Econometric Theory, **39**, 2023, 1249–1272.
doi:10.1017/S026646662100044X

STATISTICAL INFERENCE WITH *F*-STATISTICS WHEN FITTING SIMPLE MODELS TO HIGH-DIMENSIONAL DATA

Hannes Leeb
and
Lukas Steinberger
University of Vienna

We study linear subset regression in the context of the high-dimensional overall model $y = \vartheta + \theta' z + \epsilon$ with univariate response y and a d-vector of random regressors z, independent of ϵ. Here, "high-dimensional" means that the number d of available explanatory variables is much larger than the number n of observations. We consider simple linear submodels where y is regressed on a set of p regressors given by $x = M'z$, for some $d \times p$ matrix M of full rank $p < n$. The corresponding simple model, that is, $y = \alpha + \beta' x + e$, is usually justified by imposing appropriate restrictions on the unknown parameter θ in the overall model; otherwise, this simple model can be grossly misspecified in the sense that relevant variables may have been omitted. In this paper, we establish asymptotic validity of the standard F-test on the surrogate parameter β, in an appropriate sense, even when the simple model is misspecified, that is, without any restrictions on θ whatsoever and without assuming Gaussian data.

1. INTRODUCTION

The F-test is a staple tool of applied statistical analyses. It is widely used, sometimes also in situations where its applicability is debatable because underlying assumptions may not be met. We study a situation of this kind: An F-test after fitting a (possibly misspecified) working model. We focus, in particular, on a scenario where the fitted model has p explanatory variables while the true model is non-Gaussian, has d explanatory variables, with $p \ll d$, and where sample size n is of the same order as p. Scenarios like this occur, for example, in quality control studies, such as Souders and Stenbakken (1991), where a model with 18 explanatory variables (out of a total of about 8,000) is fit based on a sample of size 50; in time series forecasting with principal components as in Stock and Watson (2002), who extract a handful of factors from 149 explanatory variables based on

The first author's research was partially supported by FWF projects P 26354-N26 and P 28233-N32. Address correspondence to Hannes Leeb, Department of Statistics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria; e-mail: hannes.leeb@univie.ac.at

480 monthly observations; or in genetic analyses, such as van't Veer et al. (2002), who select and fit a model with 70 genes (out of a total of about 25,000) based on a sample of size 78.

In situations like these, the question whether the fitted model has any explanatory value is of particular interest. We show that, approximately, the usual F-statistic is F-distributed under a corresponding null-hypothesis, and that it is noncentral F-distributed in a local neighborhood of the null. Approximation errors go to zero as $n \to \infty$ if $n^2/\log d \to 0$ and if, at the same time, p is of the same order as n; cf. Theorem 4.1 and Remark 4.3, respectively. The crucial point here is the following: Our results are uniform over a large region of the parameter space. In particular, our results are completely uniform in the true unknown high-dimensional regression coefficient vector and therefore also cover situations where the fitted model is grossly misspecified in the sense that important variables have been omitted. Notice that, in general, it cannot be expected that the classical F-test in such a misspecified working model is valid, because the omitted variable bias usually leads to both mean and variance misspecification of the working model. The setting of our analysis is nonstandard in that we require a particular constellation of d, p, and n. This is a challenging setting of practical relevance, for which few theoretical results are available so far. Our findings, which are given for independent observations, also prompt the question whether similar results can be obtained under serial correlation.

The F-statistic is exactly F-distributed in a correctly specified linear model with Gaussian errors; and it is asymptotically F-distributed under the strong Gauss–Markov condition on the errors if $n \to \infty$ while the model dimension stays fixed; cf. Anderson (1958). F-tests in correctly specified models in settings where p is allowed to increase with n are studied, among others, by Portnoy (1984, 1985), Boos and Brownie (1995), Akritas and Arnold (2000), Bathke and Lankowski (2005), Harrar and Bathke (2008), and Wang and Cui (2013). In addition, there are several viable alternatives to the F-test in potentially misspecified settings; see, for example, Eicker (1967), Huber (1967), White (1980a, 1980b), Chen and Qin (2010), and Zhong and Chen (2011). For further results on hypothesis testing and marginal screening in misspecified models, see, for example, Jensen and Ramirez (1991), Ramirez and Jensen (1991), Fomby and Hill (2003), Choi and Kiefer (2011), Boos and Stefanski (2013), and the references therein.

