution	23
B Technical details for Section 3	26
C Technical details for Section 4	33
References	55
Abstract	59
Zusammenfassung	61
Curriculum Vitae	63

1 Introduction

The problem of in-sample prediction, i.e., estimating the regression function at the observed design points, is arguably among the most extensively studied topics in regression analysis. But methods designed to perform well for in-sample prediction need not perform well for out-of-sample prediction, i.e., for estimating the regression function at a new point. We study the out-of-sample predictive performance of the James–Stein estimator, which dominates the maximum likelihood estimator in the in-sample scenario; see Stein (1956) or the comprehensive monograph Judge and Bock (1978). We focus on the James–Stein estimator because of its conceptual importance (and because it is amenable to a detailed analytical analysis). The James–Stein estimator is the first method that was found to dominate maximum likelihood through shrinkage, a discovery that helped to spark the development of many of the powerful estimation methods available today that rely on some sort of shrinkage through, e.g., regularization, model selection or model averaging; see Leeb and Pötscher (2008) for a survey. We find that the James–Stein estimator can perform poorly compared to the maximum likelihood estimator in out-of-sample prediction, and we analyze and explain this phenomenon. Moreover, we show how to select a ‘good’ estimator for prediction out-of-sample among a family of James–Stein-type shrinkage estimators and discuss some properties of the selected estimator. Part of this section, Section 2 and Section 3 as well as the proofs of the results therein have been published in Huber and Leeb (2013). For notational consistency, the prediction errors will be denoted by ρ_1^2 and ρ_2^2 rather than by ρ_1 and ρ_2 as in Huber and Leeb (2013). This is because in this thesis, we consider also a third prediction error where the notation as ρ_3^2 turns out to be convenient.

Consider the Gaussian linear regression model

$$Y = X\beta + u, \quad (1)$$

where X is a fixed $n \times p$ matrix of rank p , $\beta \in \mathbb{R}^p$, $n \geq p \geq 3$ and $u \sim N(0, \sigma^2 I_n)$. For simplicity, we focus on the known variance case and we assume that $\sigma^2 = 1$. Given an estimator $\tilde{\beta}$ for β , the corresponding in-sample prediction error, i.e., the mean squared error when estimating $X\beta$ by $X\tilde{\beta}$, will be denoted by $\rho_1^2(\tilde{\beta}, \beta, X)$ and is defined by

$$\rho_1^2(\tilde{\beta}, \beta, X) = \frac{1}{n} \mathbb{E} \left[(X\tilde{\beta} - X\beta)'(X\tilde{\beta} - X\beta) \right] = \mathbb{E} \left[(\tilde{\beta} - \beta)' \frac{X'X}{n} (\tilde{\beta} - \beta) \right]. \quad (2)$$

For out-of-sample prediction, consider a new set of explanatory variables, i.e., a p -vector x_0 , that is independent of Y , and hence also independent of $\tilde{\beta}$, and that satisfies $\mathbb{E}[x_0] = 0$ and $\mathbb{E}[x_0 x_0'] = \Sigma$, where Σ is positive definite and is regarded as a nuisance parameter. (In case of fixed x_0 , e.g., $x_0 = x_0^* \in \mathbb{R}^p$, we end up with the one-dimensional estimation target $x_0^{*'}\beta$, and it is well known that the maximum likelihood estimator for $x_0^{*'}\beta$ is unique admissible minimax; see Lehmann and Casella (1998).) The out-of-sample prediction error is the mean squared error when $x_0'\tilde{\beta}$ is used to predict $x_0'\beta$, where now the mean is taken with respect to both Y and x_0 . This error will be denoted by $\rho_2^2(\tilde{\beta}, \beta, X)$ and is defined by

$$\rho_2^2(\tilde{\beta}, \beta, X) = \mathbb{E} \left[(x_0'\tilde{\beta} - x_0'\beta)^2 \right] = \mathbb{E} \left[(\tilde{\beta} - \beta)' \Sigma (\tilde{\beta} - \beta) \right]. \quad (3)$$

(Of course, the out-of-sample prediction error $\rho_2^2(\tilde{\beta}, \beta, X)$ also depends on the matrix Σ , although this dependence is not explicitly shown in our notation.) We note that the existing results of Baranchik (1973) and Dicker (2012) consider prediction errors by assuming that X is random and by taking expectations as in (2) and (3) also with respect to X . We, on the other hand, compute prediction errors by treating X as fixed, i.e., we condition on the design. The expressions on the far right-hand sides of (2) and (3) differ in the matrices $X'X/n$ and Σ . If $X'X/n$ is very close to Σ , then the in-sample prediction error will be close to the out-of-sample prediction error. But if $X'X/n$ is not very close to Σ , then the in-sample predictive performance as measured by $\rho_1^2(\tilde{\beta}, \beta, X)$ can be markedly different from the out-of-sample predictive performance as measured by $\rho_2^2(\tilde{\beta}, \beta, X)$.

We compare the maximum likelihood estimator, the James–Stein estimator and related shrinkage-type estimators by their performance as out-of-sample predictors. For in-sample prediction, i.e., in terms of the risk $\rho_1^2(\cdot, \cdot, X)$, it is well known that the James–Stein estimator dominates the maximum likelihood estimator. But for out-of-sample prediction, i.e., in terms of $\rho_2^2(\cdot, \cdot, X)$, we find that the James–Stein estimator can perform quite poorly compared to the maximum likelihood estimator; see Figure 1, relation (10), and also relation (16) in Theorem 11. (This finding contrasts a result of Baranchik (1973) as discussed at the end of Section 2 and after Theorem 10 in Section 3.) But we also find that such disappointing worst-case performance of the James–Stein estimator is atypical, in a certain sense; see relation (17) in Theorem 11.

For the case where Σ in (3) is known, estimators that dominate the maximum likelihood estimator in terms of the risk (3) are well known, and we refer to Strawderman (2003) and the references given therein. The case where Σ is unknown but estimable is studied by Baranchik (1970), Berger and Bock (1976), Berger et al. (1977) and Copas (1983). For the challenging case where Σ is unknown and not estimable (in the sense that no further structural restrictions are imposed on Σ and that p/n is large, a scenario that we study in Section 3), we are not aware of further relevant existing results.

Moreover, we consider a linear regression model with random design, where the number of explanatory variables can be infinite, together with a family of James–Stein-type shrinkage estimators. Under the assumption of Gaussianity, we show how to select a ‘good’ estimator for prediction out-of-sample. We focus on the situation where the number of candidate estimators is larger than sample size and where the dimension of the estimators is of the same order as sample size. The actual performance of an estimator is measured by the conditional out-of-sample prediction error, i.e., the expression in (3) conditional on X and Y , that is introduced and described in detail in Section 4. We show that the selected estimator is asymptotically as good as the truly best (oracle) estimator, uniformly over a large class of data-generating processes, and we show how to estimate the performance of this estimator in a uniformly consistent fashion. In the same setting, Leeb (2008) considered a family of least-squares estimators. In that sense, the findings in Section 4 extend the results in Leeb (2008).

Explicit finite-sample formulae for the out-of-sample prediction errors of the estimators in question are derived in Section 2. In Section 3, we present approximations to the out-of-sample prediction errors; our approximations become accurate as $n \rightarrow \infty$, uniformly in the underlying parameters. In Section 4, we give the details of how to select a ‘good’ out-of-sample predictor. Conclusions

are drawn in Section 5, and the more technical derivations are collected in the appendices.

2 Explicit finite-sample results

Recall that the maximum likelihood estimator of β is $\hat{\beta}_{ML} = (X'X)^{-1}X'Y \sim N(\beta, (X'X)^{-1})$. As pointed out by Stein (1956), James–Stein-type shrinkage estimators here correspond to estimators $\hat{\beta}(c)$ of β with $\hat{\beta}(c) = [1 - cp/(\hat{\beta}'_{ML}X'X\hat{\beta}_{ML})]\hat{\beta}_{ML}$, where $c \geq 0$ is a tuning parameter. More precisely, define Z and ζ as $Z = (X'X)^{1/2}\hat{\beta}_{ML}$ and $\zeta = (X'X)^{1/2}\beta$, respectively, where $(X'X)^{1/2}$ denotes a symmetric square root of $X'X$. Note that we have $Z \sim N(\zeta, I_p)$, that the unknown parameter $\beta \in \mathbb{R}^p$ corresponds to the unknown parameter $\zeta \in \mathbb{R}^p$ via $\beta = (X'X)^{-1/2}\zeta$ and that any estimator $\tilde{\beta}$ for β corresponds to an estimator $\tilde{\zeta}$ for ζ via the relation $\tilde{\beta} = (X'X)^{-1/2}\tilde{\zeta}$, and vice versa. In the Gaussian location model $Z \sim N(\zeta, I_p)$ with $\zeta \in \mathbb{R}^p$, $p \geq 3$, consider the James–Stein-type shrinkage estimator $\hat{\zeta}(c) = (1 - cp/Z'Z)Z$ with tuning parameter $c \geq 0$. The estimator $\hat{\zeta}(c)$ for ζ corresponds to the estimator $(X'X)^{-1/2}\hat{\zeta}(c)$ for β , and it is elementary to verify that $(X'X)^{-1/2}\hat{\zeta}(c)$ equals $\hat{\beta}(c)$. In particular, the traditional James–Stein estimator corresponds to the estimator $\hat{\beta}(c)$ with $c = (p - 2)/p$ and will be denoted by $\hat{\beta}_{JS}$; in the following, $\hat{\beta}_{JS}$ will also be called the James–Stein estimator (of β).

Proposition 1. *The in-sample prediction error and the out-of-sample prediction error of the James–Stein-type shrinkage estimator $\hat{\beta}(c)$ satisfy*

$$\rho_1^2(\hat{\beta}(c), \beta, X) = \rho_1^2(\hat{\beta}_{ML}, \beta, X) - \frac{1}{n} [2cp(p - 2) - c^2p^2] \mathbb{E} \left[\frac{1}{\hat{\beta}'_{ML}X'X\hat{\beta}_{ML}} \right] \quad (4)$$

with $\rho_1^2(\hat{\beta}_{ML}, \beta, X) = p/n$, and

$$\begin{aligned} \rho_2^2(\hat{\beta}(c), \beta, X) &= \rho_2^2(\hat{\beta}_{ML}, \beta, X) - 2cp \text{trace}(\Sigma(X'X)^{-1}) \mathbb{E} \left[\frac{1}{\hat{\beta}'_{ML}X'X\hat{\beta}_{ML}} \right] \\ &\quad + (c^2p^2 + 4cp) \mathbb{E} \left[\frac{\hat{\beta}'_{ML}\Sigma\hat{\beta}_{ML}}{(\hat{\beta}'_{ML}X'X\hat{\beta}_{ML})^2} \right] \end{aligned} \quad (5)$$

with $\rho_2^2(\hat{\beta}_{ML}, \beta, X) = \text{trace}(\Sigma(X'X)^{-1})$, respectively.

The formula in (4) is equivalent to the mean squared error of the James–Stein-type shrinkage estimator in the Gaussian location model that can be found in James and Stein (1961). For equation (5), we need Stein’s Lemma which we recall now.

Lemma 2 (Stein’s Lemma). *Let W be a $N(\mu, \sigma^2)$ real random variable with $0 < \sigma^2 < \infty$ and let $f : \mathbb{R} \rightarrow \mathbb{R}$ be an indefinite integral of the Lebesgue measurable function f' , essentially the derivative of f . Suppose also that $\mathbb{E}|f'(W)| < \infty$. Then*

$$\mathbb{E}[(W - \mu)f(W)] = \sigma^2 \mathbb{E}[f'(W)].$$

For a proof of this result, see Lemma 1 in Stein (1981) and the subsequent discussion.

Proof of Proposition 1: Define Z , ζ and $\hat{\zeta}(c)$ as in the paragraph preceding Proposition 1. For the in-sample prediction error of the maximum likelihood estimator, note that $\rho_1^2(\hat{\beta}_{ML}, \beta, X) = (1/n)\mathbb{E}[(Z - \zeta)'(Z - \zeta)] = (1/n)\text{trace}(\mathbb{E}[(Z - \zeta)(Z - \zeta)']) = p/n$. For relation (4), recall that $\hat{\zeta}(c)$ satisfies $\mathbb{E}[(\hat{\zeta}(c) - \zeta)'(\hat{\zeta}(c) - \zeta)] = p - [2cp(p-2) - c^2p^2]\mathbb{E}[1/Z'Z]$; cf. James and Stein (1961). To verify that this equality is equivalent to (4), first note that the expression on the left-hand side of this equality satisfies $\mathbb{E}[(\hat{\zeta}(c) - \zeta)'(\hat{\zeta}(c) - \zeta)] = n\rho_1^2(\hat{\beta}(c), \beta, X)$ (by plugging the definitions of $\hat{\zeta}(c)$, Z and ζ into the left-hand side and by recalling the definitions of $\hat{\beta}(c)$ and $\rho_1^2(\hat{\beta}(c), \beta, X)$). And, in a similar fashion, the right-hand side of that equality is equal to the expression on the right-hand side of (4), multiplied by n .

For the statement in (5), we first use the formula of the prediction error as in (3) together with the definition of the estimator, i.e., $\hat{\beta}(c) = [1 - cp/(\hat{\beta}_{ML}'X'X\hat{\beta}_{ML})]\hat{\beta}_{ML}$, to obtain that

$$\begin{aligned} \rho_2^2(\hat{\beta}(c), \beta, X) &= \rho_2^2(\hat{\beta}_{ML}, \beta, X) - 2cp \mathbb{E} \left[\frac{\hat{\beta}_{ML}'\Sigma(\hat{\beta}_{ML} - \beta)}{\hat{\beta}_{ML}'X'X\hat{\beta}_{ML}} \right] \\ &\quad + c^2p^2 \mathbb{E} \left[\frac{\hat{\beta}_{ML}'\Sigma\hat{\beta}_{ML}}{(\hat{\beta}_{ML}'X'X\hat{\beta}_{ML})^2} \right]. \end{aligned} \quad (6)$$

Write T for the $p \times p$ matrix $(X'X)^{-1/2}\Sigma(X'X)^{-1/2}$ where $(X'X)^{-1/2}$ is the inverse of a symmetric square root of $X'X$. The second expected value in (6) can then be written as

$$\mathbb{E} \left[\frac{Z'T(Z - \zeta)}{Z'Z} \right] = \sum_{i=1}^p \mathbb{E} \left[\frac{Z_i T_{i,i}(Z_i - \zeta_i)}{\sum_{k=1}^p Z_k^2} \right] + \sum_{\substack{i,j=1 \\ i \neq j}}^p \mathbb{E} \left[\frac{Z_j T_{i,j}(Z_i - \zeta_i)}{\sum_{k=1}^p Z_k^2} \right]. \quad (7)$$

For every term in the first sum on the right-hand side in (7), set, using the notation of Lemma 2, $W := Z_i$ and $f(W) = f(Z_i) := Z_i T_{i,i}/(\sum_{k=1}^p Z_k^2)$. Note that $f'(Z_i) = (T_{i,i} \sum_{k=1}^p Z_k^2 - 2Z_i^2 T_{i,i})/(\sum_{k=1}^p Z_k^2)^2$ and that

$$\mathbb{E}[|f'(Z_i)|] \leq \mathbb{E} \left[\sum_{i=1}^p |f'(Z_i)| \right] \leq M \mathbb{E} \left[\frac{p+2}{\sum_{k=1}^p Z_k^2} \right] \leq M \frac{p+2}{p-2} \leq 5M,$$

where $M = \max_{1 \leq i \leq p, 1 \leq k \leq p} \{T_{i,k}\}$, where the third inequality is obtained by (31) in Appendix A (i.e., the fact that a central chi-square distributed random variable is stochastically smaller than a noncentral chi-square distributed random variable) together with Lemma A.1, and where the last inequality follows from the fact that $p \geq 3$. Using Stein's Lemma conditional on the Z_k 's with $k \neq i$, gives

$$\mathbb{E} \left[\frac{Z_i T_{i,i}(Z_i - \zeta_i)}{\sum_{k=1}^p Z_k^2} \right] = \mathbb{E} \left[\frac{T_{i,i}}{\sum_{k=1}^p Z_k^2} \right] - 2 \mathbb{E} \left[\frac{Z_i^2 T_{i,i}}{(\sum_{k=1}^p Z_k^2)^2} \right]. \quad (8)$$

For every term in the second (double) sum on the right-hand side of (7), set $W := Z_i$ and $f(W) = f(Z_i) = Z_j T_{i,j} / \sum_{k=1}^p Z_k^2$ with $j \neq i$. Note that $f'(Z_i) = -2Z_i Z_j T_{i,j} / (\sum_{k=1}^p Z_k^2)^2$ and that

$$\begin{aligned} \mathbb{E}[|f'(Z_i)|] &\leq 2M \mathbb{E} \left[\frac{|Z_i Z_j|}{(\sum_{k=1}^p Z_k^2)^2} \right] \leq 2M \mathbb{E} \left[\frac{(\sum_{j=1}^p |Z_j|)^2}{(\sum_{k=1}^p Z_k^2)^2} \right] \\ &\leq 2M \mathbb{E} \left[\frac{p}{\sum_{k=1}^p Z_k^2} \right] \leq 2M \frac{p}{p-2} \leq 6M, \end{aligned}$$

where the third inequality is obtained by the Cauchy-Schwarz inequality, the fourth inequality by (31) in Appendix A together with Lemma A.1 and where the last inequality follows from the fact that $p \geq 3$. Use Stein's Lemma conditional on the Z_k 's with $k \neq i$ to get

$$\mathbb{E} \left[\frac{Z_j T_{i,j} (Z_i - \zeta_i)}{\sum_{k=1}^p Z_k^2} \right] = -2 \mathbb{E} \left[\frac{Z_i Z_j T_{i,j}}{(\sum_{k=1}^p Z_k^2)^2} \right]. \quad (9)$$

Using the results in (8) and (9), we can write the right-hand side of (7) as

$$\begin{aligned} \text{trace}(T) \mathbb{E} \left[\frac{1}{Z'Z} \right] - 2 \mathbb{E} \left[\frac{Z'TZ}{(Z'Z)^2} \right] &= \text{trace}(\Sigma(X'X)^{-1}) \mathbb{E} \left[\frac{1}{\hat{\beta}'_{ML} X'X \hat{\beta}_{ML}} \right] \\ &\quad - 2 \mathbb{E} \left[\frac{\hat{\beta}'_{ML} \Sigma \hat{\beta}_{ML}}{(\hat{\beta}'_{ML} X'X \hat{\beta}_{ML})^2} \right], \end{aligned}$$

where the equality is obtained by using the definitions of T and Z together with the fact that $\text{trace}(AB) = \text{trace}(BA)$ for matrices A and B of appropriate dimensions. Recall that (7) or, equivalently, the right-hand side of the preceding display, is equal to the expected value in the second term on the right-hand side of (6). Plugging this into (6) and simplifying, we obtain (5). For the out-of-sample prediction error of the maximum likelihood estimator, note that $\rho_2^2(\hat{\beta}_{ML}, \beta, X) = \text{trace}(\Sigma \mathbb{E}[(\hat{\beta}_{ML} - \beta)(\hat{\beta}_{ML} - \beta)']) = \text{trace}(\Sigma(X'X)^{-1})$. $\square$

The next result shows how the expressions in (4) and (5) depend on β and on X ; moreover, that result together with Lemma A.2 in Appendix A can be used to compute these expressions numerically. Throughout the rest of this thesis, we use symbols like $\chi_k^2(\lambda)$ to denote a random variable that is chi-square distributed with $k \geq 1$ degrees of freedom and noncentrality parameter $\lambda \geq 0$. For more details, see Appendix A.

Proposition 3. *Assume that the linear model in (1) holds and that Σ is a positive definite $p \times p$ matrix. Then the expected values in (4) and (5) can be written as*

$$\begin{aligned} \mathbb{E} \left[\frac{1}{\hat{\beta}'_{ML} X'X \hat{\beta}_{ML}} \right] &= \mathbb{E} \left[\frac{1}{\chi_p^2(\beta' X' X \beta)} \right], \\ \mathbb{E} \left[\frac{\hat{\beta}'_{ML} \Sigma \hat{\beta}_{ML}}{(\hat{\beta}'_{ML} X'X \hat{\beta}_{ML})^2} \right] &= \mathbb{E} \left[\frac{\text{trace}(\Sigma(X'X)^{-1})}{(\chi_{p+2}^2(\beta' X' X \beta))^2} \right] + \mathbb{E} \left[\frac{\beta' \Sigma \beta}{(\chi_{p+4}^2(\beta' X' X \beta))^2} \right]. \end{aligned}$$

Proof. The first equality follows upon recalling that $(X'X)^{1/2}\hat{\beta}_{ML}$ follows a Gaussian distribution with mean $(X'X)^{1/2}\beta$ and covariance matrix I_p . The second equality is obtained by applying Corollary 2 from the appendix of Bock (1975). $\square$

The well-known formula on the right-hand side of (4) shows that the in-sample prediction error of $\hat{\beta}(c)$ equals the in-sample prediction error of $\hat{\beta}_{ML}$ minus the product of a positive expected value and a polynomial in c and p . In particular, $\rho_1^2(\hat{\beta}(c), \beta, X)$ is smaller than $\rho_1^2(\hat{\beta}_{ML}, \beta, X)$ if c satisfies $0 < c < 2(p-2)/p$ and is minimized for $c = (p-2)/p$, which is the tuning parameter used by $\hat{\beta}_{JS}$. Moreover, it is easy to see that $\rho_1^2(\hat{\beta}(c), \beta, X)$ depends on β and X only through $\beta'(X'X/n)\beta$. Unlike the formula of $\rho_1^2(\hat{\beta}(c), \beta, X)$ in (4), display (5) shows that $\rho_2^2(\hat{\beta}(c), \beta, X)$ is obtained from $\rho_2^2(\hat{\beta}_{ML}, \beta, X)$ by subtracting a positive term and then adding another positive term, i.e., the second and the third term on the right-hand side of (5), that depend on c , on p and on the unknown parameters in a more complicated fashion. Proposition 3 shows that (5) depends on β and X through $\beta'(X'X/n)\beta$ and through $\beta'\Sigma\beta$ (which can be viewed as a kind of signal-to-noise ratio).

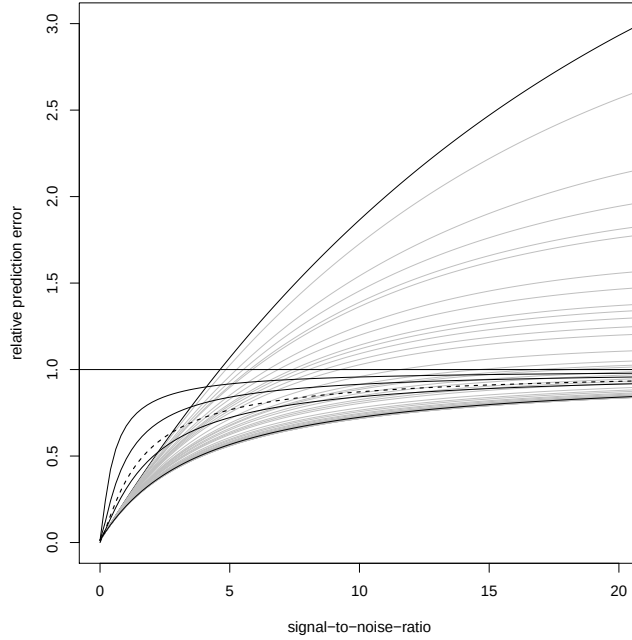


Figure 1: The solid curves in black and gray show finite-sample relative prediction errors as a function of $\beta'\Sigma\beta$ and are explained in the next paragraph. The dashed curve shows the approximation to the relative prediction error that is obtained from Theorem 4 in Section 3. The constant solid line at 1 is for reference.

Figure 1 exemplifies the relative out-of-sample prediction error of the James–Stein estimator and of the maximum likelihood estimator, i.e., $\rho_2^2(\hat{\beta}_{JS}, \beta, X)/\rho_2^2(\hat{\beta}_{ML}, \beta, X)$, as a function of $\beta'\Sigma\beta$ for various configurations in parameter space. For the figure, we selected a scenario where $X'X/n$ is not very close to Σ , so that the in-sample and the out-of-sample predictive performance of estimators differ from each other (as noted in the discussion after (3)). In particular, we took $n = 200$, $p = 160$, $\Sigma = I_p$, and X was obtained by sampling i.i.d. standard normals. The solid curves show $\rho_2^2(\hat{\beta}_{JS}, \beta, X)/\rho_2^2(\hat{\beta}_{ML}, \beta, X)$ as a function of $\beta'\Sigma\beta$, for β parallel to various eigenvectors of $X'X/n$. Let w_i be

the eigenvector corresponding to the eigenvalue ν_i of $X'X/n$, and assume that $\nu_1 \leq \dots \leq \nu_{160}$. Four solid black curves stay below 1 and appear to be ordered; starting from the top, these correspond to β parallel to w_{160} , w_{120} , w_{80} , and w_{40} . The fifth solid black curve, that exceeds 1, corresponds to β parallel to w_1 , i.e., the eigenvector of the smallest eigenvalue. This curve attains a maximum of 4.27 at 75.12 (which is off the chart) and then recedes back towards 1 as $\beta'\Sigma\beta \rightarrow \infty$. The gray curves are obtained in the same way but for eigenvectors corresponding to the remaining smallest 25% of eigenvalues. The curves in Figure 1 differ dramatically depending on whether they correspond to small eigenvalues like ν_1 on the one hand, and moderate-to-large eigenvalues like ν_{40} , ν_{80} , ν_{120} and ν_{160} on the other. Repeating these computations with X replaced by a new independent sample, we obtained essentially the same results. And for other choices of p and n , we obtained results that are qualitatively similar, including maxima above 1 corresponding to small eigenvalues. This phenomenon becomes less pronounced as p/n decreases, and it disappears completely for very small values of p/n . The results in Section 3 entail that this is no surprise.

