The Annals of Statistics
2019, Vol. 47, No. 3, 1408–1442
https://doi.org/10.1214/18-AOS1719

PREDICTION WHEN FITTING SIMPLE MODELS TO HIGH-DIMENSIONAL DATA[1]

BY LUKAS STEINBERGER AND HANNES LEEB

Albert Ludwig University of Freiburg and University of Vienna

We study linear subset regression in the context of a high-dimensional linear model. Consider $y = \vartheta + \theta' z + \epsilon$ with univariate response y and a d-vector of random regressors z, and a submodel where y is regressed on a set of p explanatory variables that are given by $x = M'z$, for some $d \times p$ matrix M. Here, "high-dimensional" means that the number d of available explanatory variables in the overall model is much larger than the number p of variables in the submodel. In this paper, we present Pinsker-type results for prediction of y given x. In particular, we show that the mean squared prediction error of the best linear predictor of y given x is close to the mean squared prediction error of the corresponding Bayes predictor $\mathbb{E}[y\|x]$, provided only that $p/\log d$ is small. We also show that the mean squared prediction error of the (feasible) least-squares predictor computed from n independent observations of (y, x) is close to that of the Bayes predictor, provided only that both $p/\log d$ and p/n are small. Our results hold uniformly in the regression parameters and over large collections of distributions for the design variables z.

1. Introduction. Fitting simple models to complex, high-dimensional data is often motivated by the belief, or assumption, that the data does indeed admit a simple representation. Theoretical analyses of simple, or sparse, modeling methods, in particular, typically rely on the assumption that the true data-generating process can be described, or at least be closely approximated, by a simple model. In situations where the sample size is relatively small, however, it is often infeasible to substantiate such beliefs or assumptions from data. For prediction, the results in this paper justify searching for, and working with, parsimonious representations without relying on the assumption that the "truth" is sparse or simple.

Given an overall linear regression model with d explanatory variables and a submodel with p explanatory variables, we consider the best linear predictor for the response given those explanatory variables that are "active" in the submodel, and the corresponding Bayes predictor. We show that the best linear predictor is comparable to the Bayes predictor in terms of relative mean squared prediction error, irrespective of, that is, uniformly in, the regression parameters, provided only that $p/\log d$ is small. This statement moreover holds uniformly over a large

Received May 2016; revised April 2017.

[1]Supported in part by FWF projects P 26354-N26 and P 28233-N32.

MSC2010 subject classifications. Primary 62H99; secondary 62F99, 62G99.

Key words and phrases. Pinsker theorem, best linear predictor, Bayes predictor, linear subset regression, non-Gaussian data, high-dimensional models, small sample size.

collection of distributions for the explanatory variables and the error term in the model. We also provide similar results for the case where the coefficients of the best linear predictor are estimated from a data sample of size n, provided that both $p/\log d$ and p/n are small.

The best linear predictor in a possibly misspecified (sub) model and its coefficients are well-studied objects in the statistics literature, certainly since Huber (1967), and recently gained new popularity as witnessed by, for example, Abadie, Imbens and Zheng (2014), Bachoc, Leeb and Pötscher (2015), Berk et al. (2013), Brannath and Scharpenberg (2014), Buja et al. (2014), Greenshtein and Ritov (2004), Leeb, Pötscher and Ewald (2015), Leeb (2008, 2009), Lee et al. (2016), Taylor et al. (2014). Near equivalence of the best linear predictor and the Bayes predictor, as we establish here, is related to the celebrated result of Pinsker (1980); the relation of our result to Pinsker's theorem will be discussed in detail later. The present paper is based on the PhD thesis of Steinberger (2015). On a technical level, we expand and further analyze findings of Steinberger and Leeb (2018), who in turn rely on Leeb (2013) and Hall and Li (1993) (see also [Diaconis and Freedman (1984)] as well as [Dümbgen and Del Conte-Zerial (2013)]). We also rely on results about extreme eigenvalues of large sample-covariance matrices by Srivastava and Vershynin (2013).

The rest of the paper is organized as follows. In Section 2, we give an outline of our findings, put them in context with existing results, and discuss some immediate consequences. A detailed description of our main results and assumptions is given in Section 3. In Section 4, we provide a high-level explanation of the mechanisms that facilitate our results. Section 5 gives an explicit analysis of the prediction problem in a simple low-dimensional setting, and Section 6 outlines how our results can be used to deal with several potential candidate models. Lastly, some additional remarks are collected in Section 7. All proofs are contained in the Appendix.

2. Overview. Throughout, we consider the linear model

$$y = \vartheta + \theta' z + \epsilon \tag{1}$$

with $\vartheta \in \mathbb{R}$ and $\theta \in \mathbb{R}^d$ for some $d \in \mathbb{N}$. We assume that the error ϵ is independent of z, with mean zero and finite variance; its distribution will be denoted by $\mathcal{L}(\epsilon)$. Moreover, we assume that the vector of regressors z has mean $\mu \in \mathbb{R}^d$ and positive definite variance/covariance matrix Σ. Our model assumptions are further discussed in Remark 7.1. No additional restrictions will be placed on the regression coefficients ϑ and θ, on the moments μ and Σ, or on the error distribution $\mathcal{L}(\epsilon)$.

We do place some assumptions on the distribution of the explanatory variables. First, we assume that z can be written as an affine transformation of independent random variables (see also Remark 7.2). With this, we can represent z as

$$z = \mu + \Sigma^{1/2} R \tilde{z} \tag{2}$$

for a vector $\tilde{z}$ with independent (but not necessarily identically distributed) components so that $\mathbb{E}[\tilde{z}] = 0$ and $\mathbb{E}[\tilde{z}\tilde{z}'] = I_d$, where $\Sigma^{1/2}$ is the positive definite and symmetric square root of Σ, and where R is an orthogonal matrix. Second, we assume that $\tilde{z}$ has a Lebesgue density, which we denote by $f_{\tilde{z}}$, with bounded marginal densities and finite marginal moments of sufficiently high order; cf. Section 3 for details.

The distribution of (y, z) in (1)–(2) is characterized by ϑ and θ, by $\mathcal{L}(\epsilon)$, by Σ and μ, by $f_{\tilde{z}}$, and by R. For the expository discussion in this section, we keep all these quantities fixed except for the regression coefficients $\vartheta \in \mathbb{R}$ and $\theta \in \mathbb{R}^p$, and the orthogonal $d \times d$ matrix R. [We note that the model (1)–(2) also covers situations where z contains lagged dependent variables; cf. Remark 7.3.]

Consider a submodel where y is regressed on x, with x given by $x = M'z$ for some full-rank $d \times p$ matrix M with $p < d$. For example, M can be a selection matrix that picks out p components of the d-vector z. (See also Remark 3.5 and Remark 3.6, as well as Section 6 regarding several submodels.) Submodels with regressors of the form $x = M'z$ also occur in principal component regression, partial least squares, and certain sufficient dimension reduction methods. We are particularly interested in situations where d is *much* larger than p, that is, $p \ll d$.

Our goal is to compare linear and nonlinear predictors of y given x. In particular, we study the Bayes predictor $\mathbb{E}[y\|x]$ and the best linear predictor $\alpha + \beta' x$, where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ minimize $\mathbb{E}[(y - (\alpha + \beta' x))^2]$. Note that both predictors depend on the model parameters, although this dependence is not explicitly shown in the notation. Their corresponding mean squared errors are

$$\mathcal{R}_N(\theta, R) = \mathbb{E}\big[\big(y - \mathbb{E}[y\|x]\big)^2\big]$$

and

$$\mathcal{R}_L(\theta) = \mathbb{E}\big[\big(y - (\alpha + \beta' x)\big)^2\big].$$

It is elementary to verify that both $\mathcal{R}_N$ and $\mathcal{R}_L$ do not depend on ϑ, and that $\mathcal{R}_L$ equals $\|(I_d - P_{\Sigma^{1/2}M})\Sigma^{1/2}\theta\|^2 + \mathrm{Var}(\epsilon)$, and hence does not depend on R, where $P_{\cdot}$ denotes the orthogonal projection onto the column space of the indicated matrix. In particular, $\mathcal{R}_L$ is minimized if $\Sigma^{1/2}\theta$ is in the span of $\Sigma^{1/2}M$, that is, if $\theta = M\beta$. Similar to Pinsker (1980), we study the risk ratio $\mathcal{R}_N/\mathcal{R}_L$. Note that we always have $\mathrm{Var}(\epsilon) \le \mathcal{R}_N \le \mathcal{R}_L \le \mathrm{Var}(y)$, so that the risk ratio $\mathcal{R}_N/\mathcal{R}_L$ is always bounded by 1. In the degenerate case where $\mathcal{R}_L = 0$, and hence also $\mathcal{R}_N = 0$, we set $\mathcal{R}_N/\mathcal{R}_L = 1$. To provide some context for our results, we next discuss two conditions which guarantee that $\mathcal{R}_N/\mathcal{R}_L = 1$.

EXAMPLE 2.1. The Bayes predictor and the best linear predictor of y given x coincide if θ satisfies $\theta' z = \beta' x$ almost surely, that is, if $\theta = M\beta$ (irrespective of ϑ and of the distribution of z or $\tilde{z}$). In that case, θ is often called "sparse" if M is a selection matrix, because then β and x are subvectors of θ and z, respectively,

and the remaining $d - p$ components of θ are equal to zero. Under such a sparsity assumption, we have $\mathrm{Var}(\epsilon) = \mathcal{R}_N = \mathcal{R}_L$. For an overview of the well-developed theory on sparse modeling, including methods for inference on the true regression coefficients ϑ and θ, we refer to the monograph of Bühlmann and van de Geer (2011). In situations where d exceeds the sample size, it is typically infeasible to ascertain whether or not the true parameter vector θ is indeed sparse or, more generally, whether or not $\theta = M\beta$ holds (for the given matrix M).

EXAMPLE 2.2. The Bayes predictor and the best linear predictor also coincide (irrespective of ϑ and θ) if our distributional assumptions on z are replaced by the requirement that z is Gaussian or, more generally, by the requirement that the law of $\tilde{z}$ is spherically symmetric. In that case, for any matrix A of appropriate dimension, conditional means of the form $\mathbb{E}[\tilde{z} \| A\tilde{z}]$ are linear functions of the conditioning variable, and hence $\mathbb{E}[y \| x] = \alpha + \beta' x$. Under this assumption, we have $\mathrm{Var}(\epsilon) \leq \mathcal{R}_N = \mathcal{R}_L$, and the inequality is typically strict. This angle is further developed in Leeb (2008, 2009), with a focus on prediction. [In fact, the class of elliptically contoured distributions is characterized by the property that conditional means are linear; cf. Eaton (1986).] But, similar to before, in situations where d exceeds the sample size, it is often infeasible to judge if z is Gaussian or if the law of $\tilde{z}$ is spherically symmetric.

The conditions discussed in the two preceding examples are satisfied by a relatively small subclass of all data-generating processes as in (1)–(2), namely by those with θ satisfying $\theta = M\beta$, and by those with $\tilde{z}$ being spherically symmetric, respectively. We here show that the mean squared errors of the Bayes predictor and the best linear predictor, that is, $\mathcal{R}_N$ and $\mathcal{R}_L$, are close to each other, uniformly in θ and uniformly over a "large" collection of design distributions of the form (2), provided only that $p/\log d$ is small. To express this more formally, write $\mathcal{O}_d$ for the collection of all orthogonal $d \times d$ matrices, and write ν_d for the uniform distribution on $\mathcal{O}_d$ (i.e., ν_d is the normalized Haar measure on the orthogonal group $\mathcal{O}_d$). Our results entail that there exists a (measurable) set $\mathbb{U} \subseteq \mathcal{O}_d$, so that both

$$\sup_{R \in \mathbb{U}} \sup_{\theta \in \mathbb{R}^d} 1 - \frac{\mathcal{R}_N(\theta, R)}{\mathcal{R}_L(\theta)} = O\left(\left(\frac{p}{\log d}\right)^{5/6}\right) \tag{3}$$

and

$$\nu_d(\mathbb{U}) = 1 + O\big(d^{-\frac{1}{12}(1 - c\frac{p}{\log d})}\big) \tag{4}$$

hold, where the constant c as well as the constants implicit in the O-terms depend only on the univariate marginal densities of $f_{\tilde{z}}$; see Theorem 3.1(i) for the detailed, and stronger, formal statement. In Theorem 3.4, we also provide a similar result for the case where α and β are replaced by estimators, that is, where the infeasible best linear predictor is replaced by the feasible ordinary least-squares predictor.

The result in (3)–(4) is nonstandard in the sense that it does not explicitly characterize the set of data-generating processes for which $\mathcal{R}_N/\mathcal{R}_L$ is close to one, that is, it does not explicitly describe the set $\mathbb{U}$. The set $\mathbb{U}$ in (3)–(4) depends on the distribution of $\tilde{z}$ in a complicated way, and our method of proof does not deliver a simple explicit characterization of this set. We can, however, control the size of $\mathbb{U}$ through (4). In contrast, the conditions that guarantee that $\mathcal{R}_N/\mathcal{R}_L = 1$ discussed in Examples 2.1 and 2.2, namely sparsity and spherical symmetry, are simple to characterize, but the collection of data-generating processes meeting these conditions is comparatively small. Furthermore, even if an explicit characterization of $\mathbb{U}$ were available, it would typically be of little use in statistical practice in situations where d exceeds the sample size, as it is then difficult to judge whether or not the data were generated by a model as in (1)–(2) with $R \in \mathbb{U}$. More importantly, however, we can characterize the size of $\mathbb{U}$ with (4). In particular, $\mathbb{U}$ is guaranteed to be large provided only that $p/\log d$ is sufficiently small. See also Section 5 for a detailed analysis of $\mathcal{R}_N(\theta, R)/\mathcal{R}_L(\theta)$, and hence also of $\mathbb{U}$, in a simple low-dimensional setting.

Although technically different, our results are similar, in spirit, to those of Pinsker (1980), in the sense that they exhibit a certain equivalence of linear and nonlinear methods in high-dimensional inference problems. See Remark 3.2 for a more detailed discussion of the similarities and differences of our results compared to Pinsker (1980). Moreover, a phenomenon qualitatively related to our findings is studied by El Karoui (2010) in the context of principal components analysis (PCA), who showed that, for high-dimensional observations following classical random matrix models, nonlinear versions of PCA using kernel matrices essentially perform a standard linear PCA. Our present results add another piece to this picture of linear methods performing comparable to nonlinear methods in certain high-dimensional settings.

We stress that our results cover the *relative* size of $\mathcal{R}_N$ and $\mathcal{R}_L$, but neither the *absolute* size of either quantity nor their absolute difference: If the upper bound in (3) is small and $R \in \mathbb{U}$, then the linear predictor $\alpha + \beta' x$ is close to best possible in the class of all predictors of y given x; but this does not restrict the absolute performance of any of these predictors. For example, θ may be such that the best linear predictor given x is constant in x, in which case $\mathcal{R}_L$ equals $\mathrm{Var}(y)$, and $\mathcal{R}_N/\mathrm{Var}(y)$ is close to 1 if $R \in \mathbb{U}$ and the upper bound in (3) is small. On the other extreme, both $\mathcal{R}_L$ and $\mathcal{R}_N$ equal $\mathrm{Var}(\epsilon)$ if $\theta = M\beta$. The usefulness of x for predicting y depends on θ, that is, on a parameter that can not be estimated in situations where d exceeds the sample size and the parameter space is $\mathbb{R}^d$. Nevertheless, given that one is committed to using x to predict y, our results justify focusing on linear predictors, provided that $p/\log d$ is small. This also prompts the question for model selection procedures. In Section 6, we show how our results can be used to handle several submodels, and we briefly sketch how a model selection procedure can then be used to select one of them, for example, based on estimated predictive performance. A more comprehensive treatment of model

selection in this context, however, poses a number of challenges that are beyond the scope of this paper.

3. Main results. Throughout, we consider the model (1)–(2), such that the error ϵ has mean zero, finite variance, and is independent of z. For the vector $\tilde{z}$ with independent components in (2), we will assume that its Lebesgue density $f_{\tilde{z}}$ belongs to one of the classes $\mathcal{F}_{d,k}(D, E)$ that are defined in the next paragraph. The distribution of (y, z) is characterized by the regression coefficients $\vartheta \in \mathbb{R}$ and $\theta \in \mathbb{R}^d$; by the error distribution $\mathcal{L}(\epsilon)$ with mean zero and finite variance; by $\mu \in \mathbb{R}^d$; by the symmetric positive definite $d \times d$ matrix Σ; by the density $f_{\tilde{z}}$; and by the orthogonal $d \times d$ matrix R.

For fixed $k \in \mathbb{N}$ and positive finite constants D and E, write $\mathcal{F}_{d,k}(D, E)$ for the class of all Lebesgue densities on $\mathbb{R}^d$ that are products of univariate densities, so that each such marginal density is bounded by D from above, and so that each univariate marginal density has mean zero, variance one and absolute moments of order up to k bounded from above by E. In the results that follow, we will assume that $f_{\tilde{z}}$ belongs to $\mathcal{F}_{d,k}(D, E)$ for appropriate constants d, k, D and E.