On a technical level, this paper relies on Wang and Cui (2013), the corresponding extensions and corrections in Steinberger (2016), and also on Steinberger and Leeb (2018, 2019); all but the first of these references are based on Steinberger (2015).

The rest of the paper is structured as follows: In Section 2, we describe the true data-generating model and the underlying parameter space. The (typically misspecified) working model and the corresponding F-statistic are described in Section 3. Our main theoretical result is given in Section 4, and the strategy for its proof, including some intuitive explanations, is outlined in Section 5. A simulation study in Section 6 demonstrates that our asymptotic approximations can "kick-in" reasonably fast.

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

2. THE TRUE MODEL

Throughout, we consider the (true) linear model

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

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 $\sigma^2 > 0$; 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 Steinberger and Leeb (2019, 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. With this, we can represent the d-vector z as

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

for a d-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 (nonrandom) matrix. Notice that R governs distributional properties of z beyond the second moments. Our results will hold for most matrices R, in an appropriate sense (cf. the set $\mathbb{U}$ in Theorem 4.1). 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. In particular, we will assume that $f_{\tilde{z}}$ belongs to one of the classes $\mathcal{F}_{d,k}(D,E)$ that are defined in the next paragraph, for appropriate constants k, D, and E. Our assumptions on z are similar to those maintained by Bai and Saranadasa (1996) and Zhong and Chen (2011). For later use, note that the distribution of (y,z) in (1) and (2) is characterized by ϑ and θ, by $\mathcal{L}(\epsilon)$, by Σ and μ, by $f_{\tilde{z}}$, and by R.

Fix an integer $k \geq 1$ and positive (finite) constants D and E. With this, write $\mathcal{F}_{d,k}(D,E)$ for the class of Lebesgue densities on $\mathbb{R}^d$ that are products of univariate marginal densities such that each marginal density is bounded from above by D, and such that each univariate marginal density has absolute moments of order up to k that are bounded by E.

We refer the reader to Steinberger and Leeb (2018) for further discussion and several relaxations of this assumption on $f_{\tilde{z}}$. In particular, we point out that also d-variate distributions on $\tilde{z}$ with some dependence among components can be allowed.

3. THE SUBMODEL AND THE *F*-TEST

Consider a submodel where y is regressed on x, with x given by

$$x \quad = \quad M' z \tag{3}$$

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

for some full-rank $d \times p$ matrix M with $p < d$. The observed data $(y_i, x_i)_{i=1,\dots,n}$ are i.i.d. copies of (y, x) following (1) and (3). Notice that we do not require high-dimensional covariates z_i (i.i.d. copies of z in (2)) to be observed, although this will be the case in many applications. Also notice that since we do not put any further assumptions on $M \in \mathbb{R}^{d\times p}$ other than being full rank, M need not be known to the user. However, often M is explicitly specified by the data analyst or obtained by some model selection procedure in a data-driven way. For example, M can be a selection matrix that picks out p components of the d-vector z. The following matrix $M_{d:3}$ would, for instance, correspond to choosing the first three variables from the d available ones,

$$M_{d:3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{d\times 3}.$$

Our results can also be used when M is obtained from a data-driven model selection procedure. This is obvious when the model selection was carried out on a holdout set that is independent of the data used to conduct the F-test, since conditioning on the holdout set would take us back to the situation of a nonrandom M. Otherwise, in principle, our ideas can still be used by applying them uniformly over all models in the candidate set. However, uniformity over the candidate set requires restrictions on the size of that set. Extending our results to model selection procedures which choose from practically relevant sets of candidate models is an important but open problem (see Steinberger and Leeb, 2019, Section 6, for first ideas in that direction). Finally, submodels with regressors of the form $x = M'z$ as in (3) 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, $p \ll d$. Trivially, we can write