Figure 1 shows that the James–Stein estimator no longer dominates the maximum likelihood estimator for out-of-sample prediction, in the sense that

$$\rho_2^2(\hat{\beta}_{JS}, \beta, X) > \rho_2^2(\hat{\beta}_{ML}, \beta, X) \quad (10)$$

for some $\beta \in \mathbb{R}^p$, if X is the design matrix used to generate the figure. (Indeed, the left-hand side of the preceding display exceeds the right-hand side by a factor of 4.27 for appropriately chosen β , as noted in the preceding paragraph.) This result goes back to Brown (1975) and to Judge and Bock (1976), who discovered the result simultaneously with entirely different methods of proof. In these papers, the authors show that the James–Stein-type shrinkage estimator is minimax if and only if $\text{trace}(S^{-1})/\lambda_{\max}(S) > 2$ and $0 < c \leq (2/p)(\text{trace}(S^{-1})/\lambda_{\max}(S) - 2)$ where $S = (X'X)^{-1/2}\Sigma(X'X)^{-1/2}$. Hence, it depends on the unknown variance-covariance matrix Σ , that can not be estimated consistently when p/n is large, whether or not the James–Stein-type shrinkage estimator dominates the maximum likelihood estimator. Section 3 provides a framework dealing with this problem. There, we examine and quantify those design matrices such that the James–Stein-type shrinkage estimator dominates the maximum likelihood estimator (see the discussion after Theorem 11). The result in (10) should be compared to the findings of Baranchik (1973): The results in that paper suggest that

$$\mathbb{E} \left[\rho_2^2(\hat{\beta}_{JS}, \beta, X) \right] \leq \mathbb{E} \left[\rho_2^2(\hat{\beta}_{ML}, \beta, X) \right] \quad (11)$$

for each $\beta \in \mathbb{R}^p$, with strict inequality for some β , if X is random with i.i.d. $N(0, \Sigma)$ -distributed rows, and where the expectation in (11) is taken with respect to X (see also Dicker (2012)). Comparing the preceding two displays, we see that for X fixed, cf. (10), $\hat{\beta}_{JS}$ can perform poorly for some $\beta \in \mathbb{R}^p$. But on average with respect to X , as considered in Baranchik (1973) and Dicker (2012), the relation in (11) suggests that $\hat{\beta}_{JS}$ performs well, irrespective of β . The performance of $\hat{\beta}_{JS}$ hence depends crucially on whether we condition on the design as in (10) or average with respect to the design distribution as in (11). We give a more detailed analysis and explanation of this phenomenon in the next section. Also note that the phenomenon in (10) and (11) is related to the ancillarity paradox of Brown (1990).

3 Asymptotic approximations

In this section, we provide approximations to quantities like $\rho_2^2(\hat{\beta}_{JS}, \beta, X)$ and $\sup_{\beta} \rho_2^2(\hat{\beta}_{JS}, \beta, X)$ for ‘typical’ design matrices X . Here, ‘typical’ means ‘in probability’ when the explanatory variables in the training period, i.e., the rows of X , are taken as realizations from the same distribution as those in the prediction period (i.e., x_0). Our approximations are uniform in the unknown parameters and become accurate as $n \rightarrow \infty$, where the dimension of the model considered at sample size n , i.e., p , is allowed to depend on sample size. Note that quantities like $\rho_2^2(\hat{\beta}(c), \beta, X)$ and $\rho_2^2(\hat{\beta}_{ML}, \beta, X)$ now become random variables through their dependence on X . We emphasize that our evaluation of performance is always taken conditional on X and that the random design is used only to describe the behavior for ‘typical’ design matrices X . We also stress, with X random, that the expectations in (4) and (5) are now to be understood as conditional on X .

More formally, the following assumptions will be maintained throughout this section: For each n and p under consideration ($n \geq p \geq 3$), we assume that the model (1) holds; that x_0 and X are independent of the error u in (1); and that the rows of X and also x_0 are i.i.d. with mean zero and positive definite covariance matrix Σ . In addition, we assume that X can be written as $X = V\Sigma^{1/2}$, where $\Sigma^{1/2}$ denotes a symmetric square root of Σ and where V is the $n \times p$ matrix obtained by taking the upper left block of a double infinite array $(V_{i,j})_{i \geq 1, j \geq 1}$ of i.i.d. random variables that have mean zero, variance one and a finite fourth moment (cf. Bai and Silverstein (2010)). Finally, we also assume that the (marginal) distribution of the $V_{i,j}$ ’s is absolutely continuous with respect to Lebesgue measure. This last assumption could be dropped altogether, but at the expense of longer and technically more involved proofs. Under these assumptions, we note that $X'X$ is invertible almost surely. We set $\rho_2^2(\hat{\beta}(c), \beta, X) = 0$ on the probability zero event where $X'X$ is degenerate. Otherwise, the random variable $\rho_2^2(\hat{\beta}(c), \beta, X)$ is defined by the expression on the right-hand side of (5), where the expected values are to be understood as conditional on X . These conventions also cover $\rho_2^2(\hat{\beta}_{ML}, \beta, X)$ and $\rho_2^2(\hat{\beta}_{JS}, \beta, X)$ in view of $\hat{\beta}_{ML} = \hat{\beta}(0)$ and $\hat{\beta}_{JS} = \hat{\beta}((p-2)/p)$. The following result provides a simple asymptotic approximation to the quantity $\rho_2^2(\hat{\beta}(c), \beta, X)$.

Theorem 4. *Assume that $n \rightarrow \infty$ and that $p = p(n)$ is such that $p/n \rightarrow t \in [0, 1)$. Moreover, for each p , let β and Σ be a p -vector and a positive definite $p \times p$ matrix, respectively, so that $\beta'\Sigma\beta \rightarrow \delta^2 \in [0, \infty]$ as $n \rightarrow \infty$. Then $\rho_2^2(\hat{\beta}_{ML}, \beta, X) \rightarrow t/(1-t)$ and $\rho_2^2(\hat{\beta}(c), \beta, X) \rightarrow r(\delta^2, c, t)$ in probability, where*

$$r(\delta^2, c, t) = \frac{t}{1-t} \left(1 - c \frac{t}{t + \delta^2} \right)^2 + c^2 \frac{t^2 \delta^2}{(t + \delta^2)^2}$$

if δ^2 is finite, and where $r(\infty, c, t) = t/(1-t)$ otherwise (expressions like $t/(t + \delta^2)$ are to be interpreted as zero if t and δ^2 are both equal to zero). Convergence of $\rho_2^2(\hat{\beta}(c), \beta, X)$ is uniform in c over compact sets, in the sense that $\sup_{0 \leq c \leq C} |\rho_2^2(\hat{\beta}(c), \beta, X) - r(\delta^2, c, t)| = o_P(1)$ for any $C > 0$. Moreover, these statements continue to hold if $\rho_2^2(\hat{\beta}(c), \beta, X)$ is replaced by the estimator $r(\hat{\delta}^2, c, p/n)$ where $\hat{\delta}^2 = \max\{Y'Y/n - 1, 0\}$.

The approximations to $\rho_2^2(\hat{\beta}(c), \beta, X)$ and $\rho_2^2(\hat{\beta}_{ML}, \beta, X)$ provided by Theorem 4, i.e., $r(\delta^2, c, t)$ and $t/(1-t)$, respectively, are such that $r(\delta^2, c, t) < t/(1-t)$ whenever $t > 0$, provided only that $0 < c \leq 2$ and $\delta^2 < \infty$; cf. the dashed line in Figure 1. The results of Dicker (2012) suggest approximations to $\mathbb{E}[\rho_2^2(\hat{\beta}(c), \beta, X)]$ and $\mathbb{E}[\rho_2^2(\hat{\beta}_{ML}, \beta, X)]$ that coincide with $r(\delta^2, c, t)$ and $t/(1-t)$, respectively.

In order to prove Theorem 4, we need to study the prediction errors of the estimators in question, i.e., $\rho_2^2(\hat{\beta}_{ML}, \beta, X)$ and $\rho_2^2(\hat{\beta}(c), \beta, X)$. The out-of-sample prediction error of the maximum likelihood estimator equals $\text{trace}(\Sigma(X'X)^{-1})$ almost surely by Proposition 1, whereas the out-of-sample prediction error of the James–Stein-type shrinkage estimator is given by (5) almost surely (where the expected values are to be understood as conditional on X , denoted by $\mathbb{E}[\cdot|X]$ in the following). Using Proposition 3 conditional on X , we can rewrite $\rho_2^2(\hat{\beta}(c), \beta, X)$ also as

$$\begin{aligned} \text{trace}(\Sigma(X'X)^{-1}) & \left(1 - 2c \mathbb{E} \left[\frac{p}{\chi_p^2(\beta'X'X\beta)} \middle| X \right] \right. \\ & \quad \left. + (c^2 + 4c/p) \mathbb{E} \left[\frac{p^2}{(\chi_{p+2}^2(\beta'X'X\beta))^2} \middle| X \right] \right) \\ & \quad + (c^2 + 4c/p) \mathbb{E} \left[\frac{p^2 \beta' \Sigma \beta}{(\chi_{p+4}^2(\beta'X'X\beta))^2} \middle| X \right] \end{aligned} \quad (12)$$

almost surely. Thus, we need the limits of the random variables $\text{trace}(\Sigma(X'X)^{-1})$ and $\mathbb{E}[(p/\chi_{p+k}^2(\beta'X'X\beta))^m | X]$ for $k = 0, 2, 4$ and $m = 1, 2$. For the last statement of the theorem, we need the limit of the random variable $\hat{\delta}^2$. These quantities are treated in Lemma 5, in Lemma 6 together with Lemma 8 and in Lemma 9, respectively.

Lemma 5. *For the $n \times p$ matrix V , write $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ for the ordered eigenvalues of $V'V$. If n and p are such that $n \rightarrow \infty$ and $p/n \rightarrow t$ for some $t \in [0, 1]$, then*

$$\sum_{i=1}^p \frac{1}{\lambda_i} \xrightarrow{a.s.} \frac{t}{1-t},$$

where the limit is to be interpreted as ∞ in the case where $t = 1$ (and where the sum on the left-hand side is to be interpreted as ∞ whenever $\lambda_p = 0$). Moreover, we also have $\lambda_1/n \rightarrow (1 + \sqrt{t})^2$ and $\lambda_p/n \rightarrow (1 - \sqrt{t})^2$ almost surely.

Proof. Consider first the case where t satisfies $0 < t < 1$. Write $\nu_1 \geq \nu_2 \geq \dots \geq \nu_p$ for the ordered eigenvalues of $V'V/n$, and note that $\nu_i = \lambda_i/n$, $1 \leq i \leq p$. Write F_n for the empirical cumulative distribution function (c.d.f.) of the ν_i 's, i.e., for $x \in \mathbb{R}$, $F_n(x)$ is the fraction of eigenvalues of $V'V/n$ that do not exceed x , and note that F_n is a random c.d.f. as it depends on $V'V/n$. Under the maintained assumptions, F_n converges weakly to a nonrandom limit F , except on a set of probability zero; cf. Theorem 3.6 of Bai and Silverstein (2010). The limit c.d.f. F corresponds to the so-called Marchenko–Pastur distribution, which is supported by the interval $[a, b]$ with $a = (1 - \sqrt{t})^2$ and $b = (1 + \sqrt{t})^2$ and which has a density given by $f(x) = \sqrt{(x-a)(b-x)}/(2\pi tx)$ for $a \leq x \leq b$. For any bounded continuous function $g(\cdot)$, it follows that $\int g(x)F_n(dx) \rightarrow \int g(x)f(x)dx$

almost surely in view of the Portmanteau Theorem. And under the maintained assumptions, we also see that the smallest and the largest eigenvalue of $V'V/n$, i.e., $\nu_p = \lambda_p/n$ and $\nu_1 = \lambda_1/n$, converge almost surely to a and to b , respectively, whenever $0 < t \leq 1$; see Theorem 5.11 of Bai and Silverstein (2010). Recall that $a = (1 - \sqrt{t})^2$ is positive, and define a function $g(\cdot)$ as $g(x) = 1/x$ if $x \geq a/2$ and set $g(x) = 2/a$ otherwise. Now first note that $\int g(x)F_n(dx) \rightarrow \int g(x)f(x)dx$, except on a probability zero event, because $g(\cdot)$ is continuous and bounded. Second, we have that $\int g(x)f(x)dx = \int (1/x)f(x)dx$ because $g(x) = 1/x$ on the support of $f(\cdot)$. And third, observe that the event where $\int g(x)F_n(dx)$ differs from $\int (1/x)F_n(dx)$ is contained in the event where $\nu_p < a/2$ and recall that $\nu_p \rightarrow a$ almost surely. Hence, these observations imply that

$$\int \frac{1}{x} F_n(dx) \xrightarrow{\text{a.s.}} \int \frac{1}{x} f(x) dx.$$

But the left-hand side in the preceding display can also be written as $\frac{1}{p} \sum_{i=1}^p \frac{1}{\nu_i} = \frac{n}{p} \sum_{i=1}^p \frac{1}{\lambda_i}$; and the right-hand side is equal to $1/(1-t)$ in view of Lemma B.1. It follows that $\sum_{i=1}^p 1/\lambda_i$ converges to $t/(1-t)$ almost surely, as required.

In the case where $t = 0$, choose an arbitrary number $\tilde{t}$ with $0 < \tilde{t} < 1$, and choose numbers $\tilde{p} \geq p$ that go to infinity with n such that $\tilde{p}/n \rightarrow \tilde{t}$ as $n \rightarrow \infty$. Write $\tilde{V}$ for the $n \times \tilde{p}$ matrix that forms the upper left block of the double array $(V_{i,j})_{i \geq 1, j \geq 1}$, and denote the eigenvalues of $\tilde{V}'\tilde{V}$ by $\tilde{\lambda}_1 \geq \dots \geq \tilde{\lambda}_{\tilde{p}}$. Because V is a submatrix of $\tilde{V}$, Cauchy's interlacing theorem entails that $\tilde{\lambda}_i \geq \lambda_i \geq \tilde{\lambda}_{i+\tilde{p}-p}$ for $1 \leq i \leq p$. This implies that $\sum_{i=1}^p 1/\lambda_i \leq \sum_{i=1}^p 1/\tilde{\lambda}_{i+\tilde{p}-p} \leq \sum_{i=1}^{\tilde{p}} 1/\tilde{\lambda}_i$, and also that $\tilde{\lambda}_{\tilde{p}} \leq \lambda_p \leq \lambda_1 \leq \tilde{\lambda}_1$. And from the case considered in the preceding paragraph, it follows that $\sum_{i=1}^{\tilde{p}} 1/\tilde{\lambda}_i \rightarrow \tilde{t}/(1-\tilde{t})$, that $\tilde{\lambda}_{\tilde{p}}/n \rightarrow (1 - \sqrt{\tilde{t}})^2$, and that $\tilde{\lambda}_1/n \rightarrow (1 + \sqrt{\tilde{t}})^2$, almost surely. Taken together, we obtain that $\limsup \sum_{i=1}^p 1/\lambda_i \leq \tilde{t}/(1-\tilde{t})$, that $(1 - \sqrt{\tilde{t}})^2 \leq \liminf \lambda_p/n$, and that $\limsup \lambda_1/n \leq (1 + \sqrt{\tilde{t}})^2$, almost surely. Letting $\tilde{t} \downarrow 0$ gives the desired result.

In the case where $t = 1$, we already know that $\lambda_1/n \rightarrow 4$ and $\lambda_p/n \rightarrow 0$ almost surely. Choose an arbitrary number $\tilde{t}$ with $0 < \tilde{t} < 1$, and let $\tilde{p} \leq p$ be such that $\tilde{p}/n \rightarrow \tilde{t}$. Now repeat the argument in the preceding paragraph, with the role of p and $\tilde{p}$ exchanged, to conclude that $\tilde{t}/(1-\tilde{t}) \leq \liminf \sum_{i=1}^p 1/\lambda_i$ almost surely. The result follows by letting $\tilde{t} \uparrow 1$. $\square$

Remark. The lemma continues to hold if $S = \sum_{i=1}^p 1/\lambda_i$ is replaced by $S^* = \sum_{i:\lambda_i > 0} 1/\lambda_i$. To see this, note that $\mathbb{P}(S = S^*) = \mathbb{P}(\lambda_i > 0, \forall i = 1, \dots, p) = 1$ because $V'V$ is invertible almost surely. By considering S^* instead of S , the lemma can be adapted to cover also the case where $t \in (1, \infty]$.

The result of the subsequent lemma is well known if the noncentrality parameter equals zero and the degrees of freedom tend to infinity. It then reduces to the law of large numbers together with the uniform integrability of the random variables in question. More generally, we can show the following result.

Lemma 6. Fix an integer $m \geq 1$, and consider random variables $\chi_k^2(\lambda)$ with $k > 2m$. Then

$$\mathbb{E} \left[\left(\frac{k + \lambda}{\chi_k^2(\lambda)} \right)^m \right] \rightarrow 1$$

as $k + \lambda \rightarrow \infty$.

Proof. Fix $m \geq 1$. The mean and the variance of $\chi_k^2(\lambda)$ are given by $k + \lambda$ and $2(k + 2\lambda)$, respectively (see Lemma A.2). Using this fact and Chebyshev's inequality, we see that $\chi_k^2(\lambda)/(k + \lambda) \rightarrow 1$, and also that $[(k + \lambda)/\chi_k^2(\lambda)]^m \rightarrow 1$, in probability as $k + \lambda \rightarrow \infty$. It remains to show, for integers k_a and reals λ_a , $a \geq 1$, satisfying $k_a > 2m$, $\lambda_a \geq 0$, and $k_a + \lambda_a \rightarrow \infty$ as $a \rightarrow \infty$, that the random variables $[(k_a + \lambda_a)/\chi_{k_a}^2(\lambda_a)]^m$ are uniformly integrable. In other words, for each fixed $\varepsilon > 0$, we need to find a constant M so that

$$\mathbb{E} \left[\left(\frac{k_a + \lambda_a}{\chi_{k_a}^2(\lambda_a)} \right)^m \left\{ \left(\frac{k_a + \lambda_a}{\chi_{k_a}^2(\lambda_a)} \right)^m > M \right\} \right] < \varepsilon \quad (13)$$

holds for each a , where we use symbols like $\{A\}$ to denote the indicator function of the event A . Set $U_a = [(k_a + \lambda_a)/\chi_{k_a}^2(\lambda_a)]^m$, use the inequality in (31) in Appendix A and Lemma A.1 to show that

$$\mathbb{E}[U_a] \leq \mathbb{E} \left[\left(\frac{k_a + \lambda_a}{\chi_{k_a}^2(0)} \right)^m \right] = (k_a + \lambda_a)^m \prod_{i=1}^m \frac{1}{k_a - 2i}.$$

Because $k_a > 2m$ by assumption, the U_a 's are all integrable. Also note that $\chi_{k_a}^2(\lambda_a)$ is distributed as $\chi_{k_a+2J_a}^2(0)$ for an independent Poisson random variable J_a with mean $\lambda_a/2$ (cf. the first paragraph in Appendix A). To derive (13), i.e., to show uniform integrability of the U_a 's, we may assume that either $k_a \rightarrow \infty$ or $\lambda_a \rightarrow \infty$ as $a \rightarrow \infty$, by switching to subsequences if necessary (because $k_a + \lambda_a \rightarrow \infty$).

Assume that $k_a \rightarrow \infty$. Here, we first consider the (finitely many) a 's for which $k_a \leq 4m$. Note that the U_a 's are finite almost everywhere, that $U_a \{U_a > M\} \rightarrow 0$ almost surely as $M \rightarrow \infty$ and that $U_a \{U_a > M\} \leq U_a$. These facts together with the dominated convergence theorem entail that we can find a constant M_1 so that (13) holds whenever $M \geq M_1$ for those a 's for which $k_a \leq 4m$. It remains to consider the a 's with $k_a > 4m$. For these a 's, we derive uniform integrability by showing that the second moments of the U_a 's are bounded. Using the law of iterated expectation, $\mathbb{E}[U_a^2]$ is given by the sum of

$$\mathbb{E} \left[\prod_{i=1}^{2m} \left(\frac{k_a + \lambda_a}{k_a - 2i + 2J_{\lambda_a}} \right) \{J_{\lambda_a} \geq \lambda_a/4\} \right] \quad (14)$$

and

$$\mathbb{E} \left[\prod_{i=1}^{2m} \left(\frac{k_a + \lambda_a}{k_a - 2i + 2J_{\lambda_a}} \right) \{J_{\lambda_a} < \lambda_a/4\} \right]. \quad (15)$$

Note that $k_1/(k_1 - k_2) \leq k_2 + 1$ for integers k_1, k_2 satisfying $k_1 > k_2$ because the inequality is equivalent to $0 \leq k_1 k_2 - k_2^2 - k_2$ and also to $k_1 \geq k_2 + 1$, which holds by assumption. The expression in (14) is bounded by $(4m + 3)^{2m}$ because the i -th fraction in the integrand of (14) is bounded by $(k_a + \lambda_a)/(k_a - 2i + \lambda_a/2) \leq (k_a + \lambda_a)/(k_a - 4m + \lambda_a/2) \leq k_a/(k_a - 4m) + 2 \leq 4m + 1 + 2$ (the first inequality holds in view of $J_{\lambda_a} \geq \lambda_a/4$, and the last inequality holds by the argument after equation (15) because $k_a > 4m$ here). To bound the

expression in (15), note that the i -th fraction in the integrand in (15) is bounded by $k_a/(k_a - 2i) + \lambda_a \leq k_a/(k_a - 4m) + \lambda_a \leq 4m + 1 + \lambda_a$ and use the Chernoff bound for the Poisson distribution (see Lemma B.2 in Appendix B) to get $\mathbb{P}(J_a < \lambda_a/4) \leq e^{-\lambda_a/2}(2e)^{\lambda_a/4} = (e/2)^{-\lambda_a/4}$. This implies that (15) is bounded by $(4m + 1 + \lambda_a)^{2m}(e/2)^{-\lambda_a/4}$. This upper bound is a continuous function of $\lambda_a \geq 0$ that converges to zero as $\lambda_a \rightarrow \infty$. Hence, the upper bound is itself bounded by a constant, uniformly in $\lambda_a \geq 0$, as required.

Now assume that $\lambda_a \rightarrow \infty$. We decompose the expression on the left-hand side of (13) into the sum of $\mathbb{E}[U_a\{U_a > M\}\{J_{\lambda_a} < \lambda_a/4\}]$ and $\mathbb{E}[U_a\{U_a > M\}\{J_{\lambda_a} \geq \lambda_a/4\}]$, and find M so that each of these two terms is bounded by $\varepsilon/2$. To this end, note that $\mathbb{E}[U_a\{U_a > M\}\{J_{\lambda_a} < \lambda_a/4\}]$ can be bounded by

$$\begin{aligned} \mathbb{E} \left[\left(\frac{k_a + \lambda_a}{\chi_{k_a}^2(0)} \right)^m \{J_{\lambda_a} < \lambda_a/4\} \right] &= \prod_{i=1}^m \frac{k_a + \lambda_a}{k_a - 2i} \mathbb{E}[\{J_{\lambda_a} < \lambda_a/4\}] \\ &\leq (2m + 1 + \lambda_a)^m (e/2)^{-\lambda_a/4}. \end{aligned}$$

In the preceding display, the inequality is derived by noting that each term in the product can be bounded by $k_a/(k_a - 2i) + \lambda_a \leq 2m + 1 + \lambda_a$ because $k_a > 2m$ by assumption, and that the expected value can be bounded using the Chernoff bound for the Poisson distribution in Lemma B.2 as before. Since $\lambda_a \rightarrow \infty$ here, the upper bound in the preceding display tends to zero and hence is smaller than $\varepsilon/2$ for sufficiently large λ_a 's, e.g., $\lambda_a > \lambda_*$. And for the (finitely many) a 's for which $\lambda_a \leq \lambda_*$, we can find a constant M_2 so that $\mathbb{E}[U_a\{U_a > M\}\{J_{\lambda_a} < \lambda_a/4\}]$ is less than $\varepsilon/2$ whenever $M \geq M_2$ (by the same arguments as in the paragraph preceding equation (14)). Lastly, $\mathbb{E}[U_a\{U_a > M\}\{J_{\lambda_a} \geq \lambda_a/4\}]$ is bounded by

$$\begin{aligned} &\mathbb{E} \left[\left(\frac{k_a + \lambda_a}{\chi_{k_a + \lceil \lambda_a/2 \rceil}^2(0)} \right)^m \left\{ \left(\frac{k_a + \lambda_a}{\chi_{k_a + \lceil \lambda_a/2 \rceil}^2(0)} \right)^m > M \right\} \right] \\ &\leq 2^m \mathbb{E} \left[\left(\frac{k_a + \lceil \lambda_a/2 \rceil}{\chi_{k_a + \lceil \lambda_a/2 \rceil}^2(0)} \right)^m \left\{ \left(\frac{k_a + \lceil \lambda_a/2 \rceil}{\chi_{k_a + \lceil \lambda_a/2 \rceil}^2(0)} \right)^m > 2^{-m} M \right\} \right], \end{aligned}$$

where $\lceil x \rceil$ denotes the smallest integer not smaller than x , and where the inequality is based on the observation that $k_a + \lambda_a = 2(k_a/2 + \lambda_a/2) < 2(k_a + \lceil \lambda_a/2 \rceil)$. Set $\tilde{k}_a = k_a + \lceil \lambda_a/2 \rceil$ and set $\tilde{\lambda}_a = 0$. To find an $M_3 \geq M_2$ such that the expression on the right-hand side of the preceding display is less than $\varepsilon/2$ whenever $M \geq M_3$, it suffices to show that the random variable $[(\tilde{k}_a + \tilde{\lambda}_a)/\chi_{\tilde{k}_a}^2(\tilde{\lambda}_a)]^m$ are uniformly integrable. Since $\tilde{k}_a \rightarrow \infty$ as $a \rightarrow \infty$, this has already been established in the second paragraph of the proof. $\square$

The following lemma is an auxiliary result and is only used to shorten the proofs of Lemma 8 and 9.

Lemma 7. *Consider the $n \times p$ matrix V , and let w be a unit vector in $\mathbb{R}^p$ ($n \geq p \geq 3$). Then $\mathbb{E}[w'V'Vw/n] = 1$ and $\text{Var}[w'V'Vw/n] \rightarrow 0$ as $n \rightarrow \infty$, irrespective of the behavior of p ; in particular, we have $w'V'Vw/n \rightarrow 1$ in probability.*

Proof. Setting $W_i^{(n)} = (\sum_{j=1}^p V_{i,j} w_j)^2$, we see that $w'V'Vw/n$ can be written as $(1/n) \sum_{i=1}^n W_i^{(n)}$, i.e., as the average of n i.i.d. random variables which have mean $\sum_{j=1}^p w_j^2 = 1$. And the variance of $W_i^{(n)}$ is given by

$$\begin{aligned} \sum_{a,b,c,d=1}^p w_a w_b w_c w_d \mathbb{E}[V_{i,a} V_{i,b} V_{i,c} V_{i,d}] - 1 &= \mathbb{E}[V_{1,1}^4] \sum_{a=1}^p w_a^4 + 3 \sum_{\substack{a,b=1 \\ a \neq b}}^p w_a^2 w_b^2 - 1 \\ &= (\mathbb{E}[V_{1,1}^4] - 1) \sum_{i=1}^p w_a^4 + 2 \sum_{\substack{a,b=1 \\ a \neq b}}^p w_a^2 w_b^2 + \sum_{a=1}^p w_a^4 + \sum_{\substack{a,b=1 \\ a \neq b}}^p w_a^2 w_b^2 - 1 \\ &\leq (\mathbb{E}[V_{1,1}^4] - 1) \sum_{i=1}^p w_a^4 + (\mathbb{E}[V_{1,1}^4] + 1) \sum_{\substack{a,b=1 \\ a \neq b}}^p w_a^2 w_b^2 \\ &\leq \mathbb{E}[V_{1,1}^4] + 1, \end{aligned}$$

where the first equality is based on the fact that $\mathbb{E}[V_{i,a} V_{i,b} V_{i,c} V_{i,d}] = 0$ whenever one of the indices a, b, c, d differs from the others, while the first and the second inequality are derived from the facts that $(\sum_{j=1}^p w_j^2)^2 = 1$ and $1 \leq \mathbb{E}[V_{1,1}^4]$. We hence can bound $\text{Var}[w'V'Vw/n]$ by $(\mathbb{E}[V_{1,1}^4] + 1)/n$. $\square$

Lemma 8. *Assume that Theorem 4 applies. If $t + \delta^2 > 0$, then*

$$\frac{p}{p + \beta' X' X \beta} \rightarrow \frac{t}{t + \delta^2}, \quad \beta' \Sigma \beta \frac{p}{p + \beta' X' X \beta} \rightarrow \frac{t \delta^2}{t + \delta^2}$$

in probability as $n \rightarrow \infty$; in case $\delta^2 = \infty$, the two limits are to be interpreted as 0 and as t , respectively. These statements continue to hold if p is replaced by $p + k$ in the denominators of the preceding display, for some fixed $k \in \mathbb{N}$.

