Prediction out-of-sample using block shrinkage estimators: model selection and predictive inference

Hannes Leeb and Nina S. Senitschnig

Department of Statistics, University of Vienna

Abstract

In a linear regression model with random design, we consider a family of candidate models from which we want to select a 'good' model for prediction out-of-sample. We fit the models using block shrinkage estimators, and we focus on the challenging situation where the number of explanatory variables can be of the same order as sample size and where the number of candidate models can be much larger than sample size. We develop an estimator for the out-of-sample predictive performance, and we show that the empirically best model is asymptotically as good as the truly best model. Using the estimator corresponding to the empirically best model, we construct a prediction interval that is approximately valid and short with high probability, i.e., we show that the actual coverage probability is close to the nominal one and that the length of this prediction interval is close to the length of the shortest but infeasible prediction interval. All results hold uniformly over a large class of data-generating processes. These findings extend results of Leeb (2009, Ann. Stat. 37:2838-2876), where the models are fit using least-squares estimators, and of Huber (2013), where the models are fit using shrinkage estimators without block structure.

Keywords: (block) James–Stein estimator, prediction out-of-sample, random design
AMS subject classifications: 62M20, 62J07

1 Introduction

We study the problem of how to find a 'good' model for prediction out-of-sample and how to use this model for designing 'valid' and 'short' prediction intervals. We assume that the data are generated from a possibly infinite dimensional linear model where we impose some technical but rather innocuous assumptions. In contrast to other authors, we do not impose sparsity on the parameters, and we do not require a special structure of the covariance matrix. We allow the collection of models to be very large and the models in our collection to be very complex, i.e., we allow the number of models to exceed sample size and the number of parameters to be of the same order as sample size. The models are fit using a block James–Stein estimator. Our results are conditional on a training sample when we repeatedly predict future observations, i.e., in our analysis we fix the training data and average over future observations. All the results presented are finite sample results.

Model selection is a well studied field in statistics. It is also well known that inference after model selection needs special treatment, i.e., ignoring the selection step and doing inference as if the model was chosen a priori leads to invalid conclusions because model selection is usually data-driven and hence random (see e.g. Pötscher (1991) or Leeb and Pötscher (2005)). For an overview of model selection procedures and the properties of post-selection estimators see Leeb and Pötscher (2008).

Berk et al. (2013) propose valid confidence intervals post-model selection regardless of which model selection procedure was used. Bachoc et al. (2017a) generalize these results to post-model selection predictors. Both articles consider linear models with homoscedastic errors. Bachoc et al. (2017b) develop a general framework allowing for linear models with heteroscedastic errors or binary response models with general link functions. The results in these papers can not be compared to ours because they are covering a non-standard and model dependent target. Lee et al. (2015), Lee et al. (2016) and Tibshirani et al. (2016) (and the references in these papers) focus on the same target and discuss post-model selection inference in a 'condition on selection' framework. Post-model selection inference on the conventional parameters is covered in Pötscher and Schneider (2010) and Schneider (2016). They consider confidence sets based on thresholding estimators in Gaussian linear regression models. The papers by Belloni et al. (2013) and Belloni et al. (2014) as well as by van de Geer et al. (2014) and Zhang and Zhang (2014) develop valid inference procedures under sparsity conditions. It is crucial to emphasize that we do not assume sparsity or any special structure of the unknown parameters. With regard to form and content, this work is closely related to the paper by Leeb (2009) where the models are fit using least-squares estimators. Here, we extend these results to a larger class of estimators.

This paper is organized as follows: we give an introduction to the setting and the framework in Section 2, i.e., we describe the overall model, the collection of candidate models and how we fit the candidate models using block James–Stein-type shrinkage estimators. Furthermore, we introduce how we measure the performance of the competing models via the conditional mean-squared prediction error. In Section 3, we deal with selecting the best model and show how we can estimate its performance. Section 4 addresses the problem of constructing prediction intervals and shows that the intervals are asymptotically valid and short with high probability.

2 Framework

As data-generating process, we consider a linear regression model that can be infinite dimensional. The candidate models correspond to finite dimensional submodels of this overall model. For simplicity, we assume that we do not have an intercept and that the explanatory variables are centered.[1] We further assume that we have a block structure in the data and that we would like to estimate and shrink the parameters corresponding to the blocks differently.[2] Of course, there are different ways of shrinking in a block design, we will discuss and further describe the one strategy we pursue. We use a training sample to fit the models and measure the out-of sample predictive performance of each model using the conditional mean-squared prediction error when repeatedly predicting future observations keeping the training sample fixed.

More formally, as data-generating process we 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

$$y = \sum_{i=1}^{\infty} x_i \beta_i + u. \tag{1}$$

Throughout the paper, we will assume that the error u is centered with variance $\sigma^2 > 0$, that x is centered with variance-covariance matrix $\Sigma = \mathbb{E}[x_i x_j]_{i \geq 1, j \geq 1}$ 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. The assumptions made so far are rather standard and innocuous. We assume further that (y, x) is jointly Gaussian. We heavily rely on Gaussianity because we need the conditional mean to be linear and the conditional variance to be constant, a property that is fulfilled only for the Gaussian distribution. Results of Steinberger and Leeb (2018) and Milovič (2015) show that approximate linearity of the conditional expectation holds for a large class of distributions. Ongoing work of Milovič and Leeb deal with the conditional variance being approximately constant. We are confident that it is possible to get rid of the normality assumption using their results.

We are given a sample of size n which will be denoted 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 (1). Further, we consider a collection of candidate models $\mathcal{M}_n$ that are finite-dimensional submodels of the overall model in (1), where we restrict some components of β to zero. Each of these models can be identified by a 0-1 sequence $m = (m_i)_{i \geq 1}$, where $m_i = 0$ if the corresponding β_i is restricted to zero and where $m_i = 1$ if β_i is not restricted. For every model $m \in \mathcal{M}_n$, we denote by $|m|$ the number of unrestricted components of β, i.e., $|m| = \sum_{i \geq 1} m_i$, and we assume that $6 \leq |m| < n$.

Throughout, we consider fixed parameters β, σ^2 and Σ as in (1), a fixed sample size n and a fixed model m. Let $\hat{\beta}^B(m)$ denote the estimator of β in model m. If $m_i = 0$, then the i-th component of $\hat{\beta}^B(m)$ is defined as zero. For the remaining components, note the following: We write $x(m)$ for those entries in x where $m_i = 1$, and we write $X(m)$ for those columns of X that are included in the submodel m, i.e., $X(m)$ is a $n \times |m|$ matrix. Because (Y, X) consists of i.i.d. samples of the pair (y, x) it follows

[1] We believe that including an intercept and non-centered explanatory variables does not change the results qualitatively and that we can handle this task using similar methods as in Leeb (2009).

[2] For the unblocked case, see Leeb and Senitschnig (2015).

that $(Y, X(m))$ consists of i.i.d. samples of the pair $(y, x(m))$. Because we assumed that (y, x) is jointly Gaussian, we know that the conditional distribution of y given $x(m)$ is Gaussian where the conditional mean is linear in $x(m)$, i.e., equals $x(m)'\theta$ for some appropriate $|m|$-vector θ, and the conditional variance is constant in $x(m)$ and equals, say, $\sigma^2(m)$. Hence, we can write

$$Y = X(m)\theta + w \tag{2}$$

where $w \sim N(0, \sigma^2(m)I_n)$ for some $\sigma^2(m) > 0$ and where $X(m)$ and w are independent. We assume that $X(m)$ consists of two blocks of dimension $n \times |m_1|$ and $n \times |m_2|$ with $|m| = |m_1| + |m_2|$ such that $3 \leq |m_1|$ and $3 \leq |m_2|$, i.e., $X(m) = (X_1(m), X_2(m))$. We can rewrite the model in the preceding display as

$$\begin{aligned} Y &= X_1(m)\theta_1 + X_2(m)\theta_2 + w \\ &= X_1(m)\theta_1^* + X_2^*(m)\theta_2 + w, \end{aligned} \tag{3}$$

where $\theta_1^* = \theta_1 + (X_1(m)'X_1(m))^{-1}X_1(m)'X_2(m)\theta_2$ and $X_2^*(m) = M_1(m)X_2(m)$ with $M_1(m) = I_n - X_1(m)(X_1(m)'X_1(m))^{-1}X_1(m)'$, i.e., the projection on the orthogonal complement of the column span of $X_1(m)$. On the probability zero event where the inverse matrix does not exist, we use the Moore-Penrose inverse instead of the usual inverse. Note that the two regressors in (3) are orthogonal so that we can estimate θ_1^* and θ_2 separately. Let $\hat{\theta}_1^*$ and $\hat{\theta}_2$ be the least-squares estimators of θ_1^* and θ_2, i.e., $\hat{\theta}_1^* = (X_1(m)'X_1(m))^{-1}X_1(m)'Y$ and $\hat{\theta}_2 = (X_2(m)'M_1(m)X_2(m))^{-1}X_2(m)'M_1(m)Y$. Let $\hat{\theta}_1^{*JS}$ and $\hat{\theta}_2^{JS}$ be the positive part James–Stein-type shrinkage estimators that are obtained by shrinking the least-squares estimators $\hat{\theta}_1^*$ and $\hat{\theta}_2$, i.e.,

$$\hat{\theta}_1^{*JS} = \left(1 - c_1\hat{\sigma}^2(m)\frac{|m_1|}{\hat{\theta}_1^{*\prime}X_1(m)'X_1(m)\hat{\theta}_1^*}\right)_+ \hat{\theta}_1^*, \tag{4}$$

$$\hat{\theta}_2^{JS} = \left(1 - c_2\hat{\sigma}^2(m)\frac{|m_2|}{\hat{\theta}_2'X_2^*(m)'X_2^*(m)\hat{\theta}_2}\right)_+ \hat{\theta}_2, \tag{5}$$

where $(x)_+ = \max\{x, 0\}$, where $c_1 \geq 0$ and $c_2 \geq 0$ are tuning parameters and where $\hat{\sigma}^2(m)$ is the usual unbiased variance estimator in model (2). Note that we can rewrite the estimators as $\hat{\theta}_1^{*JS} = (1 - a_1(m))\,\hat{\theta}_1^*$ and $\hat{\theta}_2^{JS} = (1 - a_2(m))\,\hat{\theta}_2$ where the shrinkage factors are

$$\begin{aligned} a_1(m) &= \min\left\{1, c_1\hat{\sigma}^2(m)\frac{|m_1|}{\hat{\theta}_1^{*\prime}X_1(m)'X_1(m)\hat{\theta}_1^*}\right\}, \\ a_2(m) &= \min\left\{1, c_2\hat{\sigma}^2(m)\frac{|m_2|}{\hat{\theta}_2'X_2^*(m)'X_2^*(m)\hat{\theta}_2}\right\}. \end{aligned} \tag{6}$$

Because of the definition of θ_1^*, we set $\hat{\theta}_1^{JS} = \hat{\theta}_1^{*JS} - (X_1(m)'X_1(m))^{-1}X_1(m)'X_2(m)\hat{\theta}_2^{JS}$. Hence, for the remaining $|m|$ components of $\hat{\beta}^B(m)$, we use the vector $(\hat{\theta}_1^{JS\prime}, \hat{\theta}_2^{JS\prime})'$.

We assume that we have a new copy of the random variables (y, x), independent of the training sample, that we will denote by $(y^{(0)}, x^{(0)})$. We predict $y^{(0)}$ using the predictor

$$\hat{y}^{(0)}(m) = \sum_{i=1}^{\infty} x_i^{(0)}\hat{\beta}_i^B(m).$$

The conditional mean-squared prediction error corresponding to model m will be denoted by $\rho^2(m)$ and is defined as

$$\rho^2(m) = \mathbb{E}\left[\left(\hat{y}^{(0)}(m) - y^{(0)}\right)^2 \Big\| X, Y\right],$$

where the expectation is taken with respect to $(y^{(0)}, x^{(0)})$ and where the training sample is treated as fixed. We are interested in the model that performs best within our class of candidate models, i.e., we are looking for the minimizer of $\rho^2(m)$ over $m \in \mathcal{M}_n$. Because $\rho^2(m)$ depends on the unknown parameters β, Σ and σ^2 in a complicated fashion, we approximate it by an empirical counterpart that is defined as follows

$$\hat{\rho}^2(m) = w_1 \hat{\sigma}^2(m) + w_2 \frac{Y'(I_n - M_1(m))Y}{n} + w_3 \frac{Y' M_1(m) Y}{n},$$

where the weights equal

$$\begin{aligned} w_1 &= (1 - a_2(m))^2 \frac{|m|}{n - |m| + 1} + 1 - a_2(m)^2 - (a_1(m) - a_2(m))^2 \frac{|m_1|}{n - |m_1| + 1} \\ &\quad + (a_2(m) - a_1(m))(2 - a_1(m) - a_2(m)) \frac{|m_1|}{n - |m_1| + 1}, \\ w_2 &= a_1(m)^2, \\ w_3 &= a_2(m)^2 - a_1(m)^2 \frac{|m_1|}{n} + (a_2(m) - a_1(m))^2 \frac{|m_1|}{n - |m_1| + 1}. \end{aligned}$$

In Appendix A, we have a closer look on the derivation of $\hat{\rho}^2(m)$. The following result shows that for every model m the empirical mean-squared prediction error $\hat{\rho}^2(m)$ is a good approximation for the true mean-squared prediction error $\rho^2(m)$.

Theorem 1. *For every fixed $m \in \mathcal{M}_n$, we have for every $\varepsilon > 0$*

$$\mathbb{P}\left(\left|\log \frac{\hat{\rho}^2(m)}{\rho^2(m)}\right| \geq \varepsilon\right) \leq 31|m| \exp\left(-n \left(\frac{|m_1|}{n}\right)^2 \left(1 - \frac{|m|}{n}\right)^5 \frac{\varepsilon^2}{14397(1+\varepsilon)^2}\right), \tag{7}$$

where $\rho^2(m)$ and $\hat{\rho}^2(m)$ are defined above. Alternatively, we can bound the left-hand side in the preceding inequality from above by

$$31|m| \exp\left(-n \left(1 - \frac{|m|}{n}\right)^5 \frac{\varepsilon^2}{28279(1 + \mu(m))^2 (1+\varepsilon)^2}\right). \tag{8}$$

where $\mu(m) = \theta' \Sigma(m) \theta / \sigma^2(m)$ and where $\Sigma(m)$ is the variance-covariance matrix of $x(m)$.

This result shows that for any fixed model the true and the empirical mean-squared prediction error are close to each other. Noting that $\rho^2(m) \geq \sigma^2 > 0$, the term $\hat{\rho}^2(m)/\rho^2(m)$ is always well-defined. On the probability zero event where $\hat{\rho}^2(m) = 0$, $\log(\hat{\rho}^2(m)/\rho^2(m))$ should be interpreted as ∞. There are two different upper bounds in the previous result. Both upper bounds depend on known quantities like ε, n, $|m|$ and $|m_1|$. The upper bound in (7) does not depend on unknown quantities whereas the upper bound in (8) depends on the unknown signal-to-noise ratio $\mu(m)$. For every fixed model m, we can estimate the signal to noise ratio $\mu(m)$ by $(Y'Y/n)/\hat{\sigma}^2(m) - 1$. It should be noted that both upper bounds do not tend to zero as ε gets larger. We could present a smaller upper bound but at the expense of a more complicated and complex structure of the bound that does not clearly show the effect of the individual quantities.

3 Model selection

In this section, we use Bonferroni's inequality to extent Theorem 1 to hold uniformly over the whole class of candidate models $\mathcal{M}_n$ and we show how to use the result for model selection. We will not assume that one of the candidate models is the true model (because it is not the aim of the paper to find the true model). Rather we would like to find a 'good' model for prediction out-of-sample, that is a model having a small mean-squared prediction error. Because minimizing $\rho^2(m)$ is unfeasible, we minimize the empirical mean-squared prediction error $\hat{\rho}^2(m)$ instead. Lemma 2 and the subsequent corollary motivate this approach.

Lemma 2. *Consider a finite and non-empty collection of candidate models $\mathcal{M}_n$ and let $r_n = \inf_{m \in \mathcal{M}_n} |m_1|$ and $s_n = \sup_{m \in \mathcal{M}_n} |m|$. Then, we have for each $\varepsilon > 0$*

$$\begin{aligned} &\mathbb{P}\left(\sup_{m \in \mathcal{M}_n} \left| \log \frac{\hat{\rho}^2(m)}{\rho^2(m)} \right| \geq \varepsilon \right) \\ &\leq 31|\mathcal{M}_n| s_n \exp\left(-n \left(\frac{r_n}{n}\right)^2 \left(1 - \frac{s_n}{n}\right)^5 \frac{\varepsilon^2}{14397(1+\varepsilon)^2} \right), \end{aligned} \tag{9}$$

where $|\mathcal{M}_n|$ denotes the number of candidate models in collection $\mathcal{M}_n$. The result holds uniformly over all data-generating processes as in (1)*. Alternatively, we can bound the left-hand side in the preceding display from above by*

$$31|\mathcal{M}_n| s_n \exp\left(-n \left(1 - \frac{s_n}{n}\right)^5 \frac{\varepsilon^2}{28279(1+\varepsilon)^2 d^2} \right), \tag{10}$$

where the result holds uniformly over all data-generating processes as in (1) *such that* $\mathrm{Var}(y)/\sigma^2 \leq d$ *for some $d > 0$.*

We have two different types of exchangeable upper bounds. The upper bound in (9) only depends on known quantities and is exponentially small in n if only r_n/n is not too small and if s_n/n and $|\mathcal{M}_n|$ are not too large. So we need the number of unrestricted components in the first block to be large, more precisely, we need r_n/n to be bounded away from 0, i.e., $r_n/n > \eta_1$ for some $\eta_1 > 0$ and for all $n \in \mathbb{N}$. The models can also not be too complex, i.e., s_n/n should be bounded away from 1, i.e., $s_n/n < 1 - \eta_2$ for some $\eta_2 > 0$ and for all $n \in \mathbb{N}$. Furthermore, we see that the number of models in collection $\mathcal{M}_n$ can exceed sample size (and can actually be a large multiple of sample size) but it can not be too large, e.g., complete subset selection is not possible. The upper bound in (10) also depends on the unknown quantity d which is an upper bound on the signal-to-noise ratio of the data-generating process and is exponentially small in n if only s_n/n, $|\mathcal{M}_n|$ and d are not too large. Note that the factor s_n outside of the exponential term in both bounds is negligible.

A simple consequence of the preceding result is that the empirically best model is a 'good' model. For this purpose, let $\hat{m}_n^*$ and m_n^* be minimizers of $\hat{\rho}^2(m)$ and of $\rho^2(m)$, respectively, i.e.,

$$\hat{m}_n^* = \underset{m \in \mathcal{M}_n}{\arg\min} \hat{\rho}^2(m), \quad m_n^* = \underset{m \in \mathcal{M}_n}{\arg\min} \rho^2(m).$$

Corollary 3. *Let $r_n = \inf_{m \in \mathcal{M}_n} |m_1|$ and $s_n = \sup_{m \in \mathcal{M}_n} |m|$ Then, we have for each $\varepsilon > 0$*

$$\mathbb{P}\left(\left| \log \frac{\rho^2(\hat{m}_n^*)}{\rho^2(m_n^*)} \right| \geq \varepsilon \right) \leq 31|\mathcal{M}_n| s_n \exp\left(-n \left(\frac{r_n}{n}\right)^2 \left(1 - \frac{s_n}{n}\right)^5 \frac{\varepsilon^2}{14397(2+\varepsilon)^2} \right), \tag{11}$$

as well as

$$\mathbb{P}\left(\left|\log \frac{\hat{\rho}^2(\hat{m}_n^*)}{\rho^2(\hat{m}_n^*)}\right| \geq \varepsilon\right) \leq 31|\mathcal{M}_n| s_n \exp\left(-n\left(\frac{r_n}{n}\right)^2\left(1-\frac{s_n}{n}\right)^5 \frac{\varepsilon^2}{14397(1+\varepsilon)^2}\right), \quad (12)$$

where $\mathcal{M}_n$ denotes the number of candidate models in collection $\mathcal{M}_n$. Both results hold uniformly over all data-generating processes as in (1). *Alternatively, we can bound the left-hand side in* (11) *from above by*

$$31|\mathcal{M}_n| s_n \exp\left(-n\left(1-\frac{s_n}{n}\right)^5 \frac{\varepsilon^2}{28279(2+\varepsilon)^2 d^2}\right), \quad (13)$$

and the left-hand side in (12) *from above by*

$$31|\mathcal{M}_n| s_n \exp\left(-n\left(1-\frac{s_n}{n}\right)^5 \frac{\varepsilon^2}{28279(1+\varepsilon)^2 d^2}\right). \quad (14)$$

Both results hold uniformly over all data-generating processes as in (1) *such that* $\mathrm{Var}(y)/\sigma^2 \leq d$ *for some $d > 0$.*

The results in (11) and (13) show that the empirically best model is asymptotically as good as the truly best model, in the sense that the true mean-squared prediction error of the truly best model lies close to the true mean-squared prediction error of the empirically best model. The results in (12) and (14) show that the empirical mean-squared prediction error of the empirically best model lies close to its true mean-squared prediction error. This implies that we can estimate the true performance of the empirically best model just by plugging it into the empirical mean-squared prediction error.

4 Statistical inference

In this section, we construct prediction intervals and we show that these intervals have the desired properties of being asymptotically 'valid' and asymptotically 'short'.

For a fixed model m, the prediction error equals $y^{(0)} - \hat{y}^{(0)}(m)$. Conditional on the training sample this prediction error follows a centered normal distribution with variance $\rho^2(m)$. We will denote this distribution by $\mathbb{L}(m)$, i.e., $\mathbb{L}(m) \equiv N(0, \rho^2(m))$. Using this distribution to construct prediction intervals for $y^{(0)}$ is infeasible because it depends on unknown quantities via its variance $\rho^2(m)$. Let $\widehat{\mathbb{L}}(m) \equiv N(0, \hat{\rho}^2(m))$ be an approximation to the true distribution and use this distribution to construct the prediction interval. The next result shows that $\mathbb{L}(m)$ is close to $\widehat{\mathbb{L}}(m)$ in the sense that their total variation distance is small with high probability.