$$y = \alpha + \beta' x + e \tag{4}$$

with $e = y - \alpha - \beta' x$, where α and β minimize $\mathbb{E}[(y - \alpha - \beta' x)^2]$. The “error” e has mean zero (because both (1) and (4) include an intercept), and we denote its variance by $s^2 = \mathbb{E}[e^2]$. Note that $\alpha = \vartheta + \mu'\theta - \mu' M(M'\Sigma M)^{-1}M'\Sigma\theta$ and, for later use, that

$$\begin{aligned} \beta &= (M'\Sigma M)^{-1}M'\Sigma\theta \quad \text{and} \\ s^2 &= \theta'\Sigma\theta - \theta'\Sigma M(M'\Sigma M)^{-1}M'\Sigma\theta + \sigma^2. \end{aligned} \tag{5}$$

Irrespective of whether the working model is correctly specified, the “surrogate” parameters α, β, and s^2 are always well-defined. Here, β is our main object of interest, instead of the underlying true parameter θ. Such surrogate parameters

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

are well known in the statistics literature, certainly since Huber (1967), and have recently gained new popularity, as witnessed by, for example, Abadie, Imbens, and Zheng (2014), Brannath and Scharpenberg (2014), Buja et al. (2014), and Bachoc, Leeb, and Pötscher (2019). In particular, such surrogate parameters can be consistently estimated, in a standard M-estimation setting, by the ordinary least squares (OLS) estimator or by robust alternatives, provided that p is not too large relative to n (see White, 1980a, 1980b; Portnoy, 1984, 1985); cf. also Lemma A.3 in Steinberger (2015) and Lemma A.4 in Steinberger and Leeb (2019) for analyses tailored to our present setting.

The working model (4) is correct (in the usual sense) if $\mathbb{E}[y\|z] = \mathbb{E}[y\|x]$, that is, if $\vartheta + \theta' z = \alpha + \beta' x$ or, equivalently, if $\epsilon = e$. This is the case if θ lies in the column space of M; if M is a selection matrix, this means that $M'\theta$ selects all the nonzero components of θ. Here, we do not assume that the working model is correct. In particular, we stress that e may differ from ϵ, and that e may depend on x.

When working with the simple submodel (4), a natural question is whether x has any explanatory value for the response variable y. Given a sample of $n > p+1$ i.i.d. observations of y and x from (4), a classical approach to this question is to use the F-test of the hypotheses

$$H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta \neq 0. \tag{6}$$

Let Y and X denote the $n \times 1$ vector of responses and the $n \times p$ matrix of explanatory variables, respectively. Write $\hat{\beta}$ for the OLS-estimator for β when Y is regressed on X and a constant, set $\hat{s}^2 = \|(I_n - P_{\iota,X})Y\|^2/(n-p-1)$, and write $\hat{F}_n = \hat{F}_n(X,Y)$ for the usual F-statistics for testing H_0, that is,

$$\hat{F}_n = \frac{\|(I_n - P_\iota)X\hat{\beta}\|^2/p}{\hat{s}^2},$$

if the numerator is well-defined and the denominator is positive and $\hat{F}_n = 0$ otherwise. Here, P_A denotes the orthogonal projection on the space spanned by the columns of A and ι denotes the n-vector $\iota = (1,\ldots,1)'$. Note that $\hat{F}_n > 0$ with probability one by our assumptions.

H_0 may be rephrased as the hypothesis that the best linear predictor of y given x is constant. An alternative to H_0 is the hypothesis that the Bayes-estimator of y given x is constant, that is,

$$\tilde{H}_0 : \mathbb{E}[y\|x] \text{ is constant.}$$

Testing this nonparametric hypothesis is more difficult. In the asymptotic setting that we consider in the next section, however, we find that H_0 and $\tilde{H}_0$ are close to each other in the sense that the Bayes predictor and the best linear predictor (of y given x) are close in terms of mean-squared prediction error; see Remark 4.2 for details.