Proof. If $\delta^2 > 0$, then $\beta' \Sigma \beta > 0$ for n large enough. For those n , set $w = \Sigma^{1/2} \beta / (\beta' \Sigma \beta)^{1/2}$, note that $\|w\| = 1$ and that $\beta' X' X \beta / n$ can be written as $(\beta' \Sigma \beta) w' V' V w / n$. Using Lemma 7, it follows that $\beta' X' X \beta / n \rightarrow \delta^2$ in probability as $n \rightarrow \infty$. If $\delta^2 = 0$, note that $\beta' X' X \beta / n \leq \beta' \Sigma \beta \lambda_{\max}(V' V / n)$. By Lemma 5, the largest eigenvalue of the matrix $V' V / n$ tends to $(1 + \sqrt{t})^2$ almost surely as $n \rightarrow \infty$. Since $\beta' \Sigma \beta \rightarrow 0$, it follows that $\beta' X' X \beta / n \rightarrow 0 = \delta^2$ almost surely as $n \rightarrow \infty$. Because $\beta' X' X \beta / n$ and $\beta' \Sigma \beta$ both converge to δ^2 (the former in probability), and because $p/n \rightarrow t$, the continuous mapping theorem gives both limits in case $\delta^2 < \infty$, and the first limit in case $\delta^2 = \infty$. In the remaining case where $\delta^2 = \infty$, write the quantity of interest as $(p/n) / ((p/n) / \beta' \Sigma \beta + w' V' V w / n)$, which is easily seen to converge to t , as claimed. The last statement is trivial because p/n and $(p + k)/n$ converge to the same limit. $\square$

Lemma 9. *Under the assumptions of Theorem 4, set $\hat{\delta}^2 = \max\{Y'Y/n - 1, 0\}$. Then $\hat{\delta}^2 \rightarrow \delta^2$ in probability as $n \rightarrow \infty$.*

Proof. By the continuous mapping theorem, it suffices to show that $Y'Y/n \rightarrow 1 + \delta^2$ in probability. Noting that $Y'Y \|X \sim \chi_n^2(\beta' X' X \beta)$, we need to show that

$$\frac{\chi_n^2(\beta' X' X \beta)}{n} = \frac{\chi_n^2(\beta' X' X \beta)}{n + \beta' X' X \beta} \frac{n + \beta' X' X \beta}{n}$$

converges to $1 + \delta^2$ in probability. Use the same arguments as in the proof of Lemma 8 to conclude that $\beta'X'X\beta/n$ tends to δ^2 in probability as $n \rightarrow \infty$. Hence, it remains to show that the first term on the right-hand side in the preceding display converges to 1 in probability as $n \rightarrow \infty$, i.e., that $\mathbb{P}(|\chi_n^2(\beta'X'X\beta)/(n + \beta'X'X\beta) - 1| \geq \varepsilon) \rightarrow 0$ for an arbitrary $\varepsilon > 0$. By the law of iterated expectation, we have

$$\begin{aligned} \mathbb{P}\left(\left|\frac{\chi_n^2(\beta'X'X\beta)}{n + \beta'X'X\beta} - 1\right| \geq \varepsilon\right) &= \mathbb{E}\left[\mathbb{P}\left(\left|\frac{\chi_n^2(\beta'X'X\beta)}{n + \beta'X'X\beta} - 1\right| \geq \varepsilon \middle| X\right)\right] \\ &\leq \frac{1}{\varepsilon^2} \mathbb{E}\left[\frac{2(n + 2\beta'X'X\beta)}{(n + \beta'X'X\beta)^2}\right], \end{aligned}$$

where the inequality is due to Chebyshev's inequality and to Lemma A.2. The integrand of the last term in the preceding display can be bounded by $4/(n + \beta'X'X\beta) \leq 4/n$, and this shows the statement. $\square$

Using Lemmata 5 to 9, the proof of Theorem 4 is rather straightforward. While the detailed proof is long and is given in Appendix B, we next illustrate the basic idea by considering the case where c is fixed and where $t + \delta^2$ is positive. For the maximum likelihood estimator, we have $\rho_2^2(\hat{\beta}_{ML}, \beta, X) = \text{trace}(\Sigma(X'X)^{-1}) = \text{trace}((V'V)^{-1})$ almost surely, so that $\rho_2^2(\hat{\beta}_{ML}, \beta, X) \rightarrow t/(1 - t)$ in probability by Lemma 5. For $\hat{\beta}(c)$, recall the expansion of $\rho_2^2(\hat{\beta}(c), \beta, X)$ in (12). We note that either $p \rightarrow \infty$ (in case $t > 0$), or $\beta'X'X\beta \rightarrow \infty$ in probability (in case $\delta^2 > 0$ in view of Lemma 7 because $n\beta'X'X\beta/n$ can be written as $n\beta'\Sigma\beta(w'V'Vw/n)$, where the unit vector w is given by $\Sigma^{1/2}\beta/(\beta'\Sigma\beta)^{1/2}$). Rewrite the first expected value in (12) as

$$\frac{p}{p + \beta'X'X\beta} \mathbb{E}\left[\frac{p + \beta'X'X\beta}{\chi_p^2(\beta'X'X\beta)} \middle| X\right]$$

and use the assumption that $p \geq 3$ together with Lemma 8 and Lemma 6 to show that the product in the preceding display converges to $t/(t + \delta^2)$ in probability as $n \rightarrow \infty$. Treating the remaining two expected values similarly and using Lemma 5 for $\text{trace}(\Sigma(X'X)^{-1})$ shows that $\rho_2^2(\hat{\beta}(c), \beta, X)$ converges to $r(\delta^2, c, t)$ in probability. The last statement is the convergence of $r(\hat{\delta}^2, c, t)$ to $r(\delta^2, c, t)$. Note that the function $r(\cdot, c, \cdot)$ is continuous on $[0, 1) \times [0, \infty]$ for fixed $c \geq 0$. Since $p/n \rightarrow t < 1$ by assumption and $\hat{\delta}^2 \rightarrow \delta^2$ in probability by Lemma 9, the continuous mapping theorem gives the result.

Theorem 10. *Assume that $n \rightarrow \infty$, and that $p = p(n)$ is such that $p/n \rightarrow t \in [0, 1)$. Moreover, for each p , let Σ be a positive definite $p \times p$ matrix. Then $\sup_{\beta \in \mathbb{R}^p} \rho_2^2(\hat{\beta}(c), \beta, X) \rightarrow \sup_{\delta^2 \geq 0} R(\delta^2, c, t)$ in probability, where*

$$R(\delta^2, c, t) = \frac{t}{1 - t} \left(1 - c \frac{t}{t + \delta^2}\right)^2 + c^2 \frac{t^2 \delta^2}{(1 - \sqrt{t})^2 (t + \delta^2)^2}$$

(again, expressions like $t/(t + \delta^2)$ are to be interpreted as zero if t and δ^2 are both equal to zero). Convergence is uniform in c over compact sets, so that $\sup_{0 \leq c \leq C} |\sup_{\beta \in \mathbb{R}^p} \rho_2^2(\hat{\beta}(c), \beta, X) - \sup_{\delta^2 \geq 0} R(\delta^2, c, t)| = o_P(1)$ for any $C > 0$.

Note that $\sup_{\beta \in \mathbb{R}^p} \rho_2^2(\hat{\beta}(c), \beta, X)$ does not tend to $\sup_{\delta^2 \geq 0} r(\delta^2, c, t)$ because the maximizers of the former will typically depend on X . As before, we sketch the proof and postpone the technical details to Appendix B. On the almost-sure event where $X'X$ is invertible, $\rho_2^2(\hat{\beta}(c), \beta, X)$ is given by the formula in (12). For invertible $X'X$ and for β such that $\beta'(X'X/n)\beta = d^2$, the expression in (12) depends on β only through $\beta'\Sigma\beta$ which is at most $d^2/\lambda_p(V'V/n)$, where $\lambda_p(\dots)$ denotes the smallest eigenvalue of the indicated matrix. [Indeed, we have $d^2 = \|(V'V/n)^{1/2}\Sigma^{1/2}\beta\|^2$ and $\beta'\Sigma\beta = \beta'\Sigma^{1/2}(V'V/n)^{1/2}[(V'V/n)^{-1}](V'V/n)^{1/2}\Sigma^{1/2}\beta \leq d^2/\lambda_p(V'V/n)$.] It follows that $\sup_{\beta \in \mathbb{R}^p} \rho_2^2(\hat{\beta}(c), \beta, X)$ can almost surely be written as the supremum of

$$\begin{aligned} & \text{trace}((V'V)^{-1}) \left(1 - 2c \mathbb{E} \left[\frac{p}{\chi_p^2(nd^2)} \right] + (c^2 + 4c/p) \mathbb{E} \left[\frac{p^2}{(\chi_{p+2}^2(nd^2))^2} \right] \right) \\ & + \frac{c^2 + 4c/p}{\lambda_p(V'V/n)} \mathbb{E} \left[\frac{p^2 d^2}{(\chi_{p+4}^2(nd^2))^2} \right] \end{aligned}$$

over $d^2 \geq 0$. Write $R_*(d^2, c, n, p)$ for the random variable in the preceding display. Considering a fixed tuning parameter c , the claim will follow if we can show that $\sup_{d \geq 0} |R_*(d^2, c, n, p) - R(d^2, c, t)|$ converges to zero in probability. The pointwise result in the case where $t + d^2 > 0$ follows from Lemma 5 and Lemma 8 together with Lemma 7. The uniform result as well as the case $t + d^2 = 0$ is postponed to Appendix B.

Remark. The quantities $r(\delta^2, c, t)$ and $R(\delta^2, c, t)$, as defined in Theorems 4 and 10, respectively, differ by a factor of $(1 - \sqrt{t})^2$ in the denominator of the last term. The proofs reveal that this factor is caused by the fact that $\beta'(X'X/n)\beta$ can differ considerably from its expectation, i.e., from $\beta'\Sigma\beta$, if p/n is not small, because of the gap between the smallest eigenvalue of a large-dimensional random Wishart matrix (or Gram matrix) and the smallest eigenvalue of its expectation, a well-known phenomenon in the theory of random matrices; see, for example, Bai and Silverstein (2010). In particular, in the setting of Theorem 4 with $\delta^2 = 1$, $\beta'(X'X/n)\beta$ converges to one in probability, but $\inf_{b: b'\Sigma b=1} b'(X'X/n)b$ converges to $(1 - \sqrt{t})^2$.

Theorems 4 and 10 together with the preceding remark also provide us with a more precise description of the phenomenon in (10) and (11), and with a better understanding of the underlying cause. More formally, we have the following result.

Theorem 11. *Assume that $n \rightarrow \infty$, and that $p = p(n)$ is such that $p/n \rightarrow t \in [0, 1)$. Moreover, for each p , let Σ be a positive definite $p \times p$ matrix. If $t > 1/9$, then the out-of-sample prediction errors of the James–Stein estimator and of the maximum likelihood estimator are such that*

$$\mathbb{P} \left(\sup_{\beta \in \mathbb{R}^p} \rho_2^2(\hat{\beta}_{JS}, \beta, X) - \rho_2^2(\hat{\beta}_{ML}, \beta, X) > \varepsilon \right) \rightarrow 1 \quad (16)$$

for some $\varepsilon > 0$ (that is given explicitly in the proof). And if $t \leq 1/9$, then the expression on the left-hand side of (16) converges to zero as $n \rightarrow \infty$ for each $\varepsilon > 0$. Finally, irrespective of the value of $t \in [0, 1)$, we have

$$\sup_{\beta \in \mathbb{R}^p} \mathbb{P} \left(\rho_2^2(\hat{\beta}_{JS}, \beta, X) - \rho_2^2(\hat{\beta}_{ML}, \beta, X) > \varepsilon \right) \rightarrow 0 \quad (17)$$

for each $\varepsilon > 0$.

More generally, consider tuning parameters $c_n \geq 0$ that converge to a limit $c \in [0, \infty)$ as $n \rightarrow \infty$, and consider the expression on the left-hand side of (16) with $\hat{\beta}(c_n)$ replacing $\hat{\beta}_{JS}$. The resulting expression converges to one as $n \rightarrow \infty$ for some $\varepsilon > 0$ in the case where $0 \leq c \leq 2$ and $t > [(c-2)/(c+2)]^2$, and in the case where $c > 2$ and $t > 0$. In all other cases, the resulting expression converges to zero for each $\varepsilon > 0$. And (17) holds for each $\varepsilon > 0$ with $\hat{\beta}(c_n)$ replacing $\hat{\beta}_{JS}$ if $c \leq 2$, irrespective of $t \in [0, 1]$.

The proof of this theorem is included in Appendix B.

Remark. In the setting of Theorem 11, it is easy to see that relation (16) holds uniformly over all pairs of n and p subject to $1/9 + \delta \leq p/n \leq 1 - \delta$; in addition, (17) holds uniformly over all such pairs with $p/n \leq 1 - \delta$, for each $\delta > 0$, subject to $n \geq p \geq 3$ (and also uniformly over all positive definite $p \times p$ matrices Σ). Similar statements also apply with $\hat{\beta}(c_n)$ replacing $\hat{\beta}_{JS}$, mutatis mutandis.

Through relations (16) and (17), Theorem 11 provides two complementing views on the worst-case performance of James–Stein-type shrinkage estimators. If the expression on the left-hand side of (16) is large, then the James–Stein estimator $\hat{\beta}_{JS}$ is typically outperformed, from a worst-case perspective, by the maximum likelihood estimator $\hat{\beta}_{ML}$ (whose out-of-sample prediction error is constant in β). Here, ‘typically’ means that for most realizations of the design matrix X there is a parameter β for which $\hat{\beta}_{JS}$ performs worse than $\hat{\beta}_{ML}$. By Theorem 11, we see that this occurs with probability approaching one in the statistically challenging case where $t > 1/9$ (while this occurs with probability approaching zero in the case where $t \leq 1/9$). On the other hand, if the expression on the left-hand side of (17) is small, then with high probability X is such that $\hat{\beta}_{JS}$ outperforms $\hat{\beta}_{ML}$, uniformly in β . In the setting of Theorem 11, the left-hand side of (17) always converges to zero.

If $\rho_2^2(\cdot, \cdot, X)$ is used as a risk-function, then (16) entails that $\hat{\beta}_{JS}$ is outperformed by $\hat{\beta}_{ML}$ in terms of worst-case risk, for most realizations of the design matrix X . This is a worst-case perspective, as is often adopted in frequentist statistical analyses. But the most unfavorable parameter β , for which $\rho_2^2(\hat{\beta}_{JS}, \beta, X)$ is maximized, heavily depends on X . In particular, the relation in (17) entails, for any fixed parameter β , that the probability, that X is such that β is unfavorable, is small.

4 Selection of estimators

In this section, we consider a family of James–Stein-type shrinkage estimators from which we want to select a ‘good’ estimator for prediction out-of-sample. With these findings, we extend results presented in Leeb (2008) to a more general class of estimators. Hence, the framework in this section is adopted from there.

We consider a collection of candidate estimators and measure the performance of each estimator by the conditional mean squared prediction error. Because this prediction error depends on unknown parameters in a complicated fashion, we replace the actual performance by an empirical counterpart and show that those two quantities lie ‘close’ to each other. Using this result, we

show that the empirically best estimator is asymptotically as good as the truly best (oracle) estimator uniformly over a large class of data-generating processes.

More formally, consider a response variable y , a sequence of stochastic explanatory variables $x = (x_i)_{i \geq 1}$, a sequence of unknown parameters $\beta = (\beta_i)_{i \geq 1}$ and an error term u that are related via

$$y = \sum_{i=1}^{\infty} x_i \beta_i + u. \quad (18)$$

Throughout, we assume that the error u is centered with variance $\sigma^2 \geq 0$, that x is centered with variance-covariance net $\Sigma = \mathbb{E}[x_i x_j]_{i \geq 1, j \geq 1}$ such that the x_i 's are not perfectly correlated among themselves, i.e. we require for each $k \geq 1$ and integers $i_1, \dots, i_k$ that the variance-covariance matrix of $(x_{i_1}, \dots, x_{i_k})'$ is positive definite, that u and x are independent and that the series converges in squared mean. Further, we assume that (y, x) is jointly Gaussian.

Suppose that we have a sample of size n which we will denote by (Y, X) , where $Y = (y^{(1)}, \dots, y^{(n)})'$ is a n -vector and $X = (x^{(1)}, \dots, x^{(n)})'$ is a $n \times \infty$ net and where $(y^{(j)}, x^{(j)})$ are i.i.d. copies of the random variables in (18). We use this training sample (Y, X) to find a 'good' estimator for prediction out-of-sample.

To be more precise, for every n , we consider a collection of candidate estimators $\mathcal{E}_n$. The estimators under consideration correspond to finite-dimensional submodels of the overall model in (18), where we restrict some coefficients β_i to zero. Each of these estimators can be identified by a 0-1 sequence $e = (e_i)_{i \geq 1}$, where $e_i = 0$ if the corresponding β_i is restricted to zero, and $e_i = 1$ otherwise. For every estimator $e \in \mathcal{E}_n$, we write $|e|$ for the number of unrestricted components of β_i , i.e. $|e| = \sum_{i \geq 1} e_i$ and we assume that $3 \leq |e| < n$. Unless otherwise stated, we consider fixed parameters β , σ^2 and Σ as in (18), a fixed sample size n and a fixed estimator e . Let $X(e)$ be those columns of X that are included in the submodel that corresponds to the estimator e , thus $X(e)$ is a $n \times |e|$ matrix. Furthermore, let $\hat{\beta}^+(c, e) = (\hat{\beta}_i^+(c, e))_{i \geq 1}$ be the positive part James–Stein-type shrinkage estimator. This estimator is derived as follows: If $e_i = 0$, then the i -th component of the estimator is defined as 0. The remaining $|e|$ components are obtained by using the formula

$$\left(1 - \hat{\sigma}^2(e) \frac{c|e|}{\hat{\beta}_{LS}(e)' X(e)' X(e) \hat{\beta}_{LS}(e)} \right)_+ \hat{\beta}_{LS}(e), \quad (19)$$

where $(x)_+ = \max\{x, 0\}$, where $\hat{\sigma}^2(e)$ and $\hat{\beta}_{LS}(e)$ are the residual sum of squares divided by $n - |e|$ and the least squares estimator, respectively, when regressing Y on $X(e)$, and where $c \geq 0$ is a tuning parameter. On the probability zero event where $X(e)$ is not of full rank, we use the Moore–Penrose inverse in the formula for the least squares estimator. Assume that $(y^{(0)}, x^{(0)})$ is a new copy of (y, x) as in (18) that is independent of the training sample. We then predict $y^{(0)}$ using the predictor

$$\hat{y}^{(0)}(c) = \sum_{i=1}^{\infty} x_i^{(0)} \hat{\beta}_i^+(c, e).$$

The conditional mean squared prediction error corresponding to e will be denoted by $\rho_3^2(e)$ and is defined by

$$\rho_3^2(e) = \mathbb{E} \left[\left(\hat{y}^{(0)}(c) - y^{(0)} \right)^2 \middle| X, Y \right], \quad (20)$$

where the expectation is taken with respect to $(y^{(0)}, x^{(0)})$ and the training sample is treated as fixed. Of course, the prediction error also depends on quantities like X , Y and c , but this dependence is not shown for the sake of brevity. Let a_n be the shrinkage factor of the estimator in (19), i.e.,

$$a_n = \min \left\{ \frac{c \hat{\sigma}^2(e) |e|}{\hat{\beta}_{LS}(e)' X(e) X(e)' \hat{\beta}_{LS}(e)}, 1 \right\}. \quad (21)$$

As a predictor for $\rho_3^2(e)$, we use

$$\begin{aligned} \hat{\rho}_3^2(e) &= \hat{\sigma}^2(e) \frac{|e|}{n - |e| + 1} - 2a_n \hat{\sigma}^2(e) \frac{|e|}{n - |e| + 1} \\ &\quad + a_n^2 \left(\hat{\sigma}^2(e) \frac{|e|}{n - |e| + 1} + \frac{Y'Y}{n} - \hat{\sigma}^2(e) \right) + \hat{\sigma}^2(e). \end{aligned} \quad (22)$$

Theorem 12. *For every $\varepsilon > 0$, we have*

$$\begin{aligned} &\mathbb{P} \left(\left| \rho_3^2(e) - \hat{\rho}_3^2(e) \right| \geq \varepsilon \right) \\ &\leq 16 \exp \left(-n \left(1 - \frac{|e|}{n} \right)^2 f \left(\varepsilon \left(1 - \frac{|e|}{n} \right), \sigma_y^2 \right) \right), \end{aligned} \quad (23)$$

where $\sigma_y^2 = \text{Var}(y)$ for y as in (18) and where $f(a, b) = a^2 / (16(a + 4b)^2)$.

Proof. Recall the linear model in (18) and write $x(e)$ for the $|e|$ vector of those explanatory variables that are included in the submodel that corresponds to the estimator e . Because y and $x(e)$ are jointly Gaussian, the conditional distribution of y given $x(e)$ is again Gaussian where the conditional mean is linear in $x(e)$ and the conditional variance, denoted by $\sigma^2(e)$, is constant in $x(e)$, i.e. $y|x(e) \sim N(x(e)'\theta, \sigma^2(e))$ for some appropriate $|e|$ -vector θ . This means that we can rewrite (18) as

$$y = x(e)'\theta + v, \quad (24)$$

where $v \sim N(0, \sigma^2(e))$ is independent of $x(e)$. Write again $X(e)$ for the $n \times |e|$ matrix of those explanatory variables in the sample that are included in the submodel that corresponds to the estimator e . Note that the i -th entry of Y and the i -th row of $X(e)$ are independent copies of the random variables y and $x(e)$ in (24) and that we can rewrite the linear model in matrix form as

$$Y = X(e)\theta + w, \quad (25)$$

where $w \sim N(0, \sigma^2(e)I_n)$ is independent of $X(e)$. Let $\hat{\theta}^+(c)$ be the positive part James–Stein-type shrinkage estimator of θ in model (25). Then the conditional

mean squared prediction error of the estimator $\hat{\beta}^+(c, e)$, i.e., the expression in (20), equals

$$\begin{aligned}\rho_3^2(e) &= \mathbb{E} \left[(x(e)' \theta - x(e)' \hat{\theta}^+(c) + v)^2 \middle| X, Y \right] \\ &= (\hat{\theta}^+(c) - \theta)' \Sigma(e) (\hat{\theta}^+(c) - \theta) + \sigma^2(e),\end{aligned}\quad (26)$$

where $\Sigma(e)$ is the variance-covariance matrix of $x(e)$. The result now follows upon using Proposition C.1 with $Z = X(e)$, $s^2 = \sigma^2(e)$, $S = \Sigma(e)$ and $p = |e|$ and noting that $\theta' \Sigma(e) \theta + \sigma^2(e) = \sigma_y^2$. $\square$

The upper bound in Theorem 12 depends on the known quantities ε , $|e|$ and n , and on the unknown quantity σ_y^2 . However, σ_y^2 can be estimated from the sample by, e.g., $(n-1)^{-1} \sum_{i=1}^n (y^{(i)} - \bar{y})^2$, where $\bar{y}$ is the mean of the training sample $y^{(1)}, \dots, y^{(n)}$. We note that the upper bound does not tend to zero as ε gets larger and that the result can be improved at the expense of a more complicated structure of the bound. But in the improved version, the influence of ε , $|e|$ and n would be hidden in the structure's complex form. The upper bound is exponentially small in n if only $|e|/n$ and σ_y^2 are not too large. These considerations together with Bonferroni's inequality entail the following corollary.

Corollary 13. *Consider a finite and non-empty collection of estimators $\mathcal{E}_n$, let $s_n = \sup_{e \in \mathcal{E}_n} |e|/n$. Then we have for each $\varepsilon > 0$ and each finite $d > 0$*

$$\begin{aligned}\sup_{\substack{\beta, \sigma^2, \Sigma \text{ as in (18)} \\ \text{Var}(y) \leq d}} \mathbb{P} \left(\sup_{e \in \mathcal{E}_n} |\rho_3^2(e) - \hat{\rho}_3^2(e)| \geq \varepsilon \right) \\ \leq 16 |\mathcal{E}_n| \exp \left(-n (1 - s_n)^2 f(\varepsilon (1 - s_n), d) \right),\end{aligned}\quad (27)$$

where $|\mathcal{E}_n|$ denotes the number of estimators in collection $\mathcal{E}_n$ and where $f(\cdot, \cdot)$ is defined as in Theorem 12.

Proof. Note that

$$\begin{aligned}\mathbb{P} \left(\sup_{e \in \mathcal{E}_n} |\rho_3^2(e) - \hat{\rho}_3^2(e)| \geq \varepsilon \right) &\leq \sum_{e \in \mathcal{E}_n} \mathbb{P} \left(|\rho_3^2(e) - \hat{\rho}_3^2(e)| \geq \varepsilon \right) \\ &\leq \sum_{e \in \mathcal{E}_n} 16 \exp \left(-n \left(1 - \frac{|e|}{n} \right)^2 f \left(\varepsilon \left(1 - \frac{|e|}{n} \right), \sigma_y^2 \right) \right),\end{aligned}$$

where the first inequality follows from Bonferroni's inequality and the second inequality follows from Theorem 12. The result is obtained by noting that the upper bound in the preceding display is increasing in $|e|/n$ and σ_y^2 . $\square$

The next result shows that $\rho_3^2(e)$ and $\hat{\rho}_3^2(e)$ are close uniformly over a large class of candidate estimators and uniformly over a large class of data-generating processes. More precisely, we show that the difference of those two quantities tend to zero in probability at a certain rate.