Theorem 4. *For a fixed model $m \in \mathcal{M}_n$, we have for all $\varepsilon > 0$*

$$\begin{aligned}&\mathbb{P}\left(\|\mathbb{L}(m) - \widehat{\mathbb{L}}(m)\|_{TV} \geq \varepsilon\right) \\ &\quad \leq 31|m| \exp\left(-n\left(\frac{|m_1|}{n}\right)^2\left(1-\frac{|m|}{n}\right)^5 \frac{\varepsilon^2}{900(1+4\varepsilon)^2}\right).\end{aligned} \quad (15)$$

Alternatively, we can bound the left-hand side in the preceding display from above by

$$31|m| \exp\left(-n\left(1-\frac{|m|}{n}\right)^5 \frac{\varepsilon^2}{1768(1+4\varepsilon)^2(1+\mu(m))^2}\right). \quad (16)$$

This result together with Bonferroni's inequality gives a uniform result over the whole class of candidate models.

Corollary 5. *Let $r_n = \inf_{m \in \mathcal{M}_n} |m_1|$ and $s_n = \sup_{m \in \mathcal{M}_n} |m|$. Then, we have for each $\varepsilon > 0$*

$$\begin{aligned} &\mathbb{P}\left(\sup_{m \in \mathcal{M}_n} \|\mathbb{L}(m) - \widehat{\mathbb{L}}(m)\|_{TV} \geq \varepsilon \right) \\ &\leq 31|\mathcal{M}_n| s_n \exp\left(-n \left(\frac{r_n}{n}\right)^2 \left(1 - \frac{s_n}{n}\right)^5 \frac{\varepsilon^2}{900(1+4\varepsilon)^2} \right). \end{aligned} \tag{17}$$

The result holds uniformly over all data generating processes as in (1)*. Alternatively, we can bound the left-hand side in the preceding display from above by*

$$31|\mathcal{M}_n| s_n \exp\left(-n \left(1 - \frac{s_n}{n}\right)^5 \frac{\varepsilon^2}{1768(1+4\varepsilon)^2 d^2} \right). \tag{18}$$

The result holds uniformly over all data generating processes as in (1) *such that* $\mathrm{Var}(y)/\sigma^2 \leq d$ *for some* $d > 0$*.*

Because $y^{(0)} - \hat{y}^{(0)}(m)$ is distributed as $\mathbb{L}(m)$, we see that an infeasible prediction interval for $y^{(0)}$ with coverage probability $1 - \alpha$ for some $\alpha \in (0, 1)$ is given by

$$\hat{y}^{(0)}(m) \pm Q_{1-\alpha/2}\rho(m),$$

where $Q_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution. The length of this 'prediction interval' equals $2Q_{1-\alpha/2}\rho(m)$ and is minimal for m_n^* with $2Q_{1-\alpha/2}\rho(m_n^*)$. Using $\hat{\rho}(m)$ instead of $\rho(m)$ in the previous display, we define the prediction interval as

$$\mathcal{I}(m) : \hat{y}^{(0)}(m) \pm Q_{1-\alpha/2}\hat{\rho}(m).$$

The length of this interval is minimized for $\hat{m}_n^*$. The next results shows that the coverage probability of $\mathcal{I}(\hat{m}_n^*)$, conditional on the training sample, is close to the nominal one, except on an event that has probability converging to zero as n increases under the same conditions we had in the previous section.

Corollary 6. *For every $\varepsilon > 0$, we have that*

$$\left| (1 - \alpha) - \mathbb{P}(y^{(0)} \in \mathcal{I}(\hat{m}_n^*)) \right| \leq \varepsilon \tag{19}$$

except on an event whose probability is not larger than

$$31|\mathcal{M}_n| s_n \exp\left(-n \left(\frac{r_n}{n}\right)^2 \left(1 - \frac{s_n}{n}\right)^5 \frac{\varepsilon^2}{900(1+4\varepsilon)^2} \right)$$

uniformly over all data generating processes as in (1)*. Alternatively, we can bound the probability of the exception event from above by*

$$31|\mathcal{M}_n| s_n \exp\left(-n \left(1 - \frac{s_n}{n}\right)^5 \frac{\varepsilon^2}{1768(1+4\varepsilon)^2 d^2} \right)$$

uniformly over all data generating processes as in (1) *such that* $\mathrm{Var}(y)/\sigma^2 \leq d$ *for some* $d > 0$*.*

The next result shows that the minimal length of the infeasbile prediction interval lies close to the length of $\mathcal{I}(\hat{m}_n^*)$.

Corollary 7. *For each $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\left|\log\frac{\hat{\rho}(\hat{m}_n^*)}{\rho(m_n^*)}\right| \geq \varepsilon\right) \leq 31|\mathcal{M}_n|s_n \exp\left(-n\left(\frac{r_n}{n}\right)^2\left(1-\frac{s_n}{n}\right)^5\frac{\varepsilon^2}{3600(1+2\varepsilon)^2}\right) \quad (20)$$

uniformly over all data generating processes as in (1)*. Alternatively, we can bound the left-hand side in* (20) *from above by*

$$31|\mathcal{M}_n|s_n \exp\left(-n\left(1-\frac{s_n}{n}\right)^5\frac{\varepsilon^2}{7070(1+2\varepsilon)^2d^2}\right) \quad (21)$$

uniformly over all data generating processes as in (1) *such that* $\mathrm{Var}(y)/\sigma^2 \leq d$ *for some* $d > 0$.

5 Conclusion

We have shown how to select a model that performs well for prediction out-of-sample and how to use this model to construct prediction-intervals. We measured the performance of a model by its mean-squared prediction error conditional on the training data (Y, X) when repeatedly predicting over future observations, i.e., in our analysis we kept the training data fixed and averaged over future observations. Fitting the models using least-squares estimators was done in Leeb (2008) and Leeb (2009) and fitting them using usual James–Stein-type shrinkage estimators was done in Huber (2013). In this work, we have considered a larger class of estimators namely blocked James-Stein-type shrinkage estimators. The models can be very complex in the sense that the number of parameters can grow with sample size but can never exceed sample size, and the collection of models can be huge, it can be much larger than sample size.

Because the true out-of-sample prediction error $\rho^2(m)$ is not known, we minimize its empirical counterpart $\hat{\rho}^2(m)$. First of all, we have shown that those two quantities lie close to each other in the sense that $\mathbb{P}(|\log(\hat{\rho}^2(m)/\rho^2(m))| \geq \varepsilon)$ is bounded from above by a bound that is exponentially small in n under some restrictions (see (7) and (8) and the discussion following that result). We have shown that the true performance of the model $\hat{m}_n^*$ that minimizes $\hat{\rho}^2(m)$ lies close to the minimal true performance (see (11) and (13)) and that we can use its empirical mean-squared prediction error to estimate its true performance (see (12) and (14)). Designing prediction intervals, we have used again the empirical counterpart $\hat{\rho}^2(m)$ instead of $\rho^2(m)$. We have shown that the interval $\mathcal{I}(\hat{m}_n^*)$ has actual coverage probability that lies close to the nominal one except on an event that has a probability converging to 0 exponentially fast in n (see Corollary 6). Furthermore, the length of this prediction interval is short in the sense that it is close to the length of an infeasible prediction interval that depends on the unknown out-of-sample prediction error $\rho^2(m)$ (see (20) and (21)). It should be noted that all our results are finite sample results that hold for every sample size n, and that hold uniformly over a large class of data-generating processes.

A Technical details for deriving $\hat{\rho}^2(m)$

To shorten the notation in the appendix, we drop the dependence on m in the notation and we keep in mind that we always consider a fixed model m. Thus, let $X(m) = Z$, $X_1(m) = Z_1$, $X_2(m) = Z_2$, $\sigma^2(m) = s^2$ and $\Sigma(m) = \mathbb{E}[x(m)x(m)'] = S$. The model in (2) then becomes

$$Y = Z\theta + w, \tag{22}$$

where $w \sim N(0, s^2 I_n)$ is independent of Z. Because Z is divided into two blocks, we also rewrite θ as $\theta = (\theta_1', \theta_2')'$ with θ_1 and θ_2 being a $|m_1|$-vector and a $|m_2|$-vector, respectively, and S as

$$S = \begin{pmatrix} S_{1,1} & S_{1,2} \\ S_{2,1} & S_{2,2} \end{pmatrix},$$

where $S_{1,1}$ is $|m_1| \times |m_1|$, $S_{1,2}$ is $|m_1| \times |m_2|$ and $S_{2,2}$ is $|m_2| \times |m_2|$. Note that $S_{2,1} = S_{1,2}'$ because S is symmetric. For motivating the formula for $\hat{\rho}^2(m)$, note that the true mean-squared prediction error equals

$$\rho^2(m) = (\hat{\theta}^{BJS} - \theta)' S (\hat{\theta}^{BJS} - \theta) + s^2,$$

where $\hat{\theta}^{BJS}$ is the blocked James–Stein-estimator for θ as defined in Section 2, i.e.,

$$\hat{\theta}^{BJS} = \begin{pmatrix} \hat{\theta}_1^{JS} \\ \hat{\theta}_2^{JS} \end{pmatrix} = \begin{pmatrix} \hat{\theta}_1^{*JS} - (Z_1'Z_1)^{-1} Z_1' Z_2 \hat{\theta}_2^{JS} \\ \hat{\theta}_2^{JS} \end{pmatrix},$$

where $\hat{\theta}_1^{*JS} = (1 - a_1)\hat{\theta}_1^*$ and $\hat{\theta}_2^{JS} = (1 - a_2)\hat{\theta}_2$ with[3]

$$a_1 = \min\left\{ c_1 \hat{s}^2 \frac{|m_1|}{\hat{\theta}_1^{*\prime} Z_1' Z_1 \hat{\theta}_1^*}, 1 \right\}, \; a_2 = \min\left\{ c_2 \hat{s}^2 \frac{|m_2|}{\hat{\theta}_2' Z_2' M_1 Z_2 \hat{\theta}_2}, 1 \right\}, \tag{23}$$

where $\hat{s}^2$ is the usual unbiased variance estimator for s^2 in model (22) (of course $\hat{s}^2 = \hat{\sigma}^2(m)$, $a_1 = a_1(m)$ and $a_2 = a_2(m)$ as in Section 2), where $\hat{\theta}_1 = (Z_1'Z_1)^{-1} Z_1' Y$ and $\hat{\theta}_2 = (Z_2' M_1 Z_2)^{-1} Z_2' M_1 Y$ with $M_1 = I_n - Z_1 (Z_1'Z_1)^{-1} Z_1'$. Let $\hat{\theta} = (\hat{\theta}_1', \hat{\theta}_2')'$ be the least-squares estimator in model (22) and note that $\hat{\theta}_1 = \hat{\theta}_1^* - (Z_1'Z_1)^{-1} Z_1' Z_2 \hat{\theta}_2$. Rewriting the James–Stein-type shrinkage estimator for θ_1 as

$$\begin{aligned}
\hat{\theta}_1^{JS} &= \hat{\theta}_1^* - a_1 \hat{\theta}_1^* - (Z_1'Z_1)^{-1} Z_1' Z_2 \hat{\theta}_2 + a_2 (Z_1'Z_1)^{-1} Z_1' Z_2 \hat{\theta}_2 \\
&= \hat{\theta}_1^* - (Z_1'Z_1)^{-1} Z_1' Z_2 \hat{\theta}_2 - a_2 \left[\hat{\theta}_1^* - (Z_1'Z_1)^{-1} Z_1' Z_2 \hat{\theta}_2 \right] - a_1 \hat{\theta}_1^* + a_2 \hat{\theta}_1^* \\
&= (1 - a_2)\hat{\theta}_1 + (a_2 - a_1)\hat{\theta}_1^*
\end{aligned}$$

we see that the blocked James–Stein-type shrinkage estimator for θ can be rewritten as

$$\hat{\theta}^{BJS} = (1 - a_2)\hat{\theta} + (a_2 - a_1) \begin{pmatrix} \hat{\theta}_1^* \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ denotes a $|m_2|$-vector of zeros. Note that the first term of this estimator has the same structure as the usual James–Stein estimator with shrinkage factor a_2 (see the last

[3] Umbenennen $a_1 = a_n^{(1)}$ und $a_2 = a_n^{(2)}$

display on page 33 in Huber (2013)). Using this, we can rewrite the true mean-squared prediction error as

$$\begin{aligned}\rho^2(m) = ((1-a_2)\hat{\theta}-\theta)'S((1-a_2)\hat{\theta}-\theta) + (a_2-a_1)^2\hat{\theta}_1^{*\prime}S\hat{\theta}_1^* + s^2 \\ + 2(a_2-a_1)\hat{\theta}_1^{*\prime}[S_{1,1}((1-a_2)\hat{\theta}_1-\theta_1) + S_{1,2}((1-a_2)\hat{\theta}_2-\theta_2)].\end{aligned} \tag{24}$$

We can rewrite the sum of the terms in the first line in (24) as

$$\begin{aligned}&(1-a_2)^2(\hat{\theta}-\theta)'S(\hat{\theta}-\theta) + 2a_2(a_2-1)(\hat{\theta}-\theta)'S\theta + a_2^2\theta'S\theta + s^2 \\ &+ (a_2-a_1)^2(\hat{\theta}_1^*-\theta_1^*)'S_{1,1}(\hat{\theta}_1^*-\theta_1^*) + 2(a_2-a_1)^2(\hat{\theta}_1^*-\theta_1^*)'S_{1,1}\theta_1^* + (a_2-a_1)^2\theta_1^{*\prime}S_{1,1}\theta_1^*.\end{aligned} \tag{25}$$

We can rewrite the sum of the terms in the second line in (24) as

$$\begin{aligned}&2(1-a_2)(a_2-a_1)(\hat{\theta}_1^*-\theta_1^*)'S_{1,1}(\hat{\theta}_1-\theta_1) \\ &\quad + 2(1-a_2)(a_2-a_1)(\hat{\theta}_1^*-\theta_1^*)'S_{1,2}(\hat{\theta}_2-\theta_2) \\ &\quad + 2(1-a_2)(a_2-a_1)\theta_1^{*\prime}S_{1,1}(\hat{\theta}_1-\theta_1) + 2(1-a_2)(a_2-a_1)\theta_1^{*\prime}S_{1,2}(\hat{\theta}_2-\theta_2) \\ &\quad - 2a_2(a_2-a_1)(\hat{\theta}_1^*-\theta_1^*)'(S_{1,1}\theta_1 + S_{1,2}\theta_2) \\ &\quad - 2a_2(a_2-a_1)\theta_1^{*\prime}(S_{1,1}\theta_1 + S_{1,2}\theta_2).\end{aligned}$$

Using that $\hat{\theta}_1-\theta_1 = \hat{\theta}_1^*-\theta_1^* - (Z_1'Z_1)^{-1}Z_1'Z_2(\hat{\theta}_2-\theta_2)$, we can rewrite the sum in the preceding display as

$$\begin{aligned}&2(1-a_2)(a_2-a_1)(\hat{\theta}_1^*-\theta_1^*)'S_{1,1}(\hat{\theta}_1^*-\theta_1^*) \\ &\quad + 2(1-a_2)(a_2-a_1)(\hat{\theta}_1^*-\theta_1^*)'[S_{1,2} - S_{1,1}(Z_1'Z_1)^{-1}Z_1'Z_2](\hat{\theta}_2-\theta_2) \\ &\quad + 2(1-a_2)(a_2-a_1)(\hat{\theta}_1^*-\theta_1^*)'S_{1,1}\theta_1^* \\ &\quad + 2(1-a_2)(a_2-a_1)\theta_1^{*\prime}[S_{1,2} - S_{1,1}(Z_1'Z_1)^{-1}Z_1'Z_2](\hat{\theta}_2-\theta_2) \\ &\quad - 2a_2(a_2-a_1)(\hat{\theta}_1^*-\theta_1^*)'(S_{1,1}\theta_1 + S_{1,2}\theta_2) \\ &\quad - 2a_2(a_2-a_1)\theta_1^{*\prime}(S_{1,1}\theta_1 + S_{1,2}\theta_2).\end{aligned} \tag{26}$$

Let $\tilde{Z}_2 = Z_2 - Z_1S_{1,1}^{-1}S_{1,2}$ and note that the quantity in squared brackets in (26) equals $-S_{1,1}(Z_1'Z_1)^{-1}Z_1'\tilde{Z}_2$. Using this and collecting the terms in (25) and (26), we can rewrite the true mean-squared prediction error as

$$\begin{aligned}\rho^2(m) = {}&(1-a_2)^2(\hat{\theta}-\theta)'S(\hat{\theta}-\theta) + s^2 \\ &+ (a_2-a_1)(2-a_1-a_2)(\hat{\theta}_1^*-\theta_1^*)'S_{1,1}(\hat{\theta}_1^*-\theta_1^*) \\ &+ a_2^2\theta'S\theta + 2a_2(a_1-a_2)\theta_1^{*\prime}(S_{1,1}\theta_1 + S_{1,2}\theta_2) + (a_2-a_1)^2\theta_1^{*\prime}S_{1,1}\theta_1^* \\ &+ 2a_2(a_2-1)(\hat{\theta}-\theta)'S\theta \\ &+ 2(1-a_1)(a_2-a_1)(\hat{\theta}_1^*-\theta_1^*)'S_{1,1}\theta_1^* \\ &+ 2a_2(a_1-a_2)(\hat{\theta}_1^*-\theta_1^*)'(S_{1,1}\theta_1 + S_{1,2}\theta_2) \\ &+ 2(1-a_2)(a_1-a_2)(\hat{\theta}_2-\theta_2)'\tilde{Z}_2'Z_1(Z_1'Z_1)^{-1}S_{1,1}\theta_1^* \\ &+ 2(1-a_2)(a_1-a_2)(\hat{\theta}_1^*-\theta_1^*)'S_{1,1}(Z_1'Z_1)^{-1}Z_1'\tilde{Z}_2(\hat{\theta}_2-\theta_2).\end{aligned} \tag{27}$$

Using the fact that $\theta_1^* = \theta_1 + S_{1,1}^{-1}S_{1,2}\theta_2 + (Z_1'Z_1)^{-1}Z_1'\tilde{Z}_2\theta_2$, we have

$$\begin{aligned}\theta_1^{*\prime}(S_{1,1}\theta_1 + S_{1,2}\theta_2) = {}&\theta_1'S_{1,1}\theta_1 + 2\theta_1'S_{1,2}\theta_2 + \theta_2'S_{2,1}S_{1,1}^{-1}S_{1,2}\theta_2 \\ &+ (S_{1,1}\theta_1 + S_{1,2}\theta_2)(Z_1'Z_1)^{-1}Z_1'\tilde{Z}_2\theta_2\end{aligned}$$

as well as

$$\begin{aligned}\theta_1^{*\prime} S_{1,1}\theta_1^* = {} & \theta_1' S_{1,1}\theta_1 + 2\theta_1' S_{1,2}\theta_2 + \theta_2' S_{2,1} S_{1,1}^{-1} S_{1,2}\theta_2 \\ & + 2(S_{1,1}\theta_1 + S_{1,2}\theta_2)(Z_1'Z_1)^{-1} Z_1'\tilde{Z}_2\theta_2 \\ & + \theta_2'\tilde{Z}_2' Z_1(Z_1'Z_1)^{-1} S_{1,1}(Z_1'Z_1)^{-1} Z_1'\tilde{Z}_2\theta_2.\end{aligned}$$

Noting that $\theta_1' S_{1,1}\theta_1 + 2\theta_1' S_{1,2}\theta_2 + \theta_2' S_{2,1} S_{1,1}^{-1} S_{1,2}\theta_2 = \theta' S\theta - \theta_2'(S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2$, the third line in (27) equals

$$\begin{aligned}& a_1^2(\theta' S\theta - \theta_2'(S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2) + a_2^2\theta_2'(S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2 \\ & \quad + 2a_1(a_1 - a_2)(S_{1,1}\theta_1 + S_{1,2}\theta_2)(Z_1'Z_1)^{-1} Z_1'\tilde{Z}_2\theta_2 \\ & \quad + (a_2 - a_1)^2\theta_2'\tilde{Z}_2' Z_1(Z_1'Z_1)^{-1} S_{1,1}(Z_1'Z_1)^{-1} Z_1'\tilde{Z}_2\theta_2.\end{aligned}$$

Using $\hat{\theta}_1 - \theta_1 = \hat{\theta}_1^* - \theta_1^* - S_{1,1}^{-1}S_{1,2}(\hat{\theta}_2 - \theta_2) - (Z_1'Z_1)^{-1}Z_1'\tilde{Z}_2(\hat{\theta}_2 - \theta_2)$, we see that

$$\begin{aligned}(\hat{\theta} - \theta)' S\theta &= (\hat{\theta}_1 - \theta_1)'(S_{1,1}\theta_1 + S_{1,2}\theta_2) + (\hat{\theta}_2 - \theta_2)'(S_{2,1}\theta_1 + S_{2,2}\theta_2) \\ &= (\hat{\theta}_1^* - \theta_1^*)'(S_{1,1}\theta_1 + S_{1,2}\theta_2) - (\hat{\theta}_2 - \theta_2)'\tilde{Z}_2' Z_1(Z_1'Z_1)^{-1}(S_{1,1}\theta_1 + S_{1,2}\theta_2) \\ &\quad + (\hat{\theta}_2 - \theta_2)'(S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2.\end{aligned}$$

Using the formula for θ_1^* as before in the fifth and seventh line in (27), we can rewrite the true mean-squared prediction error as

$$\begin{aligned}\rho^2(m) = {} & (1 - a_2)^2(\hat{\theta} - \theta)' S(\hat{\theta} - \theta) + s^2 \\ & + (a_2 - a_1)(2 - a_1 - a_2)(\hat{\theta}_1^* - \theta_1^*)' S_{1,1}(\hat{\theta}_1^* - \theta_1^*) \\ & + (a_2 - a_1)^2\theta_2'\tilde{Z}_2' Z_1(Z_1'Z_1)^{-1} S_{1,1}(Z_1'Z_1)^{-1} Z_1'\tilde{Z}_2\theta_2 \\ & + a_1^2(\theta' S\theta - \theta_2'(S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2) + a_2^2\theta_2'(S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2 \\ & + 2a_1(a_1 - a_2)(S_{1,1}\theta_1 + S_{1,2}\theta_2)'(Z_1'Z_1)^{-1} Z_1'\tilde{Z}_2\theta_2 \\ & + 2a_1(a_1 - 1)(\hat{\theta}_1^* - \theta_1^*)'(S_{1,1}\theta_1 + S_{1,2}\theta_2) \\ & + 2(1 - a_1)(a_2 - a_1)(\hat{\theta}_1^* - \theta_1^*)' S_{1,1}(Z_1'Z_1)^{-1} Z_1'\tilde{Z}_2\theta_2 \\ & + 2a_2(a_2 - 1)(\hat{\theta}_2 - \theta_2)'(S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2 \\ & + 2a_1(1 - a_2)(\hat{\theta}_2 - \theta_2)'\tilde{Z}_2' Z_1(Z_1'Z_1)^{-1}(S_{1,1}\theta_1 + S_{1,2}\theta_2) \\ & + 2(1 - a_2)(a_1 - a_2)(\hat{\theta}_2 - \theta_2)'\tilde{Z}_2' Z_1(Z_1'Z_1)^{-1} S_{1,1}(Z_1'Z_1)^{-1} Z_1'\tilde{Z}_2\theta_2 \\ & + 2(1 - a_2)(a_1 - a_2)(\hat{\theta}_1^* - \theta_1^*)' S_{1,1}(Z_1'Z_1)^{-1} Z_1'\tilde{Z}_2(\hat{\theta}_2 - \theta_2).\end{aligned} \tag{28}$$

In the preceding display, note that the terms in line five to ten follow, conditional on X_1 or X, respectively, a centered normal distribution with a variance that is bounded in probability, and it is easy to show that these terms converge to zero in probability. The term in the last line is of the form $w'Qw$ with $\operatorname{trace}(Q) = 0$ and $w \sim N(0, s^2 I_{|m_2|})$, and it is not hard to show that this term also converges to zero (see Lemma B.3 and the subsequent results). The two results in this section gives some distributional properties of the terms involved in $\rho^2(m)$ and $\hat{\rho}^2(m)$ and motivate that those two quantities lie 'close' to each other.