4. MAIN RESULT

Our main result is concerned with the asymptotic distribution of the F-statistic in a local neighborhood of the null-hypothesis. Here, the local neighborhood is defined through the requirement that

$$\Delta \quad = \quad \mathrm{Var}(\beta' x)/\mathrm{Var}(e) \quad = \quad \beta' M' \Sigma M \beta / s^2$$

is small. This quantity can be interpreted as a signal-to-noise ratio in (4) and depends on θ, M, Σ, and $\sigma^2 = \mathbb{E}[\epsilon^2]$; cf. (5).

If the error e in (4) is Gaussian and independent of x, then, conditional on X, the F-statistic $\hat{F}_n = \hat{F}_n(X,Y)$ is F-distributed with parameters p, $n-p-1$ and noncentrality parameter $n\delta$, where $\delta = \frac{1}{n}\beta' X'(I_n - P_\iota)X\beta/s^2 : \mathbb{P}(\hat{F}_n \leq t \| X) = F_{p,n-p-1,n\delta}(t)$, where $F_{p,n-p-1,n\delta}(t)$ denotes the cumulative distribution function (c.d.f.) of the F-distribution with indicated parameters. In our present setting, however, the error e in (4) need not be Gaussian and can (and typically will) depend on x.

We will show that the unconditional c.d.f of $\hat{F}_n$ can be approximated by $F_{p,n-p-1,n\Delta}(t)$, provided that Δ is small. Note that the noncentrality parameter $n\delta$ considered in the preceding paragraph is related to $n\Delta$ in the sense that $s^2\delta = \frac{1}{n}\beta' X'(I_n - P_\iota)X\beta$ is the empirical variance of the vector $X\beta$, while $s^2\Delta = \mathrm{Var}\,(x'\beta)$. Our approximations are uniform over most parameters in the model. Only for $\epsilon, f_{\tilde{z}}$ and R, that is, for the error in (1) and for the density of the standardized explanatory variables as well as the orthogonal matrix in (2), some restrictions are needed. We will require a moment restriction on ϵ/σ, and we will require that $f_{\tilde{z}}$ belongs to one of the classes $\mathcal{F}_{d,k}(D,E)$ introduced earlier. To formulate the restriction on R, write $\mathcal{O}_d$ for the collection of all orthogonal $d \times d$ matrices and write ν_d for the uniform distribution on that set; that is, ν_d is the normalized Haar measure on the d-dimensional orthogonal group. For R, we will require that it belongs to a Borel set $\mathbb{U} \subseteq \mathcal{O}_d$ that is large in terms of ν_d.

The following theorem is formulated in terms of suprema of certain functions that, by definition, depend on n. Furthermore, most of the sets over which these suprema are computed also depend on n through $p = p_n$ or $d = d_n$. Thus, the suprema themselves necessarily depend on n and are shown to converge to zero as $n \to \infty$. However, some of the sets over which the maximization is carried out also depend on constants that are assumed to be fixed when n increases. These are D, E, ρ, λ, L, and γ. For the proof, we operate in a triangular array setting where everything, except the mentioned quantities, is allowed to depend on n.

THEOREM 4.1. *Fix finite constants $D \geq 1$ and $E \geq 1$, and positive finite constants $\rho \in (0,1)$, λ, L and γ. For each full-rank $d \times p$ matrix M, each $d \times d$ variance/covariance matrix $\Sigma > 0$, and each $f_{\tilde{z}} \in \mathcal{F}_{d,20}(D,E)$ there exists a Borel set $\mathbb{U} = \mathbb{U}(M, \Sigma, f_{\tilde{z}}) \subseteq \mathcal{O}_d$ such that*

$$\sup_{M} \ \sup_{\Sigma} \ \sup_{f_{\tilde{z}} \in \mathcal{F}_{d,20}(D,E)} \nu_d(\mathbb{U}^c) \quad \xrightarrow{\frac{p}{\log d} \to 0} \quad 0$$

and such that the following holds: If Ξ_n *denotes either the quantity*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\hat{F}_n \le t\right) - F_{p,n-p-1,n\Delta}(t)\right| \tag{7}$$

or the quantity

$$\mathbb{P}\left(\hat{F}_n > F^{-1}_{p,n-p-1,0}(\eta)\right) - \Phi\left(-\Phi^{-1}(\eta) + \sqrt{n}\Delta\sqrt{\frac{1-p/n}{2p/n}}\right) \tag{8}$$