Consider a submodel where y is regressed on x, with x given by $x = M'z$ for some full-rank $d \times p$ matrix M with $p < d$. We first compare two (infeasible) predictors that are functions of x, namely the Bayes predictor $\mathbb{E}[y\|x]$ and the best linear predictor $\alpha + \beta' x$, where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ minimize $\mathbb{E}[(y - (\alpha + \beta' x))^2]$. Recall that their respective mean squared errors are

$$\mathcal{R}_N = \mathcal{R}_N\big(M, \theta, \mathcal{L}(\epsilon), \Sigma, f_{\tilde{z}}, R\big) = \mathbb{E}\big[\big(y - \mathbb{E}[y\|x]\big)^2\big]$$

and

$$\mathcal{R}_L = \mathcal{R}_L\big(M, \theta, \mathcal{L}(\epsilon), \Sigma, f_{\tilde{z}}\big) = \mathbb{E}\big[\big(y - (\alpha + \beta' x)\big)^2\big].$$

In the preceding display, the expressions in the middle stress that the risks depend on M because $x = M'z$; on $\theta \in \mathbb{R}^d$; on the distribution of ϵ through $\mathrm{Var}(\epsilon)$; on the positive definite $d \times d$ matrix Σ; and on the density $f_{\tilde{z}}$ of $\tilde{z}$; moreover, $\mathcal{R}_N$ also depends on the matrix $R \in \mathcal{O}_d$. It is elementary to verify that $\mathcal{R}_N$ and $\mathcal{R}_L$ [and also $\mathcal{R}_N(x)$ and $\mathcal{R}_L(x)$, which follow] do not depend on the mean parameters ϑ and μ. Similarly, we also consider the corresponding conditional risks given x, that is,

$$\mathcal{R}_N(x) = \mathcal{R}_N\big(M, \theta, \mathcal{L}(\epsilon), \Sigma, f_{\tilde{z}}, R|x\big) = \mathbb{E}\big[\big(y - \mathbb{E}\big[y\|x\big]\big)^2\|x\big]$$

and

$$\mathcal{R}_L(x) = \mathcal{R}_L\big(M, \theta, \mathcal{L}(\epsilon), \Sigma, f_{\tilde{z}}, R|x\big) = \mathbb{E}\big[\big(y - (\alpha + \beta' x)\big)^2\|x\big].$$

The distributions of the random variables in the preceding two displays depend on M, θ, the distribution of ϵ through $\mathrm{Var}(\epsilon)$, Σ, $f_{\tilde{z}}$ and on R. Also note that $\mathcal{R}_N \le \mathcal{R}_L$ and $\mathcal{R}_N(x) \le \mathcal{R}_L(x)$, almost surely, by construction. We adopt the convention that the ratios $\mathcal{R}_N/\mathcal{R}_L$ and $\mathcal{R}_N(x)/\mathcal{R}_L(x)$ are set equal to 1 whenever the respective denominator vanishes. Thus, these ratios are well defined and do not exceed 1.

THEOREM 3.1. *Fix positive integers p and d with $p < d$, and finite constants $D \geq 1$ and $E \geq 1$:*

(i) *For each full-rank $d \times p$ matrix M, each symmetric positive definite $d \times d$ matrix Σ, and each $f_{\tilde{z}} \in \mathcal{F}_{d,12}(D, E)$, there exists a Borel set $\mathbb{U} = \mathbb{U}(M, \Sigma, f_{\tilde{z}}) \subseteq \mathcal{O}_d$ that depends only on M, Σ and $f_{\tilde{z}}$, such that*

$$\sup_M \sup_{\theta, \mathcal{L}(\epsilon), \Sigma} \sup_{f_{\tilde{z}} \in \mathcal{F}_{d,12}(D,E)} \sup_{R \in \mathbb{U}} 1 - \frac{\mathcal{R}_N}{\mathcal{R}_L} \leq K_1 \left(\frac{p}{\log d} \right)^{5/6}$$

and such that

$$\sup_M \sup_{\Sigma} \sup_{f_{\tilde{z}} \in \mathcal{F}_{d,12}(D,E)} \nu_d\big(\mathbb{U}^c\big) \leq L_1 d^{-\frac{1}{12}(1 - c_1 \frac{p}{\log d})}.$$

(ii) *For each full-rank $d \times p$ matrix M, each symmetric positive definite $d \times d$ matrix Σ, and each $f_{\tilde{z}} \in \mathcal{F}_{d,20}(D, E)$, there exists a Borel set $\mathbb{V} = \mathbb{V}(M, \Sigma, f_{\tilde{z}}) \subseteq \mathcal{O}_d$ that depends only on M, Σ and $f_{\tilde{z}}$, such that*

$$\sup_M \sup_{\theta, \mathcal{L}(\epsilon), \Sigma} \sup_{f_{\tilde{z}} \in \mathcal{F}_{d,20}(D,E)} \sup_{R \in \mathbb{V}} \mathbb{P}\left(1 - \frac{\mathcal{R}_N(x)}{\mathcal{R}_L(x)} > t \right) \leq \sqrt{2} d^{-\frac{1}{12}} t^{-\frac{1}{2}} + K_2 \frac{p}{\log d}$$

holds for each $t > 0$ and such that

$$\sup_M \sup_{\Sigma} \sup_{f_{\tilde{z}} \in \mathcal{F}_{d,20}(D,E)} \nu_d\big(\mathbb{V}^c\big) \leq L_2 d^{-\frac{1}{20}(1 - c_2 \frac{p}{\log d})}.$$

(iii) *The sets $\mathbb{U}$ and $\mathbb{V}$ in parts* (i) *and* (ii) *satisfy $\mathbb{U}(M, \Sigma, f_{\tilde{z}}) = R_0 \mathbb{U}(M_0, I_d, f_{\tilde{z}})$ and $\mathbb{V}(M, \Sigma, f_{\tilde{z}}) = R_0 \mathbb{V}(M_0, I_d, f_{\tilde{z}})$, where M_0 consists of the first p columns of I_d, and R_0 is any orthogonal matrix whose first p columns are a basis for the column span of $\Sigma^{1/2} M$. Furthermore, both $\mathbb{U}$ and $\mathbb{V}$ are right-equivariant in the sense that $\mathbb{U}(M, \Sigma, f_{R\tilde{z}}) = \mathbb{U}(M, \Sigma, f_{\tilde{z}})R$ and $\mathbb{V}(M, \Sigma, f_{R\tilde{z}}) = \mathbb{V}(M, \Sigma, f_{\tilde{z}})R$, for every $R \in \mathcal{O}_d$.*

In the displays of part (i) *and* (ii), *$K_1 = K_1(D, E)$, $K_2 = K_2(D)$ and $L_i = L_i(E)$, $c_i = c_i(D)$, for $i = 1, 2$, are positive and finite constants that depend only on the indicated quantities, and suprema are taken over all full-rank $d \times p$ matrices M, over $\theta \in \mathbb{R}^d$, over $\mathcal{L}(\epsilon)$ so that ϵ has zero mean and finite variance, and over all symmetric positive definite $d \times d$ matrices Σ, when indicated.*

REMARK 3.2 (Regarding Pinsker's Theorem). Qualitatively, Theorem 3.1 tells a similar story as the classical linear minimax result of Pinsker (1980) insofar as linear procedures are shown to be almost best possible, in a certain sense, when the dimension of the parameter space is large. Pinsker's original contribution was to compare the linear and the overall minimax risk of estimation

in the Gaussian sequence model over ℓ^2 ellipsoids in the low-noise limit. His results also imply the asymptotic equivalence of the linear and nonlinear minimax risk of estimation in the Gaussian location model $\mathcal{N}(\theta, \sigma^2 I_d)$, over balls $\Theta_d(c) := \{\theta \in \mathbb{R}^d : \|\theta\|^2 \leq dc\}$, as $d \to \infty$ [cf. Beran and Dümbgen (1998)]. Despite the qualitative similarity between our results and Pinsker's, there are some fundamental differences: First and foremost, Pinsker's results crucially rely on Gaussianity whereas the main feature of our results is that they hold in many non-Gaussian situations (and become trivial in the Gaussian case). Second, while Pinsker studied the worst case risk, Theorem 3.1 provides bounds for the relative prediction risk that hold uniformly over the indicated parameters. [In our setting, a worst-case comparison of predictors does not make sense because risks are unbounded. Moreover, even if bounded risks are imposed, one can always choose parameters such that x and $\theta' z$ are independent, and thus $\operatorname{Var}(y) = \mathcal{R}_L = \mathcal{R}_N$; cf. Remark 7.5 for details.] Third, we here study the prediction problem in linear regression rather than the estimation of a location parameter. Lastly, Pinsker considered parameter spaces of finite diameter, whereas our parameter space for the regression parameter θ is all of $\mathbb{R}^d$.

REMARK 3.3 (Regarding tightness of bounds). Because the bounds in Theorem 3.1 hold uniformly over a large class of data-generating processes, it can occur that said bounds are very conservative for a specific data-generating process. The bounds in Theorem 3.1 are the best possible that our current technique of proof delivers. But detailed inspection of the proofs of results in Steinberger and Leeb (2018), which we rely on, suggests that our bounds are not tight. Tighter bounds can be obtained under appropriately stronger assumptions [see, e.g., Section 3.2 in Steinberger and Leeb (2018)] or possibly by an altogether different method of proof. Further results in that direction are currently work in progress.

The predictors considered so far are infeasible. Consider now a sample $(y_i, x_i)_{i=1}^n$ of n independent observations that are distributed as, and independent from, (y, x). We study the feasible linear predictor $\hat{\alpha}_n + \hat{\beta}_n' x$, where $\hat{\alpha}_n$ and $\hat{\beta}_n$ are the ordinary least-squares estimators for α and β, respectively, obtained by regressing $Y = (y_1, \ldots, y_n)'$ on $X = (x_1, \ldots, x_n)'$ (including an intercept). The corresponding prediction risk is

$$\begin{aligned}\mathcal{R}_{\mathrm{OLS}}(X, Y) &= \mathcal{R}_{\mathrm{OLS}}\big(M, \theta, \mathcal{L}(\epsilon), \Sigma, f_{\tilde{z}}, R | X, Y\big) \\ &= \mathbb{E}\big[\big(y - (\hat{\alpha}_n + \hat{\beta}_n' x)\big)^2 \| X, Y\big]\end{aligned}$$

and the corresponding conditional prediction risk given x is

$$\begin{aligned}\mathcal{R}_{\mathrm{OLS}}(X, Y, x) &= \mathcal{R}_{\mathrm{OLS}}\big(M, \theta, \mathcal{L}(\epsilon), \Sigma, f_{\tilde{z}}, R | X, Y, x\big) \\ &= \mathbb{E}\big[\big(y - (\hat{\alpha}_n + \hat{\beta}_n' x)\big)^2 \| X, Y, x\big].\end{aligned}$$

Similar to before, it is easy to see that the distributions of $\mathcal{R}_{\mathrm{OLS}}(X,Y)$ and $\mathcal{R}_{\mathrm{OLS}}(X,Y,x)$ do not depend on ϑ and μ (cf. Lemma A.1). Moreover, since $\mathcal{R}_N \le \mathcal{R}_{\mathrm{OLS}}(X,Y)$, almost surely, it makes sense to impose the convention that $\mathcal{R}_N/\mathcal{R}_{\mathrm{OLS}}(X,Y) = 1$ whenever the denominator is equal to zero, and similarly, because of $\mathcal{R}_N(x) \le \mathcal{R}_{\mathrm{OLS}}(X,Y,x)$, almost surely, the same convention is used for the ratio $\mathcal{R}_N(x)/\mathcal{R}_{\mathrm{OLS}}(X,Y,x)$.

THEOREM 3.4. *Fix positive integers n, p and d with $p < d$, and finite constants $D \ge 1$ and $E \ge 1$. There exists a finite positive constant $L_0 = L_0(E)$ that depends only on E, such that for $n > L_0 p$, the following statements hold true*:

(i) *For the same Borel set $\mathbb{U} = \mathbb{U}(M, \Sigma, f_{\tilde{z}}) \subseteq \mathcal{O}_d$ as in Theorem 3.1*(i), *and for every $t > 0$, we have*

$$\sup_M \sup_{\theta, \mathcal{L}(\epsilon), \Sigma} \sup_{f_{\tilde{z}} \in \mathcal{F}_{d,12}(D,E)} \sup_{R \in \mathbb{U}} \mathbb{P}\left(1 - \frac{\mathcal{R}_N}{\mathcal{R}_{\mathrm{OLS}}(X,Y)} > t\right) \le \left(\frac{p}{n}\right)^{1/3} L_3(t) + \left(\frac{p}{\log d}\right)^{5/6} 2K_1/t.$$

(ii) *For the same Borel set, $\mathbb{V} = \mathbb{V}(M, \Sigma, f_{\tilde{z}}) \subseteq \mathcal{O}_d$ as in Theorem 3.1*(ii), *and for every $t > 0$, we have*

$$\sup_M \sup_{\theta, \mathcal{L}(\epsilon), \Sigma} \sup_{f_{\tilde{z}} \in \mathcal{F}_{d,20}(D,E)} \sup_{R \in \mathbb{V}} \mathbb{P}\left(1 - \frac{\mathcal{R}_N(x)}{\mathcal{R}_{\mathrm{OLS}}(X,Y,x)} > t\right) \le \left(\frac{p}{n}\right)^{1/3} L_4(t) + \frac{p}{\log d} K_3(t).$$

In the preceding displays, the constant $K_1 = K_1(D,E)$ is the same as in Theorem 3.1(i), *$K_3(t) = K_3(D,t)$ and $L_j(t) = L_j(E,t)$, for $j = 3, 4$, depend only on the indicated quantities and suprema are taken over all full-rank $d \times p$ matrices M, over $\theta \in \mathbb{R}^d$, over $\mathcal{L}(\epsilon)$ so that ϵ has zero mean and finite variance, and over all symmetric positive definite $d \times d$ matrices Σ.*

Theorem 3.4 provides a comparison of the least-squares predictor of y given x with the corresponding Bayes predictor, which is infeasible, and the results are similar to Theorem 3.1: Provided that p/n and $p/\log d$ are both small, the risk of the feasible predictor is close to that of the infeasible one, uniformly over large portions of the parameter space. Note that the upper bounds in Theorem 3.4 cannot be expected to be small in statistically more challenging scenarii where p/n is not close to zero, because then the absolute estimation errors $|\hat{\alpha}_n - \alpha|$ and $\|\hat{\beta}_n - \beta\|$ are not small in probability; such scenarii will be studied elsewhere. Also, Theorem 3.4 should be compared to the results of Greenshtein and Ritov (2004), who, in essence, show that the best sparse linear predictor and a (sparse) predictor

based on the LASSO estimator are comparable, in terms of prediction risk, without requiring that the true model is sparse. In other words, no sufficiently sparse *linear* procedure can significantly outperform the LASSO. Our results suggest that this LASSO predictor can perform almost as well as the Bayes predictor in certain situations, thereby suggesting that no other sparse procedure (linear or not) can significantly outperform the LASSO in terms of relative mean squared prediction error.

REMARK 3.5. In order to use our results, the vector z in (2), that is, the vector of all explanatory variables, need not be observed in its entirety; only observations of y and $x = M'z$ (or of i.i.d. copies thereof) are needed. Also in practice, potentially influential explanatory variables may go unobserved.

REMARK 3.6. In all our results, we have assumed that $p < d$. In the case where $p = d$, Theorem 3.1 is trivial because the Bayes predictor and the best linear predictor coincide in that case, and a statement similar to Theorem 3.4 holds in view of uniform consistency of the ordinary least-squares predictor $\hat{\alpha} + \hat{\beta}'x$ for the Bayes predictor $\mathbb{E}[y\|x]$.

REMARK 3.7. All the constants $K_1, K_2, K_3, L_0, L_1, L_2, L_3, L_4$ and c_1, c_2 in Theorems 3.1 and 3.4 can be obtained explicitly upon detailed inspection of the proofs. Moreover, some of the constants appearing in Theorems 3.1 and 3.4 also depend on the threshold t, but our upper bounds do not necessarily vanish as $t \to \infty$. Note, however, that all tail probabilities under consideration are trivially bounded by zero whenever $t \geq 1$.

4. Outline of proof. Consider the setup of Section 2. It is easy to see that the risks $\mathcal{R}_L$ and $\mathcal{R}_N$ do not depend on the mean parameters ϑ and μ, and hence we set them both equal to zero throughout this section. This, in particular, implies that $\alpha = 0$. A standard computation yields the value of β, that is, $\beta = (M'\Sigma M)^{-1}M'\Sigma\theta$, and thus

$$y - \beta'x = \theta'\Sigma^{1/2}(I_d - P_{\Sigma^{1/2}M})R\tilde{z} + \epsilon.$$

Set $v := (I_d - P_{\Sigma^{1/2}M})\Sigma^{1/2}\theta \in \mathbb{R}^d$ so that $y - \beta'x = v'R\tilde{z} + \epsilon$. With this, we have $\mathbb{E}[y\|x] - \beta'x = v'R\mathbb{E}[\tilde{z}\|x]$ and $\mathcal{R}_L = \mathbb{E}[(y - \beta'x)^2] = \|v\|^2 + \mathrm{Var}(\epsilon)$. We arrive at

$$\begin{aligned} 1 - \frac{\mathcal{R}_N}{\mathcal{R}_L} &= \frac{\mathbb{E}[(\mathbb{E}[y\|x] - \beta'x)^2]}{\mathbb{E}[(y - \beta'x)^2]} \\ &= \mathbb{E}\left[\left(\frac{v'}{\sqrt{\|v\|^2 + \mathrm{Var}(\epsilon)}}R\mathbb{E}[\tilde{z}\|B'\tilde{z}]\right)^2\right], \end{aligned}$$

where $B := R'\Sigma^{1/2}M(M'\Sigma M)^{-1/2} \in \mathbb{R}^{d\times p}$. Indeed, the first equality follows by expanding the squares and noting that $\mathbb{E}[y\mathbb{E}[y\|x]] = \mathbb{E}[\mathbb{E}[y\|x]^2]$ and $\mathbb{E}[y\beta'x] = \mathbb{E}[\mathbb{E}[y\|x]\beta'x]$, and the second equality holds in view of the equalities involving v derived earlier and because conditioning on $B'\tilde{z} = (M'\Sigma M)^{-1/2}x$ is equivalent to conditioning on x.