For integers $k \geq 1$ and $d \geq 1$, let $\chi_k^2(\mu)$ denote a random variable that is chi-square distributed with k degrees of freedom and noncentrality parameter $\mu \geq 0$, and let $W_k(S, d)$ denote a random $k \times k$ matrix that follows a Wishart distribution with scale matrix S and d degrees of freedom. We will write χ^2_k as shorthand for $\chi^2_k(0)$. Further, $\lambda_i(\cdot)$ denotes the i-th eigenvalue of the indicated matrix. Because we only consider eigenvalues of symmetric matrices, all eigenvalues are real and we assume that they are sorted in increasing order, i.e., $\lambda_1(\cdot) \leq \ldots \leq \lambda_d(\cdot)$ if d is the dimension of the matrix.

Lemma A.1. *Let the assumptions of this section hold. Let* $\tilde{\theta}_1 = \theta_1 + S_{1,1}^{-1} S_{1,2}\theta_2$, $b = \theta' S\theta$ *and* $b_2 = \theta_2'(S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2$.

(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* $|m|$ *and* $n - |m| + 1$ *degrees of freedom, respectively, i.e.*

$$(\hat{\theta} - \theta)' S(\hat{\theta} - \theta) \sim s^2 \frac{\chi^2_{|m|}}{\chi^2_{n-|m|+1}}.$$

The term $(\hat{\theta}_1^* - \theta_1^*)' S_{1,1}(\hat{\theta}_1^* - \theta_1^*)$ *has the same distribution as the term in the preceding display with* $|m_1|$ *instead of* $|m|$.

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

$$\hat{s}^2 \sim s^2 \frac{\chi^2_{n-|m|}}{n - |m|}.$$

(iii) Let $\tilde{Z}_2 = Z_2 - Z_1 S_{1,1}^{-1} S_{1,2}$. *We have conditional on* Z_1,

$$\tilde{Z}_2\theta_2 \sim N(0, b_2 I_n). \tag{29}$$

(iv) Let $P_1 = Z_1(Z_1'Z_1)^{-1}Z_1$ *and* $M_1 = I_n - P_1$. *We have conditional on* Z,

$$Y'P_1Y \sim s^2\chi^2_{|m_1|}(\theta' Z' P_1 Z\theta / s^2), \tag{30}$$

as well as

$$Y'M_1Y \sim s^2\chi^2_{n-|m_1|}(\theta' Z' M_1 Z\theta / s^2). \tag{31}$$

If $b_2 > 0$, *we have conditional on* Z_1

$$\theta' Z' P_1 Z\theta \sim b_2\chi^2_{|m_1|}(\tilde{\theta}_1' Z_1' Z_1 \tilde{\theta}_1 / b_2). \tag{32}$$

If $b_2 = 0$, *we have*

$$\theta' Z' P_1 Z\theta = \theta_1' Z_1' Z_1 \theta_1 \sim \theta_1' S_{1,1}\theta_1 \chi^2_n. \tag{33}$$

Furthermore, we have

$$\theta' Z' M_1 Z\theta \sim b_2\chi^2_{n-|m_1|} \tag{34}$$

and

$$\tilde{\theta}_1' Z_1' Z_1 \tilde{\theta}_1 \sim (b - b_2)\chi^2_n. \tag{35}$$

If $\theta_1' S_{1,1}\theta_1 = 0$, $b_2 = 0$ *or* $b - b_2 = 0$, *respectively, the distributions should be understood as point mass at* 0.

(v) Conditional on Z_1, we have

$$Z_2' M_1 Z_2 \sim W_{|m_2|}(S_{2,2} - S_{2,1} S_{1,1}^{-1} S_{1,2}, n - |m_1|).$$

Proof. By assumption, the rows of Z are independent of each other and follow a normal distribution with mean vector zero and variance-covariance matrix S that is a symmetric, positive definite and unknown $|m| \times |m|$ matrix. By assumption Z_1 and Z_2 are matrices whose independently distributed rows follow a centered normal distribution with variance-covariance matrix $S_{1,1}$ and $S_{2,2}$, respectively. Recall that $S_{2,2} - S_{2,1} S_{1,1}^{-1} S_{1,2}$ is the Schur complement of $S_{1,1}$ in S and note that it is positive definite because S is positive definite.

Proofs for the statements in *(i)* and *(ii)* are well known but can be found in Lemma C.2 in Huber (2013).

(iii) Note that $\tilde{Z}_2$ has independent rows that follow a normal distribution with variance-covariance matrix $S_{2,2} - S_{2,1} S_{1,1}^{-1} S_{1,2}$ and that it is independent of Z_1. This fact is well known if Z_1 and Z_2 would be multivariate normal vectors. For the more general case of normal matrices, see for example Corollary 3.3.3.1 in Mardia et al. (1979).

(iv) First of all, note that P_1 and M_1 are idempotent matrices with trace$(P_1) = |m_1|$ and trace$(M_1) = n - |m_1|$. Conditional on Z, we have $Y \sim N(Z\theta, s^2 I_n)$ which shows the first and the second statement. For the statement in (32), note that $Z\theta = Z_1\tilde{\theta}_1 + \tilde{Z}_2\theta_2$ and use (29) to conclude that, conditional on Z_1, $Z\theta \sim N(Z_1\tilde{\theta}_1, b_2 I_n)$ for $b_2 > 0$. Note that $b_2 = 0$ implies that $\theta_2 = 0$ and hence that $Z\theta = Z_1\theta_1$. Noting that $Z_1\theta_1 \sim N(0, \theta_1' S_{1,1}\theta_1 I_n)$ completes the statement in (33). For the statement in (34), note that $\theta' Z' M_1 Z\theta = \theta_2' \tilde{Z}_2' M_1 \tilde{Z}_2 \theta_2$ and use the statement in (29). For the last statement, note that $Z_1\tilde{\theta}_1 \sim N(0, \tilde{\theta}_1' S_{1,1}\tilde{\theta}_1 I_n)$ and that $\tilde{\theta}_1 S_{1,1}\tilde{\theta}_1 = b - b_2$.

(v) Recall that $Z_2' M_1 Z_2 = \tilde{Z}_2' M_1 \tilde{Z}_2$ and the distribution of $\tilde{Z}_2$ discussed in the proof of (iii). The statement follows from Theorem 3.4.4 in Mardia et al. (1979).

□

Lemma A.2. *Let V be a $d \times k$, $d \geq k$, random matrix where the entries are i.i.d. standard normally distributed random variables. Then each diagonal element of $(V'V)^{-1}$ follows an inverse chi-square distribution with $d - k + 1$ degrees of freedom.*

Proof. Let $((V'V)^{-1})_{(i,j)}$ be the element in the i-th row and j-th column of $(V'V)^{-1}$. For the first diagonal element, partition the matrix V as $V = (v_1, V_{-1})$, where v_1 is the first column of V and V_{-1} are the remaining $k - 1$ columns of V, i.e., v_1 is a d-vector and V_{-1} is a $d \times (k-1)$ matrix. This entails that

$$V'V = \begin{pmatrix} v_1' v_1 & v_1' V_{-1} \\ V_{-1}' v_1 & V_{-1}' V_{-1} \end{pmatrix}.$$

The partitioned inversion rule (see, e.g., Appendix A.2.10 in Johnston and DiNardo (1997)) shows that the first diagonal element of the inverse of $V'V$ is of the form $(v_1'v_1 - v_1' V_{-1}(V_{-1}' V_{-1})^{-1} V_{-1}' v_1)^{-1}$. This can be rewritten as the inverse of $v_1' Q v_1$, where $Q = I_n - V_{-1}(V_{-1}' V_{-1})^{-1} V_{-1}'$. Note that Q is symmetric, idempotent and has rank $d - k + 1$. By assumption $v_1 \sim N(0, I_d)$ and hence $v_1' Q v_1 \| V_{-1} \sim \chi^2_{d-k+1}$ (see, e.g, Appendix B.8 in Johnston and DiNardo (1997)). Since the conditional distribution is independent of V_{-1}, the unconditional distribution coincides with the conditional distribution. This shows the statement for the first diagonal element. For the remaining diagonal elements, let $I_k^{i,j}$ be the $k \times k$ identity matrix with the i-th and the j-th column interchanged. Premultiplying a $k \times d$-matrix by $I_k^{i,j}$ interchanges the i-th and j-th row while postmultiplying a $d \times k$-matrix with $I_k^{i,j}$ interchanges the i-th and j-th column. Hence,

$$((V'V)^{-1})_{(a,a)} = (I_k^{1,a}(V'V)^{-1}I_k^{1,a})_{(1,1)} = ((I_k^{1,a}V'VI_k^{1,a})^{-1})_{(1,1)}$$

and the result follows from the first part of the proof. □

B Proof of Theorem 1

In this section, we first give auxiliary results that are needed for proofing Theorem 1. The proof of the theorem is given at the end of this section.

Lemma B.1. *For a fixed integer $d \geq 1$, let $v \sim N(0, I_d)$, and let A be a symmetric, positive semidefinite and nonrandom $d \times d$ matrix. Then, we have for any $\varepsilon > 0$ that*

$$\mathbb{P}(v'Av - \operatorname{trace}(A) \geq \varepsilon) \leq \begin{cases} e^{-\frac{d}{2}G(d\lambda_d(A),\varepsilon)} & \text{if } \lambda_d(A) > 0 \\ 0 & \text{else} \end{cases} \tag{36}$$

as well as

$$\mathbb{P}(v'Av - \operatorname{trace}(A) \leq -\varepsilon) \leq \begin{cases} e^{-\frac{d}{2}G(d\lambda_d(A),-\varepsilon)} & \text{if } \varepsilon < \operatorname{trace}(A) \\ 0 & \text{else,} \end{cases} \tag{37}$$

where the function $G : (0,\infty) \times (-\operatorname{trace}(A), \infty) \to \mathbb{R}$ is given by $G(x,y) = y/x - \log((x+y)/x)$.

Proof. Throughout the proof, we denote the eigenvalues of A shorthand as $\lambda_1 \leq \ldots \leq \lambda_d$. Let $U\Lambda U'$ be the eigenvalue decomposition of A, where U is the matrix whose columns are the normed eigenvectors of A and where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$. If $\lambda_d = 0$, then A is the zero matrix and the statements of the lemma are trivially fulfilled. Hence, from now on we assume that $\lambda_d > 0$. Because U is orthonormal, $U'w$ has the same distribution as w, i.e., $U'w \sim N(0, I_d)$. The left-hand side in (36) can be rewritten as

$$\begin{aligned} \mathbb{P}\left(w'\Lambda w - \operatorname{trace}(\Lambda) \geq \varepsilon\right) &= \mathbb{P}\left(\sum_{i=1}^d \lambda_i w_i^2 - \sum_{i=1}^d \lambda_i \geq \varepsilon\right) \\ &= \mathbb{P}\left(\mathrm{e}^{\sum_{i=1}^d s\lambda_i w_i^2} \geq \mathrm{e}^{s\left(\varepsilon + \sum_{i=1}^d \lambda_i\right)}\right), \end{aligned} \tag{38}$$

where $s > 0$ is arbitrary. Use Markov's inequality to bound the last term in the preceding display from above by

$$\mathrm{e}^{-s\left(\varepsilon+\sum_{i=1}^d \lambda_i\right)}\mathbb{E}\left[\mathrm{e}^{\sum_{i=1}^d s\lambda_i w_i^2}\right] = \mathrm{e}^{-s\left(\varepsilon+\sum_{i=1}^d \lambda_i\right)}\prod_{i=1}^d \mathbb{E}\left[\mathrm{e}^{s\lambda_i w_i^2}\right].$$

Note that for every i, $\mathbb{E}[\exp(s\lambda_i w_i^2)]$ is the moment generating function of the chi-square distribution with one degree of freedom. The moment generating function is finite for

$s\lambda_i < 1/2$ and equals $(1-2s\lambda_i)^{-1/2} = \exp(-1/2\log(1-2s\lambda_i))$. Hence, let $s>0$ be such that $s\lambda_i < 1/2$ for all $1 \leq i \leq d$ or, equivalently, such that $0 < s < 1/(2\lambda_d)$. Then, the last term in the preceding display equals

$$\exp\left(-\frac{1}{2}\sum_{i=1}^{d}\left[2s(\varepsilon/d+\lambda_i)+\log(1-2s\lambda_i)\right]\right) \tag{39}$$

$$\leq \exp\left(-\frac{d}{2}\left[2s(\varepsilon/d+\lambda_d)+\log(1-2s\lambda_d)\right]\right), \tag{40}$$

where the inequality follows upon noting that the function in squared brackets in (39) is nonincreasing in λ_i for all i. (Indeed, note that the first derivative with respect to λ_i equals $2s - 2s/(1-2s\lambda_i)$. The last term is nonpositive because $0 < 1-2\lambda_i s \leq 1$ as noted in the paragraph preceding the last display.) For choosing the optimal s in (40), we need to maximize the function $f_{\lambda_d,\varepsilon}(s) = 2s(\varepsilon/d+\lambda_d)+\log(1-2s\lambda_d)$. The first derivative with respect to s equals $2(\varepsilon/d+\lambda_d) - 2\lambda_d/(1-2s\lambda_d)$. Setting the derivative equal zero and rearranging gives $1-2s^*\lambda_d = \lambda_d/(\varepsilon/d+\lambda_d)$ which is equivalent to

$$s^* = \frac{1}{2\lambda_d} - \frac{1}{2(\varepsilon/d+\lambda_d)} = \frac{\varepsilon/d}{2\lambda_d(\varepsilon/d+\lambda_d)}.$$

Noting that $0 < s^* < 1/(2\lambda_d)$ and that $\partial^2 f_{\lambda_d,\varepsilon}(s)/\partial s^2 = -4\lambda_d^2/(1-2s\lambda_d)^2$, we see that indeed s^* optimizes the upper bound in (40). This shows the statement in (36). For the lower tail bound in (37), use the same arguments as before to show that

$$\mathbb{P}\left(v'Av - \operatorname{trace}(A) \leq -\varepsilon\right) = \mathbb{P}\left(\sum_{i=1}^{d}\lambda_i v_i^2 \leq \sum_{i=1}^{d}\lambda_i - \varepsilon\right).$$

The term on the left-hand side in the preceding display equals zero if $\sum_{i=1}^{d}\lambda_i - \varepsilon = \operatorname{trace}(A) - \varepsilon \leq 0$. Hence, the second statement in (37) is fulfilled. Thus, consider the case where $\sum_{i=1}^{d}\lambda_i - \varepsilon = \operatorname{trace}(A) - \varepsilon > 0$. Let $s>0$ be arbitrary and use Markov's inequality to show that the last term in the preceding display can be rewritten as and bounded from above by

$$\begin{aligned}\mathbb{P}\left(\mathrm{e}^{-s\sum_{i=1}^{d}\lambda_i v_i^2} \geq \mathrm{e}^{-s\left(-\varepsilon+\sum_{i=1}^{d}\lambda_i\right)}\right) &\leq \mathrm{e}^{s\left(-\varepsilon+\sum_{i=1}^{d}\lambda_i\right)}\mathbb{E}\left[\mathrm{e}^{-\sum_{i=1}^{d}s\lambda_i v_i^2}\right]\\ &= \mathrm{e}^{s\left(\sum_{i=1}^{d}(\lambda_i-\varepsilon/d)\right)}\prod_{i=1}^{d}\mathbb{E}\left[\mathrm{e}^{-s\lambda_i v_i^2}\right].\end{aligned}$$

As before $\mathbb{E}[\exp(-s\lambda_i\chi_1^2)]$ is the moment generating function of the chi-square distribution with one degree of freedom which is finite for $-s\lambda_i < 1/2$ (which is fulfilled because $s>0$ and $\lambda_i \geq 0$ for all $i=1,\ldots,n$) and equals $(1+2s\lambda_i)^{-1/2} = \exp(-1/2\log(1+2s\lambda_i))$. We can rewrite the last term in the preceding display as

$$\exp\left(-\frac{1}{2}\sum_{i=1}^{d}\left[2s(\varepsilon/d-\lambda_i)+\log(1+2s\lambda_i)\right]\right) \tag{41}$$

$$\leq \exp\left(-\frac{d}{2}\left[2s(\varepsilon/d-\lambda_d)+\log(1+2s\lambda_d)\right]\right), \tag{42}$$

where the inequality follows upon noting that the function in squared brackets in (41) is nonincreasing in λ_i. (Note that the first derivative with respect to λ_i equals $-2s+$

$2s/(1+2s\lambda_i)$. This is nonpositive as $1 \leq 1+2s\lambda_i$.) In order to choose the optimal $s > 0$, we need to maximize the function $f_{-\lambda_d,\varepsilon}(s)$, where f was defined before. Setting the first derivative with respect to s equal zero and solving for s gives by the same arguments used in the derivation of the upper tail bound

$$s^{**} = \frac{\varepsilon/d}{2\lambda_d(\lambda_d - \varepsilon/d)}.$$

Noting that $\lambda_d - \varepsilon/d$ is positive because $d\lambda_d \geq \sum_{i=1}^d \lambda_i > \varepsilon$ here, we see that $s^{**} > 0$. The second derivative of $f_{-\lambda_d,\varepsilon}(s)$ with respect to s equals $-4\lambda_d^2/(1+2\lambda_d s)^2$ which shows that $s^{**} > 0$ is a maximizer and ends the proof. □

Corollary B.2. *Under the assumptions of Lemma B.1, we have for each $\varepsilon > 0$*

$$\mathbb{P}\left(|v'Av - \operatorname{trace}(A)| \geq \varepsilon\right) \leq \begin{cases} 2\exp\left(-d\frac{\varepsilon^2}{4d\lambda_d(A)(\varepsilon + d\lambda_d(A))}\right) & \textit{if } \lambda_d(A) > 0 \\ 0 & \textit{else.} \end{cases} \tag{43}$$

Proof. The statement follows from Lemma B.1 by noting that $G(d\lambda_d(A), \varepsilon)$ equals $M(0, \varepsilon/(d\lambda_d(A)))$, where $M(\cdot,\cdot)$ is defined in Lemma C.3 in Huber (2013), together with Lemma C.4 in Huber (2013). The later lemma shows that $M(x,-y) \geq M(x,y)$ for $x \geq 0$ and $0 \leq y < 1$ and that $M(0,y) \geq y^2/(2(y+1))$ for $y \geq 0$. Hence, we have for $\varepsilon < \operatorname{trace}(A)$ that $G(d\lambda_d(A), -\varepsilon) = M(0, -\varepsilon/(d\lambda_d(A))) \geq M(0, \varepsilon/(d\lambda_d(A))) = G(d\lambda_d(A), \varepsilon)$ because $\varepsilon/(d\lambda_d(A)) \leq \varepsilon/\operatorname{trace}(A) < 1$ here, and that $M(0, \varepsilon/(d\lambda_d(A))) \geq \varepsilon^2/(2d\lambda_d(A)(\varepsilon + d\lambda_d(A)))$. □

Lemma B.3. *Let $v \sim N(0, I_d)$, and let A be a symmetric $d \times d$ matrix with* $\operatorname{trace}(A) = 0$. *Then for each $\varepsilon > 0$, we have*

$$\begin{aligned} &\mathbb{P}\left(|v'Av| \geq \varepsilon\right) \\ &\quad\leq \begin{cases} 2\exp\left(-\frac{d}{2}G(d\lambda_d(A), \varepsilon/2)\right) + 2\exp\left(-\frac{d}{2}G(-d\lambda_1(A), \varepsilon/2)\right) & \textit{if } \lambda_d(A) > 0 \\ 0 & \textit{else,} \end{cases} \end{aligned}$$

where $G(\cdot,\cdot)$ is defined in Lemma B.1.