for some fixed $\eta \in [0,1]$, *then*

$$\sup_{M} \sup_{\substack{\vartheta,\theta,\mathcal{L}(\epsilon),\mu,\sigma^2,\Sigma \\ \mathbb{E}|\epsilon/\sigma|^{8+\lambda} \le L \\ \Delta < \gamma/\sqrt{n}}} \sup_{f_{\tilde{z}}\in\mathcal{F}_{d,20}(D,E)} \sup_{R\in\mathbb{U}} \Xi_n \quad \underset{\frac{n^2}{\log d}\to 0,\, \frac{p}{n}\to\rho}{\xrightarrow{n\to\infty}} \quad 0.$$

This statement continues to hold if the restriction $\Delta < \gamma/\sqrt{n}$ *in the last display is replaced by* $\Delta < g(n)$ *provided that* $\lim_{n\to\infty} g(n) = 0$. *(Here, the suprema are taken over all full-rank* $d \times p$ *matrices* M*, all* $\vartheta \in \mathbb{R}$*, all* d*-vectors* θ *and* μ*, all distributions* $\mathcal{L}(\epsilon)$ *so that* ϵ *has mean zero and finite positive variance, all positive finite* σ^2 *and all symmetric and positive definite* $d \times d$ *matrices* Σ*, subject to the indicated restrictions.)*

Remark 4.2. Write $\mathcal{R}_N$ and $\mathcal{R}_L$ for the prediction risk of the Bayes predictor and of the best linear predictor, respectively, of y given x. That is, $\mathcal{R}_N = \mathbb{E}[(y - \mathbb{E}[y\|x])^2]$ and $\mathcal{R}_L = \mathbb{E}[(y - (\alpha + \beta' x))^2]$. The results of Steinberger and Leeb (2019) then entail that, in the setting of Theorem 4.1, $\mathcal{R}_N/\mathcal{R}_L$ converges to one, uniformly over all the parameters indicated in the last display of that theorem. In fact, the risk-ratio converges to one uniformly even if the restriction on Δ is removed altogether, and a similar statement holds for the ratio of conditional risks given x, that is, $\mathbb{E}[(y - \mathbb{E}[y\|x])^2\|x]/\mathbb{E}[(y - (\alpha + \beta' x))^2\|x]$. See Theorem 3.1 of Steinberger and Leeb (2019) for a more general form of this statement under weaker assumptions.

Remark 4.3. In Theorem 4.1, we require that the limit ρ of p/n is strictly positive and less than one. It seems natural, though, that the result would still hold if we only require $\rho \in [0,1)$. However, the assumption that $\rho > 0$ is needed for both a conceptual as well as a purely technical reason.

- From a technical point of view, we rely on the results of Steinberger (2016) and need to verify Assumption (A2) in that reference. In particular, we need to show that (see the proof of Theorem 4.1 in Appendix A)

$$\frac{1}{\sqrt{p}} \max_{i=1,\ldots,n} \mathbb{E}[(e_i/s)^4\|x_i] \xrightarrow[n\to\infty]{i.p.} 0,$$

where $e_i = y_i - \alpha - \beta' x_i$ is the error term corresponding to the ith observation pair in the working model (4). Under the $(8+\lambda)$ moment bound on $|\epsilon/\sigma|$ and the bounded

20th moments of $\tilde{z}$ imposed in Theorem 4.1, the maximum in the previous display is easily seen to be $O_{\mathbb{P}}(n^{\frac{1}{2+\lambda/4}})$. Thus, in order to get the required convergence for arbitrary $\lambda > 0$, we need p to be of the same order as n. We conjecture that the result still holds for $\rho = 0$ (and $p \to \infty$, cf. the next bullet point), but this would require an entirely different proof strategy while providing only limited additional statistical insight.