The dependence of the vector v and the matrix B on the model M and the parameters θ and Σ is simple to describe. In particular, $v = 0$ if, and only if, $\theta \in \operatorname{span} M$, that is, iff the model M is correct, in which case $\mathcal{R}_N = \mathcal{R}_L$. In contrast, the dependence of $\mathbb{E}[\tilde{z}\|B'\tilde{z}]$ on B and on the density $f_{\tilde{z}}$ does not admit a similarly simple description. Our strategy here is to use Cauchy–Schwarz to obtain the bound

$$\begin{aligned}\left|\frac{v'}{\sqrt{\|v\|^2 + \operatorname{Var}(\epsilon)}} R\mathbb{E}[\tilde{z}\|B'\tilde{z}]\right| &= \left|\frac{v'}{\sqrt{\|v\|^2 + \operatorname{Var}(\epsilon)}} R\big(\mathbb{E}[\tilde{z}\|B'\tilde{z}] - BB'\tilde{z}\big)\right| \\ &\le \|\mathbb{E}[\tilde{z}\|B'\tilde{z}] - BB'\tilde{z}\|.\end{aligned}$$

Next, we use a result of Steinberger and Leeb (2018), namely a bound of the form

$$\sup_{B\in\mathbb{G}} \mathbb{P}\big(\|\mathbb{E}[\tilde{z}\|B'\tilde{z}] - BB'\tilde{z}\| > t\big) \le \frac{1}{t}d^{-1/12} + 4\gamma_1\frac{p}{\log d},$$

which holds for a collection $\mathbb{G}$ of $d \times p$ matrices with orthonormal columns. For the size of $\mathbb{G}$, as measured by the uniform distribution on all such matrices, that is, by the normalized Haar measure $\nu_{d,p}$ on the Stiefel manifold of dimensions d and p, Steinberger and Leeb (2018) establish a bound of the form

$$\nu_{d,p}\big(\mathbb{G}^c\big) \le \kappa_1 d^{-(1-12\gamma_1\frac{p}{\log d})/12}.$$

We are interested in matrices B of a specific form, namely

$$B = R'\Sigma^{1/2}M\big(M'\Sigma M\big)^{-1/2}.$$

Therefore, we set $\mathbb{U} := \{R \in \mathcal{O}_d : R'\Sigma^{1/2}M(M'\Sigma M)^{-1/2} \in \mathbb{G}\}$ and show that $\nu_d(\mathbb{U}) = \nu_{d,p}(\mathbb{G})$. Apart from several technical details, this is the basic argument underlying Theorem 3.1(i). The proofs of Theorem 3.1(ii) and of Theorem 3.4 follow the same basic outline but require some nontrivial additional considerations to deal, among other issues, with estimation errors.

REMARK 4.1 ($\mathbb{E}[y\|x]$ is linear in x *on average w.r.t.* R). Recall that the Bayes predictor $\mathbb{E}[y\|x]$ depends on R. Consider the model (1)–(2) as before, but now with R taken as random and, in particular, uniformly distributed on $\mathcal{O}_d$, independently of all other quantities. Expectations under this model with random R will be denoted by $\mathcal{E}[\cdot]$. For random R, the Bayes predictor that we have considered so far can be written as $\mathcal{E}[y\|x, R]$. Clearly, for a fixed value of R, the function

$\mathcal{E}[y\|x, R]$ is typically nonlinear in x. Integrating out R, that is, taking the average with respect to the conditional distribution of R given x, we obtain $\mathcal{E}[y\|x]$. But $\mathcal{E}[y\|x]$ is linear in x, because the distribution of $R\tilde{z}$ is spherically symmetric. In view of this, it is not surprising that $\mathcal{E}[y\|x, R]$ is close to a linear function in x for some R's. Theorem 3.1(i) shows that the size of the collection of such R's can be controlled through $p/\log d$.

REMARK 4.2 (A crucial change of perspective). In Steinberger and Leeb (2018), for a fixed distribution of the design variables z, the existence of a "large" collection of candidate models M is proved, which all have the property that the conditional moment $\mathbb{E}[z\|M'z]$ is almost linear in $M'z$, provided that $d \gg p$. Here, on the other hand, we fix a candidate model M of interest and exhibit a "large" collection of design distributions for z, for which the conditional mean is almost linear. This has several important advantages. First, it seems to better reflect statistical practice, where usually a certain candidate model of interest, or a whole collection of such candidate models, is fixed a priori. Second, it allows a more elegant treatment of general covariance matrices Σ of the design z, whereas the analogous discussion in the framework of Steinberger and Leeb (2018) is much more involved (cf. Section 3.3 in that reference). Finally, and most importantly, it facilitates the consideration of several candidate models of potentially different dimensions, by a simple union bound argument (cf. Section 6).

5. Explicit analysis of a simple setting. Consider a uniformly distributed random vector $\tilde{z} = (\tilde{z}_1, \tilde{z}_2)' \sim \text{Unif}[-\sqrt{3}, \sqrt{3}]^2$, so $d = 2$ here, and $\mathbb{E}[\tilde{z}] = 0$, $\mathbb{E}[\tilde{z}\tilde{z}'] = I_2$. For $p = 1$, the matrix $B \in \mathbb{R}^{d\times p} = \mathbb{R}^2$ is just a two-dimensional unit vector, and we write $b := B = (b_1, b_2)' \in \mathcal{S}^1$ to emphasize this fact. Due to symmetry of the uniform distribution on the square, it suffices to consider b on a 45° segment of the unit circle, and we chose the one between the horizontal axis and the first main diagonal, that is, $b \in \mathcal{S}^1$ is such that $b_1 \geq b_2 > 0$. For such b, we consider the conditional mean and standard deviation curves only on the support of the distribution of $b'\tilde{z}$, that is, $x \in [-\sqrt{3}(b_1 + b_2), \sqrt{3}(b_1 + b_2)]$. Again, due to symmetry, it actually suffices to consider $x \geq 0$. It is straight forward but somewhat tedious to show that

$$f_{b'\tilde{z}}(x) = \begin{cases} \dfrac{1}{2\sqrt{3}b_1} & \text{if } |x| \leq \sqrt{3}(b_1 - b_2), \\ \dfrac{\sqrt{3}(b_1 - b_2) - |x|}{12b_1b_2} & \text{if } |x| \in \big[\sqrt{3}(b_1 - b_2), \sqrt{3}(b_1 + b_2)\big], \end{cases}$$

$$\mathbb{E}[\tilde{z}\|b'\tilde{z}=x]=\begin{cases}\begin{pmatrix}x/b_1\\0\end{pmatrix} & \text{if } x\in[0,\sqrt{3}(b_1-b_2)],\\ \begin{pmatrix}\dfrac{x+\sqrt{3}(b_1-b_2)}{2b_1}\\ \dfrac{x-\sqrt{3}(b_1-b_2)}{2b_2}\end{pmatrix} & \text{if } x\in[\sqrt{3}(b_1-b_2),\sqrt{3}(b_1+b_2)],\end{cases}$$

and

$$\sqrt{\operatorname{Var}[c'\tilde{z}\|b'\tilde{z}=x]}=\begin{cases}1/b_1 & \text{if } x\in[0,\sqrt{3}(b_1-b_2)],\\ \dfrac{\sqrt{3}(b_1+b_2)-x}{2\sqrt{3}b_1b_2} & \text{if } x\in[\sqrt{3}(b_1-b_2),\sqrt{3}(b_1+b_2)]\end{cases}$$

almost surely, where $c\in\mathbb{R}^2$ is one of the two unit vectors orthogonal to b. In Figure 1, we show continuous versions of the conditional expectation curve $x\mapsto\mathbb{E}[\tilde{z}\|b'\tilde{z}=x]$ (dashed lines) and the conditional standard deviation function $x\mapsto\sqrt{\operatorname{Var}[c'\tilde{z}\|b'\tilde{z}=x]}$ (dash dotted lines) as functions of $x\in\mathbb{R}$, for different values of b. The standard deviation function is superimposed in the plots with a coordinate system whose x-axis is given by the one-dimensional subspace spanned by b.

By inspection of Figure 1, the equations presented in the preceding paragraph are geometrically quite obvious. We also nicely see the nonlinearity of the conditional mean and the nonconstancy of the conditional variance. In fact, the only choices for b such that both linearity and constancy holds, are those where b is parallel to one of the coordinate axes (d). Linearity of the conditional mean is exactly satisfied also if b is parallel to one of the two main diagonals (a). From the two other panels (b) and (c), at least for the conditional expectation, we also see that approximate linearity will hold for a much larger collection of directions b. The deviation from the linear function $x\mapsto xb$ becomes even uniformly small as b approaches a main diagonal (b), and the deviation also gets small if b is almost parallel to one of the coordinate axes (c), at least if the deviation is measured in an L^2 sense with respect to the distribution of $b'\tilde{z}$.

Given the explicit formulae for the conditional moments, we can also compute the risk ratio $\mathcal{R}_N/\mathcal{R}_L$. For simplicity, we consider only the worst case in terms of error variance $\operatorname{Var}(\epsilon)$, that is, $\operatorname{Var}(\epsilon)=0$, and we assume that the model M is not correct, that is, $v\neq 0$ (where v has been defined in Section 4). Therefore, $c:=R'v/\|v\|=R'(I_d-P_{\Sigma^{1/2}M})\Sigma^{1/2}\theta/\|v\|$ is a unit vector that is orthogonal to $b=R'\Sigma^{1/2}M(M'\Sigma M)^{-1/2}$. By the variance decomposition formula, we may compute $\mathcal{R}_N/\mathcal{R}_L$ as

$$\mathbb{E}\big[\big(\mathbb{E}[c'\tilde{z}\|b'\tilde{z}]\big)^2\big]=1-\mathbb{E}[\operatorname{Var}[c'\tilde{z}\|b'\tilde{z}]],$$

and, using the formula for the density of $b'\tilde{z}$ and the conditional variance above, one easily arrives at the expression

$$\frac{\mathcal{R}_N}{\mathcal{R}_L}=\frac{b_1-b_2/2}{b_1^3},$$

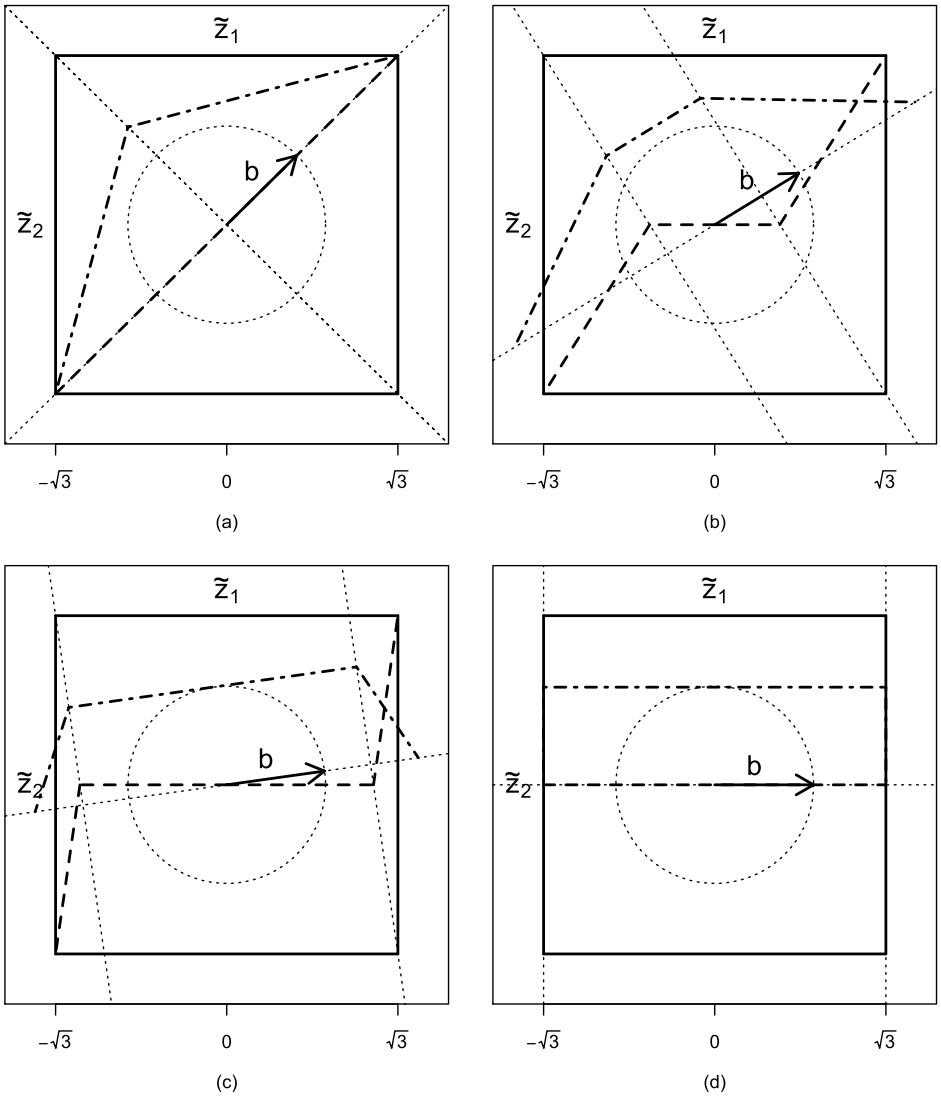

FIG. 1. *Conditional expectation curves* $x \mapsto \mathbb{E}[\tilde{z}|b'\tilde{z} = x]$ (*dashed lines*) *and superimposed conditional standard deviation function* $x \mapsto \sqrt{\mathrm{Var}[c'\tilde{z}|b'\tilde{z} = x]}$ (*dash dotted lines*) *for the bivariate uniform distribution* $\tilde{z} \sim Unif[-\sqrt{3}, \sqrt{3}]^2$, $c'b = 0$ *and different values for* $b \in \mathcal{S}^1$.

where $b = (b_1, b_2)' \in \mathcal{S}^1$ is considered only in the range $b_1 \geq b_2 > 0$, as above. Using symmetry, we obtain the general formula for the risk ratio

$$\frac{\mathcal{R}_N}{\mathcal{R}_L} = \frac{\max(|b_1|, |b_2|) - \min(|b_1|, |b_2|)/2}{\max(|b_1|, |b_2|)^3}.$$

To visualize this function in dependence on the original model parameters R, Σ and M, we fix $M = (1, 0)'$ and parameterize R as

$$R(\alpha) = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \quad \text{and} \quad \Sigma \text{ as } \Sigma(\rho) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

so that

$$\Sigma(\rho)^{1/2} = \frac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{1+\sqrt{1-\rho^2}} & \operatorname{sign}(\rho)\sqrt{1-\sqrt{1-\rho^2}} \\ \operatorname{sign}(\rho)\sqrt{1-\sqrt{1-\rho^2}} & \sqrt{1+\sqrt{1-\rho^2}} \end{pmatrix}.$$

If $\rho \in [-1, 1]$ is fixed and α runs from 0 to 2π, then $b = b(\alpha, \rho) = R(\alpha)\Sigma^{1/2}(\rho)M$ runs around the unit circle in counter clockwise direction starting at $\Sigma^{1/2}(\rho)M$. Note that $b(\alpha, 0) = (\cos(\alpha), \sin(\alpha))'$. The top panel of Figure 2 shows the risk ratio as a function of α on the domain $\alpha \in [0, \pi]$ and for $\rho = 0$. Clearly, a nonzero value of ρ simply shifts the whole plot by a certain amount. The other four panels of Figure 2 show the risk ratio as a function of $\rho \in [0, 1]$ for four fixed values of α, namely, $\alpha_1 = 0, \alpha_2 = \pi/5, \alpha_3 = 9\pi/20, \alpha_4 = 3\pi/5$. Note that for fixed α, $b(\alpha, \rho)$ runs through a 45° segment of the unit circle in counter clockwise direction as ρ runs through $[0, 1]$. There are four such segments indicated in the top panel of Figure 2, corresponding to $\alpha_1, \ldots, \alpha_4$, each of which is bordered by two vertical dashed lines that are conjoined by a horizontal dotted line. The four lower panels in Figure 2 correspond to the four indicated segments in the top panel.