Theorem 14. For each sample size n , let $\mathcal{E}_n$ be a non-empty and finite collection of candidate estimators and let $s_n = \sup_{e \in \mathcal{E}_n} |e|/n$. Define r_n as

$$r_n = \sqrt{\frac{\log(|\mathcal{E}_n| + 1)}{n(1 - s_n)^4}}.$$

Assume that $r_n \rightarrow 0$ as $n \rightarrow \infty$. Then we have

$$\sup_{e \in \mathcal{E}_n} |\hat{\rho}_3^2(e) - \rho_3^2(e)| = O_p(r_n), \quad (28)$$

where $O_p(r_n)/r_n$ denotes a sequence of random variables that is bounded in probability uniformly over all data-generating processes as in (18) that satisfy $\text{Var}(y) \leq d$, where $d > 0$ is an arbitrary fixed finite constant.

Proof. We have to show that for every $\delta > 0$ there exists a real number $K > 0$ such that

$$\sup_{\substack{\beta, \sigma^2, \Sigma \text{ as in (18)} \\ \text{Var}(y) \leq d}} \mathbb{P} \left(\sup_{e \in \mathcal{E}_n} |\hat{\rho}_3^2(e) - \rho_3^2(e)| \geq Kr_n \right) < \delta$$

holds for all n . Assume that β , σ^2 and Σ are such that $\text{Var}(y) \leq d$ and let K be some fixed constant. Recall that $f(a, b) = a^2/(16(a + 4b)^2)$ and note that $f(a, b)$ is decreasing in its second argument. Now, we can use Corollary 13 to bound the left-hand side of the preceding display from above by

$$\begin{aligned} & 16|\mathcal{E}_n| \exp \left(-n(1 - s_n)^2 f(Kr_n(1 - s_n), d) \right) \\ &= 16 \exp \left(-n(1 - s_n)^2 \frac{K^2 r_n^2 (1 - s_n)^2}{16(Kr_n(1 - s_n) + 4d)^2} + \log(|\mathcal{E}_n|) \right) \\ &\leq 16 \exp \left(-\frac{K^2}{16} \frac{\log(|\mathcal{E}_n| + 1)}{(Kr_n(1 - s_n) + 4d)^2} + \log(|\mathcal{E}_n| + 1) \right) \\ &\leq 16 \exp \left(-\log(|\mathcal{E}_n| + 1) \left(\frac{K^2}{16(1 + 4d)^2} - 1 \right) \right). \end{aligned}$$

The second inequality is valid for $n \geq n(K)$ and is obtained from the fact that $Kr_n(1 - s_n) \leq Kr_n \rightarrow 0$ as $n \rightarrow \infty$ and is hence smaller than 1 for sufficiently large n . Hence, the term in the last line of the preceding display can be made small by choosing K large enough. $\square$

If s_n is bounded away from 1, the rate in the previous theorem is essentially the same as the rate in Theorem 3.4 in Leeb (2008). Here, we have a factor of $(1 - s_n)^4$ in the denominator of r_n whereas in Leeb (2008) there is a factor of $(1 - s_n)^3$. But note that the rate in Theorem 14 is worse if s_n is close to 1. The additional factor $(1 - s_n)$ stems from the convergence rate of the smallest eigenvalue of a Wishart matrix (cf. the remark after Corollary C.12). To ensure that $r_n \rightarrow 0$ as $n \rightarrow \infty$, we need s_n to be bounded away from 1, i.e., we need to exclude models that are too complex, and we need that $\log(|\mathcal{E}_n|) = o(n(1 - r_n)^4)$. The result of the previous theorem immediately implies the following corollary that shows how to select a ‘good’ estimator.

Corollary 15. *Let the assumptions of Theorem 14 hold, and consider measurable minimizers of $\hat{\rho}_3^2(e)$ and $\rho_3^2(e)$ over the collection $\mathcal{E}_n$, more precisely*

$$\hat{e}_n^* = \arg \min_{e \in \mathcal{E}_n} \hat{\rho}_3^2(e) \quad \text{and} \quad e_n^* = \arg \min_{e \in \mathcal{E}_n} \rho_3^2(e).$$

Then

$$|\rho_3^2(\hat{e}_n^*) - \rho_3^2(e_n^*)| = O_p(r_n) \quad (29)$$

and

$$|\hat{\rho}_3^2(\hat{e}_n^*) - \rho_3^2(\hat{e}_n^*)| = O_p(r_n), \quad (30)$$

where $O_p(r_n)/r_n$ denotes a sequence of random variables that is bounded in probability uniformly over all data-generating processes as considered in (18) that satisfy $\text{Var}(y) \leq d$ for some finite constant $d > 0$.

Proof. Note that $\rho_3^2(\hat{e}_n^*) - \rho_3^2(e_n^*)$ is nonnegative as e_n^* is a minimizer of $\rho_3^2(\cdot)$. Furthermore, this difference can be rewritten as

$$\left[\rho_3^2(\hat{e}_n^*) - \hat{\rho}_3^2(\hat{e}_n^*) \right] + \left[\hat{\rho}_3^2(\hat{e}_n^*) - \hat{\rho}_3^2(e_n^*) \right] + \left[\hat{\rho}_3^2(e_n^*) - \rho_3^2(e_n^*) \right].$$

By Theorem 14, the first and the third term in the preceding display are $O_p(r_n)$ uniformly over the set of parameters β , σ^2 and Σ as in (18) satisfying $\text{Var}(y) \leq d$. The middle term in the preceding display is nonpositive because $\hat{e}_n^*$ is a minimizer of $\hat{\rho}_3^2(\cdot)$, which shows the result. The second statement is a direct consequence of Theorem 14. $\square$

The first statement of the preceding corollary shows that the empirically best estimator, i.e., the minimizer of $\hat{\rho}_3^2(e)$, is asymptotically as good as the truly best estimator, i.e., the minimizer of $\rho_3^2(e)$. The second statement gives us a way to estimate the true predictive performance of the selected estimator.

5 Discussion

We have derived explicit finite-sample formulae for the out-of-sample prediction error of the James–Stein estimator and of related James–Stein-type shrinkage estimators in a linear regression model with Gaussian errors and fixed design. In an example with a particular design matrix X , we have found that the James–Stein estimator no longer dominates the maximum likelihood estimator. We have shown that this phenomenon generally occurs for most design matrices X if the ratio of the number of explanatory variables in the model (p) and the sample size (n) exceeds $1/9$, in the sense of statement (16) of Theorem 11. At the same time, we have also shown that the James–Stein estimator outperforms the maximum likelihood estimator for most design matrices X , uniformly in the underlying parameters, in the sense of statement (17) of Theorem 11. Our findings suggest that the James–Stein estimator can perform poorly for prediction out-of-sample from a frequentist worst-case perspective. But our findings also suggest, in the setting considered here, that the worst-case performance does not properly reflect the performance in the typical case, and that the James–Stein

estimator performs favorably compared to maximum likelihood in the typical case, uniformly in the underlying parameters.

The phenomenon that the James–Stein estimator dominates the maximum likelihood estimator for in-sample prediction but can fail to do so for prediction out-of-sample is linked to the fact that the eigenvalues of $X'X/n$ can differ from the eigenvalues of Σ ; see relations (2) and (3), as well as the remark following Theorem 10. We therefore expect that other estimators, that are designed to perform well for in-sample prediction, can exhibit similar phenomena when used for prediction out-of-sample. This includes various shrinkage estimators like estimators based on model selection, penalized maximum likelihood, and other forms of regularization. Although beyond the scope of this work, it would be particularly interesting to study bridge estimators (Frank and Friedman, 1993), in particular, the LASSO (Tibshirani, 1996) and ridge regression (Hörl and Kennard, 1970), as well as the Dantzig selector (Candes and Tao, 2007) with regards to their performance as predictors out-of-sample when p/n is not small.

Furthermore, we have considered a linear regression model, possibly infinite dimensional, together with a family of James–Stein-type shrinkage estimators, and we have shown how to select an estimator that performs well for prediction out-of-sample. We have focused on the challenging case where the number of explanatory variables can be of the same order as sample size, i.e., $p = p(n)$ such that $p/n \rightarrow t \in [0, 1)$ and where the number of candidate estimators can be much larger than sample size (cf. (28) in Theorem 14). Assuming that the sample is jointly Gaussian, we have derived a finite-sample performance bound, i.e., we have replaced the actual performance of an estimator by an empirical counterpart that converges to the true prediction error exponentially fast as $n \rightarrow \infty$ (cf. (23) in Theorem 12). Using this, we have shown that the empirically best estimator, i.e., the estimator that minimizes the estimated prediction error, is asymptotically as good as the truly best estimator, in the sense of statement (29) in Corollary 15, and that we can estimate the performance of the selected estimator in a consistent fashion; cf. (30) in Corollary 15. All results hold uniformly over a large class of data-generating processes as in (18) such that the variance of the response is bounded by some finite real number. It would be interesting to investigate how the selected estimator behaves for constructing prediction intervals or, more generally, for other inference procedures, i.e., results paralleling those in Leeb (2009). In that paper, the author shows how to obtain prediction intervals, that are approximately valid with high probability, when the underlying estimators are least squares estimators.

Appendix

A Some results for the central and the noncentral chi-square distribution

Let $\chi_k^2(\lambda)$ denote a random variable that is chi-square distributed with $k \geq 1$ degrees of freedom and noncentrality parameter $\lambda \geq 0$. We only give an overview of some results. For more details, see Chapter 29 in Johnson et al. (1995). We write χ_k^2 as shorthand for $\chi_k^2(0)$. Let J_λ be a random variable that follows a Poisson distribution with mean $\lambda/2$ and let $\chi_{k+2J_\lambda}^2$ be a random variable that follows conditionally on J_λ a central chi-square distribution with $k+2J_\lambda$ degrees of freedom. We note that $\chi_{k+2J_\lambda}^2$ has the same distribution as $\chi_k^2(\lambda)$. In that sense, the law of $\chi_k^2(\lambda)$ can be viewed as a central chi-square distribution with random degrees of freedom equal to $k+2J_\lambda$. From this, it follows that $\chi_k^2(\lambda)$ is stochastically larger than χ_k^2 , whence

$$\mathbb{E} \left[\frac{1}{(\chi_k^2(\lambda))^m} \right] \leq \mathbb{E} \left[\frac{1}{(\chi_k^2)^m} \right] \quad (31)$$

for each $m \geq 1$ (where the expected values can be infinite). The density of a central chi-square distribution with k degrees of freedom equals

$$p_{\chi_k^2}(x) = \frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} e^{-x/2}. \quad (32)$$

More generally, the density of a noncentral chi-square distribution with k degrees of freedom and noncentrality parameter $\lambda \geq 0$ can be found in Johnson et al. (1995) and equals

$$p_{\chi_k^2(\lambda)}(x) = \sum_{j=0}^{\infty} \frac{e^{-\lambda/2} (\lambda/2)^j}{j!} p_{\chi_{k+2j}^2}(x). \quad (33)$$

The formula in (33) reveals the relation to the Poisson distribution. In the following results, we compute some moments of the central and the noncentral chi-square distribution as well as the moment generating function of the noncentral chi-square distribution. Of course, these results are well-known and are stated here only for the sake of completeness.

Lemma A.1. *Fix an integer $k \geq 1$. Then*

$$\mathbb{E} \left[(\chi_k^2)^{d/2} \right] = 2^{d/2} \frac{\Gamma((k+d)/2)}{\Gamma(k/2)}$$

for each integer $d \geq 1$. Furthermore,

$$\mathbb{E} \left[(\chi_k^2)^{-d} \right] = \prod_{i=1}^d \frac{1}{k-2i}$$

for each integer d satisfying $1 \leq d < k/2$.

Proof. Use (32) to obtain

$$\begin{aligned}\mathbb{E}[(\chi_k^2)^{d/2}] &= \int_0^\infty \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k+d)/2-1} e^{-x/2} dx \\ &= 2^{d/2} \frac{\Gamma((k+d)/2)}{\Gamma(k/2)} \int_0^\infty \frac{1}{2^{(k+d)/2}\Gamma((k+d)/2)} x^{(k+d)/2-1} e^{-x/2} dx.\end{aligned}$$

For the second statement, use again (32) to obtain

$$\begin{aligned}\mathbb{E}\left[\frac{1}{(\chi_k^2)^d}\right] &= \int_0^\infty \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k-2d)/2-1} e^{-x/2} dx = (1/2^d) \frac{\Gamma((k-2d)/2)}{\Gamma(k/2)} \\ &= \frac{1}{(k-2)(k-4)\dots(k-2d)},\end{aligned}$$

where the third equality is obtained by using the recursion formula for the gamma function. More precisely, iterate the recursion $\Gamma(x+1) = x\Gamma(x)$, that is valid for $x > 0$, to get

$$\begin{aligned}\Gamma(k/2) &= (k/2-1)(k/2-2)\dots(k/2-d)\Gamma(k/2-d) \\ &= 1/2^d (k-2)(k-4)\dots(k-2d)\Gamma((k-2d)/2).\end{aligned}$$

□

Lemma A.2. Fix an integer $k \geq 1$. Then $\mathbb{E}[\chi_k^2(\lambda)] = k + \lambda$ and $\mathbb{E}[(\chi_k^2(\lambda))^2] = (k + \lambda)^2 + 2(k + 2\lambda)$. Furthermore, we have

$$\mathbb{E}\left[\frac{1}{(\chi_k^2(\lambda))^d}\right] = \sum_{j=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} \prod_{i=1}^d \frac{1}{k+2j-2i}$$

for each integer d satisfying $1 \leq d < k/2$.

Proof. For the first two statements, use the first statement in Lemma A.1 to obtain for any integer $m \geq 1$

$$\begin{aligned}\mathbb{E}[(\chi_k^2(\lambda))^m] &= \int_0^\infty \sum_{j=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} x^m p_{\chi_{k+2j}^2}(x) dx \\ &= \sum_{j=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} 2^m \frac{\Gamma((k+2j)/2 + m)}{\Gamma((k+2j)/2)} \\ &= \sum_{j=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} 2^m \prod_{i=0}^{m-1} (k/2 + j + i),\end{aligned}\tag{34}$$

where the last equality is obtained upon using the recursion formula for the gamma function, i.e.,

$$\Gamma((k+2j)/2 + m) = [(k+2j)/2 + m - 1] \dots [(k+2j)/2] \Gamma((k+2j)/2),$$

and where interchanging summation and integration is allowed by the monotone convergence theorem. For $m = 1$, (34) reduces to

$$k \sum_{j=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} + 2 \sum_{j=0}^{\infty} j \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} = k + \lambda,$$

where the equality is obtained by recalling that the second sum is the expected value of a Poisson distribution with intensity $\lambda/2$. For $m = 2$, (34) reduces to

$$\begin{aligned} & \sum_{j=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} (k+2j)(k+2j+2) \\ &= (k^2 + 2k) \sum_{j=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} + 4(k+1) \sum_{j=0}^{\infty} j \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} \\ & \quad + 4 \sum_{j=0}^{\infty} j^2 \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} \\ &= k^2 + 2k + 2(k+1)\lambda + \lambda(2+\lambda), \end{aligned}$$

where the equality is obtained by recalling that the second moment of a Poisson distribution with intensity $\lambda/2$ is $\lambda/2 + \lambda^2/4$. Rearranging the terms on the right-hand side in the preceding display gives the statement. For the remaining statement, note that

$$\begin{aligned} \mathbb{E}[(\chi_k^2(\lambda))^{-d}] &= \int_0^{\infty} \sum_{j=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} x^{-d} p_{\chi_{k+2j}^2} dx \\ &= \sum_{j=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} \mathbb{E} \left[\frac{1}{(\chi_{k+2j}^2)^d} \right] \\ &= \sum_{j=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} \prod_{i=1}^d \frac{1}{k+2j-2i}, \end{aligned}$$

where the last equality is obtained from the second statement in Lemma A.1 and where again interchanging summation and integration is allowed by the monotone convergence theorem. $\square$

Lemma A.3. *Let $M_{\chi_k^2(\lambda)}(\cdot)$ denote the moment generating function of the non-central chi-square distribution with $k \geq 1$ degrees of freedom and noncentrality parameter $\lambda \geq 0$. The moment generating function is finite if $t < 1/2$ and is then given by*

$$M_{\chi_k^2(\lambda)}(t) = \exp \left(\frac{\lambda t}{1-2t} - \frac{k}{2} \log(1-2t) \right). \quad (35)$$

Proof. Fix an arbitrary $t < 1/2$. The moment generating function of the central chi-square distribution is

$$\begin{aligned} M_{\chi_k^2}(t) &= \int_0^{\infty} e^{tx} p_{\chi_k^2}(x) dx \\ &= \int_0^{\infty} \frac{1}{2^{k/2} \Gamma(k/2)} x^{k/2-1} e^{-x(1-2t)/2} dx = (1-2t)^{-k/2}, \end{aligned}$$

where we integrated by substituting $z = x(1 - 2t)$ in the last equation. Using (33) and the formula in the preceding display, we obtain

$$\begin{aligned}
M_{\chi_k^2(\lambda)}(t) &= \sum_{j=0}^{\infty} \frac{e^{-\lambda/2}(\lambda/2)^j}{j!} \int_0^{\infty} e^{tx} p_{\chi_{k+2j}^2}(x) dx \\
&= e^{-\lambda/2} \sum_{j=0}^{\infty} \frac{(\lambda/2)^j}{j!} (1 - 2t)^{-(k+2j)/2} \\
&= e^{-\lambda/2} (1 - 2t)^{-k/2} \sum_{j=0}^{\infty} \left(\frac{\lambda}{2(1 - 2t)} \right)^j \frac{1}{j!} \\
&= (1 - 2t)^{-k/2} \exp(-\lambda/2 + \lambda/(2(1 - 2t))) \\
&= (1 - 2t)^{-k/2} \exp(\lambda t/(1 - 2t)),
\end{aligned}$$

where interchanging limes and integral in the first line is allowed by the monotone convergence theorem. $\square$

B Technical details for Section 3

We start with the auxiliary results that were used in the proofs of Lemma 5 and of Lemma 6. Then we give detailed proofs of Theorems 4 and 10, and finally, we prove Theorem 11.

Lemma B.1. *For $t \geq 0$, $a = (1 - \sqrt{t})^2$ and $b = (1 + \sqrt{t})^2$, we have*

$$\int_a^b \frac{1}{x^2} \sqrt{(b-x)(x-a)} dx = 2\pi \frac{\min\{1, t\}}{|1-t|},$$

where the expression on the right-hand side of the preceding display is to be interpreted as ∞ in case $t = 1$.

Proof. The case $t = 0$ is trivial. Now, consider the case where $t > 0$. The substitution $z = \sqrt{(b-x)/(x-a)}$ implies that $x = (az^2 + b)/(z^2 + 1)$, $\sqrt{(b-x)(x-a)} = z(b-a)/(z^2 + 1)$ and that $dx/dz = 2z(a-b)/(z^2 + 1)^2$. With this, the integral of interest can be written as

$$\begin{aligned}
&2(b-a)^2 \int_0^{\infty} \frac{z^2}{(az^2 + b)^2(z^2 + 1)} dz \\
&= 2(b-a)^2 \int_0^{\infty} \left[\frac{a}{(b-a)^2} \frac{1}{(az^2 + b)} + \frac{b}{b-a} \frac{1}{(az^2 + b)^2} - \frac{1}{(b-a)^2} \frac{1}{z^2 + 1} \right] dz,
\end{aligned} \tag{36}$$

where the equality follows from a partial fraction decomposition. More precisely, note that the integrand in (36) can be written as

$$\begin{aligned}
&\frac{a(az^2 + b)(z^2 + 1) + b(b-a)(z^2 + 1) - (az^2 + b)^2}{(b-a)^2(az^2 + b)^2(z^2 + 1)} \\
&= \frac{(az^2 + b)(a-b) + b(b-a)(z^2 + 1)}{(b-a)^2(az^2 + b)^2(z^2 + 1)}.
\end{aligned}$$

After factoring out $(b - a)$ and simplifying, we see that the term on the right-hand side of the preceding display equals the integrand of the first line in the second-to-last display. Recall that $\int_0^w (u^2 + 1)^{-1} du = \arctan(w)$, that $\lim_{w \rightarrow \infty} \arctan(w) = \pi/2$ and that $2k \int_0^w (u^2 + 1)^{-(1+k)} du = w(w^2 + 1)^{-k} + (2k - 1) \int_0^w (u^2 + 1)^{-k} du$ whenever $k > 0$. For the case where $t \neq 1$, this entails that $\int_0^\infty (az^2 + b)^{-1} dz = (1/\sqrt{ab}) \int_0^\infty (u^2 + 1)^{-1} du = \pi/(2\sqrt{a}\sqrt{b})$ as well as $\int_0^\infty (az^2 + b)^{-2} dz = (1/\sqrt{ab^3}) \int_0^\infty (u^2 + 1)^{-2} du = \pi/(4\sqrt{a}\sqrt{b^3})$, where the first equality in the preceding two equations is obtained by the substitution $u = z\sqrt{a/b}$. Hence, the integral in (36) reduces to

$$\frac{\pi}{2\sqrt{a}\sqrt{b}} \left[2a + (b - a) - 2\sqrt{a}\sqrt{b} \right] = \frac{\pi}{2\sqrt{a}\sqrt{b}} (\sqrt{b} - \sqrt{a})^2.$$

In case $t < 1$, note that $\sqrt{a} = 1 - \sqrt{t}$. Hence, the right-hand side of the preceding display reduces to $4t\pi/[2(1 - \sqrt{t})(1 + \sqrt{t})] = 2t\pi/(1 - t)$. In case $t > 1$, we have $\sqrt{a} = \sqrt{t} - 1$. With this the right-hand side of the preceding display simplifies to $4\pi/[2(\sqrt{t} - 1)(\sqrt{t} + 1)] = 2\pi/(t - 1)$, as required. And in case $t = 1$, we have $a = 0$ and $b = 4$, so that the integrand in (36) reduces to $[1 - 1/(z^2 + 1)]/16$ and hence integrates to ∞ . $\square$

Lemma B.2. *Let Z have a Poisson distribution with expected value λ . Then we have for every $z < \lambda$*

$$\mathbb{P}(Z \leq z) \leq e^{-\lambda} (e\lambda/z)^z.$$

Proof. Recall that the moment generating function of the Poisson distribution exists for all $s \in \mathbb{R}$ and equals $M_Z(s) = e^{\lambda(e^s - 1)}$. Hence, we have for every $s > 0$

$$\mathbb{P}(Z \leq z) = \mathbb{P}(e^{-sZ} \geq e^{-sz}) \leq M_Z(-s)e^{sz} = e^{\lambda(e^{-s} - 1) + sz},$$

where we used Markov's inequality. The upper bound is minimized for $s = \log(\lambda/z)$. Plugging this into the far right-hand side of the preceding display gives $e^{z - \lambda + \log(\lambda/z)z} = e^{-\lambda} (e\lambda/z)^z$ as required. $\square$

Proof of Theorem 4. For the maximum likelihood estimator, we have $\rho_2^2(\hat{\beta}_{ML}, \beta, X) = \text{trace}(\Sigma(X'X)^{-1}) = \text{trace}((V'V)^{-1})$ almost surely, so that $\rho_2^2(\hat{\beta}_{ML}, \beta, X) \rightarrow t/(1 - t)$ in probability by Lemma 5. Recalling (12), $\rho_2^2(\hat{\beta}(c), \beta, X)$ is given by

$$\begin{aligned} \text{trace}((V'V)^{-1}) & \left(1 - 2c \mathbb{E} \left[\frac{p}{\chi_p^2(\beta'X'X\beta)} \middle\| X \right] \right. \\ & \quad \left. + (c^2 + 4c/p) \mathbb{E} \left[\frac{p^2}{(\chi_{p+2}^2(\beta'X'X\beta))^2} \middle\| X \right] \right) \\ & \quad + (c^2 + 4c/p) \mathbb{E} \left[\frac{p^2 \beta' \Sigma \beta}{(\chi_{p+4}^2(\beta'X'X\beta))^2} \middle\| X \right] \end{aligned} \quad (37)$$

almost surely. Our first goal is to show that $\sup_{0 \leq c \leq C} |\rho_2^2(\hat{\beta}(c), \beta, X) - r(\delta^2, c, t)| = o_P(1)$.

In the case where $t + \delta^2 = 0$, note that $r(0, c, 0) = 0$, so that $\sup_{0 \leq c \leq C} |\rho_2^2(\hat{\beta}(c), \beta, X) - r(\delta^2, c, t)|$ is bounded from above by

$$\begin{aligned} & \text{trace}((V'V)^{-1}) \left(1 + 2C \mathbb{E} \left[\frac{p}{\chi_p^2(0)} \right] + (C^2 + 4C/p) \mathbb{E} \left[\frac{p^2}{(\chi_{p+2}^2(0))^2} \right] \right) \\ & + (C^2 + 4C/p) \beta' \Sigma \beta \mathbb{E} \left[\frac{p^2}{(\chi_{p+4}^2(0))^2} \right] \end{aligned}$$

because $\chi_{p+k}^2(\beta' X' X \beta)$ is stochastically larger than $\chi_{p+k}^2(0)$ for each $k \geq 0$. Using Lemma A.1 and the fact that $p \geq 3$, we see that $\mathbb{E}[p/\chi_p^2(0)] = p/(p-2) \leq 3$, that $\mathbb{E}[(p/\chi_{p+2}^2(0))^2] = p^2/(p(p-2)) \leq 3$ and that $\mathbb{E}[(p/\chi_{p+4}^2(0))^2] = p^2/(p(p+2)) \leq 1$. Using this and because of the fact that both $\text{trace}((V'V)^{-1})$ and $\beta' \Sigma \beta$ converge to zero (the former in probability by Lemma 5), it follows that the expression in the preceding display converges to zero in probability.