Proof. Because $\operatorname{trace}(A) = \sum_{i=1}^d \lambda_i(A) = 0$, we either have $\lambda_d(A) = 0$ or $\lambda_d(A) > 0$. If $\lambda_d(A) = 0$, all eigenvalues are zero which implies that A is the zero matrix (because it is symmetric) and the statement holds trivially. From now on, we assume that $\lambda_d(A) > 0$ implying that $\lambda_1(A) < 0$. Let $U\Lambda U'$ be the eigenvalue decomposition of A. Let Λ^+ be the matrix Λ where all negative entries are substituted by 0, and let Λ^- be the matrix Λ where all nonnegative entries are substituted by 0. Note that $\Lambda = \Lambda^+ + \Lambda^-$. Hence, we have

$$v'Av - \operatorname{trace}(A) = v'U\Lambda^+U'v - \operatorname{trace}(\Lambda^+) + v'U\Lambda^-U'v - \operatorname{trace}(\Lambda^-).$$

Because U is orthonormal, we see that $U'v$ has the same distribution as v, i.e., $U'v \sim N(0, I_d)$. The left-hand side of the statement can then be rewritten as

$$\begin{aligned} &\mathbb{P}\left(|v'Av - \operatorname{trace}(A)| \geq \varepsilon\right) \\ &\quad\leq \mathbb{P}\left(\left|v'\Lambda^+v - \operatorname{trace}(\Lambda^+)\right| \geq \varepsilon/2\right) + \mathbb{P}\left(\left|v'\Lambda^-v - \operatorname{trace}(\Lambda^-)\right| \geq \varepsilon/2\right) \\ &\quad= \mathbb{P}\left(\left|v'\Lambda^+v - \operatorname{trace}(\Lambda^+)\right| \geq \varepsilon/2\right) + \mathbb{P}\left(\left|v'(-\Lambda^-)v - \operatorname{trace}(-\Lambda^-)\right| \geq \varepsilon/2\right). \end{aligned}$$

Because Λ^+ as well as $-\Lambda^-$ are positive semidefinite by construction and because $\lambda_d(\Lambda^+) = \lambda_d(A) > 0$ and $\lambda_d(-\Lambda^-) = -\lambda_1(A) > 0$, the result follows from Lemma B.1. □

Lemma B.4. *Let $v \sim N(0, I_d)$, and let A be a $d \times d$ matrix such that* $\operatorname{trace}(A) = 0$ *and $AA = \mathbf{0}$, i.e., the $d \times d$ matrix consisting of only zeros. Then, we have for every $\varepsilon > 0$*

$$\mathbb{P}\left(|v'Av| \geq \varepsilon\right) \leq \begin{cases} 4\exp\left(-\frac{d}{2}G(d\sqrt{\lambda_d(A'A)/2}, \varepsilon/2)\right) & \textit{if } \lambda_d(A + A') > 0 \\ 0 & \textit{else} \end{cases}$$

where $G(\cdot, \cdot)$ is defined in Lemma B.1.

Proof. Let A' be the transpose of A and set $B = (A + A')/2$. Then, B is symmetric, $\operatorname{trace}(B) = 0$ and $v'Av = w'Bw$. Note that $\lambda_i(B) = \lambda_i(A + A')/2$ for all $i = 1, \ldots, d$. If $\lambda_d(B) = 0$, then all eigenvalues of B are 0 (because $\operatorname{trace}(B) = 0$) implying that B is the zero matrix (because B is symmetric by construction). Noting that $\lambda_d(B) = 0$ if and only if $\lambda_d(A + A') = 0$, the second line of the statement is trivially fulfilled. If $\lambda_d(B) = \lambda_d(A + A')/2 > 0$, we can use Lemma B.3 to bound the quantity of interest from above by

$$2\exp\left(-\frac{d}{2}G(d\lambda_d(B), \varepsilon/2)\right) + 2\exp\left(-\frac{d}{2}G(-d\lambda_1(B), \varepsilon/2)\right).$$

Note that $(\lambda_i(A + A'))^2 = \lambda_j((A + A')^2)$ for some i and j, that $(A + A')^2 = A'A + AA'$ by assumption and that $A'A$ and AA' have the same eigenvalues. Hence $0 < \lambda_d(B) \leq 1/2\sqrt{\lambda_d(A'A + AA')} \leq 1/2\sqrt{2\lambda_d(A'A)}$ by Weyl's inequality. Using the same arguments as for $\lambda_d(B)$, we have $0 < -\lambda_1(B) \leq 1/2\sqrt{2\lambda_d(A'A)}$. The result follows from the fact that the function $G(\cdot, \cdot)$ is nonincreasing in its first argument. □

Corollary B.5. *Under the assumptions of Lemma B.4, we have for each $\varepsilon > 0$*

$$\mathbb{P}\left(|v'Av| \geq \varepsilon\right) \leq \begin{cases} 4\exp\left(-d\frac{\varepsilon^2}{4d\sqrt{2\lambda_d(A'A)}(\varepsilon + d\sqrt{2\lambda_d(A'A)})}\right) & \textit{if } \lambda_d(A + A') > 0 \\ 0 & \textit{else.} \end{cases}$$

Proof. The statement follows from Lemma B.4 by the same arguments as Corollary B.2 follows from Lemma B.1. □

The next two results are standard. We state it only for the sake of completeness.

Lemma B.6. *Let $w \sim N(0, \tau^2)$ with $\tau^2 > 0$. Then, we have for every $\varepsilon > 0$*

$$\mathbb{P}\left(|w| \geq \varepsilon\right) \leq 2\exp\left(-\varepsilon^2/(2\tau^2)\right).$$

Lemma B.7. *We have for every $k \in \mathbb{N}$ and every $\varepsilon > 0$*

$$\mathbb{P}\left(\chi_k^2/k - 1 \geq \varepsilon\right) \leq \exp\left(-k\varepsilon^2/(4(\varepsilon + 1))\right)$$

as well as

$$\mathbb{P}\left(\left|\chi_k^2/k - 1\right| \geq \varepsilon\right) \leq 2\exp\left(-k\varepsilon^2/(4(\varepsilon + 1))\right).$$

Lemma B.8. *Fix integers $d \geq k \geq 1$. Let $\chi_k^2(B_d)$ and $B_d > 0$ be real random variables such that $\chi_k^2(B_d)$ follows conditional on B_d a noncentral chi-square distribution with k degrees of freedom and noncentrality parameter B_d. Then for any $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\left|\frac{\chi_k^2(B_d)}{d} - \frac{k + B_d}{d}\right| \geq \varepsilon\right) \leq 2\mathbb{E}\left[\exp\left(-d\frac{\varepsilon^2}{4(\varepsilon + 2(k/d + B_d/d))}\right)\right] \tag{44}$$

as well as

$$\mathbb{P}\left(\frac{\chi_k^2(B_d)}{d} - \frac{k+B_d}{d} \geq \varepsilon\right) \leq \mathbb{E}\left[\exp\left(-d\frac{\varepsilon^2}{4(\varepsilon + 2(k/d + B_d/d))}\right)\right]. \tag{45}$$

Proof. Rewrite the left-hand side in (44) as

$$\mathbb{E}\left[\mathbb{P}\left(\left|\frac{\chi_k^2(B_d)}{k+B_d} - 1\right| \geq \varepsilon\frac{d}{k+B_d}\middle\| B_d\right)\right]$$

and use the Chernoff bound for the noncentral chi-square distribution as in Corollary C.5 in Huber (2013), conditional on B_d, to bound the term in the preceding display from above by

$$2\mathbb{E}\left[\exp\left(-k\frac{\varepsilon^2 d^2/(k(k+B_d))}{4\left(\varepsilon d/(k+B_d)+2\right)}\right)\right] = 2\mathbb{E}\left[\exp\left(-d\frac{\varepsilon^2 d}{4(\varepsilon d + 2(k+B_d))}\right)\right]. \tag{46}$$

The proof of the second statement is the same but using the one-sided version of the tail bound of the noncentral chi-square distribution. □

Lemma B.9. *Let W follow a Wishart distribution with scale matrix I_k and $d \geq k \geq 2$ degrees of freedom, i.e., $W \sim W_k(I_k, d)$. Then, we have for every $\gamma_1 \in [0,1)$*

$$\mathbb{P}\left(\lambda_1(W/d) \leq \gamma_1^2(1-\sqrt{k/d})^2\right) \leq \exp\left(-d(1-\gamma_1)^2(1-\sqrt{k/d})^2/2\right).$$

Furthermore, we have for every $\varepsilon > 0$

$$\mathbb{P}\left(\lambda_d(W/d) \geq (1+\sqrt{k/d}+\varepsilon)^2\right) \leq \exp\left(-d\varepsilon^2/2\right).$$

Especially, we have for every $\gamma_2 > 0$

$$\mathbb{P}\left(\lambda_d(W/d) \geq (1+\gamma_2)^2(1+\sqrt{k/d})^2\right) \leq \exp\left(-d\gamma_2^2(1+\sqrt{k/d})/2\right).$$

Proof. The lemma follows immediately from Theorem 2.13 in Davidson and Szarek (2001). A detailed proof of the first statement can be found in Huber (2013) (see Lemma C.11, Corollary C.12 and the following remark). □

Lemma B.10. *Let V be a random $d \times k$ matrix, $d \geq k$, with i.i.d. standard normally distributed entries. Then for $\varepsilon > 0$, we have*

$$\mathbb{P}\left(\left|\frac{\operatorname{trace}\left((V'V)^{-1}\right)}{k/(d-k+1)} - 1\right| \geq \varepsilon\right) \leq 2k\exp\left(-(d-k)\frac{\varepsilon^2}{8(\varepsilon+1)^2}\right), \tag{47}$$

as well as

$$\mathbb{P}\left(\left|\operatorname{trace}\left((V'V)^{-1}\right) - \frac{k}{d-k+1}\right| \geq \varepsilon\right) \leq 2k\exp\left(-(d-k)\frac{\varepsilon^2(1-k/d)^2}{8(\varepsilon(1-k/d)+1)^2}\right), \tag{48}$$

where $\operatorname{trace}((V'V)^{-1})$ *is to be interpreted as 1 if $V'V$ is not invertible.*

Proof. It suffices to consider only the case where $V'V$ has full rank as this happens with probability 1. Lemma A.2 shows that the trace of $(V'V)^{-1}$ is the sum of k random variables that are not independent and where each follows an inverse χ^2_{d-k+1}-distribution. Hence, the left-hand side in (47) equals

$$\mathbb{P}\left(\left|\sum_{i=1}^{k}\left(\frac{d-k+1}{\chi^2_{d-k+1}}-1\right)\right| \geq \varepsilon k\right)$$

Use the triangle inequality to bound the sum in the preceding display from above by

$$k\mathbb{P}\left(\left|\frac{d-k+1}{\chi^2_{d-k+1}}-1\right| \geq \varepsilon\right) = k\mathbb{P}\left(\frac{d-k+1}{\chi^2_{d-k+1}} \geq 1+\varepsilon\right) + k\mathbb{P}\left(\frac{d-k+1}{\chi^2_{d-k+1}} \leq 1-\varepsilon\right).$$

Note that the term on the far right-hand side equals zero if $\varepsilon \geq 1$. The sum on the right hand-side in the preceding display can be rewritten as

$$k\mathbb{P}\left(\frac{\chi^2_{d-k+1}}{d-k+1}-1 \leq -\frac{\varepsilon}{1+\varepsilon}\right) + k\mathbb{P}\left(\frac{\chi^2_{d-k+1}}{d-k+1}-1 \geq \frac{\varepsilon}{1-\varepsilon}\right),$$

where the second term is to be interpreted as zero if $\varepsilon \geq 1$. Noting that $\varepsilon/(1-\varepsilon) \geq \varepsilon/(1+\varepsilon)$ for $\varepsilon \in (0,1)$, we can bound the sum in the preceding display from above by

$$k\mathbb{P}\left(\left|\frac{\chi^2_{d-k+1}}{d-k+1}-1\right| > \frac{\varepsilon}{1+\varepsilon}\right).$$

Use the tail bound for the chi-square distribution as in Lemma B.7 to bound the term in the preceding display from above by

$$2k \exp\left(-(d-k+1)\frac{\varepsilon^2}{4(1+\varepsilon)(2\varepsilon+1)}\right),$$

which is further bounded from above by the right-hand side in (47). The second statement follows from the first statement by rewriting it as

$$\mathbb{P}\left(\left|\frac{\operatorname{trace}\left((V'V)^{-1}\right)}{k/(d-k+1)}-1\right| \geq \varepsilon\frac{d-k+1}{k}\right)$$

and noting that $\varepsilon(d-k+1)/k \geq \varepsilon(1-k/d)$. □

The next result is rather technical and is separated from the proof of the theorem for the better readability.

Lemma B.11. *(i) Let $x_1, x_2 \in [0,1]$ and $y_1, y_2 \geq 0$ and let $k_1, k_2, k, d \in \mathbb{N}$ such that $k_1 + k_2 = k < d$. Let*

$$q = (1-x_2)^2\frac{k}{d-k+1} + (x_2-x_1)(2-x_1-x_2)\frac{k_1}{d-k_1+1} + 1 + x_1^2y_1 + x_2^2y_2 \\ + (x_1-x_2)^2\frac{k_1}{d-k_1+1}y_2.$$

Then, we have the following inequalities:

$$\left|(1-x_2)^2\frac{k}{d-k+1}+(x_2-x_1)(2-x_1-x_2)\frac{k_1}{d-k_1+1}\right.$$
$$\left.+1-x_2^2-(x_1-x_2)^2\frac{k_1}{d-k_1+1}\right|/q\le\frac{1}{1-k/d} \tag{49}$$
$$x_1^2/q\le\frac{1}{1+y_1+y_2k_1/d(1-k_1/d)}\le 1 \tag{50}$$
$$\left|x_2^2-x_1^2\frac{k_1}{d}+(x_1-x_2)^2\frac{k_1}{d-k_1+1}\right|/q\le\frac{1}{(1-k_1/d)(1+y_2)}\le\frac{1}{1-k_1/d} \tag{51}$$

(ii) For $x\in[0,1]$, the following inequalities hold true:

$$2(1-x)^5/9\le(1-\sqrt{x})^4 \tag{52}$$
$$3(1-x)^5/4\le(1-\sqrt{x})^2. \tag{53}$$

For $x,y\in(0,1)$ such that $x+y\le 1$, we have the following inequalities:

$$3(1-x-y)^5/4\le(1-\sqrt{y/(1-x)})^2(1-x), \tag{54}$$
$$(1-x-y)^5/2\le(1-\sqrt{y/(1-x)})^2(1-\sqrt{x})^2(1-x), \tag{55}$$
$$8(1-x-y)^5/45\le(1-\sqrt{y/(1-x)})^2(1-\sqrt{x})^4, \tag{56}$$
$$8(1-x-y)^5/165\le(1-\sqrt{y/(1-x)})^2(1-\sqrt{x})^4(1-x), \tag{57}$$
$$3(1-x-y)^5/4\le(1-\sqrt{y/(1-x)})^2. \tag{58}$$

Proof. (i) We will show that the left hand-side in (49) is bounded from above by $d/(d-k+1)$ which implies the statement. Bound q from below by

$$q_1=(1-x_2)^2\frac{k}{d-k+1}+(x_2-x_1)(2-x_1-x_2)\frac{k_1}{d-k_1+1}+1.$$

We need to show that the term in absolute value on the left hand-side in (49) multiplied by $(d-k+1)/d$ is bounded from above by q_1. Hence, we have to show that

$$\left((1-x_2)^2\frac{k}{d-k+1}+(x_2-x_1)(2-x_1-x_2)\frac{k_1}{d-k_1+1}\right)\left(1-\frac{d-k+1}{d}\right)$$
$$+1-\frac{d-k+1}{d}+x_2^2\frac{d-k+1}{d}+(x_1-x_2)^2\frac{k_1}{d}\frac{d-k+1}{d-k_1+1}$$

as well as

$$(1-x_2)^2\frac{k}{d-k+1}+(x_2-x_1)(2-x_1-x_2)\frac{k_1}{d-k_1+1}$$
$$+(1-x_2)^2\frac{k}{d}+(x_2-x_1)(2-x_1-x_2)\frac{k_1}{d}\frac{d-k+1}{d-k_1+1}$$
$$+1-(x_1-x_2)^2\frac{k_1}{d}\frac{d-k+1}{d-k_1+1}+(1-x_2^2)\frac{d-k+1}{d}$$

is nonnegative. The term in the second-to-last display is nonnegative because $k/(d-k+1)\ge k_1/(d-k_1+1)$, $(1-x_2)^2+(x_2-x_1)(2-x_1-x_2)=(1-x_1)^2\ge 0$ and

$1-(d-k+1)/d \geq 0$. To show that the term in the preceding display is nonnegative, use $k/(d-k+1) \geq k_1/(d-k_1+1)(1-k_2/d)$ for the first term in the first line, that $k/d \geq k_1/d$ for the first term in the second line and that $(d-k+1)/(d-k_1+1) = 1-k_2/(d-k_1+1)$ for the second term in the second line. Rearranging the terms, it is enough to show that

$$\begin{aligned}
&\frac{k_1}{d-k_1+1}\left(1-\frac{k_2}{d}\right)\left((1-x_2)^2+(x_2-x_1)(2-x_1-x_2)\right)\\
&+\frac{k_1}{d}\left((1-x_2)^2+(x_2-x_1)(2-x_1-x_2)\right)+(1-x_2^2)\frac{d-k+1}{d}\\
&+1-(x_1-x_2)^2\frac{k_1}{d}\frac{d-k+1}{d-k_1+1}
\end{aligned}$$

is nonegative. The inequality in the third line in the preceding display holds because $k_1/d \leq 1$, $(d-k+1)/(d-k_1+1) \leq 1$ and $x_1, x_2 \in [0,1]$.
To show the inequality in (50), note that q is bounded from below by $1+x_1^2y_1+x_2^2y_2+(x_1-x_2)^2y_2k_1/d$ because $k/(d-k+1) \geq k_1/(d-k_1+1) \geq k_1/d$ and $(1-x_2)^2+(x_2-x_1)(2-x_1-x_2) = (1-x_1)^2 \geq 0$. Hence, we need to show that

$$1-x_1^2+y_2\left[x_2^2-x_1^2\frac{k_1}{d}\left(1-\frac{k_1}{d}\right)+(x_1-x_2)^2\frac{k_1}{d}\right] \geq 0.$$

Rewriting the sum in squared brackets as $x_2^2k_1/d+(x_2-x_1k_1/d)^2$ shows the statement.
The second inequality in (51) is clear. To show the first inequality in (51), we will show that the term on the far left-hand side is bounded from above by $1/(1-k_1/d+y_2)$ which implies the first inequality. The second inequality is immediate by recalling that $y_2 \geq 0$. Note that q is bounded from below by

$$q_2 = (1-x_1)^2\frac{k_1}{d-k_1+1}+1+x_2^2y_2+(x_1-x_2)^2\frac{k_1}{d-k_1+1}y_2$$

because $k/(d-k+1) \geq k_1/(d-k_1+1) \geq 0$. Thus, it suffices to show that

$$-q_2 \leq \left(x_2^2-x_1^2\frac{k_1}{d}+(x_1-x_2)^2\frac{k_1}{d-k_1+1}\right)\left(1-\frac{k_1}{d}+y_2\right) \leq q_2$$

or, equivalently, that

$$\begin{aligned}
&(1-x_1)^2\frac{k_1}{d-k_1+1}+\left(x_2^2+(x_1-x_2)^2\frac{k_1}{d-k_1+1}\right)\left(1-\frac{k_1}{d}\right)\\
&+1-x_1^2\frac{k_1}{d}\left(1-\frac{k_1}{d}\right)+y_2\left[2x_2^2+2(x_1-x_2)^2\frac{k_1}{d-k_1+1}-x_1^2\frac{k_1}{d}\right] \geq 0
\end{aligned} \tag{59}$$

as well as

$$\begin{aligned}
1-x_2^2\left(1-\frac{k_1}{d}\right)+x_1^2\frac{k_1}{d}\left(1-\frac{k_1}{d}\right)-(x_1-x_2)^2\frac{k_1}{d-k_1+1}\left(1-\frac{k_1}{d}\right)&\\
+(1-x_1)^2\frac{k_1}{d-k_1+1}+x_1^2\frac{k_1}{d}y_2 \geq 0&
\end{aligned} \tag{60}$$

holds. To show the inequality in (59), note that the sum of the first four terms is nonnegative because $x_1, k_1/d \in [0,1]$. Hence, we are left with showing that the sum of the terms in squared brackets is nonnegative. Using that $k_1/(d-k_1+1) \geq k_1/d$ and that $1 \geq k_1/d$, it is enough to show that $2x_2^2 + 2(x_1 - x_2)^2 - x_1^2 \geq 0$. But the left hand-side of the preceding inequality equals $4x_2^2 - 4x_1x_2 + x_1^2 = (2x_2 - x_1)^2 \geq 0$. To show the inequality in (60), note that the sum in the second line is nonnegative. Using $(1 - k_1/d)k_1/(d - k_1 + 1) \leq k_1/d$, we can bound the sum in the first line from below by $1 - x_2^2 - x_1^2(k_1/d)^2 + 2x_1x_2k_1/d = 1 - (x_1k_1/d - x_2)^2$. But this is nonegative because $x_1, x_2, k_1/d \in [0, 1]$.

(ii) Define $f_{a,b}(s,t) = (s - \sqrt{t})^a(s + \sqrt{t})^b$ for $a, b \in \mathbb{N}$ with $a < b$ and $s, t \in \mathbb{R}$ with $s > 0$ and $0 < t \leq s^2$. Note that the function is nondecreasing in s and that

$$\frac{\partial f_{a,b}(s,t)}{\partial t} = \frac{(s - \sqrt{t})^{a-1}(s + \sqrt{t})^{b-1}}{2\sqrt{t}}(s(b - a) - \sqrt{t}(a + b))$$

and that

$$\begin{aligned}\frac{\partial^2 f_{a,b}(s,t)}{\partial t^2} = \frac{(s - \sqrt{t})^{a-2}(s + \sqrt{t})^{b-2}}{4t\sqrt{t}} \Big[& s^3(a - b) + s^2\sqrt{t}(a - b)^2 \\ & + st(a - b)(2a + 2b - 3) + t\sqrt{t}(a + b)(a + b - 2)\Big].\end{aligned}$$

The functions $f_{a,b}(s,t)$ have an extremum at $t^* = s^2(b - a)^2/(a + b)^2$ with the value

$$f_{a,b}(s, t^*) = a^a b^b (2s/(a + b))^{a+b}. \tag{61}$$

This a maximizer because the term inside the squared brackets in the second-to-last display evaluated at t^* equals $-4ab(b - a)s^3/(a + b)^2$ which is negative because $a < b$ by assumption.

The inequalities in (52) and (53) are equivalent to $9/2 \geq (1 - \sqrt{x})(1 + \sqrt{x})^5 = f_{1,5}(1, x)$ and to $4/3 \geq (1 - \sqrt{x})^3(1 + \sqrt{x})^5 = f_{3,5}(1, x)$. Using the formula in (61), we see that $f_{1,5}(1, x) \leq 5^5/3^6$ and that $f_{3,5}(1, x) \leq 3^3 5^5/4^8$ which shows the statement.