We conclude that in the present case the risk ratio never drops below 0.92. Moreover, it is not possible to identify a set of values for ρ such that the risk ratio is always close to one irrespective of α. Indeed, whether or not a certain amount of correlation between the components of z leads to a high or a low risk ratio depends, in a fundamental way, on the geometry of the distribution of z, that is, on the value of α. Finally, note that this simple example cannot provide any insight into the dependence of the risk ratio on dimension, which is, of course, the main point of the general theory developed in this paper.

6. Regarding several submodels and model selection. Our results can easily be adapted to cover more than one submodel, that is, more than one matrix M. Fix $m \in \mathbb{N}$ and for $i = 1, \ldots, m$, let M_i be a full-rank $d \times p_i$ matrix with $p_i < d$. Then Theorem 3.1(i) entails that

$$\max_{1 \le i \le m} 1 - \frac{\mathcal{R}_N(M_i, \theta, \mathcal{L}(\epsilon), \Sigma, f_{\tilde{z}}, R)}{\mathcal{R}_L(M_i, \theta, \mathcal{L}(\epsilon), \Sigma, f_{\tilde{z}})} \le K_1 \left(\frac{\max_{1 \le i \le m} p_i}{\log d} \right)^{5/6}$$

provided only that $R \in \bigcap_{i=1}^m \mathbb{U}(M_i, \Sigma, f_{\tilde{z}})$, and the union bound gives

$$\nu_d \left(\left(\bigcap_{i=1}^m \mathbb{U}(M_i, \Sigma, f_{\tilde{z}}) \right)^c \right) \le L_1 \sum_{i=1}^m d^{-\frac{1}{12}(1 - c_1 \frac{p_i}{\log d})} \le m L_1 d^{-\frac{1}{12}(1 - c_1 \frac{\max_{1 \le i \le m} p_i}{\log d})};$$

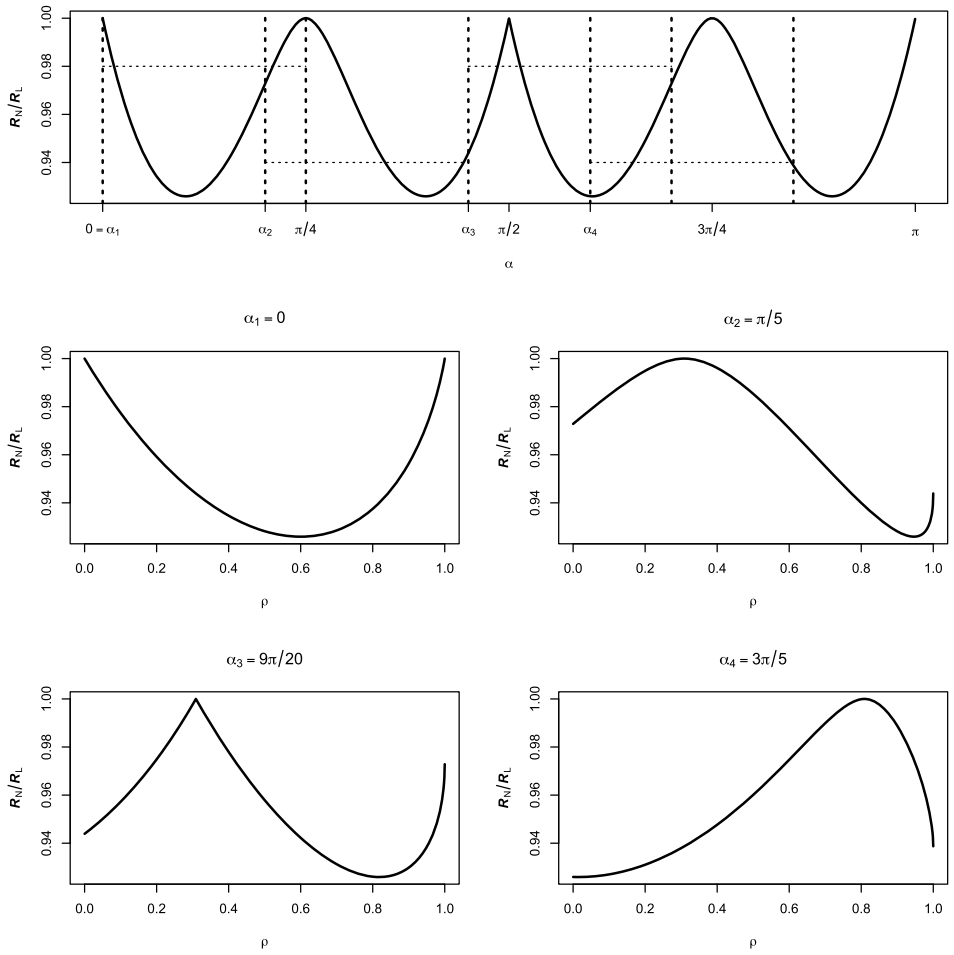

FIG. 2. Top panel: *Risk ratio* $\mathcal{R}_N/\mathcal{R}_L$ *as a function of the angle* $\alpha \in [0, \pi]$ *that governs the rotation* $R(\alpha)$ *of the design* $z = R(\alpha)\tilde{z}$. Lower panels: *Risk ratio* $\mathcal{R}_N/\mathcal{R}_L$ *as a function of the correlation coefficient* $\rho \in [0, 1]$ *of the design* $z = \Sigma(\rho)^{1/2}R(\alpha)\tilde{z}$, *for different values of* α.

see also Remark 7.4. These upper bounds hold uniformly over the M_i's and over θ, $\mathcal{L}(\epsilon)$, Σ and $f_{\tilde{z}} \in \mathcal{F}_{d,12}(D, E)$ as in Theorem 3.1(i). Similar considerations apply, mutatis mutandis, to the results in Theorem 3.1(ii) and Theorem 3.4. The (crude) union bound used in the preceding display limits the number of models, that is, m, for which the upper bound is small for fixed $\max_i p_i$ and d; see also Remark 3.3.

These considerations also suggest that our results are useful for selecting a model from a set of candidates $\mathcal{M} = \{M_i : i \leq m\}$. For convenience, write $\mathcal{R}_L(M)$, $\mathcal{R}_N(M)$ and $\mathcal{R}_{\mathrm{OLS}}(M|X, Y)$ for the risk of the best linear predictor, the Bayes predictor and the ordinary least squares predictor, respectively, corresponding to model $M \in \mathcal{M}$, as introduced prior to Theorems 3.1 and 3.4, and recall that

$\mathcal{R}_{\mathrm{OLS}}(M|X,Y)$ is a random quantity depending on the data X and Y. In Leeb (2009), candidate models are evaluated in terms of $\mathcal{R}_{\mathrm{OLS}}(M|X,Y)$. However, since this quantity depends on unknown parameters, it has to be estimated, for example, by some type of cross-validation procedure, say, $\hat{\mathcal{R}}_{\mathrm{OLS}}(M|X,Y)$. If this estimation is successful in the sense that

$$\sup_{M\in\mathcal{M}}\left|1-\frac{\hat{\mathcal{R}}_{\mathrm{OLS}}(M|X,Y)}{\mathcal{R}_{\mathrm{OLS}}(M|X,Y)}\right|\approx 0,$$

with high probability, then Theorem 3.4(i) can be used to show that $\hat{\mathcal{R}}_{\mathrm{OLS}}(M|X,Y)$ actually even estimates $\mathcal{R}_N(M)$ uniformly over $\mathcal{M}$, provided that $R\in\bigcap_{i=1}^m \mathbb{U}(M_i,\Sigma,f_z)$, that $\max_i p_i/n$ and $\max_i p_i/\log d$ are small, and that the cardinality of $\mathcal{M}$ is not too large. Consequently, if these conditions are satisfied, then the feasible model selector $\hat{M}=\operatorname{argmin}_{M\in\mathcal{M}}\hat{\mathcal{R}}_{\mathrm{OLS}}(M|X,Y)$ mimics the infeasible best candidate model $M^*=\operatorname{argmin}_{M\in\mathcal{M}}\mathcal{R}_N(M)$.

7. Additional remarks.

REMARK 7.1. (i) The linear model (1) is widely used in statistical theory and practice. It is also the natural starting point for investigating high-dimensional problems. The vast majority of the high-dimensional regression literature restricts model (1) further, for example, by imposing conditions on the sparsity or on the decay of the coefficients of the regression parameter θ and/or conditions on the covariance matrix Σ. Here, we abstain from any such restrictions and argue that simple linear submodels are still useful in many nonsparse situations.

(ii) Some parameters in (1) might seem superfluous at first glance. For example, the mean of y is $\vartheta+\theta'\mu$, so one might be tempted to absorb ϑ into $\theta'\mu$ or vice versa. But it is easy to see that if the first two (joint) moments of y and z are to be unrestricted, then the free parameters ϑ, θ, $\mathrm{Var}(\epsilon)$, μ and Σ are required.

(iii) The assumption of a true high-dimensional linear model is essential for the theoretical results of the present article. If the true model is nonlinear in the regressors z, then possibly other, nonlinear, approximations for the Bayes predictor $\mathbb{E}[y\|x]$ can be established.[2] However, the usefulness of such alternative approximations will be limited by the size of the class of true regression functions under consideration. Take, for instance, the extreme case where the true data generating model is given by $y=f(z)+\epsilon$, where $f:\mathbb{R}^d\to\mathbb{R}$ is unrestricted, that is, arbitrary, except for the requirement that $f(z)$ is square-integrable. Here, the Bayes predictor $g(M'z):=\mathbb{E}[y\|M'z]$ is unrestricted, that is, arbitrary, except for the requirement that $g(M'z)$ is square-integrable. In this model, one can not relate the Bayes predictor $\mathbb{E}[y\|M'z]$ to a linear function in $M'z$, not even approximately.

[2]For example, Theorem 2.1(ii) of Steinberger and Leeb (2018) suggests that if the true regression function is a quadratic polynomial in z, then also the Bayes predictor $\mathbb{E}[y\|x]$ will be approximately quadratic in $x=M'z$, if $p/\log d$ is small.

REMARK 7.2. Note that the results of Steinberger and Leeb (2018), that we rely on, also allow for distributions of the random d-vector $\tilde{z}$ that exhibit some dependence among the components. An extension of the present results in that direction will be considered elsewhere.

REMARK 7.3. Consider an autoregressive process of order q of the form $y_t = \sum_{j=1}^{q} \rho_j y_{t-j} + \varepsilon_t$, $1 \le t \le T$, with nonrandom starting values $y_0 = \cdots = y_{-q+1} = 0$, say, and uncorrelated innovations ε_t with mean zero and variance $\sigma^2 > 0$. Suppose we want to predict the value of y_T based on p lagged variables $(y_{j_1}, \ldots, y_{j_p})$, $1 \le j_k \le T-1$. Is the risk of the best linear predictor $\sum_{k=1}^{p} \beta_k y_{j_k}$, say, here also comparable to the risk of the Bayes predictor $\mathbb{E}[y_T \| y_{j_1}, \ldots, y_{j_p}]$? In this context, our results can be used under the additional assumption that the innovations are linear functions of independent random variables. With this, we can write the innovations as $(\varepsilon_1, \ldots, \varepsilon_T)' = \sigma^2 R\tilde{z}$ for a T-vector $\tilde{z}$ whose components are independent with mean zero and variance one, and for some orthogonal matrix R. The model equation for y_T can be brought into the form (1)–(2) by expanding the response as $y_T = \sum_{i=1}^{T} \alpha_i \varepsilon_i$, where the coefficients α_i depend on T and on the ρ_j's. The last equation is of the form (1)–(2) with $y = y_T$, $\vartheta = 0$, $\theta = (\alpha_1, \ldots, \alpha_T)'$, $z = (\varepsilon_1, \ldots, \varepsilon_T)'$, $\epsilon = 0$, $\Sigma = \sigma^2 I_T$ and $d = T$. And with these conventions, also any set of lagged dependent variables, that is, any subvector of $(y_1, \ldots, y_{T-1})'$, can be written as $M'z$ for an appropriate matrix M that depends on which components are retained in the subvector, and which also depends on the ρ_j's. We note that, in this case, the vector z is typically not observable; see also Remark 3.5. We also note that the matrix M corresponding to a submodel here depends on unknown parameters. Our results can still be used in this setting, because they are uniform in M; cf. Remark 7.4. We also point out that the formulation of the autoregressive process considered above differs from the classical formulation of time series analysis only insofar as we here assume also that the innovations are linear functions of independent random variables instead of merely being white noise.

REMARK 7.4. In the first display in Theorem 3.1(i), the order in which the suprema are taken is irrelevant except for the supremum over $R \in \mathbb{U}$, which must be taken after the suprema over M, Σ and $f_{\tilde{z}}$. This is because $\mathbb{U} = \mathbb{U}(M, \Sigma, f_{\tilde{z}})$ depends on the indicated quantities. Similar considerations also apply to Theorem 3.1(ii) and to the results in Theorem 3.4.

REMARK 7.5. Consider the relation (2), with a fixed choice of μ, Σ and $f_{\tilde{z}}$. Then, for any full-rank $d \times p$ matrix M with $p < d$, one can always choose R so that the Bayes predictor of any linear function of z given $x = M'z$ is linear in x. (The statement is trivial if $p = d$.) Moreover, for any such M, for the corresponding R as before, and for any $s \ge 0$, one can furthermore choose $\theta \in \mathbb{R}^d$ so that $\theta' z$ is independent of $x = M'z$ and so that $\mathrm{Var}(\theta' z) = s^2$. [Indeed, set $\tilde{M} = \Sigma^{1/2} M$ and

$\bar{M} = \tilde{M}(\tilde{M}'\tilde{M})^{-1/2}$, so that $\bar{M}'\bar{M} = I_p$, and write $x = M'\mu + (\tilde{M}'\tilde{M})^{1/2}\bar{M}'R\tilde{z}$. Now choose a $d \times (d-p)$ matrix $\bar{N}$ such that $R := (\bar{M} : \bar{N}) \in \mathcal{O}_d$. For this R, the regressor vector x is given by a regular affine transformation of the first p components of $\tilde{z}$. Thus, conditioning on x is equivalent to conditioning on the first p components of $\tilde{z}$, which, by independence of the components of $\tilde{z}$, implies the claim about the Bayes predictor. For the second claim, set $\theta = \Sigma^{-1/2}\bar{N}vs$, where v denotes some unit vector in $\mathbb{R}^{d-p}$. It immediately follows that $\theta'z = sv'\bar{N}'R\tilde{z}$ is a function of the last $d-p$ components of $\tilde{z}$ and hence independent of x. And we have $\mathrm{Var}(\theta'z) = s^2$, as desired.]

APPENDIX: PROOFS OF SECTION 3

Throughout the Appendix, we consider independent and identically distributed (i.i.d.) pairs (y,z), $(y_i,z_i)_{i=1}^n$ following the model (1)–(2) of Section 2 and we set $x = M'z$, $x_i = M'z_i$, $i = 1,\dots,n$, for some full rank $d \times p$ matrix M with $p \le \min(d, n-1)$. We abbreviate $\Sigma_x := \mathrm{Cov}[x] = M'\Sigma M$, $\mu_x := \mathbb{E}[x] = M'\mu$, $\tilde{M} := \Sigma^{1/2}M$, $\tilde{\theta} := \Sigma^{1/2}\theta$, $E_k := \sup_{\|w\|\le 1}\mathbb{E}[|w'\tilde{z}|^k]$, $\xi := y - \alpha - \beta'x$, $\sigma^2 := \mathbb{E}[\xi^2]$, and $\Xi := (\xi_1,\dots,\xi_n)' := Y - U\gamma$, where $U = [\iota : X]$, $Y = (y_1,\dots,y_n)'$, $X = [x_1,\dots,x_n]'$, $\iota = (1,\dots,1)' \in \mathbb{R}^n$ and $\gamma = (\alpha,\beta')' \in \mathbb{R}^{p+1}$ are the coefficients that minimize $(a,b) \mapsto \mathbb{E}[(y-(a+b'x))^2]$. Moreover, we will study the OLS estimator $\hat{\gamma}_n = (\hat{\alpha}_n, \hat{\beta}_n')' = (U'U)^{-1}U'Y$. Finally, in Section 3, we have introduced the prediction risks $\mathcal{R}_L = \mathbb{E}[(y-(\alpha+\beta'x))^2] = \sigma^2$, $\mathcal{R}_L(x) = \mathbb{E}[(y-(\alpha+\beta'x))^2\|x]$, $\mathcal{R}_N = \mathbb{E}[(y-\mathbb{E}[y\|x])^2]$, $\mathcal{R}_N(x) = \mathbb{E}[(y-\mathbb{E}[y\|x])^2\|x]$, $\mathcal{R}_{\mathrm{OLS}}(X,Y) = \mathbb{E}[(y-(\hat{\alpha}_n+\hat{\beta}_n'x))^2\|X,Y]$ and $\mathcal{R}_{\mathrm{OLS}}(X,Y,x) = \mathbb{E}[(y-(\hat{\alpha}_n+\hat{\beta}_n'x))^2\|X,Y,x]$.