In the case where $t + \delta^2 > 0$, we use the formula for $\rho_2^2(\hat{\beta}(c), \beta, X)$ in (37), the formula for $r(\delta^2, c, t)$, and the triangle inequality to bound $\sup_{0 \leq c \leq C} |\rho_2^2(\hat{\beta}(c), \beta, X) - r(\delta^2, c, t)|$ from above by

$$\begin{aligned} & \left| \text{trace}((V'V)^{-1}) - \frac{t}{1-t} \right| \\ & + 2C \left| \text{trace}((V'V)^{-1}) \mathbb{E} \left[\frac{p}{\chi_p^2(\beta' X' X \beta)} \middle| X \right] - \frac{t^2}{(1-t)(t+\delta^2)} \right| \\ & + C^2 \left| \text{trace}((V'V)^{-1}) \mathbb{E} \left[\frac{p^2}{(\chi_{p+2}^2(\beta' X' X \beta))^2} \middle| X \right] - \frac{t^3}{(1-t)(t+\delta^2)^2} \right| \\ & + C^2 \left| \mathbb{E} \left[\frac{p^2 \beta' \Sigma \beta}{(\chi_{p+4}^2(\beta' X' X \beta))^2} \middle| X \right] - \frac{t^2 \delta^2}{(t+\delta^2)^2} \right| \\ & + 4(C/p) \left(\text{trace}((V'V)^{-1}) \mathbb{E} \left[\frac{p^2}{(\chi_{p+2}^2(\beta' X' X \beta))^2} \middle| X \right] \right. \\ & \quad \left. + \beta' \Sigma \beta \mathbb{E} \left[\frac{p^2}{(\chi_{p+4}^2(\beta' X' X \beta))^2} \middle| X \right] \right). \end{aligned}$$

We already know that the first term of the preceding display converges to zero in probability in view of Lemma 5. For the remaining four terms, we note that either $p \rightarrow \infty$ (in case $t > 0$), or $\beta' X' X \beta \rightarrow \infty$ in probability (in case $\delta^2 > 0$ in view of Lemma 7 because $n\beta' X' X \beta/n$ can be written as $n\beta' \Sigma \beta (w' V' V w/n)$, where the unit vector w is given by $\Sigma^{1/2} \beta / (\beta' \Sigma \beta)^{1/2}$ at least for n large enough). Using the assumption that $p \geq 3$, Lemma 6 can be used to deal with the expected values in the preceding display. Together with the fact that $\text{trace}((V'V)^{-1}) \rightarrow$

$t/(1-t)$ in probability, this shows that the sum of the four remaining terms converges to the same limit as

$$\begin{aligned} & 2C \frac{t}{1-t} \left| \frac{p}{p + \beta' X' X \beta} - \frac{t}{t + \delta^2} \right| + C^2 \frac{t}{1-t} \left| \frac{p^2}{(p + 2 + \beta' X' X \beta)^2} - \frac{t^2}{(t + \delta^2)^2} \right| \\ & + C^2 \left| \beta' \Sigma \beta \frac{p^2}{(p + 4 + \beta' X' X \beta)^2} - \frac{t^2 \delta^2}{(t + \delta^2)^2} \right| \\ & + 4(C/p) \left(\frac{t}{1-t} \frac{p^2}{(p + 2 + \beta' X' X \beta)^2} + \beta' \Sigma \beta \frac{p^2}{(p + 4 + \beta' X' X \beta)^2} \right). \end{aligned}$$

Lemma 8 now entails that the first three terms in the preceding display converge to zero in probability. For the last term, note that we either have $p \rightarrow \infty$ if $t > 0$, or $t = 0$. In the case where $t > 0$, the sum in the brackets converges in probability to some finite limit by Lemma 8. Premultiplying by $1/p$ shows that the term converges to zero. In the case where $t = 0$, the sum in the brackets converges to zero in probability by Lemma 8.

Our second goal is to show that $\sup_{0 \leq c \leq C} |r(\hat{\delta}^2, c, p/n) - r(\delta^2, c, t)|$ also converges to zero in probability. We can bound the supremum in question by

$$\begin{aligned} & \left| \frac{p/n}{1-p/n} - \frac{t}{1-t} \right| + 2C \left| \frac{(p/n)^2}{(1-p/n)(p/n + \hat{\delta}^2)} - \frac{t^2}{(1-t)(t + \delta^2)} \right| \\ & + C^2 \left| \frac{(p/n)^3}{(1-p/n)(p/n + \hat{\delta}^2)^2} - \frac{t^3}{(1-t)(t + \delta^2)^2} \right| \\ & + C^2 \left| \hat{\delta}^2 \frac{(p/n)^2}{(p/n + \hat{\delta}^2)^2} - \delta^2 \frac{t^2}{(t + \delta^2)^2} \right|. \end{aligned}$$

The expression in the preceding display is the sum of four terms, where each term is of the form $|f(p/n, \hat{\delta}^2) - f(t, \delta^2)|$ for some function $f(\cdot, \cdot)$ which is continuous on $[0, 1) \times [0, \infty]$. Since $p/n \rightarrow t < 1$ by assumption and $\hat{\delta}^2 \rightarrow \delta^2$ in probability by Lemma 9, the continuous mapping theorem entails that each of the four terms in the preceding display converges to zero in probability. $\square$

Proof of Theorem 10. On the almost-sure event where $X'X$ is invertible, $\rho_2^2(\hat{\beta}(c), \beta, X)$ is given by the formula (37). For invertible $X'X$ and for β such that $\beta'(X'X/n)\beta = d^2$, the expression in (37) depends on β only through $\beta'\Sigma\beta$ which is at most $d^2/\lambda_p(V'V/n)$, where $\lambda_p(\dots)$ denotes the smallest eigenvalue of the indicated matrix. [Indeed, we have $d^2 = \|(V'V/n)^{1/2}\Sigma^{1/2}\beta\|^2$ and $\beta'\Sigma\beta = \beta'\Sigma^{1/2}(V'V/n)^{1/2}[(V'V/n)^{-1}](V'V/n)^{1/2}\Sigma^{1/2}\beta \leq d^2/\lambda_p(V'V/n)$.] It follows that $\sup_{\beta \in \mathbb{R}^p} \rho_2^2(\hat{\beta}(c), \beta, X)$ can almost surely be written as the supremum of

$$\begin{aligned} & \text{trace}((V'V)^{-1}) \left(1 - 2c \mathbb{E} \left[\frac{p}{\chi_p^2(nd^2)} \right] + (c^2 + 4c/p) \mathbb{E} \left[\frac{p^2}{(\chi_{p+2}^2(nd^2))^2} \right] \right) \\ & + \frac{c^2 + 4c/p}{\lambda_p(V'V/n)} \mathbb{E} \left[\frac{p^2 d^2}{(\chi_{p+4}^2(nd^2))^2} \right] \end{aligned}$$

over $d^2 \geq 0$ or, equivalently, over $d \geq 0$. Write $R_*(d^2, c, n, p)$ for the random variable in the preceding display. The claim will follow if we can show that

$$\sup_{0 \leq c \leq C} \sup_{d \geq 0} |R_*(d^2, c, n, p) - R(d^2, c, t)|$$

converges to zero in probability. Now use the triangle inequality to bound the expression in the preceding display by the supremum of

$$\begin{aligned} & \left| \text{trace}((V'V)^{-1}) - \frac{t}{1-t} \right| + 2C \left| \text{trace}((V'V)^{-1}) \mathbb{E} \left[\frac{p}{\chi_p^2(nd^2)} \right] - \frac{t^2}{(1-t)(t+d^2)} \right| \\ & + C^2 \left| \text{trace}((V'V)^{-1}) \mathbb{E} \left[\frac{p^2}{(\chi_{p+2}^2(nd^2))^2} \right] - \frac{t^3}{(1-t)(t+d^2)^2} \right| \\ & + C^2 \left| \frac{1}{\lambda_p(V'V/n)} \mathbb{E} \left[\frac{p^2 d^2}{(\chi_{p+4}^2(nd^2))^2} \right] - \frac{t^2 d^2}{(1-\sqrt{t})^2(t+d^2)^2} \right| \\ & + 4C \text{trace}((V'V)^{-1}) \mathbb{E} \left[\frac{p}{(\chi_{p+2}^2(nd^2))^2} \right] + 4C \frac{1}{\lambda_p(V'V/n)} \mathbb{E} \left[\frac{p d^2}{(\chi_{p+4}^2(nd^2))^2} \right] \end{aligned}$$

over $d \geq 0$. The first term of the preceding display converges to zero in probability and each of the remaining five terms is of the form $|U_n f_n(d) - u f(d)|$ multiplied by a constant, where U_n is a random variable that converges to $u \in \mathbb{R}$ in probability (cf. Lemma 5), and where $f_n(d)$ and $f(d)$ are functions of $d \in [0, \infty)$. [The two possible cases for U_n and u , respectively, are $U_n = \text{trace}((V'V)^{-1})$ and $u = t/(1-t) \geq 0$, as well as $U_n = 1/\lambda_p(V'V/n)$ and $u = 1/(1-\sqrt{t})^2 > 0$.] For each of the remaining five terms, we now proceed as follows: To show that $\sup_{d \geq 0} |U_n f_n(d) - u f(d)| \rightarrow 0$ in probability, it suffices to show that $|U_n| \sup_{d \geq 0} |f_n(d) - f(d)| \rightarrow 0$ in probability. To see this, note that $|U_n f_n(d) - u f(d)| \leq |U_n| |f_n(d) - f(d)| + |f(d)| |U_n - u|$ and that $|f(d)| \leq 1$ for all five possible cases. The later statement is obvious for the first, second, fourth and fifth term. For the third term, note that the inequality is equivalent to $t^2 d^2 \leq 2t d^2 + t^2 + d^4$ and that this is fulfilled as $t^2 \leq t \leq 1$. Let d_n , $n \geq 1$, be a sequence of maximizers (or near maximizers) of $|f_n - f|$, i.e., so that $\sup_{d \geq 0} |f_n(d) - f(d)| \leq |f_n(d_n) - f(d_n)| + o(1)$ where $o(1)$ is a term that converges to zero as $n \rightarrow \infty$. Because $[0, \infty]$ is compact, we may assume that $d_n \rightarrow \delta \in [0, \infty]$ (replacing the original sequence by a convergent subsequence if necessary). Moreover, it is easy to see that $\lim_{d \rightarrow \infty} f(d) = 0$ (for each of the five remaining terms in the preceding display). In particular, we can extend $f(\cdot)$ to a continuous function on $[0, \infty]$ by setting $f(\infty) = 0$. To complete the proof, we now show that either (a) $f_n(d_n) - f(\delta) \rightarrow 0$ as $n \rightarrow \infty$, or (b) $u = 0$ and $|f_n(d_n)|$ is bounded. To see (a), note that $|U_n| \sup_{d \geq 0} |f_n(d) - f(d)|$ can be bounded by $|U_n| |f_n(d_n) - f(d_n)| + |U_n| o(1) \leq |U_n| |f_n(d_n) - f(\delta)| + |U_n| |f(d_n) - f(\delta)| + |U_n| o(1)$ and that $|U_n| \rightarrow |u| < \infty$ in probability. For (b), note that $|U_n| \sup_{d \geq 0} |f_n(d) - f(d)| \leq |U_n| [f_n(d_n) + f(d_n) + o(1)]$ and recall that we already know that $|f(d_n)| \leq 1$. In the case where $t + \delta^2 > 0$, we see that either $p \rightarrow \infty$ or $nd_n \rightarrow \infty$. For the first term of the five remaining terms, note that

$$f_n(d_n) = \mathbb{E} \left[\frac{p}{\chi_p^2(nd_n^2)} \right] = \mathbb{E} \left[\frac{p + nd_n^2}{\chi_p^2(nd_n^2)} \right] \frac{p}{p + nd_n^2}.$$

With Lemma 8 and the fact that $p \geq 3$, we can conclude that the expected value tends to 1 and that hence $f_n(d_n)$ tends to $f(\delta)$, which shows that case (a) occurs for the first term. Arguing in a similar fashion shows that case (a) occurs also for the four remaining terms. And in the case where $t + \delta^2 = 0$, we see for the first, second and fourth of the five remaining terms in the preceding display that $u = 0$ and by (31) together with Lemma A.1 that $\mathbb{E}[p/\chi_p^2(nd_n^2)] \leq \mathbb{E}[p/\chi_p^2(0)] = p/(p-2) \leq 3$, that $\mathbb{E}[(p/\chi_{p+2}^2(nd_n^2))^2] \leq \mathbb{E}[(p/\chi_{p+2}^2(0))^2] = p/(p-2) \leq 3$ as well as $\mathbb{E}[p/(\chi_{p+2}^2(nd_n^2))^2] \leq \mathbb{E}[p/(\chi_{p+2}^2(0))^2] = 1/(p-2) \leq 1$. Hence, case (b) occurs for those three terms. For the third term, note that

$$\begin{aligned} f_n(d_n) - f(\delta) &= \mathbb{E} \left[\frac{d_n^2 p^2}{(\chi_{p+4}^2(nd_n^2))^2} \right] - \frac{\delta^2 t^2}{(t + \delta^2)^2} \leq d_n^2 \mathbb{E} \left[\frac{p}{(\chi_{p+4}^2(0))^2} \right] - 0 \\ &= d_n^2 \frac{p}{p+2} \leq d_n^2 \end{aligned}$$

and that $d_n \rightarrow 0 = \delta^2$. Arguing in a similar way for the fifth term, we see that for the third and the fifth term case (a) occurs. $\square$

Proof of Theorem 11. It suffices to show the statements for $\hat{\beta}(c_n)$. Set $R_* = \sup_{\delta^2 \geq 0} R(\delta^2, c, t)$ for $R(\cdot, \cdot, \cdot)$ as in Theorem 10, and note that $\sup_{\beta \in \mathbb{R}^p} \rho_2^2(\hat{\beta}(c_n), \beta, X) - \rho_2^2(\hat{\beta}_{ML}, \beta, X)$ converges to $R_* - t/(1-t)$ in probability by Theorems 4 and 10. For the first statement, assume either that $0 < c \leq 2$ and $t > [(c-2)/(c+2)]^2$ or that $c > 2$ and $t > 0$ hold (the case $c = 0$ can not occur there because $t < 1$). The statement now follows by observing, for such c and t , that $R_* - t/(1-t) > 0$, i.e., that $R(\delta^2, c, t) - t/(1-t) > 0$ for some $\delta^2 \geq 0$ (which is shown in Lemma B.3 (i)), and by setting ε equal to, say, $(R_* - t/(1-t))/2$. For the second statement, assume that $0 \leq c \leq 2$ and $t \leq [(c-2)/(c+2)]^2$, or that $c > 2$ and $t = 0$, and note, for such t and c , that $R_* - t/(1-t) \leq 0$, i.e., that $R(\delta^2, c, t) - t/(1-t) \leq 0$ for each $\delta^2 \geq 0$, which is shown in Lemma B.3 (ii).

For the last statement, let $c \in [0, 2]$, and take $\varepsilon > 0$. We need to show that $\mathbb{P}(\rho_2^2(\hat{\beta}(c_n), \beta, X) - \rho_2^2(\hat{\beta}_{ML}, \beta, X) > \varepsilon)$ converges to zero for arbitrary sequences of parameters $\beta \in \mathbb{R}^p$ and Σ . Because the set $[0, \infty]$ is compact, we may assume that $\beta' \Sigma \beta$ converges to a limit $\delta^2 \in [0, \infty]$ (by considering convergent subsequences if necessary). It now follows from Theorem 4 that $\rho_2^2(\hat{\beta}(c_n), \beta, X)$ converges in probability to the limit $r(\delta^2, c, t)$ as is given in the theorem, while $\rho_2^2(\hat{\beta}_{ML}, \beta, X)$ converges in probability to $t/(1-t)$. The claim now follows from the fact that $r(\delta^2, c, t) \leq t/(1-t)$, irrespective of $\delta^2 \in [0, \infty]$ (which is shown in Lemma B.3 (iii)). More precisely, assume that the statement does not hold. Then there exists a sequence β_n with $\beta_n' \Sigma \beta_n \rightarrow \delta^2 \in [0, \infty]$ such that $\mathbb{P}(\rho_2^2(\hat{\beta}(c), \beta, X) - \rho_2^2(\hat{\beta}_{ML}, \beta, X) > \varepsilon)$ does not converge to zero for some $\varepsilon > 0$. But this contradicts the statement of Theorem 4 because the preceding term is bounded from above by $\mathbb{P}(\rho_2^2(\hat{\beta}(c), \beta, X) - r(\delta^2, c, t) - (\rho_2^2(\hat{\beta}_{ML}, \beta, X) - t/(1-t)) > \varepsilon)$, which tends to zero for each $\varepsilon > 0$. $\square$

Lemma B.3. Let $r(\delta^2, c, t)$ and $R(\delta^2, c, t)$ be as in Theorems 4 and 10, respectively. Then the following properties hold:

(i) For $0 < c \leq 2$ and $t > [(c-2)/(c+2)]^2$ or for $c > 2$ and $t > 0$, we have

$$R(\delta^2, c, t) > t/(1-t) \quad (38)$$

for some $\delta^2 \geq 0$.

(ii) For $0 \leq c \leq 2$ and $t \leq [(c-2)/(c+2)]^2$ or for $c > 2$ and $t = 0$, we have

$$R(\delta^2, c, t) \leq t/(1-t) \quad (39)$$

for each $\delta^2 \geq 0$.

(iii) For $0 \leq c \leq 2$, we have

$$r(\delta^2, c, t) \leq t/(1-t) \quad (40)$$

for each $\delta^2 \geq 0$.

Proof. (i) For $\delta^2 < \infty$ (if $\delta^2 = \infty$, the strict inequality cannot hold), we can rewrite the difference $R(\delta^2, c, t) - t/(1-t)$ as

$$\begin{aligned} & \frac{t}{1-t} \left[-2c \frac{t}{t+\delta^2} + c^2 \frac{t^2}{(t+\delta^2)^2} + c^2 \frac{t\delta^2(1+\sqrt{t})}{(1-\sqrt{t})(t+\delta^2)^2} \right] \\ &= c \frac{t}{1-t} \frac{t}{(t+\delta^2)^2} \left[-2(t+\delta^2) + ct + c\delta^2 \frac{1+\sqrt{t}}{1-\sqrt{t}} \right] \\ &= c \frac{t}{1-t} \frac{t}{(t+\delta^2)^2} \left[\delta^2 \frac{c(1+\sqrt{t}) - 2(1-\sqrt{t})}{1-\sqrt{t}} + t(c-2) \right]. \end{aligned} \quad (41)$$

The term in the last line of the preceding display is of the form $A \cdot B$, where A denotes the term outside the square brackets and B denotes the term inside the square brackets. In the case where $t > 0$ and $c > 2$, we see that $A > 0$ as well as $B > 0$ for all $\delta^2 \in [0, \infty)$. Now, assume that $0 < c \leq 2$ (if $c = 0$, the strict inequality can not hold) and note that $B > 0$ if

$$\delta^2 \left[c(1+\sqrt{t}) - 2(1-\sqrt{t}) \right] > t(1-\sqrt{t})(2-c).$$

The remaining statement in (i) cannot hold if the term in square brackets is negative. The inequality in (38) will hold for some δ^2 if the term in square brackets is positive, i.e., if $c > 2(1-\sqrt{t})/(1+\sqrt{t})$, or, equivalently, $t > [(c-2)/(c+2)]^2$, as claimed.

(ii) The inequality holds trivially for all t and c if $\delta^2 = \infty$. Now, we can assume that $\delta^2 \geq 0$ is finite. The difference $R(\delta^2, c, t) - t/(1-t)$ can then be rewritten as in (41). Using the notation in the proof of (i), we see that $A \geq 0$ for all $\delta^2 \in [0, \infty)$, $c \geq 0$ and $t \in [0, 1)$. Note that $B \leq 0$ if

$$\delta^2 \left[c(1+\sqrt{t}) - 2(1-\sqrt{t}) \right] \leq t(1-\sqrt{t})(2-c).$$

The remaining statement in (ii) cannot hold if the term in square brackets is positive. If the term in square brackets is nonpositive, i.e., if $c \leq 2(1-\sqrt{t})/(1+\sqrt{t})$, the inequality in (39) can only hold for all $\delta^2 \in [0, \infty)$ if the right-hand side in the preceding display is nonnegative, i.e., if $0 \leq c \leq 2$.

(iii) The inequality in (40) is trivially fulfilled if $\delta^2 = \infty$. Hence, we can assume that $\delta^2 \in [0, \infty)$ from now on. Rewrite the difference $t/(1-t) - r(\delta^2, c, t)$ as

$$\begin{aligned} & \frac{t}{1-t} \left[2c \frac{t}{t+\delta^2} - c^2 \frac{t^2}{(t+\delta^2)^2} - c^2 \frac{\delta^2 t(1-t)}{(t+\delta^2)^2} \right] \\ &= c \frac{t}{1-t} \frac{t}{(t+\delta^2)^2} [2(t+\delta^2) - ct - c\delta^2(1-t)] \\ &= c \frac{t}{1-t} \frac{t}{(t+\delta^2)^2} [\delta^2(2-c(1-t)) + t(2-c)]. \end{aligned}$$

The term in front of the square brackets is nonnegative for all $c \geq 0$, $t \in [0, 1)$ and $\delta^2 \in [0, \infty)$. The term inside the square brackets is nonnegative if

$$\delta^2[2-c(1-t)] \geq t(c-2).$$

If $2-c(1-t) < 0$, the inequality in (40) cannot hold for all finite $\delta^2 \geq 0$. If $2-c(1-t) \geq 0$ or, equivalently, $c \leq 2/(1-t)$, then (40) is fulfilled for all $\delta^2 \in [0, \infty)$ if the right-hand side in the preceding display is nonpositive, i.e., if $0 \leq c \leq 2$. Noting that $2 \leq 2/(1-t)$ shows the statement. $\square$

C Technical details for Section 4

In this section, we give the result that we used to prove Theorem 12. Let Y and w be two n -vectors, Z a stochastic $n \times p$ matrix with $n \geq p \geq 3$ and θ an unknown p -vector. Assume that these quantities are related via the linear model

$$Y = Z\theta + w, \tag{42}$$

where $w \sim N(0, s^2 I_n)$ for some unknown $s^2 \geq 0$ and where w is independent of Z . We further assume that the rows of Z are independent of each other and follow a normal distribution with mean vector zero and variance-covariance matrix S , where S is a positive definite and unknown $p \times p$ matrix. More precisely, for $i = 1, \dots, n$, let $Z_{i,\cdot}$ denote the rows of the matrix Z . By assumption $Z'_{i,\cdot} \sim N(0, S)$ and $Z'_{i,\cdot}$ is independent of $Z'_{j,\cdot}$ for $i \neq j$. Let $S^{1/2}$ be a symmetric square root of S . Then $V'_{i,\cdot} := S^{-1/2} Z'_{i,\cdot} \sim N(0, I_p)$. Aligning the $V_{i,\cdot}$ by row and denoting the matrix by V , we see that V is an $n \times p$ matrix that has i.i.d. standard normally distributed entries and that we can write $Z = VS^{1/2}$. Let $\hat{\theta}$ be the least squares estimator of θ and $\hat{s}^2 = RSS/(n-p)$ where RSS is the residual sum of squares obtained by regressing Y on Z . Let $\hat{\theta}^+(c)$ be the positive part James–Stein-type shrinkage estimator in model (42), i.e.,

$$\hat{\theta}^+(c) = \left(1 - c \hat{s}^2 \frac{p}{\hat{\theta}' Z' Z \hat{\theta}} \right)_+ \hat{\theta} = \left(1 - \min \left\{ c \hat{s}^2 \frac{p}{\hat{\theta}' Z' Z \hat{\theta}}, 1 \right\} \right) \hat{\theta},$$

where $(x)_+ = \max\{x, 0\}$ and where $c \geq 0$ is a tuning parameter. Noting the equivalence of the models in (25) and in (42), using (26) and the definition of the shrinkage estimator in the preceding display, we have

$$\begin{aligned}\rho_3^2 &= (\hat{\theta}^+(c) - \theta)' S (\hat{\theta}^+(c) - \theta) + s^2 \\ &= (\hat{\theta} - \theta)' S (\hat{\theta} - \theta) - 2a_n \left[(\hat{\theta} - \theta)' S (\hat{\theta} - \theta) + (\hat{\theta} - \theta)' S \theta \right] \\ &\quad + a_n^2 \left[(\hat{\theta} - \theta)' S (\hat{\theta} - \theta) + 2(\hat{\theta} - \theta)' S \theta + \theta' S \theta \right] + s^2,\end{aligned}$$

and recall the predictor for ρ_3^2 from (22), i.e.,

$$\hat{\rho}_3^2 = \hat{s}^2 \frac{p}{n-p+1} - 2a_n \hat{s}^2 \frac{p}{n-p+1} + a_n^2 \left(\hat{s}^2 \frac{p}{n-p+1} + \frac{Y'Y}{n} - \hat{s}^2 \right) + \hat{s}^2,$$

where $a_n = \min\{cp\hat{s}^2/\hat{\theta}'Z'Z\hat{\theta}, 1\}$. Then we can show the following result.

Proposition C.1. *For each $\varepsilon > 0$, we have*

$$\begin{aligned}\mathbb{P}(|\rho_3^2 - \hat{\rho}_3^2| \geq \varepsilon) \\ \leq 16 \exp\left(-n(1-p/n)^2 f(\varepsilon(1-p/n), \theta' S \theta + s^2)\right),\end{aligned}\tag{43}$$

where ρ_3^2 and $\hat{\rho}_3^2$ are defined above and where $f(a, b) = a^2/(16(a+4b)^2)$.

We postpone the proof of Proposition C.1 to the end of this section and first prove some auxiliary results starting with specifying the distribution of the quantities involved. First of all, note that the term $(\hat{\theta} - \theta)' S \theta$ follows, conditional on Z , a centered normal distribution with variance $s^2 \theta' S (Z'Z)^{-1} S \theta$. The distribution of the remaining terms are found easily using standard methods. The proofs are included for the sake of completeness.

Lemma C.2. *Let the assumptions of this section hold.*

- (i) *The term $(\hat{\theta} - \theta)' S (\hat{\theta} - \theta)$ has the same distribution as s^2 times the ratio of two independent chi-square distributed random variables with p and $n-p+1$ degrees of freedom, respectively, i.e.,*

$$(\hat{\theta} - \theta)' S (\hat{\theta} - \theta) \sim s^2 \frac{\chi_p^2}{\chi_{n-p+1}^2}.$$

- (ii) *The estimator $\hat{s}^2$ has the same distribution as a chi-square distributed random variable with $n-p$ degrees of freedom multiplied by s^2 and divided by $n-p$, i.e.,*

$$\hat{s}^2 \sim s^2 \frac{\chi_{n-p}^2}{n-p}.$$

- (iii) *If $s^2 > 0$, the term $Y'Y$ follows conditional on Z a noncentral chi-square distribution with n degrees of freedom and noncentrality parameter $\theta' Z' Z \theta / s^2$ multiplied by s^2 , i.e., conditional on Z*

$$Y'Y \sim s^2 \chi_p^2(\theta' Z' Z \theta / s^2).$$

- (iv) If $s^2 > 0$, the term $\theta' Z' Z \theta / s^2$ follows a central chi-square distribution with n degrees of freedom multiplied by $\theta' S \theta / s^2$, i.e.,

$$\frac{\theta' Z' Z \theta}{s^2} \sim \frac{\theta' S \theta}{s^2} \chi_n^2.$$

Proof. (i) If $s^2 = 0$, the quadratic form is 0 almost surely. If $s > 0$, let $S^{1/2}$ be a symmetric square root of S and rewrite the quadratic form as

$$s^2 [(V'V)^{1/2} S^{1/2} (\hat{\theta} - \theta) / s]' (V'V)^{-1} [(V'V)^{1/2} S^{1/2} (\hat{\theta} - \theta) / s].$$

Recall that V is a $n \times p$ matrix with i.i.d. standard normally distributed entries and that $W = (V'VS)^{1/2}(\hat{\theta} - \theta)/s$ follows conditionally on V a p -dimensional normal distribution with mean vector zero and variance-covariance matrix I_p . Because the conditional distribution of W does not depend on V , the unconditional distribution coincides with the conditional distribution. Hence, the term in the preceding display has the same distribution as $s^2 W' (V'V)^{-1} W$ where $W \sim N(0, I_p)$. Furthermore, the matrix $V'V$ follows a Wishart distribution with scale matrix I_p and n degrees of freedom (see Chapter 38 in Johnson and Kotz (1972)). This shows that the quantity of interest follows the Hotelling's T-squared distribution multiplied by $n s^2$ (see Hotelling (1931)) and has consequently the same distribution as the F-distribution with p and $n - p + 1$ degrees of freedom multiplied by $p/(n - p + 1)$. Recalling the definition of the F-distribution shows the statement.