In order to show the inequalities in (54), (55), (56) and (57), it is enough to show that $(\sqrt{1 - x} - \sqrt{y})^3(\sqrt{1 - x} + \sqrt{y})^5 = f_{3,5}(\sqrt{1 - x}, y)$ is bounded from above by $4/3$, $2(1 - \sqrt{x})^2$, $45(1 - \sqrt{x})^4/(8(1 - x))$ and by $165(1 - \sqrt{x})^4/8$, respectively. Note that by (61) we have $f_{3,5}(\sqrt{1 - x}, y) \leq 3^3 5^5(1 - x)^4/4^8$. For the inequality in (54), we need to show that $3^3 5^5(1 - x)^4/4^8 \leq 4/3$ which holds because $x \in [0, 1)$. For (55), we are left with showing that $2^{17}/(3^3 5^5) \geq (1 - \sqrt{x})^2(1 + \sqrt{x})^4 = f_{2,4}(1, x)$. The right hand-side of the preceding inequality is bounded from above by $2^{10}/3^6$ by (61) which shows the statement. For (56), it suffices to show that $2^{13}/(3 \cdot 5^4) \geq (1 - \sqrt{x})(1 + \sqrt{x})^5 = f_{1,5}(1, x)$. By (61), $f_{1,5}(1, x)$ is bounded from above by $5^5/3^6$. For inequality in (57), we have to show that $(1 + \sqrt{x})^4 \leq 11 \cdot 2^{13}/(3^2 5^4)$ which is true because $(1 + \sqrt{x})^4 \leq 2^4$ by assumption.

The inequality in (58) follows immediately from (54) because the right-hand side in (58) is bounded from below by the right hand-side in (54).

□

The next two results collect the convergence rates of the main terms in the proof of Theorem 1. In both results, we look at one fixed model m.

Lemma B.12. *For each $\delta > 0$, we have*

$$\mathbb{P}(\hat{\rho}^2(m)/r > \exp(\delta)) \leq 22 \exp\left(-|m_1|\left(1 - \frac{|m|}{n}\right)^3 \frac{(\exp(\delta) - 1)^2}{210 \exp(\delta)}\right) \tag{62}$$

as well as

$$\mathbb{P}(\hat{\rho}^2(m)/r < \exp(-\delta)) \leq 22 \exp\left(-|m_1|\left(1-\frac{|m|}{n}\right)^3 \frac{(\exp(\delta)-1)^2}{840\exp(2\delta)}\right), \tag{63}$$

where

$$\begin{aligned} r =& s^2\Bigg((1-a_2)^2\frac{|m|}{n-|m|+1} + (a_2-a_1)(2-a_1-a_2)\frac{|m_1|}{n-|m_1|+1} + 1 \\ &+ a_1^2(\mu-\mu_2) + a_2^2\mu_2 + (a_1-a_2)^2\frac{|m_1|}{n-|m_1|+1}\mu_2\Bigg), \end{aligned} \tag{64}$$

where a_1 and a_2 are the shrinkage factors as in (23) *and where $\mu = \theta' S\theta/s^2$ and $\mu_2 = \theta_2'(S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2/s^2$.*

Proof. Let $P_1 = Z_1(Z_1'Z_1)^{-1}Z_1$ and $M_1 = I_n - P_1$ and rewrite $\hat{\rho}^2(m)/r$ as

$$\begin{aligned} & r_1\left(\frac{\hat{s}^2}{s^2}-1\right) + r_2\left(\frac{Y'P_1Y/s^2}{n} - \frac{|m_1|}{n} - \mu_2\frac{|m_1|}{n} - (\mu-\mu_2)\right) \\ &+ r_3\left(\frac{Y'M_1Y/s^2}{n-|m_1|} - 1 - \mu_2\right) + 1 \end{aligned}$$

with

$$\begin{aligned} r_1 &= s^2\Bigg((1-a_2)^2\frac{|m|}{n-|m|+1} + (a_2-a_1)(2-a_1-a_2)\frac{|m_1|}{n-|m_1|+1} \\ &\qquad + 1 - a_2^2 - (a_2-a_1)^2\frac{|m_1|}{n-|m_1|+1}\Bigg)/r, \\ r_2 &= s^2a_1^2/r, \\ r_3 &= s^2\left(a_2^2 - a_1^2\frac{m_1}{n} + (a_1-a_2)^2\frac{|m_1|}{n-|m_1|+1}\right)/r. \end{aligned}$$

Use Lemma B.11 (i) with $x_1 = a_1$, $x_2 = a_2$, $y_1 = \mu - \mu_2$, $y_2 = \mu_2$, $k_1 = |m_1|$, $k = |m|$ and $d = n$ to conclude that $|r_1| \leq (1-|m|/n)^{-1}$, that $|r_2| \leq (1 + \mu_2|m|_1/n(1-|m_1|/n) + \mu - \mu_2)^{-1}$ and that $|r_3| \leq (1-|m_1|/n)^{-1}(1+\mu_2)^{-1}$. Let $\tau_1 = (1+\mu_2|m_1|/n(1-|m_1|/n)+\mu-\mu_2)\,\tau$ and $\tau_2 = (1-|m_1|/n)\,(1+\mu_2)\tau$ with $\tau = \exp(\delta)-1$ and note that we can bound the left-hand side in (62) from above by

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\hat{s}^2}{s^2}-1\right|>\alpha_1\left(1-\frac{|m|}{n}\right)\tau\right)\\
&+\mathbb{P}\left(\left|\frac{Y'P_1Y/s^2}{n}-\frac{|m_1|+\theta'Z'P_1Z\theta/s^2}{n}\right|>\alpha_2\tau_1\right)\\
&+\mathbb{P}\left(\left|\frac{\theta'Z'P_1Z\theta/s^2}{n}-\mu_2\frac{|m_1|}{n}-\frac{\tilde{\theta}_1'Z_1'Z_1\tilde{\theta}_1/s^2}{n}\right|>\alpha_3\tau_1\right).\\
&+\mathbb{P}\left(\left|\frac{\tilde{\theta}_1'Z_1'Z_1\tilde{\theta}_1/s^2}{n}-(\mu-\mu_2)\right|>\alpha_4\tau_1\right)\\
&+\mathbb{P}\left(\left|\frac{Y'M_1Y/s^2}{n-|m_1|}-\frac{n-|m_1|+\theta'Z'M_1Z\theta/s^2}{n-|m_1|}\right|>\alpha_5\tau_2\right)\\
&+\mathbb{P}\left(\left|\frac{\theta'Z'M_1Z\theta/s^2}{n-|m_1|}-\mu_2\right|>\alpha_6\tau_2\right),
\end{aligned} \tag{65}$$

where $\alpha_i\in(0,1)$, $i=1,\dots,6$ are such that $\sum_{i=1}^6\alpha_i=1$ and where $\tilde{\theta}_1=\theta_1+S_{1,1}^{-1}S_{1,2}\theta_2$. The left-hand side in (63) is bounded from above by the same upper bound as in (65) with $\tau=1-\exp(-\delta)$. Lemma A.1 shows that $\hat{s}^2/s^2\sim\chi^2_{n-|m|}/(n-|m|)$ and we can bound the first term in (65) using Lemma B.7 from above by

$$2\exp\left(-(n-|m|)\frac{\alpha_1^2(1-|m|/n)^2\tau^2}{4(\alpha_1(1-|m|/n)\tau+1)}\right). \tag{66}$$

For the remaining terms, consider first the case where $\mu_2>0$ and $\mu-\mu_2>0$. Recall from (30) in Lemma A.1 that $Y'P_1Y/s^2\|Z\sim\chi^2_{|m_1|}(\theta'Z'P_1Z\theta/s^2)$ and use Lemma B.8 to bound the second term in (65) from above by

$$2\mathbb{E}\left[\exp\left(-n\frac{\alpha_2^2\tau_1^2}{4\left(\alpha_2\tau_1+2(|m_1|/n+\theta'Z'P_1Z\theta/(s^2n))\right)}\right)\right].$$

Integrating over the intersection of $\{\theta'Z'P_1Z\theta/(s^2n)\le\alpha_3\tau_1+\mu_2|m_1|/n+\tilde{\theta}_1'Z_1'Z_1\tilde{\theta}_1/(s^2n)\}$ and $\{\tilde{\theta}_1'Z_1'Z_1\tilde{\theta}_1/(s^2n)\le\alpha_4\tau_1+\mu-\mu_2\}$ and its complement, we can bound the term in the preceding display from above by

$$\begin{aligned}
&2\exp\left(-n\frac{\alpha_2^2\tau_1^2}{4\left((\alpha_2+2\alpha_3+2\alpha_4)\tau_1+2(|m_1|/n+\mu_2|m_1|/n+\mu-\mu_2)\right)}\right)\\
&+2\mathbb{P}\left(\frac{\theta'Z'P_1Z\theta/s^2}{n}-\mu_2\frac{|m_1|}{n}-\frac{\tilde{\theta}_1'Z_1'Z_1\tilde{\theta}_1/s^2}{n}>\alpha_3\tau_1\right)\\
&+2\mathbb{P}\left(\frac{\tilde{\theta}_1'Z_1'Z_1\tilde{\theta}_1/s^2}{n}-(\mu-\mu_2)>\alpha_4\tau_1\right).
\end{aligned} \tag{67}$$

For the first term in the preceding display, use $\tau_1\le(1+\mu_2|m_1|/n+\mu-\mu_2)\tau$ and $|m_1|/n+\mu_2|m_1|/n+\mu-\mu_2\le 1+\mu_2|m_1|/n+\mu-\mu_2$ in the denominator followed by $\tau_1^2/(1+\mu_2|m_1|/n+\mu-\mu_2)\ge(1-|m_1|/n)\tau^2$ in the numerator to bound this term from above by

$$2\exp\left(-n\frac{\alpha_2^2(1-|m_1|/n)\tau^2}{4\left((\alpha_2+2\alpha_3+2\alpha_4)\tau+2\right)}\right). \tag{68}$$

Recall from (32) in Lemma A.1 that $\theta' Z' P_1' Z\theta/s^2 \| Z_1 \sim \mu_2 \chi^2_{|m_1|}(\tilde{\theta}_1' Z_1' Z_1 \tilde{\theta}_1/(s^2\mu_2))$ and use Lemma B.8 to bound the sum of the third term in (65) and the second term in (67) from above by

$$4\mathbb{E}\left[\exp\left(-n\frac{\alpha_3^2\tau_1^2/\mu_2}{4(\alpha_3\tau_1 + 2(\mu_2|m_1|/n + \tilde{\theta}_1' Z_1' Z_1\tilde{\theta}_1/(s^2 n)))}\right)\right]. \tag{69}$$

Integrating over the event $\{\tilde{\theta}_1' Z_1' Z_1 \tilde{\theta}_1/(s^2 n) \le \alpha_4\tau_1 + \mu - \mu_2\}$ and its complement, using $\mu_2|m_1|/n + \mu - \mu_2 \le 1 + \mu_2|m_1|/n + \mu - \mu_2$ and $\tau_1 \le (1 + \mu_2|m_1|/n + \mu - \mu_2)\tau$ in the denominator followed by $\tau_1^2/(\mu_2(1 + \mu_2|m_1|/n + \mu - \mu_2)) \ge |m_1|/n(1 - |m_1|/n)^2\tau^2$ in the numerator, we can bound the term in the preceding display from above by

$$4\exp\left(-|m_1|\frac{\alpha_3^2(1-|m_1|/n)^2\tau^2}{4((\alpha_3 + 2\alpha_4)\tau + 2)}\right) + 4\mathbb{P}\left(\frac{\tilde{\theta}_1' Z_1' Z_1\tilde{\theta}_1/s^2}{n} - (\mu - \mu_2) > \alpha_4\tau_1\right). \tag{70}$$

By (35) in Lemma A.1 we have that $\tilde{\theta}_1' Z_1' Z_1 \tilde{\theta}_1/s^2 \sim (\mu - \mu_2)\chi^2_n$. Using Lemma B.7 together with the fact that $\tau_1/(\mu - \mu_2) \ge \tau$ for the fourth term in (65), the third term in (67) and the second term in (70) and collecting the terms in (68) and (70), we can finally bound the sum of the second, third and fourth term in (65) from above by

$$\begin{aligned} 2\exp&\left(-n\frac{\alpha_2^2(1-|m_1|/n)\tau^2}{4\left((\alpha_2 + 2\alpha_3 + 2\alpha_4)\tau + 2\right)}\right) + 4\exp\left(-|m_1|\frac{\alpha_3^2(1-|m_1|/n)^2\tau^2}{4((\alpha_3+2\alpha_4)\tau + 2)}\right) \\ &+ 8\exp\left(-n\frac{\alpha_4^2\tau^2}{4(\alpha_4\tau + 1)}\right). \end{aligned} \tag{71}$$

Recall from Lemma A.1 that $Y' M_1 Y/s^2 \| Z \sim \chi^2_{n-|m_1|}(\theta' Z' M_1 Z\theta/s^2)$ and use Lemma B.8 followed by integrating separately over the event $\{\theta' Z' M_1 Z\theta/(s^2(n - |m_1|)) \le \alpha_6\tau_2 + \mu_2\}$ and its complement, we can bound the sum of the fifth and sixth term in (65) from above by

$$\begin{aligned} 2\exp&\left(-(n-|m_1|)\frac{\alpha_5^2\tau_2^2}{4((\alpha_5 + 2\alpha_6)\tau_2 + 2(1+\mu_2))}\right) \\ &+ 2\mathbb{P}\left(\frac{\theta' Z' M_1 Z\theta/s^2}{n - |m_1|} - \mu_2 > \alpha_6\tau_2\right) + \mathbb{P}\left(\left|\frac{\theta' Z' M_1 Z\theta/s^2}{n-|m_1|} - \mu_2\right| > \alpha_6\tau_2\right). \end{aligned}$$

Plugging in the definition for τ_2 and using $1+\mu_2 \ge 1$ for the first term, recalling from (34) in Lemma A.1 that $\theta' Z' M_1 Z\theta/s^2 \sim \mu_2\chi^2_{n-|m_1|}$ and using that $\alpha_6\tau_2/\mu_2 \ge \alpha_6(1-|m_1|/n)\tau$ we can use Lemma B.7 to bound the sum in the preceding display from above by

$$\begin{aligned} 2\exp&\left(-(n-|m_1|)\frac{\alpha_5^2(1-|m_1|/n)^2\tau^2}{4((\alpha_5+2\alpha_6)(1-|m_1|/n)\tau + 2)}\right) \\ &+ 4\exp\left(-(n-|m_1|)\frac{\alpha_6^2(1-|m_1|/n)^2\tau^2}{4(\alpha_6(1-|m_1|/n)\tau + 1)}\right). \end{aligned} \tag{72}$$

Collecting the terms in (66), (71) and (72), we can bound the sum in (65) in the case where $\mu_2 > 0$ and $\mu - \mu_2 > 0$ from above by

$$\begin{aligned}
&2\exp\left(-(n-|m|)\frac{\alpha_1^2(1-|m|/n)^2\tau^2}{4(\alpha_1(1-|m|/n)\tau+1)}\right)\\
&+2\exp\left(-n\frac{\alpha_2^2(1-|m_1|/n)\tau^2}{4\left((\alpha_2+2\alpha_3+2\alpha_4)\tau+2\right)}\right)+4\exp\left(-|m_1|\frac{\alpha_3^2(1-|m_1|/n)^2\tau^2}{4((\alpha_3+2\alpha_4)\tau+2)}\right)\\
&+8\exp\left(-n\frac{\alpha_4^2\tau^2}{4(\alpha_4\tau+1)}\right)+2\exp\left(-(n-|m_1|)\frac{\alpha_5^2(1-|m_1|/n)^2\tau^2}{4((\alpha_5+2\alpha_6)(1-|m_1|/n)\tau+2)}\right)\\
&+4\exp\left(-(n-|m_1|)\frac{\alpha_6^2(1-|m_1|/n)^2\tau^2}{4(\alpha_6(1-|m_1|/n)\tau+1)}\right).
\end{aligned} \tag{73}$$

In the case where $\mu_2 > 0$ and $\mu - \mu_2 = 0$, we have that $\tilde{\theta}_1 = \mathbf{0}$ and that $\theta' Z' P_1 Z\theta/s^2 \| Z_1 \sim \mu_2 \chi^2_{|m_1|}$. Hence, the fourth term in (65) equals zero and we can bound the fifth and sixth term in (65) as before. For the sum of the second and the third term in (65), we can use the bound in (68) and the tail bound of the central chi-square distribution as in Lemma B.7 together with the fact that $\tau_1/(\mu_2|m_1|/n) \geq (1-|m_1|/n)\tau$ to bound the sum from above by

$$2\exp\left(-n\frac{\alpha_2^2(1-|m_1|/n)\tau^2}{4\left((\alpha_2+2\alpha_3+2\alpha_4)\tau+2\right)}\right)+4\exp\left(-|m_1|\frac{\alpha_3^2(1-|m_1|/n)^2\tau^2}{4(\alpha_3(1-|m_1|/n)\tau+1)}\right).$$

This shows that we can use the bound in (73) also in the case where $\mu_2 > 0$ and $\mu - \mu_2 = 0$. In the case where $\mu_2 = 0$ and $\mu - \mu_2 > 0$, we have that $\theta_2 = \mathbf{0}$, that $\theta' Z' P_1 Z\theta/s^2 = \theta_1' Z_1' Z_1 \theta_1/s^2 = \tilde{\theta}_1' Z_1' Z_1 \tilde{\theta}_1/s^2 \sim \mu\chi^2_n$, $\theta' Z' M_1 Z\theta = 0$ and that $Y' M_1 Y/s^2 \| Z \sim \chi^2_{n-|m_1|}$. Hence, the third and the sixth term in (65) equals 0 and we can bound the sum of the second and fourth term in (65) as before. Noting that $\tau_2 = (1-|m_1|/n)\tau$ here, we can use Lemma B.7 to bound the fifth term in (65) from above by

$$2\exp\left(-(n-|m_1|)\frac{\alpha_5^2(1-|m_1|/n)^2\tau^2}{4(\alpha_5(1-|m_1|/n)\tau+1)}\right).$$

This shows that we can use the bound in (73) also in the case where $\mu_2 = 0$ and $\mu - \mu_2 > 0$. In the remaining case where $\mu_2 = \mu - \mu_2 = 0$, we have that $\theta = \mathbf{0}$ as well as $Y' P_1 Y/s^2 \| Z \sim \chi^2_{|m_1|}$ and $Y' M_1 Y/s^2 \| Z \sim \chi^2_{n-|m_1|}$. The third, fourth and sixth term in (65) equal 0 and we can bound the fifth term using the bound in the preceding display. Bound the second term in (65) using the tail bound of the chi-square distribution as in Lemma B.7 from above by

$$2\exp\left(-n\frac{\alpha_2^2(n/|m_1|)\tau^2}{4(\alpha_2(n/|m_1|)\tau+1)}\right).$$

Hence, we can use the bound in (73) also in this case. Use the sum in (73) with $\tau = \exp(\delta) - 1$ and that $\gamma(\exp(\delta)-1)+1 \leq \exp(\delta)$ for any $\gamma \in (0,1)$ and $(\alpha_i + 2\alpha_j)(\exp(\delta) - 1) + 2 \leq (\alpha_i + 2\alpha_j + 2\alpha_k)(\exp(\delta)-1) + 2 \leq 2\exp(\delta)$ for any $i \neq j \neq k \in \{1,2,3,4,5,6\}$

and that $1 \geq 1 - |m_1|/n \geq 1 - |m|/n$ to bound the left-hand side in (62) from above by

$$\begin{aligned}
&2\exp\left(-n\left(1-\frac{|m|}{n}\right)^3 \frac{\alpha_1^2(\exp(\delta)-1)^2}{4\exp(\delta)}\right)\\
&\quad + 2\exp\left(-n\left(1-\frac{|m|}{n}\right)^3 \frac{\alpha_2^2(\exp(\delta)-1)^2}{8\exp(\delta)}\right)\\
&\quad + 4\exp\left(-|m_1|\left(1-\frac{|m|}{n}\right)^3 \frac{\alpha_3^2(\exp(\delta)-1)^2}{8\exp(\delta)}\right)\\
&\quad + 8\exp\left(-n\left(1-\frac{|m|}{n}\right)^3 \frac{\alpha_4^2(\exp(\delta)-1)^2}{4\exp(\delta)}\right)\\
&\quad + 2\exp\left(-n\left(1-\frac{|m|}{n}\right)^3 \frac{\alpha_5^2(\exp(\delta)-1)^2}{8\exp(\delta)}\right)\\
&\quad + 4\exp\left(-n\left(1-\frac{|m|}{n}\right)^3 \frac{\alpha_6^2(\exp(\delta)-1)^2}{4\exp(\delta)}\right).
\end{aligned} \tag{74}$$

To balance the terms in the preceding display, we choose $\alpha_1 = \alpha_4 = \alpha_6 = 1/(3(1+\sqrt{2}))$ and $\alpha_2 = \alpha_3 = \alpha_5 = \sqrt{2}/(3(1+\sqrt{2}))$ which finally shows the statement in (62).

To show the statement in (63), use the bound in (73) with $\tau = 1 - \exp(-\delta) = (\exp(\delta)-1)/\exp(\delta)$ and note that $\gamma(\exp(\delta)-1) + \exp(\delta) \leq 2\exp(\delta)$ for any $\gamma \in (0,1)$ and $(\alpha_i + 2\alpha_j)(\exp(\delta)-1) + 2\exp(\delta) \leq (\alpha_i + 2\alpha_j + 2\alpha_k)(\exp(\delta)-1) + 2\exp(\delta) \leq 4\exp(\delta)$ for any $i \neq j \neq k \in \{1,2,3,4,5,6\}$ and that $1 \geq 1 - |m_1|/n \geq 1 - |m|/n$. Hence, we can bound the left hand-side in (63) from above by

$$\begin{aligned}
&2\exp\left(-n\left(1-\frac{|m|}{n}\right)^3 \frac{\alpha_1^2(\exp(\delta)-1)^2}{8\exp(2\delta)}\right)\\
&\quad + 2\exp\left(-n\left(1-\frac{|m|}{n}\right)^3 \frac{\alpha_2^2(\exp(\delta)-1)^2}{16\exp(2\delta)}\right)\\
&\quad + 4\exp\left(-|m_1|\left(1-\frac{|m|}{n}\right)^3 \frac{\alpha_3^2(\exp(\delta)-1)^2}{16\exp(2\delta)}\right)\\
&\quad + 8\exp\left(-n\left(1-\frac{|m|}{n}\right)^3 \frac{\alpha_4^2(\exp(\delta)-1)^2}{8\exp(2\delta)}\right)\\
&\quad + 2\exp\left(-n\left(1-\frac{|m|}{n}\right)^3 \frac{\alpha_5^2(\exp(\delta)-1)^2}{16\exp(2\delta)}\right)\\
&\quad + 4\exp\left(-n\left(1-\frac{|m|}{n}\right)^3 \frac{\alpha_6^2(\exp(\delta)-1)^2}{8\exp(2\delta)}\right).
\end{aligned} \tag{75}$$

To balance the terms in the preceding display, we use the same weights as above which finally shows the statement in (63).