As a preliminary consideration for the entire Appendix, we recall that the best linear predictor for y based on x is given by

$$\begin{aligned}\alpha + \beta'x &= \mathbb{E}[y] + \mathrm{Cov}\big[y,x'\big]\mathrm{Cov}[x]^{-1}\big(x - \mathbb{E}[x]\big)\\ &= \vartheta + \theta'\mu + \theta'\Sigma M\big(M'\Sigma M\big)^{-1}M'\Sigma^{1/2}R\tilde{z}\\ &= \vartheta + \theta'\mu + \tilde{\theta}'P_{\tilde{M}}R\tilde{z},\end{aligned}$$

and thus $\xi = y - \alpha - \beta'x = \tilde{\theta}'(I_d - P_{\tilde{M}})R\tilde{z} + \epsilon$, and the corresponding linear prediction risks are given by $\mathcal{R}_L = \|(I_d - P_{\tilde{M}})\tilde{\theta}\|^2 + \mathrm{Var}[\epsilon]$ and $\mathcal{R}_L(x) = \tilde{\theta}'(I_d - P_{\tilde{M}})R\mathbb{E}[\tilde{z}\tilde{z}'\|x]R'(I_d - P_{\tilde{M}})\tilde{\theta} + \mathrm{Var}[\epsilon]$. Moreover, it is easy to see that $\mathcal{R}_N = \mathbb{E}[(\tilde{\theta}'R\tilde{z} - \mathbb{E}[\tilde{\theta}'R\tilde{z}\|x])^2] + \mathrm{Var}[\epsilon]$ and $\mathcal{R}_N(x) = \mathbb{E}[(\tilde{\theta}'R\tilde{z} - \mathbb{E}[\tilde{\theta}'R\tilde{z}\|x])^2\|x] + \mathrm{Var}[\epsilon]$.

A.1. Auxiliary results.

LEMMA A.1. *The distributions of $\mathcal{R}_L$, $\mathcal{R}_L(x)$, $\mathcal{R}_N$, $\mathcal{R}_N(x)$, $\mathcal{R}_{\mathrm{OLS}}(X,Y)$ and $\mathcal{R}_{\mathrm{OLS}}(X,Y,x)$ do not depend on ϑ and μ.*

PROOF. Of course the distributions of $\mathcal{R}_L$ and $\mathcal{R}_N$ are degenerate. The claim about the distributions of the linear and nonlinear prediction risks follows immediately from the preliminary consideration and the fact that conditioning on a random variable and on the corresponding centered random variable is equivalent. For the OLS risks, recall that

$$y - \hat{\alpha}_n - \hat{\beta}_n' x = y - \frac{\iota' Y}{n} - Y'(I_n - P_\iota)X\big[X'(I_n - P_\iota)X\big]^{-1}\Big(x - \frac{X'\iota}{n}\Big),$$

the distribution of which does not depend on the mean parameters ϑ and μ. $\square$

LEMMA A.2. *If $\mathcal{R}_L > 0$, then*

$$\mathbb{P}\Big(\Big|\frac{\mathcal{R}_L(x)}{\mathcal{R}_L} - 1\Big| > t\Big) \le \mathbb{P}\big(\big\|\mathbb{E}[\tilde{z}\tilde{z}'\|B_0'\tilde{z}] - \big(I_d - B_0B_0' + B_0B_0'\tilde{z}\tilde{z}'B_0B_0'\big)\big\| > t\big),$$

for every $t > 0$, where $B_0 = R'\tilde{M}(\tilde{M}'\tilde{M})^{-1/2}$.

PROOF. Simply observe that

$$\begin{aligned}
&\big|\mathcal{R}_L(x) - \mathcal{R}_L\big| \\
&\quad = \big|\tilde{\theta}'(I_d - P_{\tilde{M}})R\mathbb{E}\big[\tilde{z}\tilde{z}'\|x\big]R'(I_d - P_{\tilde{M}})\tilde{\theta} \\
&\qquad - \tilde{\theta}'(I_d - P_{\tilde{M}})RR'(I_d - P_{\tilde{M}})\tilde{\theta}\big| \\
&\quad = \big|\tilde{\theta}'(I_d - P_{\tilde{M}})R\big[\mathbb{E}\big[\tilde{z}\tilde{z}'\|M'\mu + \tilde{M}'R\tilde{z}\big] \\
&\qquad - \big(I_d - B_0B_0' + B_0B_0'\tilde{z}\tilde{z}'B_0B_0'\big)\big]R'(I_d - P_{\tilde{M}})\tilde{\theta}\big| \\
&\quad \le \big\|(I_d - P_{\tilde{M}})\tilde{\theta}\big\|^2\big\|\mathbb{E}\big[\tilde{z}\tilde{z}'\|B_0'\tilde{z}\big] - \big(I_d - B_0B_0' + B_0B_0'\tilde{z}\tilde{z}'B_0B_0'\big)\big\|,
\end{aligned}$$

and $\mathcal{R}_L = \|(I_d - P_{\tilde{M}})\tilde{\theta}\|^2 + \mathrm{Var}[\epsilon]$. $\square$

LEMMA A.3. *The following holds true*:

(i) *For every $v \in \mathbb{R}^p$ and $k \in \mathbb{N}$, we have*

$$\begin{aligned}
\mathbb{E}\big[\big|v'\Sigma_x^{-1/2}x\xi\big|^k\big] &\le 4^{k-1}\big(\|v\|^k + \big|v'\Sigma_x^{-1/2}\mu_x\big|^k\big) \\
&\quad \times \big(\big\|(I_d - P_{\tilde{M}})\tilde{\theta}\big\|^k + \mathbb{E}\big[|\epsilon|^k\big]\big)E_{2k}.
\end{aligned}$$

(ii) $\mathbb{E}[\|\Sigma_x^{-1/2}X'\Xi/n\|^2] \le \frac{p}{n}4(1 + \|\Sigma_x^{-1/2}\mu_x\|^2/p)\sigma^2E_4$.

PROOF. For (i), write $\xi = y - \alpha - \beta'x = \tilde{\theta}'(I_d - P_{\tilde{M}})R\tilde{z} + \epsilon$ and $v'\Sigma_x^{-1/2}x\xi = v'\Sigma_x^{-1/2}(x - \mu_x)\xi + v'\Sigma_x^{-1/2}\mu_x\xi$. Now, using the triangle inequality and the elementary inequality $(a + b)^k \le 2^{k-1}(a^k + b^k)$, for $a, b \ge 0$, $k \ge 1$, we obtain

$$\mathbb{E}\big[\big|v'\Sigma_x^{-1/2}x\xi\big|^k\big] \le 2^{k-1}\big(\mathbb{E}\big[\big|v'\Sigma_x^{-1/2}\tilde{M}'R\tilde{z}\xi\big|^k\big] + \big|v'\Sigma_x^{-1/2}\mu_x\big|^k\mathbb{E}\big[|\xi|^k\big]\big).$$

Note that since $1 \le E_{2k}$ by Jensen's inequality, we have $E_k \le E_{2k}$ in view of Lyapunov's inequality. For the first expectation on the right-hand side of the previous display, using the Cauchy–Schwarz inequality, we find that

$$\begin{aligned}
&\mathbb{E}\big[\big|v'\Sigma_x^{-1/2}\tilde{M}'R\tilde{z}\xi\big|^k\big] \\
&\quad\le 2^{k-1}\big(\mathbb{E}\big[\big|v'\Sigma_x^{-1/2}\tilde{M}'R\tilde{z}\tilde{\theta}'(I_d - P_{\tilde{M}})R\tilde{z}\big|^k\big] + \mathbb{E}\big[\big|v'\Sigma_x^{-1/2}\tilde{M}'R\tilde{z}\epsilon\big|^k\big]\big) \\
&\quad\le 2^{k-1}\big(\big\|R'\tilde{M}\Sigma_x^{-1/2}v\big\|^k\big\|R'(I_d - P_{\tilde{M}})\tilde{\theta}\big\|^k E_{2k} \\
&\qquad + \big\|R'\tilde{M}\Sigma_x^{-1/2}v\big\|^k\mathbb{E}\big[|\epsilon|^k\big]E_k\big) \\
&\quad\le 2^{k-1}\|v\|^k\big(\big\|(I_d - P_{\tilde{M}})\tilde{\theta}\big\|^k + \mathbb{E}\big[|\epsilon|^k\big]\big)E_{2k},
\end{aligned}$$

where we have used that $R'\tilde{M}\Sigma_x^{-1/2} = R'\tilde{M}(\tilde{M}'RR'\tilde{M})^{-1/2}$ has orthonormal columns, and thus $\|R'\tilde{M}\Sigma_x^{-1/2}v\| = \|v\|$. Similarly, we get

$$\mathbb{E}\big[|\xi|^k\big] \le 2^{k-1}\big(\big\|(I_d - P_{\tilde{M}})\tilde{\theta}\big\|^k + \mathbb{E}\big[|\epsilon|^k\big]\big)E_{2k},$$

and thus we obtain the final bound

$$\begin{aligned}
&\mathbb{E}\big[\big|v'\Sigma_x^{-1/2}x\xi\big|^k\big] \\
&\quad\le 2^{2(k-1)}\big(\|v\|^k + \big|v'\Sigma_x^{-1/2}\mu_x\big|^k\big)\big(\big\|(I_d - P_{\tilde{M}})\tilde{\theta}\big\|^k + \mathbb{E}\big[|\epsilon|^k\big]\big)E_{2k}.
\end{aligned}$$

For part (ii), note that the $x_i\xi_i$ are i.i.d. with mean $\mathbb{E}[x_i\xi_i] = 0$, let $e_j \in \mathbb{R}^p$ denote the jth element of the standard basis in $\mathbb{R}^p$ and apply part (i) with $k = 2$, $v = e_j$ and $j = 1, \dots, p$, to get

$$\begin{aligned}
&\mathbb{E}\big[\big\|\Sigma_x^{-1/2}X'\Xi/n\big\|^2\big] \\
&\quad= \frac{1}{n^2}\sum_{i,l=1}^{n}\mathbb{E}\big[\big(\Sigma_x^{-1/2}x_i\xi_i\big)'\big(\Sigma_x^{-1/2}x_l\xi_l\big)\big] \\
&\quad= \frac{1}{n}\sum_{j=1}^{p}\mathbb{E}\big[\big|e_j'\Sigma_x^{-1/2}x_1\xi_1\big|^2\big] \\
&\quad\le \frac{1}{n}\sum_{j=1}^{p}4\big(1 + \big|e_j'\Sigma_x^{-1/2}\mu_x\big|^2\big)\big(\big\|(I_d - P_{\tilde{M}})\tilde{\theta}\big\|^2 + \mathbb{E}\big[\epsilon^2\big]\big)E_4 \\
&\quad= \frac{p}{n}4\big(1 + \big\|\Sigma_x^{-1/2}\mu_x\big\|^2/p\big)\sigma^2E_4,
\end{aligned}$$

where we have used the abbreviation $\sigma^2 = \|(I_d - P_{\tilde{M}})\tilde{\theta}\|^2 + \mathrm{Var}[\epsilon]$. This completes the proof of the lemma. □

LEMMA A.4. *Suppose that $\sigma^2 > 0$ and that the standardized regressors $\tilde{z}$ satisfy the following regularity condition*: *There exist positive and finite constants*

C, η, such that for every orthogonal projection matrix P in $\mathbb{R}^d$ and every $t > C \operatorname{rank} P$, one has $\mathbb{P}(\|P\tilde{z}\|^2 > t) \le Ct^{-1-\eta}$.[3] *Then there exists a positive finite constant $C_0 = C_0(C, \eta)$, such that if $n > C_0 p$, we have*

$$\mathbb{P}\big(\|\Sigma_x^{1/2}(\hat{\beta}_n - \beta)\|/\sigma > t\big) \le C_1\left(\frac{p}{n}\right)^{\frac{1}{2}\frac{\eta}{1+\eta}} + \frac{64E_4}{t^2}\frac{p}{n} + \frac{8}{t}\frac{\sqrt{p}}{n}, \tag{5}$$

and

$$\begin{aligned}&\mathbb{P}\big(|\hat{\alpha}_n - \alpha|/\sigma > t\big)\\ &\quad\le 2C_1\left(\frac{p}{n}\right)^{\frac{1}{2}\frac{\eta}{1+\eta}}\\ &\qquad + 24\left(\frac{1}{nt^2} + \left(\frac{1}{nt^2}\right)^{1/3} + \left(\frac{p}{nt^2}\right)^{1/2}\right)\big(1 + \|\Sigma_x^{-1/2}\mu_x\|^2\big)\sqrt{E_4},\end{aligned} \tag{6}$$

for all $t > 0$. Here, the constants C_0 and C_1 are given by $C_0 = 512(48C)^{2+2/\eta}(6 + 6/\eta)^{1+4/\eta}$ and $C_1 = 4(C_0 \vee 1)^{\eta/(2+2\eta)}$. (This result continues to hold for general $\tilde{z}$ with $\mathbb{E}[\tilde{z}] = 0$, $\mathbb{E}[\tilde{z}\tilde{z}'] = I_d$, that has a Lebesgue density and that satisfies the additional tail condition of the lemma. No independence of components is needed.)

PROOF. We begin with a few preliminary considerations. First, note that since the design matrix X has a Lebesgue density and $p < n$, we have $\mathbb{P}(\det U'U = 0) = 0$, so the OLS estimator $\hat{\gamma}_n$ exists and is unique, almost surely. Next, recall $\Xi = (\xi_1, \dots, \xi_n)' = Y - U\gamma$ and use the Frish–Waugh–Lovell theorem to obtain

$$\begin{aligned}\begin{pmatrix}\hat{\alpha}_n - \alpha\\ \hat{\beta}_n - \beta\end{pmatrix} &= \hat{\gamma}_n - \gamma = (U'U)^{-1}U'\Xi\\ &= \begin{bmatrix}\iota'(I_n - P_X)\Xi/\iota'(I_n - P_X)\iota\\ [X'(I_n - P_\iota)X]^{-1}X'(I_n - P_\iota)\Xi\end{bmatrix}.\end{aligned} \tag{7}$$

We also use the abbreviations $\bar{x}_i = \Sigma_x^{-1/2}x_i$, $\bar{X} = X\Sigma_x^{-1/2}$, $\mu_{\bar{x}} = \mathbb{E}[\bar{x}_1] = \Sigma_x^{-1/2}\mu_x$, and $\hat{\Sigma}_{\bar{x}} = \bar{X}'(I_n - P_\iota)\bar{X}/n$.

Now, to establish the statement in (5), the estimation error of β can be written as $\hat{\beta}_n - \beta = \Sigma_x^{-1/2}[\bar{X}'(I_n - P_\iota)\bar{X}/n]^{-1}\bar{X}'(I_n - P_\iota)\Xi/n$. Together with the representation $\xi_i = \tilde{\theta}'(I_d - P_{\tilde{M}})R\tilde{z}_i + \epsilon_i$, it is apparent that the distribution of $\hat{\beta}_n - \beta$ does not depend on $\mu \in \mathbb{R}^d$, which is why we may restrict to the case $\mu = 0$, for this part. This also entails that $\mu_x = 0 = \mu_{\bar{x}}$. We bound the scaled estimation error as follows:

$$\|\Sigma_x^{1/2}(\hat{\beta}_n - \beta)\|/\sigma \le \|\hat{\Sigma}_{\bar{x}}^{-1}\|\big(\|\bar{X}'\Xi/n\|/\sigma + \|\bar{X}'\iota/n\||\iota'\Xi/n|/\sigma\big). \tag{8}$$

[3] This is the condition (SR) of Srivastava and Vershynin (2013), Theorem 1.1.