- From a conceptual point of view, we need to impose at least that $p \to \infty$ as $n \to \infty$, or otherwise the conclusions of Theorem 4.1, with Ξ_n as in (7) and as in (8), are at odds with one another: Simply note that the F-distribution with p and $n-p-1$ degrees of freedom converges to $1/p$ times a χ^2_p-distribution with p degrees of freedom if $n \to \infty$ and p is fixed. But this is well approximated by a normal distribution as in (8) only if p is large. However, in the case where Ξ_n is equal to the expression in (7), the theorem also remains true in the classical situation where p is fixed and only n diverges to ∞, although we do not present the details here. It can be established using Lemma A.2 (which shows that the effect of misspecification vanishes if $np/\log d$ is small; see also Section 5) and asymptotic normality of the OLS estimator (e.g., White, 1980a, Lemma 2) in the correctly specified case, which leads to an asymptotic χ^2_p-distribution of the F-statistic. In that sense, the conclusion with Ξ_n equal to (7) is more generally true than that for (8). We include this part of the theorem nonetheless, because the Gaussian approximation in (8) has the advantage that it is easier to interpret than the more complicated distribution function of the noncentral F-distribution in (7); see also the discussion in Steinberger (2016), Remark 2.4. Also note that the statement regarding (8) in Theorem 4.1 coincides with the conclusion of Theorem 1 in Zhong and Chen (2011) for the correctly specified Gaussian error case.

Remark 4.4 (On heteroskedasticity). The critical reader will observe that if the working model omits relevant variables, the error term e in (4) will depend on x. In particular, the conditional variance $\mathrm{Var}[y_i \| x_i]$ may depend on the working regressors x_i and the submodel may be heteroskedastic. Thus, the classical F-test as studied in Theorem 4.1 does not appear to be the appropriate method of choice here. The paradox is resolved by noticing that our results imply $\mathrm{Var}[y_i \| x_i] = \mathrm{Var}[e_i \| x_i] \approx \mathrm{Var}[e_i] = s^2$ if $np/\log d$ is sufficiently small (cf. Lemma A.1 for the precise statement). If, on the other hand, the true innovations ϵ_i in model (1) are already heteroskedastic, then this will typically propagate also to the submodel and will invalidate the F-test. The dimension effect mentioned above will not alleviate this problem.

In general, and irrespective of the source of heteroskedasticity in the working model (4), instead of conducting a standard F-test, one may follow the classical Eicker–Huber–White (Eicker, 1967; Huber, 1967; White, 1980a, 1980b) approach to heteroskedasticity robust testing (see also Preinerstorfer and Pötscher, 2016). In the context of model (4), since $\mathbb{E}[e_i x_i] = 0$ (and under additional boundedness assumptions, cf. White, 1980a, Assumptions 1–4), the heteroskedasticity corrected test for the regression parameter $\gamma = (\alpha, \beta')' \in \mathbb{R}^{p+1}$ suggested in Eicker (1967, Theorem 3.2) and White (1980a, Theorem 1) is asymptotically valid (as $n \to \infty$ and d, p fixed) in the sense that its asymptotic rejection probability is equal to the nominal significance level. Of course, this adjusted test will have less power than the standard F-test if the model (4) is linear and homoskedastic, which is why in practice one might be inclined to simply use the standard F-test if there is no strong indication of a violation of linearity and homoskedasticity. Our results justify the use of the standard F-test under omitted variables when $d \gg np$. We also point out that the classical Eicker–Huber–White heteroskedasticity robust test is no longer statistically valid in the asymptotic framework $p/n \to \rho \in (0,1)$ we consider here (cf. Cattaneo, Jansson, and Newey, 2018; Dobriban and Su, 2018; Li and Yao, 2019; Jochmans, 2020), but valid alternatives for the high-dimensional regime have been suggested in these references.