□

Alternatively, we have the following result where the upper bound depends on the quantity μ!

Lemma B.13. *For each $\delta > 0$, we have*

$$\mathbb{P}(\hat{\rho}^2(m)/r > \exp(\delta)) \leq 22 \exp\left(-n\left(1 - \frac{|m|}{n}\right)^3 \frac{(\exp(\delta) - 1)^2}{210 \exp(\delta)(1 + \mu)}\right) \tag{76}$$

as well as

$$\mathbb{P}(\hat{\rho}^2(m)/r < \exp(-\delta)) \leq 22 \exp\left(-n\left(1 - \frac{|m|}{n}\right)^3 \frac{(\exp(\delta) - 1)^2}{840 \exp(2\delta)(1 + \mu)}\right), \tag{77}$$

where

$$\begin{aligned} r =& s^2 \Bigg((1 - a_2)^2 \frac{|m|}{n - |m| + 1} + (a_2 - a_1)(2 - a_1 - a_2)\frac{|m_1|}{n - |m_1| + 1} + 1 \\ &+ a_1^2(\mu - \mu_2) + a_2^2 \mu_2 + (a_1 - a_2)^2 \frac{|m_1|}{n - |m_1| + 1}\mu_2\Bigg), \end{aligned} \tag{78}$$

where a_1 and a_2 are the shrinkage factors as in (23) *and where $\mu = \theta' S \theta / s^2$ and $\mu_2 = \theta_2'(S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2/s^2$.*

Proof. Use the bound in (73) in the proof of Lemma B.12 except for the third term. This term originates from bounding the sum

$$\begin{aligned} &\mathbb{P}\left(\left|\frac{\theta' Z' P_1 Z \theta / s^2}{n} - \mu_2 \frac{|m_1|}{n} - \frac{\tilde{\theta}_1' Z_1' Z_1 \tilde{\theta}_1 / s^2}{n}\right| > \alpha_3 \tau_1\right) \\ &+ 2\mathbb{P}\left(\frac{\theta' Z' P_1 Z \theta / s^2}{n} - \mu_2 \frac{|m_1|}{n} - \frac{\tilde{\theta}_1' Z_1' Z_1 \tilde{\theta}_1 / s^2}{n} > \alpha_3 \tau_1\right). \end{aligned} \tag{79}$$

In the case where $\mu_2 > 0$ and $\mu - \mu_2 > 0$, bound the sum in the preceding display from above by using the bound in (69) followed by integrating over the event $\{\tilde{\theta}_1' Z_1' Z_1 \tilde{\theta}_1 / (s^2 n) \leq \alpha_4 \tau_1 + \mu - \mu_2\}$ and its complement to get

$$\begin{aligned} &4 \exp\left(-n \frac{\alpha_3^2 \tau_1^2 / \mu_2}{4((\alpha_3 + 2\alpha_4)\tau_1 + 2(\mu_2 |m_1|/n + \mu - \mu_2))}\right) \\ &+ 4\mathbb{P}\left(\frac{\tilde{\theta}_1' Z_1' Z_1 \tilde{\theta}_1 / s^2}{n} - (\mu - \mu_2) > \alpha_4 \tau_1\right). \end{aligned}$$

Bound the second term in the preceding display as in the proof of Lemma B.12 (cf. the bound in (70)). For the first term use $\tau_1 \leq (1 + \mu_2 |m_1|/n + \mu - \mu_2)\tau$ and $\mu_2|m_1|/n + \mu - \mu_2 \leq 1 + \mu_2|m_1|/n + \mu - \mu_2$ in the denominator followed by $\tau_1^2/(\mu_2(1 + \mu_2|m_1|/n + \mu - \mu_2)) \geq \tau^2(1 - |m_1|/n)/(1 + \mu_2)$ in the numerator to bound it from above by

$$4 \exp\left(-n \frac{\alpha_3^2 (1 - |m_1|/n)\tau^2}{4((\alpha_3 + 2\alpha_4)\tau + 2)(1 + \mu_2)}\right). \tag{80}$$

Bound the sum in (79) in the case where $\mu_2 > 0$ and $\mu - \mu_2 = 0$ recalling that $\tilde{\theta}_1 = \mathbf{0}$ and that $\theta' Z' P_1 Z \theta / s^2 \| Z_1 \sim \mu_2 \chi^2_{|m_1|}$ and using Lemma B.7 from above by

$$4 \exp\left(-n \frac{\alpha_3 \tau_1^2 / \mu_2}{4(\alpha_3 \tau_1 + \mu_2 |m_1|/n)}\right).$$

Using $\alpha_3\tau_1 + \mu_2|m_1|/n \leq (\alpha_3\tau + 1)(1 + \mu_2|m_1|/n) \leq ((\alpha_3 + 2\alpha_4)\tau + 2)(1 + \mu_2|m_1|/n)$ in the denominator followed by $\tau_1^2/(\mu_2(1 + \mu_2|m_1|/n)) \geq \tau^2(1 - |m_1|/n)/(1 + \mu_2)$ in the numerator, we can bound the term in the preceding display from above by the term in (80). In the case where $\mu_2 = 0$, the sum in (79) equals 0 and can trivially by bounded from above by the term in (80).

To show the statements in (76) and (77), we can use the bound in (73) where we replace the third term by the term in (80). Setting $\tau = \exp(\delta) - 1$ and using $(\alpha_3 + 2\alpha_4)(\exp(\delta) - 1) + 2 \leq 2\exp(\delta)$ and $1 - |m_1|/n \geq (1 - |m|/n)^3$ for the term in (80), we can bound the left-hand side in (76) from above by the sum in (73) where we replace the third term by

$$4\exp\left(-n\left(1 - \frac{|m|}{n}\right)^3 \frac{\alpha_3^2(\exp(\delta) - 1)^2}{8\exp(\delta)(1 + \mu_2)}\right).$$

Choosing the weights as in the proof of Lemma B.12 and using the fact that $1 \leq 1 + \mu_2 \leq 1 + \mu$ gives the result. To bound the left-hand side in (77), set $\tau = 1 - \exp(-\delta) = (\exp(\delta) - 1)/\exp(\delta)$ and use $(\alpha_3 + 2\alpha_4)(\exp(\delta) - 1) + 2\exp(\delta) \leq 4\exp(\delta)$ and $1 - |m_1|/n \geq (1 - |m|/n)^3$ in (80). Hence, we can use the upper bound in (75) where we replace the third term by

$$4\exp\left(-n\left(1 - \frac{|m|}{n}\right)^3 \frac{\alpha_3^2(\exp(\delta) - 1)^2}{16\exp(\delta)(1 + \mu_2)}\right).$$

Using the same weights as in the proof of Lemma B.12 together with the fact that $1 \leq 1 + \mu_2 \leq 1 + \mu$ gives the result. □

Lemma B.14. *For each $\delta > 0$, we have*

$$\mathbb{P}(\rho^2(m)/r > \exp(\delta)) \leq (58 + 2|m_1|)\exp\left(-|m_1|\frac{|m_1|}{n}\left(1 - \frac{|m|}{n}\right)^5 \frac{(\exp(\delta) - 1)^2}{10477\exp(2\delta)}\right) \tag{81}$$

as well as

$$\mathbb{P}(\rho^2(m)/r < \exp(-\delta)) \leq (58 + 2|m_1|)\exp\left(-|m_1|\frac{|m_1|}{n}\left(1 - \frac{|m|}{n}\right)^5 \frac{(\exp(\delta) - 1)^2}{13089\exp(2\delta)}\right), \tag{82}$$

where r was defined in (64) *in Lemma B.12.*

Proof. Recall the expansion of the true prediction error

$$\begin{aligned}\rho^2(m) = (1-a_2)^2(\hat\theta-\theta)'S(\hat\theta-\theta)\\
&+ (a_2-a_1)(2-a_1-a_2)(\hat\theta_1^*-\theta_1^*)'S_{1,1}(\hat\theta_1^*-\theta_1^*)\\
&+ (a_2-a_1)^2\theta_2'\tilde Z_2'Z_1(Z_1'Z_1)^{-1}S_{1,1}(Z_1'Z_1)^{-1}Z_1'\tilde Z_2\theta_2\\
&+ 2a_1(a_1-a_2)(S_{1,1}\theta_1+S_{1,2}\theta_2)'(Z_1'Z_1)^{-1}Z_1'\tilde Z_2\theta_2\\
&+ 2a_1(a_1-1)(\hat\theta_1^*-\theta_1^*)'(S_{1,1}\theta_1+S_{1,2}\theta_2)\\
&+ 2(1-a_1)(a_2-a_1)(\hat\theta_1^*-\theta_1^*)'S_{1,1}(Z_1'Z_1)^{-1}Z_1'\tilde Z_2\theta_2\\
&+ 2a_2(a_2-1)(\hat\theta_2-\theta_2)'(S_{2,2}-S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2\\
&+ 2a_1(1-a_2)(\hat\theta_2-\theta_2)'\tilde Z_2'Z_1(Z_1'Z_1)^{-1}(S_{1,1}\theta_1+S_{1,2}\theta_2)\\
&+ 2(1-a_2)(a_1-a_2)(\hat\theta_2-\theta_2)'\tilde Z_2'Z_1(Z_1'Z_1)^{-1}S_{1,1}(Z_1'Z_1)^{-1}Z_1'\tilde Z_2\theta_2\\
&+ 2(1-a_2)(a_1-a_2)(\hat\theta_1^*-\theta_1^*)'S_{1,1}(Z_1'Z_1)^{-1}Z_1'\tilde Z_2(\hat\theta_2-\theta_2)\\
&+ s^2 + a_1^2(\theta'S\theta-\theta_2'(S_{2,2}-S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2) + a_2^2\theta_2'(S_{2,2}-S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2,\end{aligned} \tag{83}$$

where a_1 and a_2 are the shrinkage factors as in (23). As before let $\mu = \theta'S\theta/s^2$ and $\mu_2 = \theta_2(S_{2,2}-S_{2,1}S_{1,1}^{-1}S_{1,2})\theta_2/s^2$ and recall that $\mu-\mu_2 = \tilde\theta_1'S_{1,1}\tilde\theta_1$ with $\tilde\theta_1 = \theta_1 + S_{1,1}^{-1}S_{1,2}\theta_2$. Rewrite the sum of the first three lines and the last line in the previous display as

$$\begin{aligned}&(1-a_2)^2\left[(\hat\theta-\theta)'S(\hat\theta-\theta) - s^2\frac{|m|}{n-|m|+1}\right]\\
&+ (a_2-a_1)(2-a_1-a_2)\left[(\hat\theta_1^*-\theta_1^*)'S_{1,1}(\hat\theta_1^*-\theta_1^*) - s^2\frac{|m_1|}{n-|m_1|+1}\right]\\
&+ (a_2-a_1)^2\left[\theta_2'\tilde Z_2'Z_1(Z_1'Z_1)^{-1}S_{1,1}(Z_1'Z_1)^{-1}Z_1'\tilde Z_2\theta_2 - s^2\mu_2\operatorname{trace}\left(S_{1,1}(Z_1'Z_1)^{-1}\right)\right]\\
&+ (a_2-a_1)^2\left[s^2\mu_2\operatorname{trace}\left(S_{1,1}(Z_1'Z_1)^{-1}\right) - s^2\mu_2\frac{|m_1|}{n-|m_1|+1}\right] + r.\end{aligned} \tag{84}$$

Hence, the term $\rho^2(m)/r$ is of the form

$$\sum_{i=i}^{11} r_iT_i + 1,$$

where the r_i's involve the terms a_1, a_2 and r, e.g., $r_1 = (1-a_2)^2/r$, where $T_1,\dots,T_4$ are the random variables in squared brackets in display (84) and $T_5,\dots,T_{11}$ are the random variables in the fourth up to the tenth line in (83), i.e., the part involving θ, Z and S, e.g., $T_5 = (S_{1,1}\theta_1+S_{1,2}\theta_2)'(Z_1'Z_1)^{-1}Z_1'\tilde Z_2\theta_2$. Using this notation and setting $\tau = \exp(\delta)-1$, we can bound the left-hand side in (81) from above by

$$\sum_{i=i}^{11}\mathbb{P}\left(|r_i||T_i| > \alpha_i\tau\right), \tag{85}$$

where $\alpha_i \in (0,1)$, $i = 1,\dots,11$ with $\sum_{i=1}^{11}\alpha_i = 1$. The left-hand side in (82) is bounded from above by the same upper bound with $\tau = 1 - \exp(-\delta)$. Note that

$$\begin{aligned} r &\geq s^2 \left(1 + a_1^2(\mu - \mu_2) + a_2^2\mu_2 + (a_1 - a_2)^2 \frac{|m_1|}{n - |m_1| + 1}\mu_2\right) \\ &\geq s^2 \left(1 + (a_1 - a_2)^2 \frac{|m_1|}{n - |m_1| + 1}\mu_2\right) \geq s^2 \end{aligned} \tag{86}$$

Using the third inequality for r and recalling that $a_1 \in [0,1]$ and $a_2 \in [0,1]$, we see that $|r_1| \leq 1/s^2$, $|r_2| \leq 1/s^2$ as well as $|r_{11}| \leq 2/s^2$. For $\mu_2 > 0$ and $\mu - \mu_2 > 0$, use the second inequality for r to conclude that $|r_3| = |r_4| \leq (s^2\mu_2|m_1|/(n - |m|_1 + 1))^{-1}$, and use the first inequality for r to conclude that $|r_5| \leq (s^2\sqrt{\mu_2(\mu - \mu_2)|m_1|/(n - |m_1| + 1)})^{-1}$, that $|r_6| \leq (s^2\sqrt{\mu - \mu_2})^{-1}$, that $|r_7| \leq (s^2\sqrt{\mu_2|m_1|/(n - |m_1| + 1)})^{-1}$, that $|r_8| \leq (s^2\sqrt{\mu_2})^{-1}$, that $|r_9| \leq (s^2\sqrt{\mu - \mu_2})^{-1}$ and that $|r_{10}| \leq (s^2\sqrt{\mu_2|m_1|/(n - |m_1| + 1)})^{-1}$.[4] In the case where $\mu_2 = 0$ it follows that $\theta_2 = \mathbf{0}$, i.e., the zero vector, which implies that $T_3 = T_4 = T_5 = T_7 = T_8 = T_{10} = 0$. In the case where $\mu - \mu_2 = 0$, it follows that $\tilde{\theta}_1 = \mathbf{0}$ as well as $S_{1,1}\tilde{\theta}_1 = S_{1,1}\theta_1 + S_{1,2}\theta_2 = \mathbf{0}$ and hence that $T_5 = T_6 = T_9 = 0$. In these cases the corresponding terms in (85) equal 0 and can be trivially bounded from above by the bounds for the case $\mu_2 > 0$ and $\mu - \mu_2 > 0$, respectively, that are derived in the next paragraphs. Hence, from now on we assume that $\mu_2 > 0$ and $\mu - \mu_2 > 0$.

Use Lemma A.1 together with the upper bounds of $|r_1|$ and $|r_2|$ and the Chernoff bounds as in Lemma A.1 and Lemma A.3 (i) in Leeb (2008) and the bound in Lemma C.8 (i) in Huber (2013) to bound the sum of the first and the second term in (85) from above by

$$\begin{aligned} &2\exp\left(-(n - |m| + 1)\frac{\alpha_1^2(1 - |m|/(n+1))^2\tau^2}{4(\alpha_1(1 - |m|/(n+1))\tau + 1)^2}\right) \\ &+ 2\exp\left(-(n - |m_1| + 1)\frac{\alpha_2^2(1 - |m_1|/(n+1))^2\tau^2}{4(\alpha_2(1 - |m_1|/(n+1))\tau + 1)^2}\right). \end{aligned} \tag{87}$$

Lemma A.1 shows that $\tilde{Z}_2\theta_2/(s\sqrt{\mu_2})\|Z_1 \sim N(0, I_n)$. Set $V_1 = Z_1S_{1,1}^{-1/2}$ and note that $n\lambda_n(Z_1(Z_1'Z_1)^{-1}S_{1,1}(Z_1'Z_1)^{-1}Z_1') = \lambda_1^{-1}(V_1'V_1/n)$ and that $\operatorname{trace}(S_{1,1}(Z_1'Z_1)^{-1}) = \operatorname{trace}((V_1'V_1)^{-1})$. Hence, we can use Corollary B.2 conditional on Z_1 together with the upper bound on $|r_3|$ to bound the third term in (85) from above by

$$2\mathbb{E}\left[\exp\left(-n\frac{\alpha_3^2|m_1|^2/(n - |m_1| + 1)^2\tau^2}{4\lambda_1^{-1}(V_1'V_1/n)(\alpha_3|m_1|/(n - |m_1| + 1)\tau + \lambda_1^{-1}(V_1'V_1/n))}\right)\right].$$

Note that the term in the preceding display is nonincreasing in $\lambda_1(V_1'V_1/n)$ and that V_1 is a $n \times |m_1|$ matrix that has i.i.d. standard normally distributed entries. Let

$$A_1 = \{\lambda_1(V_1'V_1/n) > \gamma_1^2(1 - \sqrt{|m_1|/n})^2\}$$

for some $\gamma_1 \in (0,1)$. Integrate the term in the second-to-last display separately over A_1 and its complement, note that the integrand is bounded from above by 1 and use Lemma B.9 on the complement of A_1 to bound the third term in (85) finally from above by

$$\begin{aligned} &2\exp\left(-|m_1|\frac{|m_1|}{n}\frac{\alpha_3^2\gamma_1^4n^2/(n - |m_1| + 1)^2(1 - \sqrt{|m_1|/n})^4\tau^2}{4(\alpha_3\gamma_1^2|m_1|/(n - |m_1| + 1)(1 - \sqrt{|m_1|/n})^2\tau + 1)}\right) \\ &+ 2\exp\left(-n\frac{(1 - \gamma_1)^2(1 - \sqrt{|m_1|/n})^2}{2}\right). \end{aligned} \tag{88}$$

[4] The inequalities can be rewritten to have the form $2xy \leq 1 + x^2 + y^2 + z$ where $z \geq 0$.

Bound the fourth term in (85) noting that $\operatorname{trace}(S_{1,1}(Z_1'Z_1)^{-1}) = \operatorname{trace}((V_1'V_1)^{-1})$ and using (47) in Lemma B.10 together with the upper bound of $|r_4|$ from above by

$$2|m_1| \exp\left(-(n-|m_1|)\frac{\alpha_4^2\tau^2}{8(\alpha_4\tau+1)^2}\right).$$

Use (29) in Lemma A.1 to show that $T_5 \| Z_1 \sim N(0, s^2\mu_2(S_{1,1}\theta_1 + S_{1,2}\theta_2)'(Z_1'Z_1)^{-1}(S_{1,1}\theta_1 + S_{1,2}\theta_2))$. Use Lemma B.6, conditional on Z_1, together with the upper bound of $|r_5|$ to bound the fifth term in (85) from above by

$$2\mathbb{E}\left[\exp\left(-|m_1|\frac{\alpha_5^2 s^2(\mu-\mu_2)n/(n-|m_1|+1)\tau^2}{2(S_{1,1}\theta_1+S_{1,2}\theta_2)'(Z_1'Z_1/n)^{-1}(S_{1,1}\theta_1+S_{1,2}\theta_2)}\right)\right].$$

We can bound the term in the denominator in the preceding display from above by $2s^2(\mu-\mu_2)/\lambda_1(V_1'V_1/n)$. Using additionally that $n/(n-|m_1|+1) \geq 1$, we can bound the term in the preceding display from above by

$$2\mathbb{E}\left[\exp\left(-|m_1|\frac{\alpha_5^2\lambda_1(V_1'V_1/n)\tau^2}{2}\right)\right].$$

Because also the terms T_6, T_7, T_8, T_9 and T_{10} follow, conditional on Z, a centered normal distribution, we will use the same arguments as for the fifth term, mutatis mutandis, for bounding the summands involving these terms. Define $\tilde{V}_2 = \tilde{Z}_2(S_{2,2} - S_{2,1}S_{1,1}^{-1}S_{1,2})^{-1/2}$ and note that $\tilde{V}_2$ is a $n \times |m_2|$-matrix that has i.i.d. entries that follow a standard normal distribution. Note that the conditional variance of T_6 is bounded from above by $s^4(\mu-\mu_2)/\lambda_1(V_1'V_1)$, that the conditional variance of T_7 is bounded from above by $s^4\mu_2\lambda_{|m_2|}(\tilde{V}_2'\tilde{V}_2)/\lambda_1^2(V_1'V_1)$, that the conditional variance of T_8 is bounded from above by $s^4\mu_2/\lambda_1(\tilde{V}_2'M_1\tilde{V}_2)$, that the conditional variance of T_9 is bounded from above by $s^4(\mu-\mu_2)\lambda_{|m_2|}(\tilde{V}_2'\tilde{V}_2)/(\lambda_1(\tilde{V}_2'M_1\tilde{V}_2)\lambda_1(V_1'V_1))$ and that the conditional variance of T_{10} is bounded from above by $s^4\mu_2\lambda_{|m_2|}^2(\tilde{V}_2'\tilde{V}_2)/(\lambda_1(\tilde{V}_2'M_1\tilde{V}_2)\lambda_1^2(V_1'V_1))$. Using the upper bounds of the variances, the upper bounds on $|r_i|$ together with Lemma B.6 conditional on Z and the facts that $n/(n-|m_1|+1) \geq 1$ as well as $(n-|m_1|)/(n-|m_1|+1) \geq 1/2$, we can bound the sum of the sixth up to the tenth term in (85) from above by

$$\begin{aligned}
&2\mathbb{E}\left[\exp\left(-n\frac{\alpha_6^2\lambda_1(V_1'V_1/n)\tau^2}{2}\right)\right] \\
&+ 2\mathbb{E}\left[\exp\left(-|m_1|\frac{\alpha_7^2\lambda_1^2(V_1'V_1/n)\tau^2}{2\lambda_{|m_2|}(\tilde{V}_2'\tilde{V}_2/n)}\right)\right] \\
&+ 2\mathbb{E}\left[\exp\left(-(n-|m_1|)\frac{\alpha_8^2\lambda_1(\tilde{V}_2'M_1\tilde{V}_2/(n-|m_1|))\tau^2}{2}\right)\right] \\
&+ 2\mathbb{E}\left[\exp\left(-(n-|m_1|)\frac{\alpha_9^2\lambda_1(\tilde{V}_2'M_1\tilde{V}_2/(n-|m_1|))\lambda_1(V_1'V_1/n)\tau^2}{2\lambda_{|m_2|}(\tilde{V}_2'\tilde{V}_2/n)}\right)\right] \\
&+ 2\mathbb{E}\left[\exp\left(-|m_1|\frac{\alpha_{10}^2\lambda_1(\tilde{V}_2'M_1\tilde{V}_2/(n-|m_1|))\lambda_1^2(V_1'V_1/n)\tau^2}{4\lambda_{|m_2|}^2(\tilde{V}_2'\tilde{V}_2/n)}\right)\right].
\end{aligned}$$