The term $\|\bar{X}'\Xi/n\|/\sigma$ is treated by Lemma A.3(ii) and Markov's inequality, which yields $\mathbb{P}(\|\bar{X}'\Xi/n\|/\sigma > t) \le t^{-2}4E_4 p/n$. For the second term in parentheses, using the fact that the rows of $\bar{X}$ are i.i.d. with mean zero and covariance matrix I_p and that the $(\xi_i)_{i=1}^n$ are also i.i.d. with mean zero and variance σ^2, a standard argument involving Markov's inequality shows that

$$\begin{aligned}\mathbb{P}\big(\|\bar{X}'\iota/n\|\,|\iota'\Xi/n|/\sigma > t\big) &\le \mathbb{P}\big(\|\bar{X}'\iota/n\| > \delta\big) + \mathbb{P}\big(|\iota'\Xi/n|/\sigma > t/\delta\big)\\ &\le \delta^{-2}p/n + \delta^2 t^{-2}/n,\end{aligned}$$

for all $t, \delta > 0$. Optimizing the upper bound over $\delta > 0$, we obtain for every $t > 0$,

$$\mathbb{P}\big(\|\bar{X}'\iota/n\|\,|\iota'\Xi/n|/\sigma > t\big) \le \frac{2}{t}\frac{\sqrt{p}}{n}.$$

For the inverse sample covariance term, we get, for $\delta > 1$,

$$\begin{aligned}\mathbb{P}\big(\|\hat{\Sigma}_{\bar{x}}^{-1}\| > \delta\big) &= \mathbb{P}\big(\lambda_{\min}(\hat{\Sigma}_{\bar{x}}) < 1/\delta\big) \le \mathbb{P}\big(|\lambda_{\min}(\hat{\Sigma}_{\bar{x}}) - 1| > (\delta-1)/\delta\big)\\ &\le \mathbb{P}\big(\|\hat{\Sigma}_{\bar{x}} - I_p\| > (\delta-1)/\delta\big)\\ &\le \frac{\delta}{\delta-1}\left(\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \bar{x}_i\bar{x}_i' - I_p\right\|\right] + \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \bar{x}_i\right\|^2\right]\right)\\ &= \frac{\delta}{\delta-1}\left(\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \bar{x}_i\bar{x}_i' - I_p\right\|\right] + \frac{p}{n}\right).\end{aligned}$$

The remaining expectation can be bounded using Theorem 1.1 of Srivastava and Vershynin (2013). To apply this result, we have to verify the (SR) condition in that reference for the standardized regressor $\bar{x}_1$. Recall that $\mu = 0$ and thus $\bar{x}_1 = \Sigma_x^{-1/2}M'\Sigma^{1/2}R\tilde{z}_1 = (\tilde{M}'\tilde{M})^{-1/2}\tilde{M}'R\tilde{z}_1$, since $\tilde{M}'\tilde{M} = M'\Sigma M = \Sigma_x$. Fix a projection matrix P in $\mathbb{R}^p$ and observe that $\|P\bar{x}_1\|^2 = \|\tilde{M}(\tilde{M}'\tilde{M})^{-1/2}P(\tilde{M}'\tilde{M})^{-1/2} \times \tilde{M}'R\tilde{z}_1\|^2$, where $\tilde{M}(\tilde{M}'\tilde{M})^{-1/2}P(\tilde{M}'\tilde{M})^{-1/2}\tilde{M}'$ is a projection matrix in $\mathbb{R}^d$ of the same rank as P. Invoking our assumption on the distribution of $\tilde{z}_1$, and noting that this assumption is invariant under orthogonal transformations of $\tilde{z}_1$, establishes the validity of the (SR) condition. Thus, Theorem 1.1 of Srivastava and Vershynin (2013) applies with $\varepsilon = (C_0 p/n)^{\frac{1}{2}\frac{\eta}{1+\eta}}$ provided that $p/n < 1/C_0$ [so that $\varepsilon \in (0,1)$], and we obtain

$$\begin{aligned}\mathbb{P}\big(\|\hat{\Sigma}_{\bar{x}}^{-1}\| > \delta\big) &\le \frac{\delta}{\delta-1}\left(\left[C_0\frac{p}{n}\right]^{\frac{1}{2}\frac{\eta}{1+\eta}} + \frac{p}{n}\right)\\ &\le \frac{2\delta}{\delta-1}\left((C_0 \vee 1)\frac{p}{n}\right)^{\frac{1}{2}\frac{\eta}{1+\eta}}.\end{aligned} \tag{9}$$

Here, $C_0 := 512(48C)^{2+2/\eta}(6+6/\eta)^{1+4/\eta}$. Returning to (8), we arrive at

$$\begin{aligned}\mathbb{P}\big(\|\Sigma_x^{1/2}(\hat\beta_n-\beta)\|/\sigma>t\big) &\le \mathbb{P}\big(\|\hat\Sigma_{\bar x}^{-1}\|>2\big)+\mathbb{P}\big(\|\bar X'\Xi/n\|/\sigma>t/4\big)\\ &\quad+\mathbb{P}\big(\|\bar X'\iota/n\|\,|\iota'\Xi/n|/\sigma>t/4\big)\\ &\le 4\bigg((C_0\vee 1)\frac{p}{n}\bigg)^{\frac12\frac{\eta}{1+\eta}}+\frac{64E_4}{t^2}\frac{p}{n}+\frac{8}{t}\frac{\sqrt{p}}{n},\end{aligned}$$

which completes the proof of (5), upon defining $C_1 = 4(C_0\vee 1)^{\eta/(2+2\eta)}$.

For the inequality in (6), we first note that the distribution of the estimation error $\hat\alpha_n-\alpha$ does depend on the mean parameter $\mu\in\mathbb{R}^d$ which is generally unrestricted. Next, we use the Sherman–Morrison formula to rewrite

$$\begin{aligned}\frac1n\iota'(I_n-P_X)\iota &= \frac1n\iota'(I_n-P_{\bar X})\iota\\ &= 1-\frac{\iota'\bar X}{n}\big(\hat\Sigma_{\bar x}+\bar X'\iota\iota'\bar X/n^2\big)^{-1}\frac{\bar X'\iota}{n}\\ &= 1-\frac{\frac{\iota'\bar X}{n}\hat\Sigma_{\bar x}^{-1}\frac{\bar X'\iota}{n}}{1+\frac{\iota'\bar X}{n}\hat\Sigma_{\bar x}^{-1}\frac{\bar X'\iota}{n}}=\frac{1}{1+\frac{\iota'\bar X}{n}\hat\Sigma_{\bar x}^{-1}\frac{\bar X'\iota}{n}},\end{aligned}$$

and similarly,

$$\frac1n\iota'(I_n-P_X)\Xi=\frac{\iota'\Xi}{n}-\frac{\frac{\iota'\bar X}{n}\hat\Sigma_{\bar x}^{-1}\frac{\bar X'\Xi}{n}}{1+\frac{\iota'\bar X}{n}\hat\Sigma_{\bar x}^{-1}\frac{\bar X'\iota}{n}},$$

to arrive at

$$\hat\alpha_n-\alpha=\frac{\frac1n\iota'(I_n-P_X)\Xi}{\frac1n\iota'(I_n-P_X)\iota}=\frac{\iota'\Xi}{n}\bigg(1+\frac{\iota'\bar X}{n}\hat\Sigma_{\bar x}^{-1}\frac{\bar X'\iota}{n}\bigg)-\frac{\iota'\bar X}{n}\hat\Sigma_{\bar x}^{-1}\frac{\bar X'\Xi}{n},$$

and, in turn, at the bound

$$|\hat\alpha_n-\alpha|\le\bigg|\frac{\iota'\Xi}{n}\bigg|+\|\hat\Sigma_{\bar x}^{-1}\|\bigg\|\frac{\bar X'\iota}{n}\bigg\|^2\bigg|\frac{\iota'\Xi}{n}\bigg|+\|\hat\Sigma_{\bar x}^{-1}\|\bigg\|\frac{\bar X'\iota}{n}\bigg\|\bigg\|\frac{\bar X'\Xi}{n}\bigg\|.$$

Next, since $\mathbb{E}[\|\bar X'\iota/n\|^2]=p/n+\|\mu_{\bar x}\|^2$, we have for $t,\delta>0$,

$$\begin{aligned}\mathbb{P}\bigg(\bigg\|\frac{\bar X'\iota}{n}\bigg\|^2\bigg|\frac{\iota'\Xi}{n}\bigg|\Big/\sigma>t\bigg)&\le\mathbb{P}\bigg(\bigg|\frac{\iota'\Xi}{n}\bigg|\Big/\sigma>t/\delta\bigg)+\mathbb{P}\bigg(\bigg\|\frac{\bar X'\iota}{n}\bigg\|^2>\delta\bigg)\\ &\le\frac{\delta^2}{t^2}\frac1n+\frac{p/n+\|\mu_{\bar x}\|^2}{\delta},\end{aligned}$$

and this upper bound is minimized at $\delta_0=(nt^2[p/n+\|\mu_{\bar x}\|^2]/2)^{1/3}$, which leads to the optimized bound

$$\mathbb{P}\bigg(\bigg\|\frac{\bar X'\iota}{n}\bigg\|^2\bigg|\frac{\iota'\Xi}{n}\bigg|\Big/\sigma>t\bigg)\le\bigg(\frac{1}{nt^2}\bigg)^{1/3}(1+\|\mu_{\bar x}\|^2)^{2/3}(4^{-1/3}+2^{1/3}).\tag{10}$$

Similarly, using Lemma A.3(ii), we obtain

$$\begin{aligned}\mathbb{P}\Big(\Big\|\frac{\bar{X}'\iota}{n}\Big\|\Big\|\frac{\bar{X}'\Xi}{n}\Big\|\Big/\sigma > t\Big) &\le \mathbb{P}\Big(\Big\|\frac{\bar{X}'\Xi}{n}\Big\|\Big/\sigma > t/\delta\Big) + \mathbb{P}\Big(\Big\|\frac{\bar{X}'\iota}{n}\Big\| > \delta\Big)\\ &\le \frac{\delta^2}{t^2}\frac{p}{n}4(1+\|\mu_{\bar{x}}\|^2/p)E_4 + \frac{p/n+\|\mu_{\bar{x}}\|^2}{\delta^2},\end{aligned}$$

for $t, \delta > 0$. Optimizing this over $\delta > 0$ is easy and yields the bound

$$\begin{aligned}\mathbb{P}\Big(\Big\|\frac{\bar{X}'\iota}{n}\Big\|\Big\|\frac{\bar{X}'\Xi}{n}\Big\|\Big/\sigma > t\Big) &\le \frac{2}{t}\sqrt{\Big(\frac{p}{n}+\|\mu_{\bar{x}}\|^2\Big)\frac{p}{n}4(1+\|\mu_{\bar{x}}\|^2/p)E_4}\\ &\le \frac{4}{t}\sqrt{\frac{p}{n}}(1+\|\mu_{\bar{x}}\|^2)\sqrt{E_4}.\end{aligned} \tag{11}$$

Now we return to the scaled absolute estimation error and combine (9), (10) and (11) to get

$$\begin{aligned}&\mathbb{P}(|\hat{\alpha}_n - \alpha|/\sigma > t)\\ &\quad\le \mathbb{P}(|\iota'\Xi/n|/\sigma > t/3) + \mathbb{P}\Big(\|\hat{\Sigma}_{\bar{x}}^{-1}\|\Big\|\frac{\bar{X}'\iota}{n}\Big\|^2\Big|\frac{\iota'\Xi}{n}\Big|\Big/\sigma > t/3\Big)\\ &\qquad+ \mathbb{P}\Big(\|\hat{\Sigma}_{\bar{x}}^{-1}\|\Big\|\frac{\bar{X}'\iota}{n}\Big\|\Big\|\frac{\bar{X}'\Xi}{n}\Big\|\Big/\sigma > t/3\Big)\\ &\quad\le \frac{9}{t^2}\frac{1}{n} + \mathbb{P}\Big(\Big\|\frac{\bar{X}'\iota}{n}\Big\|^2\Big|\frac{\iota'\Xi}{n}\Big|\Big/\sigma > t/6\Big)\\ &\qquad+ \mathbb{P}\Big(\Big\|\frac{\bar{X}'\iota}{n}\Big\|\Big\|\frac{\bar{X}'\Xi}{n}\Big\|\Big/\sigma > t/6\Big) + 2\mathbb{P}(\|\hat{\Sigma}_{\bar{x}}^{-1}\| > 2)\\ &\quad\le \frac{9}{t^2}\frac{1}{n} + \Big(\frac{36}{nt^2}\Big)^{1/3}(1+\|\mu_{\bar{x}}\|^2)^{2/3}(4^{-1/3}+2^{1/3})\\ &\qquad+ \frac{24}{t}\sqrt{\frac{p}{n}}(1+\|\mu_{\bar{x}}\|^2)\sqrt{E_4} + 8\Big((C_0\vee 1)\frac{p}{n}\Big)^{\frac{1}{2}\frac{\eta}{1+\eta}}\\ &\quad\le 2C_1\Big(\frac{p}{n}\Big)^{\frac{1}{2}\frac{\eta}{1+\eta}} + 24\Big(\frac{1}{nt^2}+\Big(\frac{1}{nt^2}\Big)^{1/3}+\Big(\frac{p}{nt^2}\Big)^{1/2}\Big)(1+\|\mu_{\bar{x}}\|^2)\sqrt{E_4},\end{aligned}$$

which concludes the proof. □

A.2. Proof of Theorems 3.1 and 3.4. The proof of Theorem 3.1(i) is based on Theorem 2.1(i) of Steinberger and Leeb (2018) with $Z = \tilde{z}$ and $\tau = 1/2$. Note that if $f_{\tilde{z}} \in \mathcal{F}_{d,12}(D, E)$, then the assumptions of that result are satisfied [cf. Steinberger and Leeb (2018), Example 3.1], and we obtain existence of a Borel

subset $\mathbb{G}_1 = \mathbb{G}_1(f_{\tilde{z}}) \subseteq \mathcal{V}_{d,p}$ of the Stiefel manifold $\mathcal{V}_{d,p}$ of order $d \times p$, that depends on the density $f_{\tilde{z}}$, such that for all $t > 0$,

$$\sup_{B \in \mathbb{G}_1} \mathbb{P}\big(\big\|\mathbb{E}[\tilde{z}\|B'\tilde{z}] - BB'\tilde{z}\big\| > t\big) \le \frac{1}{t} d^{-1/12} + 4\gamma_1 \frac{p}{\log d}, \tag{12}$$

and

$$\nu_{d,p}\big(\mathbb{G}_1^c\big) \le \kappa_1 d^{-(1-12\gamma_1 \frac{p}{\log d})/12}, \tag{13}$$

where $\nu_{d,p}$ denotes the uniform distribution on the Stiefel manifold. Here, the constant $\gamma_1 = \gamma_1(D)$ depends only on D, and the constant $\kappa_1 = \kappa_1(E)$ depends only on E. Moreover, it is easy to see from (12) [cf. also Theorem 2.1(iii) of Steinberger and Leeb (2018)] that the set $\mathbb{G}_1$ is right-invariant under the action of $\mathcal{O}_p$ and that it is left-equivariant in the sense that $\mathbb{G}_1(f_{R\tilde{z}}) = R\mathbb{G}_1(f_{\tilde{z}})$, for every $R \in \mathcal{O}_d$. For any full rank $d \times p$ matrix M, any symmetric positive definite $d \times d$ matrix Σ and $f_{\tilde{z}} \in \mathcal{F}_{d,12}(D, E)$, we define the set

$$\mathbb{U} := \mathbb{U}(M, \Sigma, f_{\tilde{z}}) := \big\{R \in \mathcal{O}_d : R'\Sigma^{1/2}M\big(M'\Sigma M\big)^{-1/2} \in \mathbb{G}_1(f_{\tilde{z}})\big\}.$$

From the properties of $\mathbb{G}_1$ mentioned above, it is easy to deduce the claim of Theorem 3.1(iii) about the set $\mathbb{U}$. Now take a random matrix U that is uniformly distributed on $\mathcal{O}_d$ and another random matrix V that is uniformly distributed on $\mathcal{O}_p$, such that U and V are independent, and note that by right-invariance of $\mathbb{G}_1$,

$$\begin{aligned}
\nu_d(\mathbb{U}) &= \mathbb{P}\big(U\Sigma^{1/2}M\big(M'\Sigma M\big)^{-1/2} \in \mathbb{G}_1\big) \\
&= \mathbb{P}\big(U\Sigma^{1/2}M\big(M'\Sigma M\big)^{-1/2}V \in \mathbb{G}_1\big) \\
&= \nu_{d,p}(\mathbb{G}_1),
\end{aligned}$$

because $\Sigma^{1/2}M(M'\Sigma M)^{-1/2} \in \mathcal{V}_{d,p}$ and $\nu_{d,p}$ is characterized by left and right invariance under the appropriate orthogonal groups. This establishes the desired bound on $\nu_d(\mathbb{U}^c)$, upon choosing $L_1 = \kappa_1$ and $c_1 = 12\gamma_1$.