- (ii) Note that $\text{RSS} = (Y - Z\hat{\theta})'(Y - Z\hat{\theta})$. If $s^2 = 0$, then RSS is zero almost surely. If $s^2 > 0$, then rewrite RSS as $w'(I_n - Z(Z'Z)^{-1}Z')w$. Note that the matrix $I_n - Z(Z'Z)^{-1}Z'$ is idempotent with rank $n - p$ and recall that w/s follows a centered n -dimensional normal distribution with variance-covariance matrix I_n . The result that $w'(I_n - Z(Z'Z)^{-1}Z')w$ is then chi-square distributed with degrees of freedom equal to $\text{rank}(I_n - Z(Z'Z)^{-1}Z')$ is well known (see e.g. Appendix B.8 in Johnston and DiNardo (1997)).
- (iii) Note that, conditional on Z , Y/s follows a n -dimensional normal distribution with mean vector $Z\theta/s$ and variance-covariance matrix equal to I_n .
- (iv) If θ is the zero vector, the statement is trivial. Otherwise, note that $\theta' S \theta$ is larger than zero as S is positive definite by assumption and that $Z\theta = VS^{1/2}\theta$ is a n -vector with independent entries that follow a central normal distribution with variance $\theta' S \theta$. Hence, $\theta' Z' Z \theta \sim \theta' S \theta \chi_n^2$, and this shows the statement. $\square$

The previous lemma shows that we need tail bounds for a central as well as a noncentral chi-square distributed random variable and for the ratio of two independent central chi-square distributed random variables. The tail bound for the noncentral chi-square distribution needed here is given in Lemma C.7. Lemma C.3 and Corollary C.5 are auxiliary results to show Lemma C.7. As a special case, we treat the tail bound for the central chi-square distribution in Lemma C.3 and Corollary C.5. The tail bound for the ratio is given in Appendix A of Leeb (2008).

Lemma C.3. For each $\varepsilon > 0$, we have

$$\mathbb{P}\left(\frac{\chi_k^2(\lambda)}{k+\lambda} - 1 \geq \varepsilon\right) \leq e^{-\frac{k}{2}M(\frac{\lambda}{k}, \varepsilon)} \quad (44)$$

and

$$\mathbb{P}\left(\frac{\chi_k^2(\lambda)}{k+\lambda} - 1 \leq -\varepsilon\right) \leq \begin{cases} e^{-\frac{k}{2}M(\frac{\lambda}{k}, -\varepsilon)} & \text{if } \varepsilon < 1 \\ 0 & \text{else,} \end{cases} \quad (45)$$

where the function $M : [0, \infty) \times (-1, \infty) \rightarrow \mathbb{R}$ is given by

$$M(x, y) = \log\left(\frac{1 + \sqrt{4x(1+x)(1+y)} + 1}{2(1+x)(1+y)}\right) + x + (1+x)(1+y) - \sqrt{4x(1+x)(1+y)} + 1.$$

Proof. Let $\varepsilon > 0$ be arbitrary, and let s be such that $0 < s < 1/2$. Using Markov's inequality and Lemma A.3, we have

$$\begin{aligned} \mathbb{P}\left(\frac{\chi_k^2(\lambda)}{k+\lambda} - 1 \geq \varepsilon\right) &= \mathbb{P}\left(e^{s\chi_k^2(\lambda)} \geq e^{s(1+\varepsilon)(k+\lambda)}\right) \\ &\leq M_{\chi_k^2(\lambda)}(s)e^{-s(1+\varepsilon)(k+\lambda)} \\ &= \exp\left(-\frac{k}{2}\left[\log(1-2s) - \frac{2\lambda s}{k(1-2s)} + \frac{2s(1+\varepsilon)(k+\lambda)}{k}\right]\right). \end{aligned} \quad (46)$$

Setting $x := \lambda/k$ and $y := \varepsilon$, the term in square brackets in (46) can be written as $f_{x,y}(s) = \log(1-2s) - 2xs/(1-2s) + 2s(1+x)(1+y)$. To find the optimal, meaning the smallest, upper bound in (46), we need to maximize $f_{x,y}(s)$ over $0 < s < 1/2$. Setting the derivative of f , i.e., $f'_{x,y}(s) = -2/(1-2s) - 2x/(1-2s)^2 + 2(1+x)(1+y)$, equal to 0 and rearranging gives $-(1-2s) - x + (1-2s)^2(1+x)(1+y) = 0$. For finding the maximizer of $f_{x,y}(s)$ over $0 < s < 1/2$, we need to find the roots of the function $h(t) = t^2(1+x)(1+y) - t - x$, where $t = 1-2s$, that lie in the interval $(0, 1)$. Note that $h(0)$ is nonpositive and that $h(\cdot)$ is convex because the second derivative is positive. This implies that only one root lies on the positive axis, i.e.,

$$t^* = \frac{1 + \sqrt{4x(1+x)(1+y)} + 1}{2(1+x)(1+y)}.$$

Note that $0 < t^* < 1$. To see that $t^* < 1$, we need to show that $2(1+x)(1+y) - 1 > \sqrt{4x(1+x)(1+y)} + 1$. Recalling that $x \geq 0$ and $y > 0$, we see that the left-hand side of the preceding inequality is positive. Squaring both sides gives the equivalent condition $4(1+x)^2(1+y)^2 - 4(1+x)(1+y) + 1 > 4x(1+x)(1+y) + 1$. Simplifying and dividing by $4(1+x)(1+y) > 0$, we obtain $(1+x)(1+y) - 1 > x$, which holds. Using the fact that $t = 1-2s$ or, equivalently, $s = (1-t)/2$, we see that

$$s^* = \frac{2(1+x)(1+y) - 1 - \sqrt{4x(1+x)(1+y)} + 1}{4(1+x)(1+y)}$$

is a root of $f'_{x,y}$ that lies in $(0, 1/2)$ as required. Noting that $f''_{x,y}(s) = (-4(1 - 2s) - 8x)/(1 - 2s)^3$ and recalling that $0 < 1 - 2s^* = t^*$, we see that s^* is a maximizer of $f_{x,y}$ and therefore minimizes the upper bound in (46). Note that

$$\begin{aligned} \frac{s^*}{1 - 2s^*} &= \frac{2(1+x)(1+y) - 1 - \sqrt{4x(1+x)(1+y) + 1}}{2(1 + \sqrt{4x(1+x)(1+y) + 1})} \\ &= \frac{(1+x)(1+y)}{1 + \sqrt{4x(1+x)(1+y) + 1}} - \frac{1}{2}. \end{aligned}$$

Therefore,

$$\begin{aligned} f_{x,y}(s^*) &= \log \left(\frac{1 + \sqrt{4x(1+x)(1+y) + 1}}{2(1+x)(1+y)} \right) - \frac{2x(1+x)(1+y)}{1 + \sqrt{4x(1+x)(1+y) + 1}} + x \\ &\quad + \frac{2(1+x)(1+y) - 1 - \sqrt{4x(1+x)(1+y) + 1}}{2} \\ &= \log \left(\frac{1 + \sqrt{4x(1+x)(1+y) + 1}}{2(1+x)(1+y)} \right) + x + (1+x)(1+y) \\ &\quad - \frac{8x(1+x)(1+y) + 2 + 2\sqrt{4x(1+x)(1+y) + 1}}{2(1 + \sqrt{4x(1+x)(1+y) + 1})} \\ &= \log \left(\frac{1 + \sqrt{4x(1+x)(1+y) + 1}}{2(1+x)(1+y)} \right) + x + (1+x)(1+y) \\ &\quad - \sqrt{4x(1+x)(1+y) + 1}. \end{aligned}$$

Inserting this in (46) and replacing x by λ/k and y by ε , gives (44).

In order to prove (45), note that $\mathbb{P}(\chi_k^2(\lambda)/(b+k) - 1 < -\varepsilon) = 0$ if $\varepsilon \geq 1$. Therefore, the second inequality in (45) is trivially fulfilled. Assuming that $\varepsilon < 1$, we get for an arbitrary $s > 0$

$$\begin{aligned} \mathbb{P}\left(\frac{\chi_k^2(\lambda)}{k+\lambda} - 1 \leq -\varepsilon\right) &= \mathbb{P}\left(e^{-s\chi_k^2(\lambda)} \geq e^{-s(1-\varepsilon)(\lambda+k)}\right) \\ &\leq M_{\chi_k^2(\lambda)}(-s)e^{s(1-\varepsilon)(k+\lambda)} \\ &= \exp\left(-\frac{k}{2}\left[\log(1+2s) + \frac{2\lambda s}{k(1+2s)} - \frac{2s(1-\varepsilon)(k+\lambda)}{k}\right]\right). \quad (47) \end{aligned}$$

As before, set $x := \lambda/k$ and $y := \varepsilon$. In order to calculate the optimal upper bound in (47), we need to maximize the term in square brackets with respect to $s \in (0, \infty)$, i.e., $g_{x,y}(s) = \log(1+2s) + 2xs/(1+2s) - 2s(1+x)(1-y)$. Note that $g_{x,y}(s) = f_{x,-y}(-s)$ which implies that $g'_{x,y}(s) = -f'_{x,-y}(-s)$. In order to get the maximizer of $g_{x,y}(s)$, we need to find the roots of $g'_{x,y}(s)$ or, equivalently, the roots of $\tilde{h}(u) = u^2(1+x)(1-y) - u - x$, where $u = 1 + 2s$, that lie in the interval $(1, \infty)$. Because $\tilde{h}(0)$ is nonpositive and $\tilde{h}(\cdot)$ is convex, there is only one root lying on the positive axis, i.e.

$$u^* = \frac{1 + \sqrt{4x(1+x)(1-y) + 1}}{2(1+x)(1-y)}.$$

Note that $u^* > 1$ which is equivalent to $\sqrt{4x(1+x)(1-y)+1} > 2(1+x)(1-y) - 1$. If the right-hand side is nonpositive, the inequality is trivially fulfilled. If not, squaring both sides gives the equivalent condition $x(1+x)(1-y) > (1+x)^2(1-y)^2 - (1+x)(1-y)$. Rearranging and dividing by $(1+x)^2(1-y) > 0$, we obtain $1 > 1-y$, which holds as $y > 0$ by assumption. Using the fact that $u = 1 + 2s$ or, equivalently, $s = (u-1)/2$, we see that

$$\tilde{s}^* = \frac{1 - 2(1+x)(1-y) + \sqrt{4x(1+x)(1-y)+1}}{4(1+x)(1-y)}$$

is a root of $g'_{x,y}$ that lies in $(0, \infty)$. Noting that $g''_{x,y}(s) = f''_{x,-y}(-s) = (-4(1+2s) - 8x)/(1+2s)^3$ and recalling that $u^* = 1 + 2\tilde{s}^* > 1$, we see that $\tilde{s}^*$ is a maximizer of $g_{x,y}$ and minimizes the upper bound in (47). Since $\tilde{s}^*$ equals $-s^*$ with y replaced by $-y$ and $g_{x,y}(s) = f_{x,-y}(-s)$, we see that $g_{x,y}(\tilde{s}^*) = f_{x,-y}(-\tilde{s}^*) = f_{x,-y}(s^*) = M(x, -y)$. This shows the first line in (45) and ends the proof. $\square$

Lemma C.4. *Let $M(x, y)$ be defined as in Lemma C.3 and let $m(x, y) = y^2(1+x)/(2(y+2))$ and $\tilde{m}(y) = y^2/(2(y+1))$. Then the following relations hold:*

(i) *For $x \geq 0$ and $0 \leq y < 1$, we have*

$$M(x, -y) \geq M(x, y).$$

(ii) *For $x \geq 0$ and $y \geq 0$, we have*

$$M(x, y) \geq m(x, y) \geq 0.$$

For $x = 0$, we even have

$$M(0, y) \geq \tilde{m}(y) \geq 0.$$

Proof. (i) Recall that

$$M(x, y) = \log \left(\frac{1 + W(x, y)}{2(1+x)(1+y)} \right) + x + (1+x)(1+y) - W(x, y),$$

where $W(x, y) = \sqrt{4x(1+x)(1+y)+1}$, and note that $W(x, 0) = \sqrt{4x^2 + 4x + 1} = 2x+1$. This implies that $M(x, 0) = \log(1) + 2x+1 - (2x+1) = 0$ and hence $M(x, -y) - M(x, y) = 0$ for $y = 0$. Therefore, it suffices to show that the difference of $M(x, -y) - M(x, y)$ is nondecreasing in y for $y \geq 0$ or, equiv-

alently, that the derivative of the difference is nonnegative. Observing that $\partial W(x, y)/\partial y = 2x(1+x)/W(x, y)$, the derivative of the first term in $M(x, y)$ is

$$\begin{aligned}
& \frac{\partial}{\partial y} \log \left(\frac{1 + W(x, y)}{2(1+x)(1+y)} \right) \\
&= \frac{2(1+x)(1+y)}{1 + W(x, y)} \frac{2x(1+x)(1+y)/W(x, y) - (1 + W(x, y))}{2(1+x)(1+y)^2} \\
&= \frac{2x(1+x)(1+y) - W(x, y) - W(x, y)^2}{W(x, y)(1 + W(x, y))(1+y)} \\
&= \frac{4x(1+x)(1+y) + 1 - 1 - 2W(x, y) - 2W(x, y)^2}{2W(x, y)(1 + W(x, y))(1+y)} \\
&= \frac{-1 - 2W(x, y) - W(x, y)^2}{2W(x, y)(1 + W(x, y))(1+y)} \\
&= \frac{-(1 + W(x, y))}{2W(x, y)(1+y)}.
\end{aligned}$$

Therefore, the derivative of $M(x, y)$ with respect to y is

$$\begin{aligned}
\frac{\partial}{\partial y} M(x, y) &= -\frac{1 + W(x, y)}{2W(x, y)(1+y)} + 1 + x - \frac{2x(1+x)}{W(x, y)} \\
&= 1 + x - \frac{1 + W(x, y) + 4x(1+x)(1+y)}{2W(x, y)(1+y)} \\
&= 1 + x - \frac{W(x, y) + W(x, y)^2}{2W(x, y)(1+y)} \\
&= 1 + x - \frac{1 + W(x, y)}{2(1+y)}. \tag{48}
\end{aligned}$$

Using the chain rule for the derivative of $M(x, -y)$, we get

$$\frac{\partial}{\partial y} [M(x, -y) - M(x, y)] = \frac{1 + W(x, -y)}{2(1-y)} + \frac{1 + W(x, y)}{2(1+y)} - 2 - 2x.$$

Recalling that $W(x, 0) = 2x + 1$, the right-hand side of the preceding display evaluated at $y = 0$ equals 0. Using the same argument as before, it is sufficient to show that the second derivative of $M(x, -y) - M(x, y)$ with respect to y is nonnegative. Note that

$$\begin{aligned}
\frac{\partial^2}{\partial y^2} [M(x, -y) - M(x, y)] &= \frac{-4x(1+x)(1-y)/W(x, -y) + 2(1 + W(x, -y))}{4(1-y)^2} \\
&\quad + \frac{4x(1+x)(1+y)/W(x, y) - 2(1 + W(x, y))}{4(1+y)^2} \\
&= \frac{-x(1+x)}{W(x, -y)(1-y)} + \frac{1 + W(x, -y)}{2(1-y)^2} + \frac{x(1+x)}{W(x, y)(1+y)} - \frac{1 + W(x, y)}{2(1+y)^2} \\
&= \frac{1}{2(1-y)^2} - \frac{1}{2(1+y)^2} \\
&\quad + \frac{-2x(1+x)(1-y) + W(x, -y)^2}{2W(x, -y)(1-y)^2} + \frac{2x(1+x)(1+y) - W(x, y)^2}{2W(x, y)(1+y)^2}.
\end{aligned}$$

The difference of the two terms in the second-to-last line is nonnegative since $0 \leq y < 1$. Hence, it suffices to show that the sum in the last line is nonnegative. This sum can be rewritten as

$$\begin{aligned} & \frac{2x(1+x)(1-y)+1}{2W(x,-y)(1-y)^2} + \frac{-2x(1+x)(1+y)-1}{2W(x,y)(1+y)^2} \\ &= \frac{x(1+x)}{W(x,-y)(1-y)} - \frac{x(1+x)}{W(x,y)(1+y)} \\ & \quad + \frac{1}{2W(x,-y)(1-y)^2} - \frac{1}{2W(x,y)(1+y)^2}. \end{aligned}$$

This sum is nonnegative because $W(x,y) \geq W(x,-y)$ and because $1+y \geq 1-y > 0$.

(ii) For the first statement, recall that $M(x,0) - m(x,0) = 0$ for all $x \geq 0$. By the same argument as in (i), it suffices to show that the derivative of $M(x,y) - m(x,y)$ with respect to y is nonnegative. Note that

$$\frac{\partial m(x,y)}{\partial y} = \frac{2y(1+x)(2+y) - y^2(1+x)}{2(2+y)^2} = \frac{y(1+x)(4+y)}{2(2+y)^2}. \quad (49)$$

Hence, we need to show that

$$1+x - \frac{1+W(x,y)}{2(1+y)} - \frac{y(1+x)(4+y)}{2(2+y)^2} \geq 0$$

holds. Multiplying both sides by $2(1+y)(2+y)^2$, we can rewrite the sum on the left-hand side in the preceding display as

$$\begin{aligned} & 2(1+x)(1+y)(2+y)^2 - (2+y)^2(1+W(x,y)) - y(4+y)(1+x)(1+y) \\ &= (1+x)(1+y)[2(4+4y+y^2) - y(4+y)] - (2+y)^2(1+W(x,y)) \\ &= (1+x)(1+y)(8+4y+y^2) - (2+y)^2(1+W(x,y)) \\ &= (1+x)(1+y)(2+y)^2 + 4(1+x)(1+y) - (2+y)^2(1+W(x,y)). \end{aligned}$$

Thus, it remains to show that $(1+x)(1+y)(2+y)^2 + 4(1+x)(1+y) - (2+y)^2 \geq (2+y)^2 W(x,y)$ holds. Since both sides are positive, it suffices to show that the square of the left-hand side is larger or equal than the square of the right-hand side, or, equivalently, that

$$\begin{aligned} & (1+x)^2(1+y)^2(2+y)^4 + 16(1+x)^2(1+y)^2 \\ & \quad - (2+y)^4 4x(1+x)(1+y) + 8(1+x)^2(1+y)^2(2+y)^2 \\ & \quad - 2(2+y)^4(1+x)(1+y) - 8(1+x)(1+y)(2+y)^2 \end{aligned}$$

is nonnegative. After dividing by $(1+x)(1+y) > 0$, we can rewrite the sum in the preceding display as

$$\begin{aligned}
& (2+y)^4[(1+x)(1+y) - 4x - 2] + 8(2+y)^2((1+x)(1+y) - 1) \\
& \quad + 16(1+x)(1+y) \\
& = (2+y)^4(xy + y - 3x - 1) + 8(2+y)^2(x + y + xy) + 16(1+x)(1+y) \\
& = (2+y)^2[(xy + y - 3x - 1)(y^2 + 4y + 4) + 8(x + y + xy)] + 16(1+x)(1+y) \\
& \geq (2+y)^2[(xy + y)(4y + 4) - 3xy^2 - y^2 - 4xy + 4y - 4x - 4] \\
& \quad + 16(1+x)(1+y) \\
& = (2+y)^2(xy^2 + 3y^2 + 8y) - (y^2 + 4y + 4)(4x + 4) + 16(1+x)(1+y) \\
& \geq (2+y)^2(xy^2 + 3y^2) - 4y^2(x + 1) \\
& \geq 4y^2(x + 3) - 4y^2(x + 1),
\end{aligned}$$

where the inequalities are obtained from the facts that $y^2 + 4y + 4 \geq 4y + 4$, $8y \geq 0$ and $(2+y)^2 \geq 4$, respectively. The last line in the preceding display is nonnegative and shows that $M(x, y) \geq m(x, y)$. The second inequality of the first statement, i.e., $m(x, y) \geq 0$ is true since $x \geq 0$ and $y \geq 0$ by assumption.

For the case where $x = 0$, recall that $M(0, y) = y - \log(1+y)$ and that $\tilde{m}(y) = y^2/(2(y+1))$. The inequality is then trivially fulfilled for $y = 0$. Hence, it suffices to show that the derivative of $y - \log(1+y) - y^2/(2(1+y))$ is nonnegative. Note that the derivative equals

$$1 - \frac{1}{1+y} - \frac{4y(1+y) - 2y^2}{4(1+y)^2} = \frac{2(1+y)^2 - 2(1+y) - 2y(1+y) + y^2}{2(1+y)^2}$$

and simplifies to $y^2/(2(1+y)^2)$, which is clearly nonnegative. The second inequality is trivial. $\square$

Remark. Note that Lemma C.3 together with Lemma C.4 generalize the results of Lemma A.2 and of Lemma A.3 in Leeb (2008). In our notation, $\lambda = 0$ corresponds to the case treated in Leeb (2008). As $M(0, \varepsilon) = \log(1/(1+\varepsilon)) + \varepsilon = \varepsilon - \log(1+\varepsilon)$, this implies that Lemma A.2 in Leeb (2008) is indeed a special case of Lemma C.3. Note furthermore that the lower bound for $M(0, \varepsilon)$, i.e., $\tilde{m}(\varepsilon)$, is slightly better than $\varepsilon^2/4$, which is the lower bound of the corresponding function in Lemma A.3 in Leeb (2008), and that the lower bound here holds for all $\varepsilon > 0$.

Using Lemma C.4, an easy corollary to Lemma C.3 is the following result.

Corollary C.5. *Under the assumptions of Lemma C.3 and $m(\cdot, \cdot)$ and $\tilde{m}(\cdot)$ as defined in Lemma C.4, we have for each $\varepsilon > 0$*

$$\mathbb{P}\left(\frac{\chi_k^2(\lambda)}{k+\lambda} - 1 \geq \varepsilon\right) \leq \exp\left(-\frac{k}{2} m\left(\frac{\lambda}{k}, \varepsilon\right)\right) \quad (50)$$

as well as

$$\mathbb{P}\left(\left|\frac{\chi_k^2(\lambda)}{k+\lambda} - 1\right| \geq \varepsilon\right) \leq 2 \exp\left(-\frac{k}{2} M\left(\frac{\lambda}{k}, \varepsilon\right)\right) \leq 2 \exp\left(-\frac{k}{2} m\left(\frac{\lambda}{k}, \varepsilon\right)\right). \quad (51)$$

In the case of a central chi-square distribution, i.e., when $\lambda = 0$, we have

$$\mathbb{P}\left(\frac{\chi_k^2}{k} - 1 \geq \varepsilon\right) \leq \exp\left(-\frac{k}{2} \tilde{m}(\varepsilon)\right) \quad (52)$$

as well as

$$\mathbb{P}\left(\left|\frac{\chi_k^2}{k} - 1\right| \geq \varepsilon\right) \leq 2 \exp\left(-\frac{k}{2} M(0, \varepsilon)\right) \leq 2 \exp\left(-\frac{k}{2} \tilde{m}(\varepsilon)\right). \quad (53)$$

Proof. The first statement is a consequence of (44) in Lemma C.3 and of Lemma C.4 (ii). For the second statement, recall from Lemma C.3 that $\mathbb{P}(\chi_k^2(\lambda)/(k + \lambda) - 1 \leq -\varepsilon) = 0$ if $\varepsilon \geq 1$, that $\mathbb{P}(\chi_k^2(\lambda)/(k + \lambda) - 1 \leq -\varepsilon) \leq \exp(-(k/2)M(\lambda/k, -\varepsilon))$ for $0 < \varepsilon < 1$ and from Lemma C.4 (i) that $M(\lambda/k, \varepsilon) \leq M(\lambda/k, -\varepsilon)$ for $0 < \varepsilon < 1$. Together with (44) in Lemma C.3, this shows the first inequality in (51). The second inequality follows from Lemma C.4 (ii). The statements corresponding to the central chi-square distribution follow by the same arguments but using the corresponding property for $x = 0$ in Lemma C.4 (ii). $\square$

The following result shows that the main properties of the complicated function $M(x, y)$ are well approximated by the simple function $m(x, y)$, i.e., we show that we can bound the function $M(x, y)$ from above by a multiple of $m(x, y)$. We show a similar statement for the functions $M(0, y)$ and $\tilde{m}(y)$.

Lemma C.6. *For all $x \geq 0$ and all $y \geq 0$, we have*

$$m(x, y) \leq M(x, y) \leq 2m(x, y)$$

as well as

$$\tilde{m}(y) \leq M(0, y) \leq 2\tilde{m}(y),$$

where $M(x, y)$, $m(x, y)$ and $\tilde{m}(y)$ are defined as in Lemma C.3 and in Lemma C.4, respectively.