5. PROOF STRATEGY AND INTUITION

The proof of Theorem 4.1 relies mainly on the results of Steinberger (2016) and Steinberger and Leeb (2018). Steinberger (2016) (building on Wang and Cui, 2013) derives the asymptotic distribution of the F-statistic under local alternatives in a correctly specified but non-Gaussian p-parameter linear model and in the $p/n \to \rho \in (0,1)$ framework. The main result in that reference is identical to that of Theorem 4.1 in the case where Ξ_n is given by the expression in (8), but with $\mathbb{U} = \mathcal{O}_d$ and with the important additional assumption that the working model M is linear and homoskedastic, that is

$$\mathbb{E}[e\|x] = 0 \quad \text{and} \quad \mathrm{Var}[e\|x] = \mathrm{Var}[e] = s^2. \tag{9}$$

The core of the rather technical proof in Steinberger (2016) relies on a martingale central limit theorem applied to a quadratic form in the independent error terms (see Lemma 6.1 in that reference). Notice that due to the potentially omitted variables, (9) is not necessarily true in our present setting. Thus, one main innovation of the present work is to extend these existing results on the asymptotic distribution of the F-statistic to potentially misspecified submodels (i.e., models where (9) holds only approximately).

This is where the results of Steinberger and Leeb (2018) come into play. They consider the omitted variables setting that is also adopted here, with the dimension d of the true linear model far exceeding the dimension p of the working model. They essentially show that for the set $\mathbb{U}$ in Theorem 4.1 and for $R \in \mathbb{U}$, we have

$$\mathbb{E}[e\|x] \approx 0 \quad \text{and} \quad \mathrm{Var}[e\|x] \approx \mathrm{Var}[e] = s^2, \tag{10}$$

in an appropriate sense and uniformly in $\theta \in \mathbb{R}^d$, provided that $p/\log d$ is sufficiently small (see the beginning of Section A for details). To shed some light on this result, consider the noiseless case $\sigma^2 = 0 = \vartheta$, $\mu = 0$, $\|\theta\| = 1$ and $p = 1$. Thus, $y = \theta' z$ and $x = m' z$, for some unit vector $m \in \mathbb{R}^d$. Now, for most θ and m on the unit sphere in $\mathbb{R}^d$ (measured by the uniform distribution), $y = \theta' z = \sum_{j=1}^d \theta_j z_j$ and $x = m' z = \sum_{j=1}^d m_j z_j$ will be approximately jointly normally distributed if

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

d is large (provided that the distribution of z is sufficiently regular). If they were exactly jointly Gaussian, then we would, of course, have $\mathbb{E}[y\|x] = \beta x$ and $\mathrm{Var}[y\|x] = \mathrm{Var}[y]$. However, making these approximations uniform in $\theta \in \mathbb{R}^d$ is far from trivial (cf. Hall and Li, 1993; Leeb, 2013; Steinberger and Leeb, 2018).

The main task in the proof of Theorem 4.1 is therefore to make the link between (10) and (9) rigorous. To that end, we consider an "artificial" sample Y^* consisting of elements

$$y_i^* := \alpha + \beta' x_i + e_i^*$$

with

$$e_i^* := s\frac{e_i - \mathbb{E}[e_i\|x_i]}{\sqrt{\mathrm{Var}[e_i\|x_i]}}$$

(cf. Lemma A.1 in Appendix A). Clearly, (9) is satisfied for e_i^* instead of e. In Lemma A.2, using the results of Steinberger and Leeb (2018), we then prove an equivalence of experiments type result, showing that the F-statistic $\hat{F}_n(X,Y)$ computed from the original data X and Y, and that computed from X and Y^*, are close to each other in probability. More precisely, we show that

$$n^k\left(\hat{F}_n(X,Y) - \hat{F}_n(X,Y^*)\right) \quad \xrightarrow{i.p.} \quad 0,$$

for every $k \in \mathbb{N}$, provided that $np/\log d \to 0$. The additional n in $np/\log d \to 0$ comes from the fact that we need to apply the approximation in (10) to get $e_i^* \approx e_i$, for each $i = 1, \ldots, n$.

Finally, we want to point out that an alternative to relying on approximations as in (10) could be to directly analyse the asymptotic distribution of the F-statistic in the misspecified case. Building on basic observations from the fixed (nonrandom) design Gaussian linear model under omitted variable bias, one would expect a doubly noncentral F-distribution as the limiting distribution with noncentrality parameters somehow depending on the approximation errors in (10). Intuitively, these noncentrality terms should reduce to those of the singly noncentral F-distribution known from a correct Gaussian linear model, provided that the approximations in (10) are accurate. However, in the random design case, one would also have to deal with the heteroskedastic variances $\mathrm{Var}[y_i\|x_i]$, $i = 1, \ldots, n$, and it is far from obvious how exactly these will affect the limiting noncentrality parameters, or if the limit is even F-distributed at all. This approach may be technically quite involved and is not further explored here.