For the statement about the ratio of $\mathcal{R}_N$ and $\mathcal{R}_L$, we first note that it is no restriction to consider only parameter configurations such that $\mathcal{R}_L > 0$, because otherwise $\mathcal{R}_N = \mathcal{R}_L = 0$ and $1 - \mathcal{R}_N/\mathcal{R}_L = 0$ by convention, so that the desired inequality is trivially true on this portion of the parameter space. Moreover, it suffices to consider M, Σ and $f_{\tilde{z}} \in \mathcal{F}_{d,12}(D, E)$ such that $\mathbb{U}(M, \Sigma, f_{\tilde{z}}) \neq \varnothing$, because, by convention, $\sup \varnothing = -\infty$. Now fix M, θ, $\mathcal{L}(\epsilon)$, Σ, $f_{\tilde{z}} \in \mathcal{F}_{d,12}(D, E)$ and $R \in \mathbb{U}(M, \Sigma, f_{\tilde{z}})$ as above, write

$$\begin{aligned}
1 - \mathcal{R}_N/\mathcal{R}_L &= (\mathcal{R}_L - \mathcal{R}_N)/\mathcal{R}_L \\
&= \mathbb{E}\big[\big(\mathbb{E}[y\|x] - (\alpha + \beta' x)\big)^2/\mathcal{R}_L\big],
\end{aligned}$$

where the second equality is easy to verify, and define $\Delta_1 := (\mathbb{E}[y\|x] - (\alpha + \beta' x))/\sqrt{\mathcal{R}_L}$. Fix $\delta > 0$ and $a, b \geq 1$ such that $1/a + 1/b = 1$, and use Hölder's inequality to obtain the bound

$$\begin{aligned}\mathbb{E}[\Delta_1^2] &= \mathbb{E}[\Delta_1^2 \mathbf{1}_{\{|\Delta_1| > \delta\}}] + \mathbb{E}[\Delta_1^2 \mathbf{1}_{\{|\Delta_1| \leq \delta\}}] \\ &\leq \big(\mathbb{E}[\Delta_1^{2a}]\big)^{1/a} \big(\mathbb{P}(|\Delta_1| > \delta)\big)^{1/b} + \delta^2.\end{aligned}$$

In view of the preliminary considerations of the Appendix, $\mathcal{R}_L = \|(I_d - P_{\Sigma^{1/2}M})\Sigma^{1/2}\theta\|^2 + \mathrm{Var}[\epsilon] > 0$ and

$$\Delta_1 = \mathbb{E}[\theta' \Sigma^{1/2}(I_d - P_{\Sigma^{1/2}M}) R\tilde{z}/\sqrt{\mathcal{R}_L}\|x],$$

and thus, for $v = R'(I_d - P_{\Sigma^{1/2}M})\Sigma^{1/2}\theta/\sqrt{\mathcal{R}_L}$, $\mathbb{E}[\Delta_1^{2a}] \leq \mathbb{E}[|v'\tilde{z}|^{2a}] \leq E_{2a}$, in view of $\|v\| \leq 1$. Because $f_{\tilde{z}} \in \mathcal{F}_{d,12}(D, E)$ and $R \in \mathbb{U}(M, \Sigma, f_{\tilde{z}})$, we have $B_0 := R'\Sigma^{1/2}M(M'\Sigma M)^{-1/2} \in \mathbb{G}_1(f_{\tilde{z}})$, and thus inequality (12) entails that

$$\begin{aligned}\mathbb{P}(|\Delta_1| > \delta) &\leq \mathbb{P}\big(|v'(\mathbb{E}[\tilde{z}\|B_0'\tilde{z}] - B_0 B_0'\tilde{z})| > \delta\big) \\ &\leq \frac{1}{\delta} d^{-1/12} + 4\gamma_1 \frac{p}{\log d}. \qquad (14)\end{aligned}$$

Altogether, we obtain the upper bound

$$\mathbb{E}[\Delta_1^2] \leq E_{2a}^{1/a}\left(d^{-1/12}\delta^{-1} + 4\gamma_1 \frac{p}{\log d}\right)^{1/b} + \delta^2.$$

We cannot analytically optimize this upper bound in δ. But clearly, the dominating term in this bound is $(p/\log d)^{1/b}$, and we cannot hope to improve the rate as $p/\log d \to 0$ beyond this. So we choose δ such that $\delta^2 = c(p/\log d)^{1/b}$ for some $c > 0$. This also entails that

$$d^{-1/12}\delta^{-1} = c^{-1/2}\frac{p}{\log d}\left(\frac{\log d}{p d^{\frac{2b}{12(2b+1)}}}\right)^{\frac{2b+1}{2b}} \leq c^{-1/2}\frac{p}{\log d} K_0(b),$$

where $K_0(b) := \max\{(d^{-b/(12b+6)}\log d)^{1+1/(2b)}, d \geq 2\}$ depends only on b. Hence, substituting δ, we arrive at the upper bound

$$\mathbb{E}[\Delta_1^2] \leq \big[E_{2a}^{1/a}\big(c^{-1/2}K_0(b) + 4\gamma_1\big)^{1/b} + c\big]\left(\frac{p}{\log d}\right)^{1/b}.$$

Choosing $b > 1$ as small as possible optimizes the rate, while $a > 1$ should be chosen small enough to guarantee that E_{2a} is still bounded. Using Rosenthal's inequality [Rosenthal (1970), Theorem 3], we get that E_{2a} is bounded by $\max_j \mathbb{E}[|e_j'\tilde{z}|^{2a}]$ times a constant $C(a) > 0$ that depends only on a, and where e_j is the jth element of the standard basis in $\mathbb{R}^d$. Under the twelfth moment bound that we get from $f_{\tilde{z}} \in \mathcal{F}_{d,12}(D, E)$, we may thus take $a = 6$ and $b = 6/5$, and set

$$K_1 := K_1(D, E) := \inf_{c>0}\big[\big(EC(6)\big)^{1/6}\big(c^{-1/2}K_0(6/5) + 4\gamma_1(D)\big)^{5/6} + c\big]$$

to obtain the desired result.

To establish part (ii), we first construct the set $\mathbb{V}$ in a similar way as above, using Theorem 2.1(ii) of Steinberger and Leeb (2018) (see also Example 3.1 in that reference), again, with $Z = \tilde{z}$ and $\tau = 1/2$. In particular, if $f_{\tilde{z}} \in \mathcal{F}_{d,20}(D, E)$, then there exists a Borel set $\mathbb{G}_2 \subseteq \mathcal{V}_{d,p}$ of the Stiefel manifold, such that (12) holds with $\mathbb{G}_2$ replacing $\mathbb{G}_1$, because $\mathcal{F}_{d,20}(D, E) \subseteq \mathcal{F}_{d,12}(D, E^{3/5})$, such that

$$\begin{aligned}\sup_{B\in\mathbb{G}_2} \mathbb{P}\big(\big\|\mathbb{E}\big[\tilde{z}\tilde{z}'\|B'\tilde{z}\big] - \big(I_d - BB' + BB'\tilde{z}\tilde{z}'BB'\big)\big\| > t\big) \\ \leq \frac{1}{t}d^{-1/20} + 4\gamma_2\frac{p}{\log d},\end{aligned} \tag{15}$$

for every $t > 0$, and such that

$$\nu_{d,p}\big(\mathbb{G}_2^c\big) \leq \kappa_2 d^{-(1-20\gamma_2\frac{p}{\log d})/20}, \tag{16}$$

where the constant $\gamma_2 = \gamma_2(D)$ depends only on D, and the constant $\kappa_2 = \kappa_2(E)$ depends only on E. For any full rank $d \times p$ matrix M, any symmetric positive definite $d \times d$ matrix Σ and $f_{\tilde{z}} \in \mathcal{F}_{d,20}(D, E)$, we define the set

$$\mathbb{V} := \mathbb{V}(M, \Sigma, f_{\tilde{z}}) := \big\{R \in \mathcal{O}_d : R'\Sigma^{1/2}M\big(M'\Sigma M\big)^{-1/2} \in \mathbb{G}_2(f_{\tilde{z}})\big\},$$

and the same argument as above, involving right-invariance and left-equivariance of $\mathbb{G}_2$, establishes the bound on $\nu_d(\mathbb{V}^c)$ claimed by part (ii), and the properties of $\mathbb{V}$ claimed by part (iii) of Theorem 3.1.

By analogous arguments as in part (i), using also the convention that $\mathcal{R}_N(x)/\mathcal{R}_L(x) = 1$ if $\mathcal{R}_L(x) = 0$, it suffices to consider parameter choices M, Σ and $f_{\tilde{z}} \in \mathcal{F}_{d,20}(D, E)$, such that $\mathbb{V}(M, \Sigma, f_{\tilde{z}}) \neq \varnothing$ and we may restrict to the event where $\mathcal{R}_L(x) > 0$. If for a given parameter configuration the probability of the event $\{\mathcal{R}_L(x) > 0\}$ is equal to zero, then the desired upper bound is trivially true on this portion of the parameter space and it remains to consider those parameters for which $\mathbb{P}(\mathcal{R}_L(x) > 0) > 0$. On this part of the parameter space, we hence also have $\mathcal{R}_L = \mathbb{E}[\mathcal{R}_L(x)] > 0$. Therefore, we may consider

$$1 - \mathcal{R}_N(x)/\mathcal{R}_L(x) = \big(\mathcal{R}_L(x) - \mathcal{R}_N(x)\big)/\mathcal{R}_L(x) = \Delta_1^2\frac{\mathcal{R}_L}{\mathcal{R}_L(x)},$$

and conclude that for $t > 0$:

$$\begin{aligned}&\mathbb{P}\big(1 - \mathcal{R}_N(x)/\mathcal{R}_L(x) > t, \mathcal{R}_L(x) > 0\big) \\ &\quad\leq \mathbb{P}\big(\Delta_1^2 > t/2\big) + \mathbb{P}\bigg(\frac{\mathcal{R}_L}{\mathcal{R}_L(x)} > 2, \mathcal{R}_L(x) > 0\bigg) \\ &\quad\leq \sqrt{2}d^{-1/12}t^{-1/2} + 4\gamma_1\frac{p}{\log d} + \mathbb{P}\bigg(\bigg|\frac{\mathcal{R}_L(x)}{\mathcal{R}_L} - 1\bigg| > 1/2\bigg).\end{aligned}$$

To bound the probability on the last line of the previous display, recall that because of $R \in \mathbb{V}(M, \Sigma, f_{\tilde{z}})$, we have $B_0 = R'\Sigma^{1/2}M(M'\Sigma M)^{-1/2} \in \mathbb{G}_2(f_{\tilde{z}})$, and use Lemma A.2. Therefore, (15) entails that

$$\mathbb{P}\left(\left|\frac{\mathcal{R}_L(x)}{\mathcal{R}_L} - 1\right| > 1/2\right) \le 2d^{-1/20} + 4\gamma_2\frac{p}{\log d},$$

and the proof of Theorem 3.1(ii) is completed upon appropriately choosing the constant $K_2 = K_2(D)$.

Next, we prove Theorem 3.4. The constant $L_0(E)$ will be chosen at the end of the proof of part (ii) of that theorem. For the proof of Theorem 3.4(i), we take the set $\mathbb{U}$ as above and note that because of Lemma A.1, we may set $\vartheta = 0$ and $\mu = 0$. By our conventions, it suffices to consider the event $H_1 = \{\mathcal{R}_{\mathrm{OLS}}(X, Y) > 0\}$. If $\mathbb{P}(H_1) = 0$, then the result is trivially true, so we need to consider only parameters for which $\mathbb{P}(H_1) > 0$. Note that for such a choice of parameters we must also have $\mathcal{R}_L > 0$, because otherwise $0 = \mathcal{R}_L = \mathbb{E}[\xi^2]$, which implies that $\xi = \xi_i = 0$, a.s., and thus $\hat{\gamma}_n = \gamma$, a.s., so that $\mathcal{R}_{\mathrm{OLS}}(X, Y) = \mathbb{E}[(y - (1, x')\hat{\gamma}_n)^2 \| X, Y] = \mathcal{R}_L = 0$, a.s., which means that $\mathbb{P}(H_1) = 0$. Now, on H_1, we get

$$1 - \frac{\mathcal{R}_N}{\mathcal{R}_{\mathrm{OLS}}(X, Y)} = 1 - \frac{\mathcal{R}_N}{\mathcal{R}_L}\frac{\mathcal{R}_L}{\mathcal{R}_{\mathrm{OLS}}(X, Y)}.$$

We may therefore rewrite and bound the expression of interest on the event H_1 as

$$\begin{aligned}
1 - \frac{\mathcal{R}_N}{\mathcal{R}_{\mathrm{OLS}}(X, Y)} &= 1 - \frac{\mathcal{R}_N}{\mathcal{R}_L} + \frac{\mathcal{R}_N}{\mathcal{R}_L}\left(1 - \frac{\mathcal{R}_L}{\mathcal{R}_{\mathrm{OLS}}(X, Y)}\right) \\
&\le 1 - \frac{\mathcal{R}_N}{\mathcal{R}_L} + \frac{\mathcal{R}_{\mathrm{OLS}}(X, Y) - \mathcal{R}_L}{\mathcal{R}_{\mathrm{OLS}}(X, Y)} \\
&\le 1 - \frac{\mathcal{R}_N}{\mathcal{R}_L} + \frac{\mathcal{R}_{\mathrm{OLS}}(X, Y) - \mathcal{R}_L}{\mathcal{R}_L},
\end{aligned}$$

because $\mathcal{R}_N \le \mathcal{R}_L \le \mathcal{R}_{\mathrm{OLS}}(X, Y)$. Using the notation introduced at the beginning of the Appendix, it is easy to see that

$$\begin{aligned}
\mathcal{R}_{\mathrm{OLS}}(X, Y) &= \mathbb{E}\big[\big(\xi + \alpha - \hat{\alpha}_n + (\beta - \hat{\beta}_n)'x\big)^2 \| X, Y\big] \\
&= \mathcal{R}_L + \mathbb{E}\big[\big(\alpha - \hat{\alpha}_n + (\beta - \hat{\beta}_n)'x\big)^2 \| X, Y\big] \\
&= \mathcal{R}_L + (\hat{\alpha}_n - \alpha)^2 + \big\|\Sigma_x^{1/2}(\hat{\beta}_n - \beta)\big\|^2 > 0,
\end{aligned}$$

because the residual $\xi = y - (\alpha + \beta'x)$ of orthogonal projection onto the space of linear functions in x is orthogonal on everything in that space and because

$\mathbb{E}[x\|X,Y]=\mathbb{E}[x]=M'\mu=0$. Together, since $\sigma^2=\mathcal{R}_L$ and $R\in\mathbb{U}$, we get

$$
\begin{aligned}
&\mathbb{P}\biggl(1-\frac{\mathcal{R}_N}{\mathcal{R}_{\mathrm{OLS}}(X,Y)}>t\biggr)\\
&\quad\le\mathbb{P}\biggl(1-\frac{\mathcal{R}_N}{\mathcal{R}_L}>t/2\biggr)+\mathbb{P}\biggl(\frac{\mathcal{R}_{\mathrm{OLS}}(X,Y)-\mathcal{R}_L}{\sigma^2}>t/2\biggr)\\
&\quad\le\frac{2}{t}K_1\biggl(\frac{p}{\log d}\biggr)^{5/6}\\
&\qquad+\mathbb{P}\bigl((\hat\alpha_n-\alpha)^2/\sigma^2+\bigl\|\Sigma_x^{1/2}(\hat\beta_n-\beta)\bigr\|^2/\sigma^2>t/2\bigr)\\
&\quad\le\frac{2}{t}K_1\biggl(\frac{p}{\log d}\biggr)^{5/6}+\mathbb{P}\bigl(|\hat\alpha_n-\alpha|/\sigma>\sqrt{t}/2\bigr)\\
&\qquad+\mathbb{P}\bigl(\bigl\|\Sigma_x^{1/2}(\hat\beta_n-\beta)\bigr\|/\sigma>\sqrt{t}/2\bigr)
\end{aligned}
\tag{17}
$$

in view of Theorem 3.1(i), where the constant $K_1=K_1(D,E)$ depends only on D and E. In order to apply Lemma A.4, we have to verify its assumptions. But this is carried out in Section 1.5 of Srivastava and Vershynin (2013). In particular, since $f_{\tilde z}\in\mathcal{F}_{d,12}(D,E)$, the discussion in that reference shows that there exists a constant $\bar C=\bar C(E)$ that depends only on the moment bound $E>0$, such that for every projection matrix P in $\mathbb{R}^d$ and every $t>\bar C\operatorname{rank}P$,

$$\mathbb{P}\bigl(\|P\tilde z\|>t\bigr)\le\bar Ct^{-3}.$$

So we may apply Lemma A.4 with $\eta=2$. Thus, if $n/p>\bar C_0:=512(48\bar C)^3 9^3$ and for $\bar C_1=\bar C_1(E):=4(\bar C_0\vee 1)^{1/3}$, we have

$$
\begin{aligned}
\mathbb{P}\bigl(\bigl\|\Sigma_x^{1/2}(\hat\beta_n-\beta)\bigr\|/\sigma>t\bigr)&\le\bar C_1\biggl(\frac{p}{n}\biggr)^{\frac12\frac{\eta}{1+\eta}}+\frac{64E_4}{t^2}\frac{p}{n}+\frac{8}{t}\frac{\sqrt{p}}{n}\\
&\le\biggl(\frac{p}{n}\biggr)^{1/3}E_4\biggl(\bar C_1+\frac{64}{t^2}+\frac{8}{t}\biggr).
\end{aligned}
$$

Moreover, we also have

$$
\begin{aligned}
&\mathbb{P}\bigl(|\hat\alpha_n-\alpha|/\sigma>t\bigr)\\
&\quad\le2\bar C_1\biggl(\frac{p}{n}\biggr)^{\frac12\frac{\eta}{1+\eta}}\\
&\qquad+24\biggl(\frac{1}{nt^2}+\biggl(\frac{1}{nt^2}\biggr)^{1/3}+\biggl(\frac{p}{nt^2}\biggr)^{1/2}\biggr)\bigl(1+\bigl\|\Sigma_x^{-1/2}\mu_x\bigr\|^2\bigr)\sqrt{E_4}\\
&\quad\le\biggl(\frac{p}{n}\biggr)^{1/3}E_4\biggl(2\bar C_1+\frac{24}{t^2}+\frac{24}{t^{2/3}}+\frac{24}{t}\biggr),
\end{aligned}
$$

because $\mu_x = M'\mu = 0$. Finally, returning to (17), we get

$$\mathbb{P}\Big(1 - \frac{\mathcal{R}_N}{\mathcal{R}_{\mathrm{OLS}}(X,Y)} > t\Big) \le \Big(\frac{p}{\log d}\Big)^{5/6} K_1 \frac{2}{t} + \Big(\frac{p}{n}\Big)^{1/3} E_4 \tilde{L}_3,$$

where $\tilde{L}_3 = \tilde{L}_3(E,t)$ depends only on E and t. Noting that E_4 can be upper bounded by a constant that depends only on E, in view of Rosenthal's inequality [Rosenthal (1970), Theorem 3], completes the proof of Theorem 3.4(i).