Proof. The first inequalities in both displays were already shown in Lemma C.4 (ii). The second inequality in the first display is equivalent to $2m(x, y) - M(x, y)$ being nonnegative. Recall from the proof of Lemma C.4 that $M(x, 0) = 0$ as well as $m(x, 0) = 0$. Hence the statement is true for $y = 0$ and it suffices to show that $\partial[2m(x, y) - M(x, y)]/\partial y$ is nonnegative. Using (48) and (49), we have

$$\frac{\partial}{\partial y} [2m(x, y) - M(x, y)] = \frac{y(1+x)(4+y)}{(2+y)^2} - 1 - x + \frac{1+W(x, y)}{2(1+y)}.$$

The right-hand side in the preceding display is nonnegative if and only if

$$2y(1+x)(1+y)(4+y) - 2(1+x)(1+y)(2+y)^2 + (1+W(x, y))(2+y)^2 \geq 0.$$

The left-hand side in the preceding display equals

$$\begin{aligned} & 2(1+x)(1+y)(4y+y^2-4-4y-y^2) + (1+W(x, y))(2+y)^2 \\ &= (1+W(x, y))(2+y)^2 - 8(1+x)(1+y). \end{aligned}$$

Recall that $W(0, y) = 1$, hence the right-hand side in the preceding display equals $2(2 + y)^2 - 8(1 + y)$ for $x = 0$. This is nonnegative for all $y \geq 0$. Hence, we can assume that $x > 0$ from now on. To show that the right-hand side of the preceding display is nonnegative, we can multiply by $W(x, y) - 1 > 0$ which gives $4x(1 + x)(1 + y)(2 + y)^2 - 8(1 + x)(1 + y)(W(x, y) - 1)$. Dividing by $4(1 + x)(1 + y) > 0$, it suffices to show that $x(2 + y)^2 + 2 \geq 2W(x, y)$. After squaring both sides and simplifying, we obtain the equivalent statement

$$x(2 + y)^4 + 4(2 + y)^2 - 16(1 + x)(1 + y) \geq 0.$$

The statement is true for $y = 0$ and for all $x \geq 0$. Hence, it remains to show that the left-hand side of the preceding inequality is nondecreasing in y . The derivative with respect to y is $4x(2 + y)^3 + 8(2 + y) - 16(1 + x)$. Evaluating this at $y = 0$ gives $48x + 16 - 16(1 + x)$ which is nonnegative for all $x \geq 0$. Hence, it is sufficient to show that the second partial derivative of the right-hand side in the preceding display with respect to y is nonnegative. The second derivative equals $12x(2 + y)^2 + 8$ and is clearly nonnegative.

To show the second statement of the lemma, note that the second inequality is equivalent to $\log(1 + y) - y/(1 + y) \geq 0$. This inequality is fulfilled for $y = 0$. Hence, it suffices to show that the derivative of the difference is nonnegative. The derivative equals $1/(1 + y) - 1/(1 + y)^2$, and this ends the proof. $\square$

Lemma C.7. *Fix an integer $n \geq 1$. Let $b > 0$ be a real number and let B_n be a real random variable such that $B_n \sim b\chi_n^2$. Let furthermore $\chi_n^2(B_n)$ be a real random variable that follows conditional on B_n a noncentral chi-square distribution with n degrees of freedom and noncentrality parameter B_n . Then for any $\varepsilon > 0$ and any $\alpha \in (0, 1)$, we have*

$$\begin{aligned} \mathbb{P}\left(\left|\frac{\chi_n^2(B_n)}{n} - 1 - b\right| \geq \varepsilon\right) &\leq 2 \exp\left(-n \frac{\varepsilon^2(1 - \alpha)^2}{4(\varepsilon(1 + \alpha) + 2(1 + b))}\right) \\ &\quad + 4 \exp\left(-n \frac{\varepsilon^2 \alpha^2}{4b(\varepsilon \alpha + b)}\right). \end{aligned} \quad (54)$$

Proof. Let $\alpha \in (0, 1)$ be arbitrary. Then the left-hand side in (54) is bounded by

$$\begin{aligned} &\mathbb{P}\left(\left|\frac{\chi_n^2(B_n)}{n} - \frac{n + B_n}{n}\right| \geq \varepsilon(1 - \alpha)\right) + \mathbb{P}\left(\left|\frac{B_n}{n} - b\right| \geq \varepsilon \alpha\right) \\ &= \mathbb{P}\left(\left|\frac{\chi_n^2(B_n)}{n + B_n} - 1\right| \geq \varepsilon(1 - \alpha) \frac{n}{n + B_n}\right) + \mathbb{P}\left(\left|\frac{B_n/n}{b} - 1\right| \geq \frac{\varepsilon \alpha}{b}\right). \end{aligned} \quad (55)$$

Using Corollary C.5 conditional on B_n and the definition of $m(\cdot, \cdot)$, we can bound the first term in (55) by

$$\begin{aligned} &2 \mathbb{E}\left[\exp\left(-\frac{n}{2} m\left(\frac{B_n}{n}, \varepsilon(1 - \alpha) \frac{n}{n + B_n}\right)\right)\right] \\ &= 2 \mathbb{E}\left[\exp\left(-\frac{n}{2} \frac{\varepsilon^2(1 - \alpha)^2 n / (n + B_n)}{2(\varepsilon(1 - \alpha)n / (n + B_n) + 2)}\right)\right] \\ &= 2 \mathbb{E}\left[\exp\left(-n \frac{\varepsilon^2(1 - \alpha)^2 n}{4(\varepsilon(1 - \alpha)n + 2(n + B_n))}\right)\right]. \end{aligned}$$

The term in the last line of the preceding display is increasing in B_n . Integrating separately over the events $\{B_n < \delta\}$ and $\{B_n \geq \delta\}$, for some $\delta > 0$, and noting that the integrand is smaller or equal than 1, we can bound the term in the last line of the preceding display further by

$$2 \exp \left(-n \frac{\varepsilon^2(1-\alpha)^2 n}{4(\varepsilon(1-\alpha)n + 2(n+\delta))} \right) + 2\mathbb{P}(B_n \geq \delta).$$

Noting that

$$\mathbb{P}(B_n \geq \delta) = \mathbb{P} \left(\frac{B_n/n}{b} - 1 \geq \frac{\delta/n}{b} - 1 \right)$$

and in view of the second term in (55), we choose $\delta = n(\varepsilon\alpha + b)$. Recalling that $(B_n/n)/b \sim \chi_n^2/n$ by assumption, we use Corollary C.5 to bound the sum in (55) by

$$2 \exp \left(-n \frac{\varepsilon^2(1-\alpha)^2}{4(\varepsilon(1-\alpha) + 2(1+\alpha\varepsilon + b))} \right) + 4 \exp \left(-n \frac{\varepsilon^2\alpha^2}{4b(\varepsilon\alpha + b)} \right).$$

□

The next lemma is a technical result that is separated from the main results only for the sake of clarity.

Lemma C.8. *Let x and y be nonnegative real numbers. Then the following statements hold true:*

(i) *If $x > 0$, we have*

$$\frac{y^2}{2(1+x+y)^2} \leq (1+x) \log \left(\frac{1+x+y}{1+x} \right) - x \log \left(\frac{x+y}{x} \right).$$

(ii) *The function $\Gamma(x+1/2)/\Gamma(x) - \sqrt{x}$ is nondecreasing in x for $x \geq 1$.*

(iii) *For $x \geq 0$ and $0 \leq y \leq 1$, we have*

$$\frac{x^2(1-y)^4}{4(x(1-y)+1)^2} \leq \frac{(1-y)^2}{4} \leq (1-\sqrt{y})^2.$$

(iv) *For $x \geq 0$ and $y > 0$, we have*

$$\frac{x^2}{(x+1+y)^2} \leq \frac{x^2}{4y}.$$

Proof. (i) The inequality holds trivially for $y = 0$. Hence, it suffices to show that the difference of the right-hand side and the left-hand side is nondecreasing in y . The first derivative of the difference with respect to y is

$$\begin{aligned} & \frac{1+x}{1+x+y} - \frac{x}{x+y} - \frac{y(1+x+y) - y^2}{(1+x+y)^3} \\ &= \frac{y}{(1+x+y)(x+y)} - \frac{y(1+x)}{(1+x+y)^3} \\ &= \frac{y[(1+x+y)^2 - (1+x)(x+y)]}{(x+y)(1+x+y)^3}. \end{aligned}$$

The term in the last line is nonnegative, and this shows the statement.

- (ii) It is enough to show that the derivative with respect to x is nonnegative for $x \geq 1$. The first derivative equals

$$\begin{aligned} & \frac{\psi(x+1/2)\Gamma(x+1/2)\Gamma(x) - \psi(x)\Gamma(x)\Gamma(x+1/2)}{\Gamma(x)^2} - \frac{1}{2\sqrt{x}} \\ &= \frac{\Gamma(x+1/2)(\psi(x+1/2) - \psi(x))}{\Gamma(x)} - \frac{1}{2\sqrt{x}}, \end{aligned}$$

where $\psi(\cdot)$ is the derivative of the logarithm of the gamma function, the so-called digamma function. To show that the term in the preceding display is nonnegative, it suffices to show that

$$2\sqrt{x} \frac{\Gamma(x+1/2)}{\Gamma(x)} (\psi(x+1/2) - \psi(x)) \geq 1 \quad (56)$$

holds for $x \geq 1$. Kershaw's first double inequality (see (1.3) in Kershaw (1983)) states that

$$\left(y + \frac{s}{2}\right)^{1-s} < \frac{\Gamma(y+1)}{\Gamma(y+s)} < \left(y - \frac{1}{2} + \left(s + \frac{1}{4}\right)^{1/2}\right)^{1-s}$$

holds for $y > 0$ and $0 < s < 1$. Use the first inequality with $y = x - 1/2$ and $s = 1/2$ to show that $\Gamma(x+1/2)/\Gamma(x) > \sqrt{x-1/4}$. Furthermore, Theorem 1.1. in Qi (2007) shows that the function

$$\frac{\psi(x+t) - \psi(x+s)}{t-s} - \frac{2x+s+t+1}{2(x+s)(x+t)}$$

is completely monotonic in $x \in (-\min\{s, t\}, \infty)$, provided that $s \neq t$ and that $|t-s| < 1$. A function $f(z)$ is called completely monotonic if $(-1)^n f^{(n)}(z) \geq 0$ for $n = 0, 1, 2, \dots$, where $f^{(n)}(z)$ denotes the n -th derivative of $f(z)$. This implies that the function in the preceding display is nonnegative. Using this fact with $t = 1/2$ and $s = 0$, shows that $\psi(x+1/2) - \psi(x) \geq (4x+3)/(4x(2x+1))$. Hence, we can bound the left-hand side in (56) from below by

$$\sqrt{x \left(x - \frac{1}{4}\right)} \frac{4x+3}{2x(2x+1)}$$

and it is sufficient to show that the term in the preceding display is larger or equal than 1 for $x \geq 1$, or, equivalently, that $(4x+3)\sqrt{x(4x-1)} \geq 4x(2x+1)$ holds. Squaring both sides, expanding the brackets and simplifying, it remains to show that $(16x^2 + 24x + 9)(4x-1) \geq 16x(4x^2 + 4x + 1)$, or, equivalently, that $16x^2 - 4x \geq 9$. Rewriting the left-hand side as $16x^2 - 16x + 12x = 16x(x-1) + 12x$ shows the statement as $x \geq 1$ by assumption.

- (iii) The first inequality follows immediately from the fact $x^2(1-y)^2/(x(1-y)+1)^2 \leq 1$. After taking the square root and rearranging, the second inequality is equivalent to $1 - 2\sqrt{y} + y \geq 0$. The left-hand side equals $(1 - \sqrt{y})^2$, and this shows the statement.

- (iv) The statement follows when we show that $4y \leq (x+1+y)^2$. The right-hand side is increasing in x , hence it suffices to show the statement for $x = 0$. But then the preceding inequality is equivalent to $0 \leq 1-2y+y^2 = (1-y)^2$. $\square$

The next result shows the standard Chernoff bound for the normal distribution.

Lemma C.9. *Let W be a random vector of dimension k with $W \sim N(0, I_k)$, where I_k is the k -dimensional identity matrix, and let c be a nonrandom k -vector. Then we have for any $\varepsilon > 0$*

$$\mathbb{P}(|W'c| \geq \varepsilon) \leq \begin{cases} 2e^{-\varepsilon^2/(2c'c)} & \text{if } c'c > 0 \\ 0 & \text{else.} \end{cases} \quad (57)$$

Proof. If $c'c = 0$, then c is the zero vector and the statement is trivially fulfilled. Now, we assume that c is not equal to the zero vector, i.e., that $c'c > 0$. Since the standard normal distribution is symmetric around zero, i.e., W and $-W$ have the same distribution, it suffices to treat the upper tail bound, i.e., $\mathbb{P}(W'c > \varepsilon)$. For the lower tail bound, note that $\mathbb{P}(W'c \leq -\varepsilon) = \mathbb{P}((-W)'c \leq -\varepsilon) = \mathbb{P}(W'c \geq \varepsilon)$. Recall that the moment generating function of the standard normal distribution is defined for all $u \in \mathbb{R}$ and equals $\exp(u^2/2)$. Moreover, note that $W'c/\sqrt{c'c} \sim N(0, 1)$ and hence, we have for any $s > 0$ using Markov's inequality

$$\mathbb{P}(W'c \geq \varepsilon) = \mathbb{P}\left(e^{sW'c/\sqrt{c'c}} \geq e^{s\varepsilon/\sqrt{c'c}}\right) \leq e^{-s\varepsilon/\sqrt{c'c} + s^2/2}.$$

The upper bound is minimized for $s^* = \varepsilon/\sqrt{c'c}$. The last term on the far right-hand side in the preceding display then equals $\exp(-\varepsilon^2/c'c + \varepsilon^2/(2c'c)) = \exp(-\varepsilon^2/(2c'c))$. $\square$

Corollary C.10. *Under the assumptions of Lemma C.9, assume that c is of the form Dd , where D is a nonsingular $k \times k$ matrix and d is a k -vector. Then we have*

$$\mathbb{P}(|W'Dd| \geq \varepsilon) \leq \begin{cases} 2e^{-\varepsilon^2/(2\lambda_k(D'D)d'd)} = 2e^{-\varepsilon^2\lambda_1((D'D)^{-1})/(2d'd)} & \text{if } d'd > 0 \\ 0 & \text{else,} \end{cases} \quad (58)$$

where $\lambda_1(\cdot)$ denotes the smallest eigenvalue of the indicated matrix whereas $\lambda_k(\cdot)$ denotes the largest one.

Proof. Note that $D'D$ as well as $(D'D)^{-1}$ are positive definite and that $Dd = 0$ can only occur when $d = 0$ because D is nonsingular. If $d = 0$, then the statement holds trivially. If $d \neq 0$, then $d'd > 0$, $d'D'Dd > 0$ and $d'D'Dd \leq \lambda_k(D'D)d'd$. This shows together with the first line of (57) in Lemma C.9 the remaining statement. $\square$

The following lemma combines the results about logarithmic Sobolev inequalities, concentration results for Lipschitz functions and the theory of random matrices. An overview of these topics can be found in Anderson et al. (2010). The result is due to Davidson and Szarek (2001) and is stated here for completeness with a more detailed proof.

Lemma C.11. *Let V be a $n \times k$, $n \geq k \geq 2$, random matrix with entries that are i.i.d standard normally distributed random variables. Then for every $\varepsilon > 0$, we have*

$$\mathbb{P}(s_1(V) \leq \sqrt{n} - \sqrt{k} - \sqrt{n\varepsilon}) \leq \exp(-n\varepsilon^2/2), \quad (59)$$

where $s_1(\cdot)$ denotes the smallest singular value of the indicated matrix.

Proof. The singular values of V , denoted by $s_1(V) \leq \dots \leq s_k(V)$, are Lipschitz functions of the entries of the matrix V with Lipschitz constant 1 (see Corollary 7.3.8 in Horn and Johnson (1985)). Furthermore, in Gross (1975) (or in Lemma 2.3.2 in Anderson et al. (2010)), it is shown that the standard normal distribution in $\mathbb{R}^N$ fulfills a logarithmic Sobolev inequality with constant 1. Regard the matrix V as a vector in $\mathbb{R}^N$ with $N = nk$ and use the preceding two results together with Herbst's Lemma (see e.g. Lemma 2.3.3 in Anderson et al. (2010)) to show that for all $\lambda \in \mathbb{R}$

$$\mathbb{E}[e^{\lambda(s_1(V) - \mathbb{E}[s_1(V)])}] \leq e^{\lambda^2/2}.$$

Hence, we have for every $\varepsilon > 0$ and $u > 0$

$$\mathbb{P}(s_1(V) - \mathbb{E}[s_1(V)] \leq -\varepsilon) = \mathbb{P}(e^{-u(s_1(V) - \mathbb{E}[s_1(V)])} \geq e^{u\varepsilon}) \leq e^{-u\varepsilon + u^2/2},$$

where the inequality follows from Markov's inequality and the inequality in the second-to-last display. Optimizing with respect to u gives $u^* = \varepsilon$ and the upper bound then equals $\exp(-\varepsilon^2/2)$. The result of the lemma follows if we can show that $\mathbb{E}[s_1(V)] \geq \sqrt{n} - \sqrt{k}$. For this, we use Gordon's extension of Slepian's Lemma, which can be found in Gordon (1985, Theorem 1.4). Let S^{k-1} and S^{n-1} be the unit $k-1$ and $n-1$ sphere, respectively, and let $a \in S^{k-1}$ and $b \in S^{n-1}$. We use Gordon's inequality with $X_{a,b} = \langle (g, h), (a, b) \rangle = \langle g, a \rangle + \langle h, b \rangle$, where $g \sim N(0, I_k)$ and $h \sim N(0, I_n)$ are independent, and with $Y_{a,b} = \langle Va, b \rangle$. If we can show for all $a, a' \in S^{k-1}$ and $b, b' \in S^{n-1}$ that

- (i) $\mathbb{E}[(X_{a,b} - X_{a,b'})^2] \leq \mathbb{E}[(Y_{a,b} - Y_{a,b'})^2]$ and
- (ii) $\mathbb{E}[(X_{a,b} - X_{a',b'})^2] \geq \mathbb{E}[(Y_{a,b} - Y_{a',b'})^2]$ for $a \neq a'$,

it follows that

$$\mathbb{E} \left[\inf_a \sup_b X_{a,b} \right] \leq \mathbb{E} \left[\inf_a \sup_b Y_{a,b} \right]. \quad (60)$$

In fact, Gordon's Lemma is stated in terms of finitely many random variables, i.e., where the index set for a and b is finite. That we can use the theorem as stated here, is shown in the subsequent remark, which shows that we can approximate the infimum and the supremum by reducing to finite minima and maxima. First, we calculate the expected values of $\inf_a \sup_b X_{a,b}$ and of $\inf_a \sup_b Y_{a,b}$ and then we verify conditions (i) and (ii). Recalling that $\sup_{u \in S^{m-1}} \langle x, u \rangle = \|x\|_2$ and $\inf_{u \in S^{m-1}} \langle x, u \rangle = -\|x\|_2$, shows that

$$\inf_a \sup_b X_{a,b} = \inf_a \langle g, a \rangle + \sup_b \langle h, b \rangle = -\|g\|_2 + \|h\|_2$$

and that

$$\inf_a \sup_b Y_{a,b} = \inf_a \sup_b \langle Va, b \rangle = \inf_a \sqrt{a' V' V a} = s_1(V).$$

Because $g \sim N(0, I_k)$, we see that $\|g\|_2 \sim \sqrt{\chi_k^2}$ and by Lemma A.1, we have $\mathbb{E}[\|g\|_2] = \sqrt{2}\Gamma((k+1)/2)/\Gamma(k/2)$. Using the same arguments for $\mathbb{E}\|h\|_2$, we obtain $\mathbb{E}[\inf_a \sup_b X_{a,b}] = \sqrt{2}[\Gamma((n+1)/2)/\Gamma(n/2) - \Gamma((k+1)/2)/\Gamma(k/2)]$. To show that the right-hand side of this equality is bounded from below by $\sqrt{n} - \sqrt{k}$, it suffices to show that the function $\Gamma(x+1/2)/\Gamma(x) - \sqrt{x}$ is increasing in x for $x \geq 1$. But this is obtained from Lemma C.8 (ii). Now, it remains to verify conditions (i) and (ii). Note that

$$\begin{aligned} \mathbb{E}[(X_{a,b} - X_{a',b'})^2] &= \mathbb{E}[(\langle g, a - a' \rangle + \langle h, b - b' \rangle)^2] \\ &= \mathbb{E}\left[\left(\sum_{j=1}^k g_j(a_j - a'_j) + \sum_{i=1}^n h_i(b_i - b'_i)\right)^2\right] \\ &= \sum_{j=1}^k (a_j - a'_j)^2 + \sum_{i=1}^n (b_i - b'_i)^2 \end{aligned}$$

and that

$$\begin{aligned} \mathbb{E}[(Y_{a,b} - Y_{a',b'})^2] &= \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^k \sum_{l=1}^n \sum_{m=1}^k V_{i,j} V_{l,m} (a_j b_i - a'_j b'_i)(a_m b_l - a'_m b'_l)\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^n \sum_{j=1}^k V_{i,j} (a_j b_i - a'_j b'_i)\right)^2\right] \\ &= \sum_{i=1}^n \sum_{j=1}^k (a_j b_i - a'_j b'_i)^2, \end{aligned}$$

where the last equality follows from the fact that V has independent entries that have mean zero and variance 1. If $a = a'$, then the last lines in the preceding two displays simplify to $\sum_{i=1}^n (b_i - b'_i)^2$ and hence condition (i) is satisfied with an equality. If $a \neq a'$, then, we have to show that

$$\sum_{i=1}^n \sum_{j=1}^k (a_j b_i - a'_j b'_i)^2 \leq \sum_{j=1}^k (a_j - a'_j)^2 + \sum_{i=1}^n (b_i - b'_i)^2$$

for unit vectors $a \in S^{k-1}$ and $b \in S^{n-1}$. The inequality in the preceding display is equivalent to $-pq \leq 1 - p - q$ with $p = \sum_j a_j a'_j$ and $q = \sum_i b_i b'_i$. We can rewrite the inequality as $0 \leq (1-p)(1-q)$. Noting that $p \in [-1, 1]$ and $q \in [-1, 1]$ shows condition (ii). $\square$

Remark. Let $A \subseteq \mathbb{R}^k$ be compact and let $Z : A \times \Omega \rightarrow \mathbb{R}$ be a stochastic process that is continuous in its first argument for almost all $\omega \in \Omega$. For convenience, we suppress the dependence on ω . We want to show that the sequence $\max_{a \in A_m} Z(a)$, where A_m are finite subsets of A that satisfy $A_m \subseteq$

A_{m+1} such that $\cup_m A_m$ is dense in A , converges monotonically to the supremum over A , i.e., to $\sup_{a \in A} Z(a)$ as $m \rightarrow \infty$. Convergence always refers to almost sure convergence. Let $U = \{u_1, u_2, u_3, \dots\}$ be a countable dense subset of A . Such subsets exist because every subset of a separable metric space is separable (see Chapter 2 in Dudley (2004)). For all $m \in \mathbb{N}$, set $A_m = \cup_{i=1}^m u_i$. Then for all $m \in \mathbb{N}$, A_m are finite and increasing subsets of A , i.e., $A_m \subseteq A_{m+1}$ and $A_m \subseteq A$, and $\cup_m A_m = U$ is dense. Let $T_m = \max_{a \in A_m} Z(a)$. Because $Z(a)$ is continuous in a and A is compact, there exists an $a^* \in A$ such that $Z(a^*) = \sup_{a \in A} Z(a)$. By the monotonicity of the sets A_m , the sequence T_m is nondecreasing and is furthermore bounded by $Z(a^*)$. Hence, $\sup_{m \in \mathbb{N}} T_m \leq Z(a^*) < \infty$ and the sequence T_m monotonically tends to $\sup_{m \in \mathbb{N}} T_m$. To show that $T_m \rightarrow Z(a^*)$, it suffices to show that $\sup_{m \in \mathbb{N}} T_m = Z(a^*)$. This is true as Z is continuous in a . More precisely, assume that $\sup_{m \in \mathbb{N}} T_m < Z(a^*)$. Then there exists an $\varepsilon > 0$ such that $T_m + \varepsilon \leq Z(a^*)$ for all $m \in \mathbb{N}$, i.e., that $\max_{a \in A_m} Z(a) \leq Z(a^*) - \varepsilon$, or, equivalently, that for all $m \in \mathbb{N}$, $Z(a) \leq Z(a^*) - \varepsilon$ for all $a \in A_m$. Fix this ε . By continuity, we can find a $\delta > 0$ such that if $\|a' - a''\|_2 < \delta$, then $|Z(a') - Z(a'')| < \varepsilon$. Since U is a dense subset, there exists a $u \in U$ such that $\|u - a^*\|_2 < \delta$ with $u \in A_m$ for one m . By continuity, it follows that $|Z(u) - Z(a^*)| < \varepsilon$, i.e., that $Z(a^*) - \varepsilon < Z(u) < Z(a^*) + \varepsilon$, a contradiction. Hence, it follows that $\max_{a \in A_m} Z(a) \rightarrow \sup_{a \in A} Z(a)$ almost surely as $m \rightarrow \infty$.

Now, let $A \subseteq \mathbb{R}^k$ and $B \subseteq \mathbb{R}^n$ be compact and let $X : A \times B \times \Omega \rightarrow \mathbb{R}$ be a doubly indexed stochastic process that is continuous in its first and second argument for almost all $\omega \in \Omega$. Use the argument in the preceding paragraph to show that for a fixed $a \in A$, it follows that $\lim_{m \rightarrow \infty} \max_{b \in B_m} X(a, b) = \sup_{b \in B} X(a, b)$ where B_m are finite subsets of B that satisfy $B_m \subseteq B_{m+1}$ such that $\cup_m B_m$ is dense in B . Note that $\tilde{X}(a) = \sup_{b \in B} X(a, b)$ is continuous in a . To see this, note that by the compactness of B , there exists a $b^* \in B$ such that $\sup_{b \in B} X(a, b) = X(a, b^*)$ and assume that $\tilde{X}(a)$ is not continuous in a_0 , i.e. that $X(a_0, b^*)$ is not continuous in its both arguments. But this contradicts the assumption that the process is continuous in its first and its second argument. Hence, we can use the same argument as in the preceding paragraph, mutatis mutandis, to show that $\max_{a \in A_{\tilde{m}}} (-\tilde{X}(a))$ is monotonically increasing and tends to $\sup_{a \in A} (-\tilde{X}(a))$, or equivalently, that $\min_{a \in A_{\tilde{m}}} \tilde{X}(a)$ is monotonically decreasing and tends to $\inf_{a \in A} \tilde{X}(a)$ as $\tilde{m} \rightarrow \infty$. Together with the monotone convergence theorem, we conclude that $\mathbb{E}[\min_{a \in A_{\tilde{m}}} \max_{b \in B_m} X(a, b)]$ tends to $\mathbb{E}[\inf_{a \in A} \sup_{b \in B} X(a, b)]$ as both m and $\tilde{m}$ tend to infinity.

To conclude (60), use Theorem 1.4 in Gordon (1985) for every $m \in \mathbb{N}$ and $\tilde{m} \in \mathbb{N}$ to get

$$\mathbb{E} \left[\min_{a \in A_{\tilde{m}}} \max_{b \in B_m} X(a, b) \right] \leq \mathbb{E} \left[\min_{a \in A_{\tilde{m}}} \max_{b \in B_m} Y(a, b) \right].$$

The result in (60) follows by letting $m \rightarrow \infty$ and $\tilde{m} \rightarrow \infty$ and using the result in the preceding paragraph.