In order to get upper bounds for the sum of the previous two displays, we need upper bounds on $\lambda_{|m_2|}(\tilde{V}_2'\tilde{V}_2/n)$ and lower bounds on $\lambda_1(V_1'V_1/n)$ and $\lambda_1(\tilde{V}_2'M_2\tilde{V}_2/(n-|m_1|))$. Recall A_1 and define

$$\begin{aligned}
A_2 &= \{\lambda_1(\tilde{V}_2'M_1'\tilde{V}_2/(n-|m_1|)) > \gamma_2^2(1-\sqrt{|m_2|/(n-|m_1|)})^2\} \\
A_3 &= \{\lambda_{|m_2|}(\tilde{V}_2'\tilde{V}_2/n) < (1+\gamma_3)^2(1+\sqrt{|m_2|/n})^2\}
\end{aligned}$$

for $\gamma_2 \in (0,1)$ and $\gamma_3 > 0$. Integrating the corresponding terms separately over A_1, A_2, A_3 and its complements, noting that the integrand is bounded from above by 1 and using Lemma B.9, we can bound the sum of the fifth up to the tenth term in (85) finally from above by

$$\begin{aligned}
&2\exp\left(-|m_1|\frac{\alpha_5^2\gamma_1^2(1-\sqrt{|m_1|/n})^2\tau^2}{2}\right)\\
&+2\exp\left(-n\frac{\alpha_6^2\gamma_1^2\ (1-\sqrt{|m_1|/n})^2\tau^2}{2}\right)\\
&+2\exp\left(-|m_1|\frac{\alpha_7^2\gamma_1^4(1-\sqrt{|m_1|/n})^4\tau^2}{2(1+\gamma_3)^2(1+\sqrt{|m_2|/n})^2}\right)\\
&+2\exp\left(-(n-|m_1|)\frac{\alpha_8^2\gamma_2^2(1-\sqrt{|m_2|/(n-|m_1|)})^2\tau^2}{2}\right)\\
&+2\exp\left(-(n-|m_1|)\frac{\alpha_9^2\gamma_1^2\gamma_2^2(1-\sqrt{|m_2|/(n-|m_1|)})^2(1-\sqrt{|m_1|/n})^2\tau^2}{2(1+\gamma_3)^2(1+\sqrt{|m_2|/n})^2}\right)\\
&+2\exp\left(-|m_1|\frac{\alpha_{10}^2\gamma_1^4\gamma_2^2(1-\sqrt{|m_2|/(n-|m_1|)})^2(1-\sqrt{|m_1|/n})^4\tau^2}{4(1+\gamma_3)^4(1+\sqrt{|m_2|/n})^4}\right)\\
&+10\exp\left(-n\frac{(1-\gamma_1)^2(1-\sqrt{|m_1|/n})^2}{2}\right)+6\exp\left(-n\frac{\gamma_3^2(1+\sqrt{|m_2|/n})^2}{2}\right)\\
&+6\exp\left(-n\frac{(1-\gamma_2)^2(1-\sqrt{|m_2|/(n-|m_1|)})^2}{2}\right).
\end{aligned} \tag{89}$$

The term T_{11} can be rewritten as $w'Qw$ where $Q = Z_1(Z_1'Z_1)^{-1}S_{1,1}(Z_1'Z_1)^{-1}Z_1'\tilde{Z}_2(Z_2'M_1Z_2)^{-1}Z_2'M_1$ and where $w \sim N(0, s^2I_n)$ as in (22). Noting that trace$(Q) = 0$ and that $QQ = \mathbf{0}$, i.e., the zero matrix, we can use Corollary B.5, conditional on Z, together with the upper bound on $|r_{11}|$ to bound the last term in (85) if $\lambda_n(Q+Q') > 0$ from above by

$$4\mathbb{E}\left[\exp\left(-n\frac{\alpha_{11}^2\tau^2}{8n\sqrt{2\lambda_n(Q'Q)}(\alpha_{11}\tau + 2n\sqrt{2\lambda_n(Q'Q)})}\right)\right]. \tag{90}$$

Recall from the proof of Lemma B.4 that $\lambda_n(Q'Q) > 0$ in this case. In the case where $\lambda_n(Q+Q') = 0$, the corresponding upper bound equals 0. Note that $M_1Z_2 = M_1\tilde{Z}_2$, that

$$Q'Q = M_1\tilde{V}_2(\tilde{V}_2'M_1\tilde{V}_2)^{-1}\tilde{V}_2'V_1(V_1'V_1)^{-3}V_1'\tilde{V}_2(\tilde{V}_2'M_1\tilde{V}_2)^{-1}\tilde{V}_2'M_1,$$

that the term in (90) is nondecreasing in $\lambda_n(Q'Q)$ and finally that

$$\lambda_n(Q'Q) \le \frac{1}{n(n-|m_1|)}\frac{\lambda_{|m_2|}(\tilde{V}_2'\tilde{V}_2/n)}{\lambda_1(\tilde{V}_2'M_1\tilde{V}_2/(n-|m_1|))\lambda_1^2(V_1'V_1/n)}.$$

Integrating the term in (90) separately over $A_1 \cap A_2 \cap A_3$ and its complement, noting that the integrand is bounded from above by 1 and using Lemma B.9, we can bound the last term in (85) from above by

$$
\begin{aligned}
&+4\exp\left(-n\frac{\alpha_{11}^2\gamma_1^4\gamma_2^2C_1^2\tau^2}{8\sqrt{2}(1+\gamma_3)\left(\alpha_{11}\gamma_1^2\gamma_2C_1\tau+2\sqrt{2}(1+\gamma_3)\right)}\right)\\
&+4\exp\left(-n\frac{(1-\gamma_1)^2(1-\sqrt{|m_1|/n})^2}{2}\right)+4\exp\left(-n\frac{\gamma_3^2(1+\sqrt{|m_2|/n})^2}{2}\right)\\
&+4\exp\left(-n\frac{(1-\gamma_2)^2(1-\sqrt{|m_2|/(n-|m_1|)})^2}{2}\right),
\end{aligned}
\tag{91}
$$

where $C_1=(1-\sqrt{|m_2|/(n-|m_1|)})(1-\sqrt{|m_1|/n})^2\sqrt{1-|m_1|/n}/(1+\sqrt{|m_2|/n})$.

Collecting the terms in (87), (88), (89) and (91) and using the facts that $(n+1)(1-|m_1|/(n+1))^3 \geq (n+1)(1-|m|/(n+1))^3 \geq n(1-|m|/n)^3$, $n/(n-|m_1|+1) \geq 1$, $|m_1|/(n-|m_1|+1)(1-\sqrt{|m_1|/n})^2 \leq 1$ and that $1 \leq 1+\sqrt{|m_2|/n} \leq 2$, we can bound the sum in (85) finally from above by

$$
\begin{aligned}
&2\exp\left(-n\left(1-\frac{|m|}{n}\right)^3\frac{\alpha_1^2\tau^2}{4(\alpha_1(1-|m|/(n+1))\tau+1)^2}\right)\\
&\quad+2\exp\left(-n\left(1-\frac{|m|}{n}\right)^3\frac{\alpha_2^2\tau^2}{4(\alpha_2(1-|m_1|/(n+1))\tau+1)^2}\right)\\
&\quad+2\exp\left(-|m_1|\frac{|m_1|}{n}\frac{\alpha_3^2\gamma_1^4(1-\sqrt{|m_1|/n})^4\tau^2}{4(\alpha_3\gamma_1^2\tau+1)}\right)\\
&\quad+2|m_1|\exp\left(-n\left(1-\frac{|m_1|}{n}\right)\frac{\alpha_4^2\tau^2}{8(\alpha_4\tau+1)^2}\right)\\
&\quad+2\exp\left(-|m_1|\frac{\alpha_5^2\gamma_1^2(1-\sqrt{|m_1|/n})^2\tau^2}{2}\right)\\
&\quad+2\exp\left(-n\frac{\alpha_6^2\gamma_1^2\ (1-\sqrt{|m_1|/n})^2\tau^2}{2}\right)\\
&\quad+2\exp\left(-|m_1|\frac{\alpha_7^2\gamma_1^4(1-\sqrt{|m_1|/n})^4\tau^2}{8(1+\gamma_3)^2}\right) \qquad (92)\\
&\quad+2\exp\left(-n\left(1-\frac{|m_1|}{n}\right)\frac{\alpha_8^2\gamma_2^2(1-\sqrt{|m_2|/(n-|m_1|)})^2\tau^2}{2}\right)\\
&\quad+2\exp\left(-n\left(1-\frac{|m_1|}{n}\right)\frac{\alpha_9^2\gamma_1^2\gamma_2^2(1-\sqrt{|m_2|/(n-|m_1|)})^2(1-\sqrt{|m_1|/n})^2\tau^2}{8(1+\gamma_3)^2}\right)\\
&\quad+2\exp\left(-|m_1|\frac{\alpha_{10}^2\gamma_1^4\gamma_2^2(1-\sqrt{|m_2|/(n-|m_1|)})^2(1-\sqrt{|m_1|/n})^4\tau^2}{64(1+\gamma_3)^4}\right)\\
&\quad+4\exp\left(-n\frac{\alpha_{11}^2\gamma_1^4\gamma_2^2\tilde{C}_1^2\tau^2}{16\sqrt{2}(1+\gamma_3)\left(\alpha_{11}\gamma_1^2\gamma_2\tilde{C}_1\tau+4\sqrt{2}(1+\gamma_3)\right)}\right)\\
&\quad+16\exp\left(-n\frac{(1-\gamma_1)^2(1-\sqrt{|m_1|/n})^2}{2}\right)+10\exp\left(-n\frac{\gamma_3^2}{2}\right)\\
&\quad+10\exp\left(-n\frac{(1-\gamma_2)^2(1-\sqrt{|m_2|/(n-|m_1|)})^2}{2}\right).
\end{aligned}
$$

where $\tilde{C}_1=(1-\sqrt{|m_2|/(n-|m_1|)})(1-\sqrt{|m_1|/n})^2\sqrt{1-|m_1|/n}$. For the statement in (81) use the bound in the preceding display with $\tau=\exp(\delta)-1$, note that $\alpha(\exp(\delta)-1)+D\leq D\exp(\delta)\leq D\exp(2\delta)$ holds for any $D\geq 1$ and $\alpha\in(0,1)$ and that $\exp(\delta)\geq 1\geq(\exp(\delta)-1)^2/\exp(2\delta)$. Using this together with the inequalities in Lemma B.11 (ii), we can bound the left hand-side in (81) from above by

$$\begin{aligned}
&2\exp\left(-n\frac{\alpha_1^2}{4}\psi(\delta,|m|/n)\right)+2\exp\left(-n\frac{\alpha_2^2}{4}\psi(\delta,|m|/n)\right)\\
&+2\exp\left(-|m_1|\frac{|m_1|}{n}\frac{\alpha_3^2\gamma_1^4}{18}\psi(\delta,|m|/n)\right)+2|m_1|\exp\left(-n\frac{\alpha_4^2}{8}\psi(\delta,|m|/n)\right)\\
&+2\exp\left(-|m_1|\frac{3\alpha_5^2\gamma_1^2}{8}\psi(\delta,|m|/n)\right)+2\exp\left(-n\frac{3\alpha_6^2\gamma_1^2}{8}\psi(\delta,|m|/n)\right)\\
&+2\exp\left(-|m_1|\frac{\alpha_7^2\gamma_1^4}{36(1+\gamma_3)^2}\psi(\delta,|m|/n)\right)+2\exp\left(-n\frac{3\alpha_8^2\gamma_2^2}{8}\psi(\delta,|m|/n)\right)\\
&+2\exp\left(-n\frac{\alpha_9^2\gamma_1^2\gamma_2^2}{16(1+\gamma_3)^2}\psi(\delta,|m|/n)\right)+2\exp\left(-|m_1|\frac{\alpha_{10}^2\gamma_1^4\gamma_2^2}{360(1+\gamma_3)^4}\psi(\delta,|m|/n)\right)\\
&+4\exp\left(-n\frac{\alpha_{11}^2\gamma_1^4\gamma_2^2}{2640(1+\gamma_3)^2}\psi(\delta,|m|/n)\right)+16\exp\left(-n\frac{3(1-\gamma_1)^2}{8}\psi(\delta,|m|/n)\right)\\
&+10\exp\left(-n\frac{\gamma_3^2}{2}\psi(\delta,|m|/n)\right)+10\exp\left(-n\frac{3(1-\gamma_2)^2}{8}\psi(\delta,|m|/n)\right),
\end{aligned}\tag{93}$$

where $\psi(x,y)=(1-y)^5(\exp(x)-1)^2/\exp(2x)$. To balance the terms in the preceding display, choose the weights to be $\alpha_1=\alpha_2=1/51$, $\alpha_3=11/255$, $\alpha_4=43/1530$, $\alpha_5=\alpha_6=\alpha_8=5/306$, $\alpha_7=19/306$, $\alpha_9=21/510$, $\alpha_{10}=1/5$, $\alpha_{11}=137/255$ and $\gamma_1=\gamma_2=1-\alpha_1\sqrt{2/3}$, $\gamma_3=\alpha_1/\sqrt{2}$. This shows the statement in (81).

For the second statement, use the bound in (92) with $\tau=1-\exp(-\delta)=(\exp(\delta)-1)/\exp(\delta)$, note that $\alpha(\exp(\delta)-1)+D\exp(\delta)\le(D+1)\exp(\delta)$ for every $D\ge 0$ and every $\alpha\in(0,1)$ and that $1\ge(\exp(\delta)-1)^2/\exp(2\delta)$. Using this together with the inequalities in Lemma B.11 (ii), we can bound the left hand-side in (82) from above by

$$\begin{aligned}
&2\exp\left(-n\frac{\alpha_1^2}{16}\psi(\delta,|m|/n)\right)+2\exp\left(-n\frac{\alpha_2^2}{16}\psi(\delta,|m|/n)\right)\\
&+2\exp\left(-|m_1|\frac{|m_1|}{n}\frac{\alpha_3^2\gamma_1^4}{36}\psi(\delta,|m|/n)\right)+2|m_1|\exp\left(-n\frac{\alpha_4^2}{32}\psi(\delta,|m|/n)\right)\\
&+2\exp\left(-|m_1|\frac{3\alpha_5^2\gamma_1^2}{8}\psi(\delta,|m|/n)\right)+2\exp\left(-n\frac{3\alpha_6^2\gamma_1^2}{8}\psi(\delta,|m|/n)\right)\\
&+2\exp\left(-|m_1|\frac{\alpha_7^2\gamma_1^4}{36(1+\gamma_3)^2}\psi(\delta,|m|/n)\right)+2\exp\left(-n\frac{3\alpha_8^2\gamma_2^2}{8}\psi(\delta,|m|/n)\right)\\
&+2\exp\left(-n\frac{\alpha_9^2\gamma_1^2\gamma_2^2}{16(1+\gamma_3)^2}\psi(\delta,|m|/n)\right)+2\exp\left(-|m_1|\frac{\alpha_{10}^2\gamma_1^4\gamma_2^2}{360(1+\gamma_3)^4}\psi(\delta,|m|/n)\right)\\
&+4\exp\left(-n\frac{\alpha_{11}^2\gamma_1^4\gamma_2^2}{330\sqrt{2}(1+\gamma_3)(1+4\sqrt{2}(1+\gamma_3))}\psi(\delta,|m|/n)\right)\\
&+16\exp\left(-n\frac{3(1-\gamma_1)^2}{8}\psi(\delta,|m|/n)\right)+10\exp\left(-n\frac{\gamma_3^2}{2}\psi(\delta,|m|/n)\right)\\
&+10\exp\left(-n\frac{3(1-\gamma_2)^2}{8}\psi(\delta,|m|/n)\right),
\end{aligned}\tag{94}$$

To balance the terms in the preceding display, choose the weights to be $\alpha_1=\alpha_2=5/143$, $\alpha_3=541/10010$, $\alpha_4=99/2002$, $\alpha_5=\alpha_6=\alpha_8=29/2002$, $\alpha_7=547/10010$, $\alpha_9=73/2002$, $\alpha_{10}=1777/10010$, $\alpha_{11}=1030/2002$, $\gamma_1=\gamma_2=1-\alpha_1/\sqrt{6}$ and $\gamma_3=\alpha_1/\sqrt{8}$ which shows the statement in (82). □

Lemma B.15. *For each $\delta > 0$, we have*

$$\mathbb{P}(\rho^2(m)/r > \exp(\delta)) \leq (58 + 2|m_1|) \exp\left(-n\left(1 - \frac{|m|}{n}\right)^5 \frac{(\exp(\delta) - 1)^2}{19371(1 + \mu)^2 \exp(2\delta)}\right) \tag{95}$$

as well as

$$\mathbb{P}(\rho^2(m)/r < \exp(-\delta)) \leq (58 + 2|m_1|) \exp\left(-n\left(1 - \frac{|m|}{n}\right)^5 \frac{(\exp(\delta) - 1)^2}{22534(1 + \mu)^2 \exp(2\delta)}\right), \tag{96}$$

where r was defined in (64) *in Lemma B.12.*

Proof. We take over some of the upper bounds from Lemma B.14 but use different bounds for the third, the fifth up to the eighth and the tenth term in (85). As in the proof of Lemma B.14, it suffices to consider only the case where $\mu_2 > 0$ and $\mu - \mu_2 > 0$. In the case where $\mu_2 = 0$ we have that $\theta_2 = \mathbf{0}$ and hence that all of the above mentioned terms equal 0, in the case where $\mu - \mu_2 = 0$, we have that $\tilde{\theta}_1 = \mathbf{0}$ and hence that the fifth term equals 0. In these cases we can trivially bound the corresponding terms by the upper bounds that we will derive now. Use $|r_3| \leq 1/s^2$ together with Corollary B.2 conditional on Z as before to bound the third term in (85) from above by

$$2\mathbb{E}\left[\exp\left(-n\frac{\alpha_3^2\tau^2/\mu_2^2}{4\lambda_1^{-1}(V_1'V_1/n)(\alpha_3\tau/\mu_2 + \lambda_1^{-1}(V_1'V_1/n))}\right)\right].$$

As in the proof of Lemma B.14 let $A_1 = \{\lambda_1(V_1'V_1/n) > \gamma_1^2(1 - \sqrt{|m_1|/n})^2\}$ and integrate separately over the event A_1 and its complement to bound the term in the preceding display from above by (cf. the bound in (88))

$$2\exp\left(-n\frac{\alpha_3^2\gamma_1^4(1 - \sqrt{|m_1|/n})^4\tau^2}{4\mu_2(\alpha_3\gamma_1^2(1 - \sqrt{|m_1|/n})^2\tau + \mu_2)}\right) + 2\exp\left(-n\frac{(1 - \gamma_1)^2(1 - \sqrt{|m_1|/n})^2}{2}\right).$$

For the remaining four terms, note that $r \geq s^2(1 + a_1^2(\mu - \mu_2)) \geq s^2$ which implies that $|r_5| \leq 2a_1/r \leq (s^2\sqrt{\mu - \mu_2})^{-1}$, $|r_6| \leq 2a_1(1 - a_1)/s^2 \leq 1/(2s^2)$, $|r_7| \leq 2/s^2$, $|r_8| \leq 2a_2(1 - a_2)/s^2 \leq 1/(2s^2)$ and $|r_{10}| \leq 2/s^2$. Using these upper bounds together with the arguments in the proof of Lemma B.14 to bound the sum of the fifth up to the eighth and the tenth term in (85) from above by

$$
\begin{aligned}
&2\mathbb{E}\left[\exp\left(-n\frac{\alpha_5^2\lambda_1(V_1'V_1/n)\tau^2}{2\mu_2}\right)\right]\\
&+2\mathbb{E}\left[\exp\left(-n\frac{2\alpha_6^2\lambda_1(V_1'V_1/n)\tau^2}{\mu-\mu_2}\right)\right]\\
&+2\mathbb{E}\left[\exp\left(-n\frac{\alpha_7^2\lambda_1^2(V_1'V_1/n)\tau^2}{8\mu_2\lambda_{|m_2|}(\tilde{V}_2'\tilde{V}_2/n)}\right)\right]\\
&+2\mathbb{E}\left[\exp\left(-(n-|m_1|)\frac{2\alpha_8^2\lambda_1(\tilde{V}_2'M_1\tilde{V}_2/(n-|m_1|))\tau^2}{\mu_2}\right)\right]\\
&+2\mathbb{E}\left[\exp\left(-(n-|m_1|)\frac{\alpha_{10}^2\lambda_1(\tilde{V}_2'M_1\tilde{V}_2/(n-|m_1|))\lambda_1^2(V_1'V_1/n)\tau^2}{8\mu_2\lambda_{|m_2|}^2(\tilde{V}_2'\tilde{V}_2/n)}\right)\right].
\end{aligned}
$$