The set $\mathbb{V}$ in Theorem 3.4(ii) is the same as above. Next, in view of Lemma A.1, it is no restriction to set $\vartheta = 0$ and $\mu = 0$. For the statement about the OLS risk $\mathcal{R}_{\mathrm{OLS}}(X,Y,x)$, we note that by our conventions for division by zero, it suffices to consider the event $H_2 := \{\mathcal{R}_{\mathrm{OLS}}(X,Y,x) > 0\}$. Thus, we only need to consider parameter configurations for which $\mathbb{P}(H_2) > 0$. For such a choice of parameters, we also have $\mathcal{R}_L > 0$, because otherwise $0 = \mathcal{R}_L = \mathbb{E}[\xi^2]$, which implies that $\xi = \xi_i = 0$, a.s., and thus $\hat{\gamma}_n = \gamma$, a.s., so that $\mathcal{R}_{\mathrm{OLS}}(X,Y,x) = \mathbb{E}[(y - (1,x')\hat{\gamma}_n)^2 \| X,Y,x] = \mathbb{E}[\xi^2 \| x] = 0$, a.s., and we have ruled out this case already. Now, on H_2, we get

$$1 - \frac{\mathcal{R}_N(x)}{\mathcal{R}_{\mathrm{OLS}}(X,Y,x)} = 1 - \frac{\mathcal{R}_N(x)}{\mathcal{R}_L(x)} \frac{\mathcal{R}_L(x)}{\mathcal{R}_{\mathrm{OLS}}(X,Y,x)},$$

because if $\mathcal{R}_L(x) = 0$, then $\mathcal{R}_N(x) = 0$, and by convention, $\mathcal{R}_N(x)/\mathcal{R}_L(x) = 1$, and both expressions in the display above are equal to 1. We may therefore rewrite and bound the expression of interest on the event H_2 as

$$\begin{aligned} 1 - \frac{\mathcal{R}_N(x)}{\mathcal{R}_{\mathrm{OLS}}(X,Y,x)} &= \Big|1 - \frac{\mathcal{R}_N(x)}{\mathcal{R}_L(x)} + \frac{\mathcal{R}_N(x)}{\mathcal{R}_L(x)}\Big(1 - \frac{\mathcal{R}_L(x)}{\mathcal{R}_{\mathrm{OLS}}(X,Y,x)}\Big)\Big| \\ &\le 1 - \frac{\mathcal{R}_N(x)}{\mathcal{R}_L(x)} + \Big|\frac{\mathcal{R}_{\mathrm{OLS}}(X,Y,x) - \mathcal{R}_L(x)}{\mathcal{R}_{\mathrm{OLS}}(X,Y,x)}\Big|. \end{aligned}$$

Using the notation introduced at the beginning of the Appendix, it is easy to see that

$$\begin{aligned} \mathcal{R}_{\mathrm{OLS}}(X,Y,x) &= \mathbb{E}\big[\big(\xi + \alpha - \hat{\alpha}_n + (\beta - \hat{\beta}_n)'x\big)^2 \| X,Y,x\big] \\ &= \mathcal{R}_L(x) + 2\mathbb{E}[\xi \| x]\big(\alpha - \hat{\alpha}_n \\ &\quad + (\beta - \hat{\beta}_n)'x\big) + \big(\alpha - \hat{\alpha}_n + (\beta - \hat{\beta}_n)'x\big)^2 \\ &=: \mathcal{R}_L(x) + \Delta_2. \end{aligned}$$

Together we get, for $R \in \mathbb{V}$,

$$\begin{aligned} &\mathbb{P}\Big(1 - \frac{\mathcal{R}_N(x)}{\mathcal{R}_{\mathrm{OLS}}(X,Y,x)} > t\Big) \\ &\quad \le \mathbb{P}\Big(1 - \frac{\mathcal{R}_N(x)}{\mathcal{R}_L(x)} > t/2\Big) \end{aligned}$$

$$
\begin{aligned}
&+\mathbb{P}\biggl(\biggl|\frac{\mathcal{R}_{\mathrm{OLS}}(X,Y,x)-\mathcal{R}_L(x)}{\mathcal{R}_{\mathrm{OLS}}(X,Y,x)}\biggr|>t/2,H_2\biggr)\\
&\le 2d^{-\frac{1}{12}}t^{-\frac{1}{2}}+K_2\frac{p}{\log d}\\
&\quad+\mathbb{P}\bigl(|\Delta_2|/\sigma^2>t/4\bigr)+\mathbb{P}\bigl(\bigl|\mathcal{R}_L(x)+\Delta_2\bigr|/\sigma^2\le 1/2\bigr)\\
&\le 2d^{-\frac{1}{12}}t^{-\frac{1}{2}}+K_2\frac{p}{\log d}+\mathbb{P}\bigl(|\Delta_2|/\sigma^2>t/4\bigr)+\mathbb{P}\bigl(|\Delta_2|/\sigma^2\ge 1/4\bigr)\\
&\quad+\mathbb{P}\biggl(\biggl|1-\frac{\mathcal{R}_L(x)}{\mathcal{R}_L}\biggr|>1/4\biggr),
\end{aligned}
\tag{18}
$$

in view of Theorem 3.1(ii), and where we have used the reverse triangle inequality to get $1-|\mathcal{R}_L(x)+\Delta_2|/\sigma^2\le|1-\mathcal{R}_L(x)/\sigma^2|+|\Delta_2/\sigma^2|$. The constant $K_2=K_2(D)$ depends only on D. Since we again have $B_0=R'\Sigma^{1/2}M(M'\Sigma M)^{-1/2}\in\mathbb{G}_2(f_{\tilde{z}})$, as in the proof of Theorem 3.1(ii), Lemma A.2 and (15) yield

$$
\mathbb{P}\biggl(\biggl|\frac{\mathcal{R}_L(x)}{\mathcal{R}_L}-1\biggr|>1/4\biggr)\le 4d^{-1/20}+4\gamma_2\frac{p}{\log d}.
$$

It remains to study the tail probabilities of $|\Delta_2|/\sigma^2$,

$$
\begin{aligned}
&\mathbb{P}\bigl(|\Delta_2|/\sigma^2>t\bigr)\\
&\quad\le\mathbb{P}\bigl(2|\mathbb{E}[\xi\|x](\alpha-\hat{\alpha}_n+(\beta-\hat{\beta}_n)'x)|/\sigma^2>t/2\bigr)\\
&\qquad+\mathbb{P}\bigl(\bigl|(\alpha-\hat{\alpha}_n+(\beta-\hat{\beta}_n)'x)\bigr|^2/\sigma^2>t/2\bigr)\\
&\quad\le\mathbb{P}\bigl(|\alpha-\hat{\alpha}_n+(\beta-\hat{\beta}_n)'x|/\sigma>t\bigr)\\
&\qquad+\mathbb{P}\bigl(|\mathbb{E}[\xi\|x]|/\sigma>1/4\bigr)\\
&\qquad+\mathbb{P}\bigl(\bigl|(\alpha-\hat{\alpha}_n+(\beta-\hat{\beta}_n)'x)\bigr|^2/\sigma^2>t/2\bigr)\\
&\quad\le\mathbb{P}\bigl(|\alpha-\hat{\alpha}_n|/\sigma>t/2\bigr)+\mathbb{P}\bigl(|(\beta-\hat{\beta}_n)'x|/\sigma>t/2\bigr)\\
&\qquad+\mathbb{P}\bigl(|\Delta_1|>1/4\bigr)\\
&\qquad+\mathbb{P}\bigl(|\alpha-\hat{\alpha}_n|/\sigma>t^{1/2}/2^{3/2}\bigr)+\mathbb{P}\bigl(|(\beta-\hat{\beta}_n)'x|/\sigma>t^{1/2}/2^{3/2}\bigr).
\end{aligned}
$$

To bound the tails of $|(\hat{\beta}_n-\beta)'x|/\sigma$, we use the conditional Markov inequality to get

$$
\begin{aligned}
\mathbb{P}\bigl(|(\hat{\beta}_n-\beta)'x|/\sigma>t\bigr)&=\mathbb{E}\bigl[\mathbb{P}\bigl(|(\hat{\beta}_n-\beta)'\Sigma_x^{1/2}\Sigma_x^{-1/2}x|^2>t^2\sigma^2\|\hat{\beta}_n\bigr)\bigr]\\
&\le\mathbb{E}\biggl[\biggl(\frac{1}{t^2\sigma^2}\|\Sigma_x^{1/2}(\hat{\beta}_n-\beta)\|^2\biggr)\wedge 1\biggr].
\end{aligned}
$$

Splitting the integral in the part where the integrand is greater than some arbitrary $\delta > 0$ and where it is not greater than δ, using boundedness by 1, we obtain

$$\begin{aligned}&\mathbb{E}\biggl[\biggl(\frac{1}{t^2\sigma^2}\bigl\|\Sigma_x^{1/2}(\hat{\beta}_n-\beta)\bigr\|^2\biggr)\wedge 1\biggr]\\&\qquad\le \mathbb{P}\bigl(\bigl\|\Sigma_x^{1/2}(\hat{\beta}_n-\beta)\bigr\|^2/\sigma^2>\delta t^2\bigr)+\delta.\end{aligned}$$

We have already verified the assumptions of Lemma A.4 in the case where $f_{\tilde{z}} \in \mathcal{F}_{d,12}(D, E)$. By an analogous argument, using the fact that now $f_{\tilde{z}} \in \mathcal{F}_{d,20}(D, E)$, there exists a constant $\tilde{C} = \tilde{C}(E)$ that depends only on the moment bound $E > 0$, such that for every projection matrix P in $\mathbb{R}^d$ and every $t > \tilde{C}\operatorname{rank} P$,

$$\mathbb{P}\bigl(\|P\tilde{z}\|>t\bigr)\le \tilde{C}t^{-5}.$$

So we may apply Lemma A.4 with $\eta = 4$. Thus, if $n/p > \tilde{C}_0 := 512(48\tilde{C})^{5/2}(15/2)^2$, we have

$$\begin{aligned}&\mathbb{P}\bigl(|(\hat{\beta}_n-\beta)'x|/\sigma>t\bigr)\\&\quad\le \tilde{C}_1\biggl(\frac{p}{n}\biggr)^{\frac{1}{2}\frac{\eta}{1+\eta}}+\frac{64E_4}{\delta t^2}\frac{p}{n}+\frac{8}{t\sqrt{\delta}}\frac{\sqrt{p}}{n}+\delta\\&\quad\le \tilde{C}_1\biggl(\frac{p}{n}\biggr)^{2/5}+\frac{64E_4}{t^2}\biggl(\frac{p}{n}\biggr)^{1/2}+\frac{8}{t}\biggl(\frac{p}{n}\biggr)^{3/4}+\biggl(\frac{p}{n}\biggr)^{1/2}\\&\quad\le \biggl(\frac{p}{n}\biggr)^{2/5}E_4\biggl(\tilde{C}_1+1+\frac{64}{t^2}+\frac{8}{t}\biggr),\end{aligned}$$

with $\delta = \sqrt{p/n}$ and where $\tilde{C}_1 = \tilde{C}_1(E) = 4(\tilde{C}_0 \vee 1)^{2/5}$. Moreover, we also have

$$\begin{aligned}&\mathbb{P}\bigl(|\hat{\alpha}_n-\alpha|/\sigma>t\bigr)\\&\quad\le 2\tilde{C}_1\biggl(\frac{p}{n}\biggr)^{\frac{1}{2}\frac{\eta}{1+\eta}}\\&\qquad+24\biggl(\frac{1}{nt^2}+\biggl(\frac{1}{nt^2}\biggr)^{1/3}+\biggl(\frac{p}{nt^2}\biggr)^{1/2}\biggr)\bigl(1+\bigl\|\Sigma_x^{-1/2}\mu_x\bigr\|^2\bigr)\sqrt{E_4}\\&\quad\le \biggl(\frac{p}{n}\biggr)^{1/3}E_4\biggl(2\tilde{C}_1+\frac{24}{t^2}+\frac{24}{t^{2/3}}+\frac{24}{t}\biggr),\end{aligned}$$

because $\mu_x = M'\mu = 0$. Since (12) still holds with $\mathbb{G}_2$ replacing $\mathbb{G}_1$, we can use (14) and put the previous two tail bounds together to arrive at

$$\mathbb{P}\bigl(|\Delta_2|/\sigma^2>t\bigr)\le\biggl(\frac{p}{n}\biggr)^{1/3}E_4C_2+4d^{-1/12}+4\gamma_1\frac{p}{\log d},$$

where $C_2 = C_2(\tilde{C}_1(E), t)$ is a positive finite constant that depends only on E and $t > 0$ and $\gamma_1 = \gamma_1(D)$ depends only on D. Finally, returning to (18), we get

$$\begin{aligned}
&\mathbb{P}\left(1 - \frac{\mathcal{R}_N(x)}{\mathcal{R}_{\mathrm{OLS}}(X, Y, x)} > t\right) \\
&\quad \le (2t^{-1/2} + 12)d^{-1/20} + (K_2 + 4\gamma_2 + 8\gamma_1)\frac{p}{\log d} + \left(\frac{p}{n}\right)^{1/3} E_4 C_3 \\
&\quad \le \left(\frac{p}{n}\right)^{1/3} E_4 C_3 + \frac{p}{\log d} C_4,
\end{aligned}$$

where $C_3 = C_3(E, t)$ depends only on E and t, and $C_4 = C_4(D, t)$ depends only on D and t. As above, we finally note that E_4 can be upper bounded by a constant that depends only on E, in view of Rosenthal's inequality [Rosenthal (1970), Theorem 3]. Choosing $L_0 = L_0(E) = \max\{\bar{C}_0(E), \tilde{C}_0(E)\}$ completes the proof of Theorem 3.4. □

Acknowledgments. We thank the Associate Editor and the referees for thoughtful comments and constructive feedback.

REFERENCES

ABADIE, A., IMBENS, G. W. and ZHENG, F. (2014). Inference for misspecified models with fixed regressors. *J. Amer. Statist. Assoc.* **109** 1601–1614. MR3293613

BACHOC, F., LEEB, H. and PÖTSCHER, B. M. (2015). Valid confidence intervals for post-model-selection prediction. Arxiv preprint. Available at arXiv:1412.4605.

BERAN, R. and DÜMBGEN, L. (1998). Modulation of estimators and confidence sets. *Ann. Statist.* **26** 1826–1856. MR1673280

BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122

BRANNATH, W. and SCHARPENBERG, M. (2014). Interpretation of linear regression coefficients under mean model miss-specification. Arxiv preprint. Available at arXiv:1409.8544.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761

BUJA, A. R., BROWN, L. D., GEORGE, E., PITKIN, E., TRASKIN, M., ZHAN, K. and ZHAO, L. (2014). A conspiracy of random predictors and model violations against classical inference in regression. Arxiv preprint. Available at arXiv:1404.1578.

DIACONIS, P. and FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12** 793–815. MR0751274

DÜMBGEN, L. and DEL CONTE-ZERIAL, P. (2013). On low-dimensional projections of high-dimensional distributions. In *From Probability to Statistics and Back*: *High-Dimensional Models and Processes*. *Inst. Math. Stat.* (*IMS*) *Collect.* **9** 91–104. IMS, Beachwood, OH. MR3186751

EATON, M. L. (1986). A characterization of spherical distributions. *J. Multivariate Anal.* **20** 272–276. MR0866075

EL KAROUI, N. (2010). The spectrum of kernel random matrices. *Ann. Statist.* **38** 1–50. MR2589315

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

HALL, P. and LI, K.-C. (1993). On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Statist.* **21** 867–889. MR1232523

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I*: *Statistics* 221–233. Univ. California Press, Berkeley, CA. MR0216620

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948

LEEB, H. (2008). 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** 661–690. MR2537807

LEEB, H. (2009). Conditional predictive inference post model selection. *Ann. Statist.* **37** 2838–2876. MR2541449

LEEB, H. (2013). On the conditional distributions of low-dimensional projections from high-dimensional data. *Ann. Statist.* **41** 464–483. MR3099110

LEEB, H., PÖTSCHER, B. M. and EWALD, K. (2015). On various confidence intervals post-model-selection. *Statist. Sci.* **30** 216–227. MR3353104

PINSKER, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Inf. Transm.* **16** 120–133. MR0624591

ROSENTHAL, H. P. (1970). On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.* **8** 273–303. MR0271721

SRIVASTAVA, N. and VERSHYNIN, R. (2013). Covariance estimation for distributions with $2 + \varepsilon$ moments. *Ann. Probab.* **41** 3081–3111. MR3127875

STEINBERGER, L. (2015). Statistical inference in high-dimensional linear regression based on simple working models. Ph.D. thesis, Univ. Vienna.

STEINBERGER, L. and LEEB, H. (2018). On conditional moments of high-dimensional random vectors given lower-dimensional projections. *Bernoulli* **24** 565–591. MR3706769

TAYLOR, J., LOCKHART, R., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). Exact post-selection inference for forward stepwise least angle regression. Arxiv preprint. Available at arXiv:1401.3889.

L. STEINBERGER
DEPARTMENT OF MATHEMATICAL STOCHASTICS
ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG
ERNST-ZERMELO-STRASSE 1
79104 FREIBURG IM BREISGAU
GERMANY
E-MAIL: lukas.steinberger@stochastik.uni-freiburg.de

H. LEEB
DEPARTMENT OF STATISTICS
AND OPERATIONS RESEARCH
UNIVERSITY OF VIENNA
OSKAR-MORGENSTERN-PLATZ 1
A-1090 VIENNA
AUSTRIA
E-MAIL: hannes.leeb@univie.ac.at