Corollary C.12. *Under the assumptions of Lemma C.11, we have for every $0 < \varepsilon \leq 1 - \sqrt{k/n}$*

$$\mathbb{P} \left(\lambda_1(V'V/n) \leq (1 - \sqrt{k/n} - \varepsilon)^2 \right) \leq \exp(-n \varepsilon^2/2). \quad (61)$$

Proof. First of all, recall that $V'V$ has full rank almost surely and that $s_1(A) = \sqrt{\lambda_1(A'A)}$ for any $n \times k$ matrix A . If $\varepsilon = 1 - \sqrt{k/n}$, the probability on the left-hand side in (61) reduces to $\mathbb{P}(\lambda_1(V'V/n) = 0)$ which equals zero because all eigenvalues of $V'V$ are positive almost surely. Then the inequality is trivially fulfilled. In the case where $0 < \varepsilon < 1 - \sqrt{k/n}$, we have

$$\begin{aligned} \mathbb{P}\left(\lambda_1(V'V/n) \leq (1 - \sqrt{k/n} - \varepsilon)^2\right) &= \mathbb{P}\left(s_1(V/\sqrt{n}) \leq 1 - \sqrt{k/n} - \varepsilon\right) \\ &= \mathbb{P}\left(s_1(V) \leq \sqrt{n} - \sqrt{k} - \sqrt{n}\varepsilon\right) \end{aligned}$$

and the statement follows from Lemma C.11. $\square$

Remark. Let $\beta \in (0, 1]$. Then we have

$$\mathbb{P}\left(\lambda_1(V'V/n) \leq (1 - \beta)^2(1 - \sqrt{k/n})^2\right) \leq \exp\left(-n\beta^2(1 - \sqrt{k/n})^2/2\right).$$

If $\beta = 1$ or $k = n$, the left-hand side in the preceding display reduces to $\mathbb{P}(\lambda_1(V'V/n) = 0)$ which equals zero as discussed in Corollary C.12 and the statement holds trivially. If $k < n$ and $\beta \in (0, 1)$, the result follows from Corollary C.12 by setting $\varepsilon = \beta(1 - \sqrt{k/n}) > 0$.

Proof of Proposition C.1. Without loss of generality, we can assume that $Z'Z$ (and hence also $V'V$) is invertible as this happens with probability 1. Let a_n be the shrinkage factor of the estimator $\hat{\theta}^+(c)$, i.e., $a_n = \min\{c\hat{s}^2 p / \hat{\theta}'Z'\hat{\theta}, 1\}$. Recall that

$$\begin{aligned} \rho_3^2 &= (\hat{\theta} - \theta)'S(\hat{\theta} - \theta) - 2a_n \left[(\hat{\theta} - \theta)'S(\hat{\theta} - \theta) + (\hat{\theta} - \theta)'S\theta \right] \\ &\quad + a_n^2 \left[(\hat{\theta} - \theta)'S(\hat{\theta} - \theta) + 2(\hat{\theta} - \theta)'S\theta + \theta'S\theta \right] + s^2 \end{aligned}$$

and that

$$\hat{\rho}_3^2 = \hat{s}^2 \frac{p}{n-p+1} - 2a_n \hat{s}^2 \frac{p}{n-p+1} + a_n^2 \left(\hat{s}^2 \frac{p}{n-p+1} + \frac{Y'Y}{n} - \hat{s}^2 \right) + \hat{s}^2.$$

If $s^2 = 0$, Lemma C.2 shows that $(\hat{\theta} - \theta)'S(\hat{\theta} - \theta)$ and $\hat{s}^2$ equal zero almost surely. This implies that the quantities in the preceding two displays equal zero almost surely and that the statement in (43) holds trivially. Hence, we assume that $s^2 > 0$ from now on.

Rewrite the term on the left-hand side in (43) as

$$\mathbb{P}\left(|\rho_3^2 - \tilde{\rho}_3^2 + \tilde{\rho}_3^2 - \hat{\rho}_3^2| \geq \varepsilon\right),$$

where

$$\tilde{\rho}_3^2 = (1 - a_n)^2 s^2 \frac{p}{n-p+1} - a_n^2 s^2.$$

Using the expansion of ρ_3^2 and $\hat{\rho}_3^2$ as before, we can bound the term in the second-to-last display or, equivalently, the left-hand side in (43) by

$$\begin{aligned} & \mathbb{P}\left((1-a_n)^2 \left|(\hat{\theta}-\theta)'S(\hat{\theta}-\theta) - s^2 \frac{p}{n-p+1}\right| \geq \varepsilon\alpha_1\right) \\ & + \mathbb{P}\left(2|a_n^2 - a_n| \left|(\hat{\theta}-\theta)'S\theta\right| \geq \varepsilon\alpha_2\right) + \mathbb{P}\left(a_n^2 |\theta'S\theta + s^2 - Y'Y/n| \geq \varepsilon\alpha_3\right) \\ & + \mathbb{P}\left(|1-a_n^2| |s^2 - \hat{s}^2| \geq \varepsilon\alpha_4\right) + \mathbb{P}\left((1-a_n)^2 \frac{p}{n-p+1} |s^2 - \hat{s}^2| \geq \varepsilon\alpha_5\right), \end{aligned}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ and α_5 are weights lying in $(0, 1)$ that sum up to one. Because $a_n \in [0, 1]$, we have $(1-a_n)^2 \leq 1-a_n^2 \leq 1$ and $0 \leq a_n - a_n^2 \leq 1/4$. Hence, we can bound the sum in the preceding display using Lemma C.2 by

$$\begin{aligned} & \mathbb{P}\left(\left|\frac{\chi_p^2}{\chi_{n-p+1}^2} - \frac{p}{n-p+1}\right| \geq \frac{\varepsilon\alpha_1}{s^2}\right) + \mathbb{P}\left(\left|W'(V'V)^{-1/2} \frac{S^{1/2}\theta}{s}\right| \geq \frac{2\varepsilon\alpha_2}{s^2}\right) \\ & + \mathbb{P}\left(\left|\frac{\chi_n^2(\theta'Z'Z\theta/s^2)}{n} - \mu - 1\right| \geq \frac{\varepsilon\alpha_3}{s^2}\right) \\ & + \mathbb{P}\left(\left|\frac{\chi_{n-p}^2}{n-p} - 1\right| \geq \frac{\varepsilon\alpha_4}{s^2}\right) + \mathbb{P}\left(\left|\frac{\chi_{n-p}^2}{n-p} - 1\right| \geq \frac{\varepsilon\alpha_5}{s^2} \frac{n-p+1}{p}\right), \end{aligned} \quad (62)$$

where $W = (V'V)^{1/2}S^{1/2}(\hat{\theta}-\theta)/s$ and where $\mu = \theta'S\theta/s^2$. Write η as shorthand for ε/s^2 . Use Lemma A.1 and Lemma A.3 (i) in Leeb (2008) and Lemma C.8 (i) to bound the first term in the preceding display by

$$\begin{aligned} & 2 \exp\left(-\frac{n-p+1}{2} \frac{\eta^2\alpha_1^2}{2(\eta\alpha_1 + p/(n-p+1) + 1)^2}\right) \\ & = 2 \exp\left(-(n-p+1) \frac{\eta^2\alpha_1^2(n-p+1)^2/(n+1)^2}{4(\eta\alpha_1(n-p+1)/(n+1) + 1)^2}\right) \\ & \leq 2 \exp\left(-(n-p+1) \frac{\eta^2\alpha_1^2(1-p/n)^2}{4(\eta\alpha_1(1-p/n) + 1)^2}\right). \end{aligned} \quad (63)$$

Noting that $W \sim N(0, I_p)$, we can bound the second term in (62) using Corollary C.10 conditional on Z or, equivalently, V with $D = (V'V)^{-1/2}$ and $d = S^{1/2}\theta/s$ by

$$2 \mathbb{E} \left[\exp \left(-n \frac{2\eta^2\alpha_2^2}{\mu} \lambda_1(V'V/n) \right) \right],$$

where the term is to be interpreted as zero if $\mu = 0$. The integrand of this term is bounded by 1 and nonincreasing in $\lambda_1(V'V/n)$. Integrating separately over the events $\{\lambda_1(V'V/n) > (1-\beta)^2(1-\sqrt{p/n})^2\}$ and $\{\lambda_1(V'V/n) \leq (1-\beta)^2(1-\sqrt{p/n})^2\}$ for some $\beta \in (0, 1)$ and using the remark after Corollary C.12 on the second event, we can bound the term in the preceding display by

$$2 \exp \left(-n \frac{2\eta^2\alpha_2^2}{\mu} (1-\beta)^2(1-\sqrt{p/n})^2 \right) + 2 \exp \left(-n \frac{\beta^2(1-\sqrt{p/n})^2}{2} \right),$$

where the sum is to be interpreted as zero if $\mu = 0$. Note that $(1 - \sqrt{p/n})^2 \geq (1 - p/n)^2/4$ holds by the second inequality in Lemma C.8 (iii). Using this together with the inequality in Lemma C.8 (iv) with $x = \eta\alpha_2(1 - p/n)$ and $y = \mu$ for the first term and together with the first inequality in Lemma C.8 (iii) with $x = \eta$ and $y = p/n$ for the second term, we can bound the sum in the preceding display for all $\mu \geq 0$ by

$$2 \exp \left(-n \frac{2(1 - \beta)^2 \eta^2 \alpha_2^2 (1 - p/n)^2}{(\eta\alpha_2(1 - p/n) + 1 + \mu)^2} \right) + 2 \exp \left(-n \frac{\beta^2 \eta^2 (1 - p/n)^4}{8(\eta(1 - p/n) + 1)^2} \right).$$

Choosing $\beta \in (0, 1)$ such that $2(1 - \beta)^2 = \beta^2/8$ gives $\beta = 4/5$. Thus, we can bound the second term in (62) for all $\mu \geq 0$ by

$$2 \exp \left(-n \frac{2\eta^2 \alpha_2^2 (1 - p/n)^2}{25(\eta\alpha_2(1 - p/n) + 1 + \mu)^2} \right) + 2 \exp \left(-n \frac{2\eta^2 (1 - p/n)^4}{25(\eta(1 - p/n) + 1)^2} \right). \quad (64)$$

If $\mu > 0$, bound the third term in (62) using Lemma C.7 for any $\gamma \in (0, 1)$ by

$$2 \exp \left(-n \frac{\eta^2 \alpha_3^2 (1 - \gamma)^2}{4(\eta\alpha_3(1 + \gamma) + 2(1 + \mu))} \right) + 4 \exp \left(-n \frac{\eta^2 \alpha_3^2 \gamma^2}{4\mu(\eta\alpha_3\gamma + \mu)} \right).$$

Choosing $\gamma = 1/2$ for simplicity, we can bound the sum in the preceding display by

$$2 \exp \left(-n \frac{\eta^2 \alpha_3^2}{8(3\eta\alpha_3 + 4 + 4\mu)} \right) + 4 \exp \left(-n \frac{\eta^2 \alpha_3^2}{8\mu(\eta\alpha_3 + 2\mu)} \right). \quad (65)$$

Note that $\mu = 0$ if and only if $\theta = 0$. In that case, we can bound the third term in (62) using Corollary C.5 by

$$2 \exp \left(-\frac{n}{2} \tilde{m}(\eta\alpha_3) \right) = 2 \exp \left(-n \frac{\eta^2 \alpha_3^2}{4(\eta\alpha_3 + 1)} \right),$$

where the equality is obtained from the fact that $\tilde{m}(y) = y^2/(2(y + 1))$. The term in the preceding display can be bounded from above by the first term in (65). Hence, we will use the sum in (65) as an upper bound for the third term in (62) for all $\mu \geq 0$.

For the sum of the fourth and the fifth term in (62), let α_4 and α_5 be such that $\alpha_4 = \alpha_5(n - p + 1)/p$ and $\alpha_4 + \alpha_5 = \delta$ where $\delta \in (0, 1)$ is such that $\alpha_1 + \alpha_2 + \alpha_3 + \delta = 1$. Hence, we choose $\alpha_5 = \delta p/(n + 1)$ and use Corollary C.5 to bound the sum of those two terms by

$$\begin{aligned} & 4 \exp \left(-\frac{n - p}{2} \tilde{m} \left(\eta \delta \frac{n - p + 1}{n + 1} \right) \right) \\ &= 4 \exp \left(-(n - p) \frac{\eta^2 \delta^2 (n - p + 1)^2 / (n + 1)^2}{4(\eta \delta (n - p + 1) / (n + 1) + 1)} \right) \\ &\leq 4 \exp \left(-(n - p) \frac{\eta^2 \delta^2 (1 - p/n)^2}{4(\eta \delta (1 - p/n) + 1)} \right). \end{aligned} \quad (66)$$

Using the results in (63), (64), (65) and (66), we can bound the left-hand side in (43) by

$$\begin{aligned} & 2 \exp \left(-(n-p+1) \frac{\eta^2 \alpha_1^2 (1-p/n)^2}{4(\eta \alpha_1 (1-p/n) + 1)^2} \right) \\ & + 2 \exp \left(-n \frac{2\eta^2 \alpha_2^2 (1-p/n)^2}{25(\eta \alpha_2 (1-p/n) + 1 + \mu)^2} \right) \\ & + 2 \exp \left(-n \frac{2\eta^2 (1-p/n)^4}{25(\eta(1-p/n) + 1)^2} \right) + 2 \exp \left(-n \frac{\eta^2 \alpha_3^2}{8(3\eta \alpha_3 + 4 + 4\mu)} \right) \\ & + 4 \exp \left(-n \frac{\eta^2 \alpha_3^2}{8\mu(\eta \alpha_3 + 2\mu)} \right) + 4 \exp \left(-(n-p) \frac{\eta^2 \delta^2 (1-p/n)^2}{4(\eta \delta (1-p/n) + 1)} \right), \end{aligned}$$

where $\alpha_1, \alpha_2, \alpha_3$ and δ are weights lying in $(0, 1)$ that sum up to 1. Choosing $\alpha_1 = 1/8, \alpha_2 = 1/4, \alpha_3 = 1/2$ and $\delta = 1/8$, the sum in the preceding display equals

$$\begin{aligned} & 2 \exp \left(-(n-p+1) \frac{\eta^2 (1-p/n)^2}{4(\eta(1-p/n) + 8)^2} \right) \\ & + 2 \exp \left(-n \frac{2\eta^2 (1-p/n)^2}{25(\eta(1-p/n) + 4(1+\mu))^2} \right) \\ & + 2 \exp \left(-n \frac{2\eta^2 (1-p/n)^4}{25(\eta(1-p/n) + 1)^2} \right) + 2 \exp \left(-n \frac{\eta^2}{16(3\eta + 8(1+\mu))} \right) \\ & + 4 \exp \left(-n \frac{\eta^2}{16\mu(\eta + 4\mu)} \right) + 4 \exp \left(-(n-p) \frac{\eta^2 (1-p/n)^2}{32(\eta(1-p/n) + 8)} \right). \end{aligned}$$

The sum in the preceding display can be bounded from above by

$$\begin{aligned} & 2 \exp \left(-(n-p) \frac{\eta^2 (1-p/n)^2}{16(\eta(1-p/n) + 4)^2} \right) \\ & + 2 \exp \left(-n \frac{\eta^2 (1-p/n)^2}{16(\eta(1-p/n) + 4(1+\mu))^2} \right) \\ & + 2 \exp \left(-n \frac{\eta^2 (1-p/n)^4}{16(\eta(1-p/n) + 4)^2} \right) \\ & + 6 \exp \left(-n \frac{\eta^2}{16(\eta + 4(1+\mu))^2} \right) + 4 \exp \left(-(n-p) \frac{\eta^2 (1-p/n)^2}{16(\eta(1-p/n) + 4)^2} \right). \end{aligned}$$

Use the fact that $\mu \geq 0$ for the first, the third and the last term, the fact that $\eta^2/(\eta + 4(1+\mu))^2$ is increasing in η for all $\mu \geq 0$ together with the fact that $\eta \geq \eta(1-p/n)$ for the fourth term and the fact that $1 \geq 1-p/n$ for all but the third term, to bound the sum in the preceding display by

$$16 \exp \left(-n \frac{\eta^2 (1-p/n)^4}{16(\eta(1-p/n) + 4(1+\mu))^2} \right).$$

Recalling that $\eta = \varepsilon/s^2$ and $\mu = \theta'S\theta/s^2$, we can rewrite the term in the preceding display as

$$16 \exp \left(-n (1-p/n)^2 \frac{\varepsilon^2 (1-p/n)^2}{16(\varepsilon(1-p/n) + 4(\theta'S\theta + s^2))^2} \right).$$

□

References

- G. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*. Cambridge University Press, Cambridge, 1st edition, 2010.
- Z. Bai and J. W. Silverstein. *Spectral Analysis of Large Dimensional random matrices*. Springer, New York, 2nd edition, 2010.
- A. J. Baranchik. A family of minimax estimators of the mean of a multivariate normal distribution. *Ann. Math. Stat.*, 41(2):642–645, 1970.
- A. J. Baranchik. Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. *Ann. Statist.*, 1(2):312–321, 1973.
- J. Berger and M. E. Bock. Combining independent normal mean estimation problems with unknown variances. *Ann. Statist.*, 4(3):642–648, 1976.
- J. Berger, M. E. Bock, L. D. Brown, G. Casella, and L. Gleser. Minimax estimation of a normal mean vector for arbitrary quadratic loss and unknown covariance matrix. *Ann. Statist.*, 5(4):763–771, 1977.
- M. E. Bock. Minimax estimators of the mean of a multivariate normal distribution. *Ann. Statist.*, 3(1):209–218, 1975.
- L. D. Brown. Estimation with incompletely specified loss functions (the case of several location parameters). *J. Amer. Statist. Assoc.*, 70(350):417–427, 1975.
- L. D. Brown. An ancillarity paradox which appears in multiple linear regression. *Ann. Statist.*, 18(2):471–493, 1990.
- E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n . *Ann. Statist.*, 35(6):2313–2351, 2007.
- J. B. Copas. Regression, prediction and shrinkage. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 45(3):311–354, 1983.
- K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook of the Geometry of Banach Spaces*, volume 1, pages 317–366. North-Holland Publishing Company, Amsterdam, 2001.
- L. Dicker. Dense signals, linear estimators, and out-of-sample prediction for high-dimensional linear models. *arXiv:1102.2952 [math.ST]*, 2012.
- R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, Cambridge, 1st edition, 2004.
- I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.
- Y. Gordon. Some inequalities for Gaussian processes and applications. *Israel J. Math.*, 50(4):265–289, 1985.

- L. Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975.
- A. E. Hörl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1st edition, 1985. Reprinted with corrections in 1990.
- H. Hotelling. The generalization of Student’s ratio. *Ann. Math. Stat.*, 2(3):360–378, 1931.
- N. Huber and H. Leeb. Shrinkage estimators for prediction out-of-sample: Conditional performance. *Comm. Statist. Theory Methods*, 42(7):1245–1263, 2013.
- W. James and C. Stein. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:361–379, 1961.
- N. L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Multivariate Distributions*. John Wiley & Sons, New York, 1972.
- N. L. Johnson, S. Kotz, and N. Balakrishnan. *Distributions in Statistics: Continuous Univariate Distributions*, volume 2. John Wiley & Sons, New York, 2nd edition, 1995.
- J. Johnston and J. E. DiNardo. *Econometric Methods*. McGraw-Hill, New York, 4th edition, 1997.
- G. G. Judge and M. E. Bock. A comparison of traditional and Stein-rule estimators under weighted squared error loss. *Internat. Econ. Rev.*, 17(1):234–240, 1976.
- G. G. Judge and M. E. Bock. *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*. North-Holland Publishing Company, Amsterdam, 1978.
- D. Kershaw. Extensions of W. Gautschi’s inequalities for the gamma function. *Math. Comp.*, 41(164):607–611, 1983.
- H. Leeb. Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli*, 14(3):661–690, 2008.
- H. Leeb. Conditional predictive inference post model selection. *Ann. Statist.*, 37(5B):2838–2876, 2009.
- H. Leeb and B. M. Pötscher. Model selection. In T. G. Andersen, R. A. David, J.-P. Kreiß, and T. Mikosch, editors, *Handbook of Financial Time Series*, pages 785–821. Springer, New York, 2008.
- E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, New York, 2nd edition, 1998.

- F. Qi. A completely monotonic function involving the divided difference of the psi function and an equivalent inequality involving sums. *ANZIAM J.*, 48(4): 523–532, 2007.
- C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:197–206, 1956.
- C. Stein. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.
- W. E. Strawderman. On minimax estimation of a normal mean vector for general quadratic loss. *Lecture Notes-Monograph Series*, 42:3–14, 2003.
- R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58(1):267–288, 1996.

Abstract

We consider a linear regression model and study the performance of the James–Stein estimator (and of related James–Stein-type shrinkage estimators) for out-of-sample prediction. In contrast to in-sample prediction, where the regression function is estimated at the observed design points, we estimate the regression function at a new point. Further, we show how to select a ‘good’ estimator for prediction out-of-sample among a family of candidate estimators. In both problems, we focus on the challenging situation where the number of explanatory variables can be of the same order as sample size. We measure the performance of estimators by their (conditional) mean squared prediction errors, we derive simple uniform approximations to these errors, and we show that these errors can be estimated in a uniformly consistent fashion.

For prediction out-of-sample, we find that the James–Stein estimator can perform poorly compared to the maximum likelihood estimator, which is known to be dominated by the James–Stein estimator in the in-sample scenario (see Stein (1956) or the comprehensive monograph Judge and Bock (1978)). Our findings have important ramifications for the out-of-sample performance of methods that rely on some sort of shrinkage through, e.g. regularization, model selection or model averaging; see Leeb and Pötscher (2008) for a survey. When evaluating the predictive performance of estimators, we treat the regressor matrix in the training data as fixed, i.e., we condition on the design variables. Our results contrast those obtained by Baranchik (1973) and, more recently, by Dicker (2012) in an unconditional performance evaluation.

Further, we consider a linear regression model with Gaussian random design together with a family of James–Stein-type shrinkage estimators, where the number of candidate estimators can be much larger than sample size. We show that the empirically best estimator, i.e., the minimizer of the estimated prediction error, is asymptotically as good as the truly best (oracle) estimator, uniformly over a large class of data-generating processes. Moreover, we show that we can estimate the performances of both estimators in a uniformly consistent fashion. Our main results are explicit uniform finite sample performance bounds for Gaussian data. These findings extend results of Leeb (2008) where the underlying estimators are least squares estimators.

Zusammenfassung

Wir vergleichen den James-Stein-Schätzer (und ähnliche Shrinkage-Schätzer) mit dem Maximum-Likelihood-Schätzer als out-of-sample Prädiktoren. Im Gegensatz zur in-sample Prognose, bei der die Regressionsfunktion an bereits beobachteten Punkten geschätzt wird, schätzt man in der out-of-sample Prognose die Regressionsfunktion an nicht in der Stichprobe enthaltenen Punkten. Weiters betrachten wir eine Familie von Shrinkage-Schätzern, aus der wir einen "guten" out-of-sample Prädiktor auswählen möchten. In beiden Problemen betrachten wir den interessanten Fall, in dem die Anzahl der Parameter und der Stichprobenumfang von der gleichen Größenordnung sein können. Die Performance der Schätzer wird mit der (bedingten) erwarteten quadratischen Abweichung gemessen, wir leiten Näherungen für diese Abweichungen her und zeigen, dass diese gleichmäßig konsistent geschätzt werden können.

Wir zeigen, dass der James-Stein-Schätzer für die out-of-sample Prognose schlechter als der Maximum-Likelihood-Schätzer abschneiden kann, obwohl er den Maximum-Likelihood-Schätzer in der in-sample Prognose dominiert (siehe Stein (1956) oder das Buch von Judge und Bock (1978)). Dieses Resultat wirft auch ein neues Licht auf die out-of-sample Performance anderer Shrinkage-Methoden, die auf Regularisierung oder Modellselektion beruhen; für einen Überblick siehe Leeb und Pötscher (2008). In der Analyse bedingen wir auf die Design-Variablen. Unsere Ergebnisse stehen im Gegensatz zu jenen von Baranchik (1973) und von Dicker (2012), die beide über die Design-Verteilung mitteln.

Weiters betrachten wir ein lineares Regressionsmodell mit Gauß'schen Design-Variablen und eine Familie von Shrinkage-Schätzern, wobei die Anzahl an Schätzern den Stichprobenumfang übersteigen kann. Wir zeigen, dass der empirisch beste Schätzer, das ist der Minimierer der geschätzten Performance, asymptotisch gleich gut ist wie der beste Schätzer, das ist jener Schätzer, der die wahre Performance minimiert, und dass man die Performance beider Schätzer gleichmäßig schätzen kann. Beide Resultate gelten gleichmäßig über eine große Klasse von datengenerierenden Prozessen. Diese Ergebnisse erweitern die Resultate in Leeb (2008), in dem die betrachtete Familie aus Kleinst-Quadrate-Schätzern besteht.

CURRICULUM VITAE

Personal Information

Name	Nina Huber
Address	Department of Statistics and Operations Research University of Vienna Universitätsstraße 5/3/9 1010 Vienna, Austria
Email	n.huber@univie.ac.at
Phone	+43 1 4277 386 22

Research Interests

- Predictive Inference.
- Shrinkage Estimation.
- Model Selection.
- Random Matrices in High Dimensions.

Education

since October 2009	PhD studies of Statistics, University of Vienna, Shrinkage methods for prediction out-of-sample: Performance and selection of estimators Adviser: Prof. Hannes Leeb.
August 2009	M.Sc. in Statistics, University of Vienna, Master's Thesis: Analyse von Werbedaten in Österreich mittels moderner statistischer Methoden Adviser: a.o. Prof. Andreas Futschik.
February 2007	B.Sc. in Statistics, University of Vienna.

Professional Experience

- | | |
|-------------------|---|
| 2009 – | Research Assistant, Department of Statistics and Operations Research, University of Vienna. |
| 2006 – 2007, 2009 | Teaching Assistant, Department of Statistics and Operations Research, University of Vienna. |

Publications

- [1] Shrinkage methods for prediction out-of-sample: Performance and selection of estimators, PhD thesis, University of Vienna, forthcoming.
- [2] Shrinkage estimators for prediction out-of-sample: Conditional performance. (With H. Leeb) *Comm. Statist. Theory Methods*, 42(7):1245–1263, 2013.
- [3] Analyse von Werbedaten in Österreich mittels moderner statistischer Methoden, Master's Thesis, University of Vienna, 2009.

Conferences and Workshops

- | | |
|----------------|---|
| July 2012 | 8th World Congress in Probability and Statistics, Istanbul, Turkey. |
| July 2012 | Workshop Statistical Inference in Complex/High-Dimensional Problems, Vienna, Austria. |
| September 2011 | 17th European Young Statisticians Meeting, Lisbon, Portugal. |

Teaching Experience

- Exercises in Linear Algebra
- Exercises in Statistical Inference
- Exercises in Asymptotic Statistics
- Exercises in Linear Models
- Exercises in Probability Theory I