Integrating the corresponding terms over A_1, A_2, A_3 and its complements as in the proof of Lemma B.14, mutatis mutandis, we can finally bound the sum of the third, the fifth up to the eighth and the tenth term in (85) from above by

$$
\begin{aligned}
&2\exp\left(-n\frac{\alpha_3^2\gamma_1^4(1-\sqrt{|m_1|/n})^4\tau^2}{4\mu_2(\alpha_3\gamma_1^2(1-\sqrt{|m_1|/n})^2\tau+\mu_2)}\right)\\
&+2\exp\left(-n\frac{\alpha_5^2\gamma_1^2(1-\sqrt{|m_1|/n})^2\tau^2}{2\mu_2}\right)\\
&+2\exp\left(-n\frac{2\alpha_6^2\gamma_1^2(1-\sqrt{|m_1|/n})^2\tau^2}{\mu-\mu_2}\right)\\
&+2\exp\left(-n\frac{\alpha_7^2\gamma_1^4(1-\sqrt{|m_1|/n})^4\tau^2}{8\mu_2(1+\gamma_3)^2(1+\sqrt{|m_2|/n})^2}\right)\\
&+2\exp\left(-(n-|m_1|)\frac{2\alpha_8^2\gamma_2^2(1-\sqrt{|m_2|/(n-|m_1|)})^2\tau^2}{\mu_2}\right)\\
&+2\exp\left(-(n-|m_1|)\frac{\alpha_{10}^2\gamma_1^4\gamma_2^2(1-\sqrt{|m_2|/(n-|m_1|)})^2(1-\sqrt{|m_1|/n})^4\tau^2}{8\mu_2(1+\gamma_3)^4(1+\sqrt{|m_2|/n})^4}\right)\\
&+10\exp\left(-n\frac{(1-\gamma_1)^2(1-\sqrt{|m_1|/n})^2}{2}\right)+4\exp\left(-n\frac{\gamma_3^2(1+\sqrt{|m_2|/n})^2}{2}\right)\\
&+4\exp\left(-n\frac{(1-\gamma_2)^2(1-\sqrt{|m_2|/(n-|m_1|)})^2}{2}\right).
\end{aligned}
$$

Plugging in $\tau=\exp(\delta)-1$, using $\mu_2(\alpha(\exp(\delta)-1)+\mu_2)\leq(1+\mu_2)^2\exp(\delta)$ for any $\alpha\in[0,1]$, $\mu_2\leq 1+\mu_2$ as well as $1\leq 1+\sqrt{|m_2|/n}\leq 2$, $1-|m_1|/n\geq 1-|m|/n$ and $\exp(\delta)\geq 1\geq(\exp(\delta)-1)^2/\exp(2\delta)$ together with the bounds in Lemma B.11 (ii), we can bound the sum of the first six terms in the preceding display from above by

$$\begin{aligned}
&2\exp\left(-n\left(1-\frac{|m_1|}{n}\right)^5\frac{\alpha_3^2\gamma_1^4(\exp(\delta)-1)^2}{18(1+\mu_2)^2\exp(2\delta)}\right)\\
&\quad+2\exp\left(-n\left(1-\frac{|m_1|}{n}\right)^5\frac{3\alpha_5^2\gamma_1^2(\exp(\delta)-1)^2}{8(1+\mu_2)\exp(2\delta)}\right)\\
&\quad+2\exp\left(-n\left(1-\frac{|m_1|}{n}\right)^5\frac{3\alpha_6^2\gamma_1^2(\exp(\delta)-1)^2}{2(1+\mu-\mu_2)\exp(2\delta)}\right)\\
&\quad+2\exp\left(-n\left(1-\frac{|m_1|}{n}\right)^5\frac{\alpha_7^2\gamma_1^4(\exp(\delta)-1)^2}{144(1+\gamma_3)^2(1+\mu_2)\exp(2\delta)}\right)\\
&\quad+2\exp\left(-n\left(1-\frac{|m_1|}{n}\right)^5\frac{3\alpha_8^2\gamma_2^2(\exp(\delta)-1)^2}{2(1+\mu_2)\exp(2\delta)}\right)\\
&\quad+2\exp\left(-n\left(1-\frac{|m|}{n}\right)^5\frac{\alpha_{10}^2\gamma_1^4\gamma_2^2(\exp(\delta)-1)^2}{2640(1+\gamma_3)^4(1+\mu_2)\exp(2\delta)}\right)
\end{aligned}$$

Hence, we can bound the left-hand side in (95) from above by the sum in (93) where we replace the sum of the third, fifth up to the eighth and tenth term by the sum in the preceding display. Choosing the weights to be $\alpha_1 = \alpha_2 = 7/487$, $\alpha_3 = 76/2435$, $\alpha_4 = 99/4870$, $\alpha_5 = 29/2435$, $\alpha_6 = \alpha_8 = 29/4870$, $\alpha_7 = 87/974$, $\alpha_9 = 29/974$, $\alpha_{10} = 1901/4870$, $\alpha_{11} = 941/2435$, $\gamma_1 = \gamma_2 = 1 - \alpha_1\sqrt{2/3}$, $\gamma_3 = \alpha_1/\sqrt{2}$ and noting that $1 \le 1 + \mu - \mu_2 \le 1 + \mu$, $1 \le 1 + \mu_2 \le 1 + \mu$ as well as $|m_1|(58/|m_1| + 2) \le |m_1|64/3$ gives the statement. To bound the left-hand side in (96), we use the bound in the second-to-last display with $\tau = 1 - \exp(-\delta) = (\exp(\delta) - 1)/\exp(\delta)$ and the facts that $\mu_2\exp(\delta)(\alpha(\exp(\delta) - 1) + \mu_2\exp(\delta)) \le (1 + \mu_2)^2\exp(2\delta)$ for any $\alpha \in [0, 1]$, $\mu_2 \le 1 + \mu_2$ as well as $1 \le 1 + \sqrt{|m_2|/n} \le 2$, $1 - |m_1|/n \ge 1 - |m|/n$ and $1 \ge (\exp(\delta) - 1)^2/\exp(2\delta)$ together with the bounds in Lemma B.11 (ii) and get the upper bound as in the preceding display. Hence, we can bound the left-hand side in (96) from above by using the sum in (94) where we replace the sum of the third, fifth up to the eighth and tenth term by the sum in the preceding display. Choosing the weights to be $\alpha_1 = \alpha_2 = 2/75$, $\alpha_3 = 13/450$, $\alpha_4 = 17/450$, $\alpha_5 = 11/1000$, $\alpha_6 = \alpha_8 = 11/2000$, $\alpha_7 = 743/9000$, $\alpha_9 = 248/9000$, $\alpha_{10} = 649/1800$, $\alpha_{11} = 581/1500$, $\gamma_1 = \gamma_2 = 1 - \alpha_1/\sqrt{6}$, $\gamma_3 = \alpha_1/\sqrt{8}$ and noting that $1 \le 1 + \mu - \mu_2 \le 1 + \mu$ and that $1 \le 1 + \mu_2 \le 1 + \mu$ gives the statement.

□

Proof of Theorem 1. The left-hand side of (7) can be rewritten as

$$\mathbb{P}\left(\frac{\hat\rho^2(m)}{\rho^2(m)} \ge \exp(\varepsilon)\right) + \mathbb{P}\left(\frac{\hat\rho^2(m)}{\rho^2(m)} \le \exp(-\varepsilon)\right).$$

We can bound the sum in the preceding display for any $\alpha_1, \alpha_2 \in (0, 1)$ from above by

$$\begin{aligned}
&\mathbb{P}\left(\hat\rho^2(m)/r \ge \exp(\alpha_1\varepsilon)\right) + \mathbb{P}\left(\rho^2(m)/r \le \exp(-(1-\alpha_1)\varepsilon)\right)\\
&\quad+\mathbb{P}\left(\hat\rho^2(m)/r \le \exp(-\alpha_2\varepsilon)\right) + \mathbb{P}\left(\rho^2(m)/r \ge \exp((1-\alpha_2)\varepsilon)\right),
\end{aligned}$$

where r was defined in (64) in Lemma B.12. Note that $(\exp(\alpha\varepsilon) - 1)^2/\exp(\alpha\varepsilon) \ge (\exp(\alpha\varepsilon) - 1)^2/\exp(2\alpha\varepsilon) \ge \alpha^2\varepsilon^2/(1 + \varepsilon)^2$ holds for any $\alpha \in (0, 1)$ and $\varepsilon > 0$. Using this fact together with Lemma B.12 and Lemma B.14, we can bound the sum in the preceding display from above by

$$\begin{aligned}
&22\exp\left(-|m_1|\left(1-\frac{|m|}{n}\right)^3\frac{\alpha_1^2\varepsilon^2}{210(1+\varepsilon)^2}\right)\\
&\quad+(58+2|m_1|)\exp\left(-|m_1|\frac{|m_1|}{n}\left(1-\frac{|m|}{n}\right)^5\frac{(1-\alpha_1)^2\varepsilon^2}{13089(1+\varepsilon)^2}\right)\\
&\quad+22\exp\left(-|m_1|\left(1-\frac{|m|}{n}\right)^3\frac{\alpha_2^2\varepsilon^2}{840(1+\varepsilon)^2}\right)\\
&\quad+(58+2|m_1|)\exp\left(-|m_1|\frac{|m_1|}{n}\left(1-\frac{|m|}{n}\right)^5\frac{(1-\alpha_2)^2\varepsilon^2}{10477(1+\varepsilon)^2}\right).
\end{aligned}$$

Choosing $\alpha_1 \in (0,1)$ such that the exponents of the first two terms in the preceding display coincide gives $\alpha_1 = \sqrt{210}\sqrt{|m_1|/n}(1-|m|/n)/(\sqrt{210}\sqrt{|m_1|/n}(1-|m|/n)+\sqrt{13089})$. Doing the same for α_2 and noting that $\sqrt{|m_1|/n}(1-|m|/n) \leq \sqrt{|m_1|/n}(1-|m_1|/n) \leq 2/(3\sqrt{3})$ and noting that$160+4|m_1| \leq 160+4|m| = |m|(160/|m|+4) \leq 31|m|$ because $|m| \geq 6$ gives the result in (7).

To show the result in (8), we use Lemma B.13 and Lemma B.15 together with the inequalities discussed above and the facts that $1+\mu \leq (1+\mu)^2$ and that $1-|m|/n \leq 1$ to bound the sum in the second-to-last display from above by

$$\begin{aligned}
&22\exp\left(-n\left(1-\frac{|m|}{n}\right)^5\frac{\alpha_1^2\varepsilon^2}{210(1+\mu)^2(1+\varepsilon)^2}\right)\\
&\quad+(58+2|m_1|)\exp\left(-n\left(1-\frac{|m|}{n}\right)^5\frac{(1-\alpha_1)^2\varepsilon^2}{22534(1+\mu)^2(1+\varepsilon)^2}\right)\\
&\quad+22\exp\left(-n\left(1-\frac{|m|}{n}\right)^5\frac{\alpha_2^2\varepsilon^2}{840(1+\mu)^2(1+\varepsilon)^2}\right)\\
&\quad+(58+2|m_1|)\exp\left(-n\left(1-\frac{|m|}{n}\right)^5\frac{(1-\alpha_2)^2\varepsilon^2}{19371(1+\mu)^2(1+\varepsilon)^2}\right).
\end{aligned}$$

Choosing $\alpha_1, \alpha_2 \in (0,1)$ such that the exponents in the preceding display coincide gives $\alpha_1 = \sqrt{210}/(\sqrt{210}+\sqrt{22534})$ and $\alpha_2 = \sqrt{840}/(\sqrt{840}+\sqrt{19371})$. Noting that $160 + 4|m_1| \leq 31|m|$ as before gives the result.

□

C Technical details for Section 3

Proof of Lemma 2: By Bonferronis inequality the left-hand side in (9) is bounded from above by

$$\begin{aligned}
&\sum_{m\in\mathcal{M}_n}\mathbb{P}\left(\left|\log\frac{\hat{\rho}^2(m)}{\rho^2(m)}\right|\geq\varepsilon\right)\\
&\qquad\leq\sum_{m\in\mathcal{E}_n}31|m|\exp\left(-n\left(\frac{|m_1|}{n}\right)^2\left(1-\frac{|m|}{n}\right)^5\frac{\varepsilon^2}{14397(1+\varepsilon)^2}\right).
\end{aligned}$$

where the inequality follows from (7) in Theorem 1. Replacing $|m|$ by s_n and $|m_1|$ by r_n in every summand gives an upper bound for the sum on the right-hand side in the preceding display. The statement of the Lemma follows now by replacing the sum by the number of summands. Alternatively, we can use the upper bound in (8) in Theorem 1 to bound the left-hand side in (9) from above by

$$\sum_{m \in \mathcal{M}_n} 31|m| \exp\left(-n \ \left(1 - \frac{|m|}{n}\right)^5 \frac{\varepsilon^2}{28279(1+\varepsilon)^2(1+\mu(m))^2}\right),$$

where $\mu(m) = \theta'\Sigma(m)\theta/\sigma^2(m)$. We can upper bound the sum in the preceding display by replacing $|m|$ by s_n in every summand. Noting that $1 + \mu(m) = (\sigma^2(m) + \theta'\Sigma(m)\theta)/\sigma^2(m) = \mathrm{Var}(y)/\sigma^2(m) \leq \mathrm{Var}(y)/\sigma^2 \leq d$ and replacing the sum by the number of summands gives the result.

□

Proof of Corollary 3: Because m_n^* is a minimizer of $\rho^2(\cdot)$, the left-hand side in (11) equals $\mathbb{P}(\log(\rho^2(\hat{m}_n^*)/\rho^2(m_n^*)) \geq \varepsilon)$. On the event where $\hat{\rho}^2(m_n^*) > 0$ as well as $\hat{\rho}^2(\hat{m}^*) > 0$, which happens with probability 1, we have that

$$\log\frac{\rho^2(\hat{m}_n^*)}{\rho^2(m_n^*)} = \log\frac{\rho^2(\hat{m}_n^*)}{\hat{\rho}^2(\hat{m}_n^*)} + \log\frac{\hat{\rho}^2(\hat{m}_n^*)}{\hat{\rho}^2(m_n^*)} + \log\frac{\hat{\rho}^2(m_n^*)}{\rho^2(m_n^*)}.$$

Because $\hat{m}_n^*$ is a minimizer of $\hat{\rho}^2(\cdot)$, the second term is nonpositive and we can bound the left-hand side in (11) from above by

$$\mathbb{P}\left(\log\frac{\hat{\rho}^2(\hat{m}_n^*)}{\rho^2(\hat{m}_n^*)} \leq -\varepsilon/2\right) + \mathbb{P}\left(\log\frac{\hat{\rho}^2(m_n^*)}{\rho^2(m_n^*)} \geq \varepsilon/2\right). \tag{97}$$

Using Bonferroni's inequality, we can bound the sum in (97) from above by

$$\sum_{m \in \mathcal{M}_n}\left[\mathbb{P}\left(\log\frac{\hat{\rho}^2(m)}{\rho^2(m)} \geq \varepsilon/2\right) + \mathbb{P}\left(\log\frac{\hat{\rho}^2(m)}{\rho^2(m)} \leq -\varepsilon/2\right)\right].$$

Using the same arguments as in the proof of the previous lemma shows the statements in (11) and (13). The statements in (12) and (14) are direct consequences of Lemma 2. □

D Technical details for Section 4

Proof of Theorem 4: Recall that $\mathbb{L}(m)$ denotes a Gaussian distribution with zero mean and variance equal to $\rho^2(m)$, i.e., $\mathcal{N}(0, \rho^2(m))$, and that $\widehat{\mathbb{L}}(m)$ denotes a Gaussian distribution with zero mean and variance equal to $\hat{\rho}^2(m)$, i.e., $\mathcal{N}(0, \hat{\rho}^2(m))$. Using the fact that $\|\mathcal{N}(0, \hat{\rho}^2(m)) - \mathcal{N}(0, \rho^2(m))\|_{TV} = \|\mathcal{N}(0, \hat{\rho}^2(m)/\rho^2(m)) - \mathcal{N}(0,1)\|_{TV}$ together with Lemma D.1 in Leeb (2009), we have

$$\|\widehat{\mathbb{L}}(m) - \mathbb{L}(m)\|_{TV} \leq \frac{\left|\log(\hat{\rho}^2(m)/\rho^2(m))\right|}{\sqrt{2\pi\exp(1)}} \leq \frac{\left|\log(\hat{\rho}^2(m)/\rho^2(m))\right|}{4}. \tag{98}$$

Hence, the statements follow from Theorem 1 with 4ε replacing ε. □

Proof of Corollary 5: Using the inequality in (98), we can bound the left-hand side in (17) from above by

$$\mathbb{P}\left(\sup_{m\in\mathcal{M}_n} \|\widehat{\mathbb{L}}(m) - \mathbb{L}(m)\|_{TV} \geq \varepsilon\right) \leq \mathbb{P}\left(\sup_{m\in\mathcal{M}_n} \left|\log \frac{\hat{\rho}^2(m)}{\rho^2(m)}\right| \geq 4\varepsilon\right).$$

The result follows from Lemma 2 with 4ε replacing ε. □

Proof of Corollary 6: Note that we have $y^{(0)} \in \mathcal{I}(\hat{m}_n^*)$ if and only if $y^{(0)} - \hat{y}^{(0)}(m) \in \left[-Q_{(1-\alpha/2)}\hat{\rho}(\hat{m}_n^*), Q_{(1-\alpha/2)}\hat{\rho}(\hat{m}_n^*)\right]$. Denote the interval in squared brackets by A and let $\mathbb{L}(m, A)$ and $\widehat{\mathbb{L}}(m, A)$ denote the probability of A under $\mathbb{L}(m)$ and $\widehat{\mathbb{L}}(m)$, respectively. Note that the prediction interval was chosen such that $\widehat{\mathbb{L}}(\hat{m}_n^*, A) = 1 - \alpha$. Hence, the left-hand side of (19) can be rewritten as

$$\left|\widehat{\mathbb{L}}(\hat{m}_n^*, A) - \mathbb{L}(\hat{m}_n^*, A)\right|.$$

But this term is clearly bounded from above by the total variation distance of $\widehat{\mathbb{L}}(\hat{m}_n^*)$ and $\mathbb{L}(\hat{m}_n^*)$ and the result follows from Corollary 5. □

Proof of Corollary 7: Because m_n^* and $\hat{m}_n^*$ are minimizers of $\rho^2(\cdot)$ and $\hat{\rho}^2(\cdot)$, respectively, we have the following inequality

$$\log \frac{\hat{\rho}^2(\hat{m}_n^*)}{\rho^2(\hat{m}_n^*)} \leq \log \frac{\hat{\rho}^2(\hat{m}_n^*)}{\rho^2(m_n^*)} \leq \log \frac{\hat{\rho}^2(m_n^*)}{\rho^2(m_n^*)}.$$

We use the fact that $\log(\hat{\rho}(\hat{m}_n^*)/\rho(m_n^*)) = \log(\hat{\rho}^2(\hat{m}_n^*)/\rho^2(m_n^*))/2$ together with the second inequality in the preceding display followed by Bonferroni's inequality to get

$$\mathbb{P}\left(\log \frac{\hat{\rho}(\hat{m}_n^*)}{\rho(m_n^*)} \geq \varepsilon\right) \leq \sum_{m\in\mathcal{M}_n} \mathbb{P}\left(\log \frac{\hat{\rho}^2(m)}{\rho^2(m)} \geq 2\varepsilon\right).$$

Similarly but using the first inequality in the second-to-last display, we get

$$\mathbb{P}\left(\log \frac{\hat{\rho}^2(\hat{m}_n^*)}{\rho^2(m_n^*)} \leq -\varepsilon\right) \leq \sum_{m\in\mathcal{M}_n} \mathbb{P}\left(\log \frac{\hat{\rho}^2(m)}{\rho^2(m)} \leq -2\varepsilon\right).$$

Taken together, we can bound the left-hand side in (20) from above by

$$\sum_{m\in\mathcal{M}_n} \mathbb{P}\left(\left|\log \frac{\hat{\rho}^2(m)}{\rho^2(m)}\right| \geq 2\varepsilon\right)$$

and the result follows from the same arguments as in the proof of Lemma 2 with ε replaced by 2ε. □

References

F. Bachoc, H. Leeb, and B. M. Pötscher. Valid confidence intervals for post-model-selection predictors. *ArXiv e-prints*, 2017a. 1412.4605v3.

F. Bachoc, D. Preinerstorfer, and L. Steinberger. Uniformly valid confidence intervals post-model-selection. *ArXiv e-prints*, 2017b. 1611.01043v4.

A. Belloni, V. Chernozhukov, and C. B. Hansen. Inference for high-dimensional sparse econometric models. In D. Acemoglu, M. Arellano, and E. Dekel, editors, *Advances in Economics and Econometrics: Tenth World Congress*, volume 3 of *Econometric Society Monographs*, pages 245–295. Cambridge University Press, 2013.

A. Belloni, V. Chernozhukov, and C. B. Hansen. Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.*, 81:608–650, 2014.

R. Berk, L. D. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *Ann. Statist.*, 41(2):802–837, 2013.

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*, pages 317–366. North-Holland Publishing Company, Amsterdam, 2001.

N. Huber. *Shrinkage methods for prediction out-of-sample: Performance and selection of estimators.* PhD thesis, University of Vienna, Austria, 2013.

J. Johnston and J. E. DiNardo. *Econometric Methods.* McGraw-Hill, New York, 4th edition, 1997.

J. D. Lee, Y. Sun, and J. E. Taylor. On model selection consistency of regularized M-estimators. *Electron. J. Stat.*, 9:608–642, 2015.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the LASSO. *Ann. Statist.*, 44:907–927, 2016.

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 and inference: Facts and fiction. *Econometric Theory*, 21:21–59, 2005.

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.

H. Leeb and N. S. Senitschnig. Shrinkage estimators for prediction out-of-sample: selection of estimators and predictive inference. unpublished manuscript, 2015.

K. Mardia, J. Kent, and B. J.M. *Multivariate Analysis.* Academic Press, London-New York-Toronto-Sydney-San Francisco, 1979.

I. Milovič. *Conditional means of low-dimensional projections from high-dimensional data. Explicit error bounds.* PhD thesis, University of Vienna, 2015.

B. M. Pötscher. Effects of model selection on inference. *Econometric Theory*, 7:163–185, 1991.

B. M. Pötscher and U. Schneider. Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electron. J. Stat.*, 4:334–360, 2010.

U. Schneider. Confidence sets based on thresholding estimators in high-dimensional Gaussian regression models. *Econom. Rev.*, 35:1412–1455, 2016.

L. Steinberger and H. Leeb. On conditional moments of high-dimensional random vectors given lower-dimensional projections. *Bernoulli, forthcoming*, 2018.

R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.*, 111:600–620, 2016.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42:1166–1202, 2014.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 76:217–242, 2014.