6. SIMULATION ANALYSIS

Theorem 4.1 is an asymptotic result. In this section, we study a range of non-asymptotic scenarios through simulation to investigate how soon these asymptotic approximations become accurate. We consider a rather small sample size of $n = 50$ and look at different configurations of the model dimensions d and p with $p < d$, and also at different points in parameter space.

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

The theorem contains two asymptotic statements, one about the distribution of the F-statistic and one about the size of the set $\mathbb{U}$. For the distribution of the F-statistic, we compare the rejection probability of the F-test under the null-hypothesis with the nominal significance level $\alpha = 0.05$. The nominal significance level provides a natural benchmark. (Clearly, one can also investigate the power of the F-test through simulation experiments, but, unlike the significance level, it is less obvious what the right benchmark for the power should be.) In particular, we simulate 1,000 independent realizations $F_{j,r}, j = 1, \ldots, 1{,}000$ of the F-statistic at sample size $n = 50$ under the null for each point in parameter space (the index r will be explained shortly), and compare the empirical significance level $\overline{p}_r = 1{,}000^{-1} \sum_{j=1}^{1{,}000} \mathbf{1}\{F_{j,r} > F^{-1}_{p,n-p-1,0}(1-\alpha)\}$ with the nominal level α.

Gauging the size of $\mathbb{U}$ is more difficult, because that set is not given explicitly. We proceed as follows: We fix all the parameters in (1) and (2) except for the orthogonal matrix R in (2). We then simulate 100 independent realizations R_r of R, compute $\overline{p}_r$ as outlined above, $r = 1, \ldots, 100$, and finally compute $\overline{D} = 100^{-1} \sum_{r=1}^{100} |\overline{p}_r - \alpha|$. If $R_r \in \mathbb{U}$, then $\overline{p}_r$ should be close to α, in view of the last display in Theorem 4.1. We use $\overline{D}$ and the empirical distribution of the $\overline{p}_r$, $r = 1, \ldots, 100$, as indicators for the size of $\mathbb{U}$.

The remaining parameters in (1) and (2), and the submodel matrix M are chosen as follows for any fixed values of d and p: The intercept terms ϑ and μ are set to zero, for convenience. We do not include an error term in the true model, that is, we set $\sigma^2 = 0$, because the effect of misspecification becomes more pronounced when the error variance σ^2 is small.[1] (Note that the case where $\sigma^2 = 0$ is not covered by Theorem 4.1 per se, but inspection of the proof shows that our results also apply in this case; cf. Remark A.3.) For $\tilde{z}$, we consider product distributions with zero mean and i.i.d. components from the student-t distribution with 2, 3, and 5 degrees of freedom, as well as from the centered exponential, uniform, Bernoulli$\{-1, 1\}$ and Gaussian distributions. (Note that the scaling of these distributions is inconsequential, because of the scale-invariance of the F-statistic $\hat{F}(X, Y)$ in both arguments and the fact that we do not include an error term in the full model, that is, scaling of $\tilde{z}_i$ is equivalent to scaling of both $y_i = \theta' z_i$ and $x_i = M' z_i$. Similarly, also the scaling of θ and Σ has no impact on the value of the F-statistic.) For Σ, we chose a spiked covariance matrix $\Sigma = U \mathrm{diag}(\lambda_1, \ldots, \lambda_d) U'$ with eigenvalues $\lambda_1 = \lambda_2 = 400$ and $\lambda_3 = \cdots = \lambda_d = 1$ and an orthogonal matrix of eigenvectors U chosen randomly from the uniform distribution on the orthogonal group.[2] For the matrix M, which describes the working model, we take M equal

[1] Note that if the error variance $\sigma^2 = \mathrm{Var}[\epsilon_i]$ in the true model $y_i = \theta' z_i + \epsilon_i$ is overly large, that is, much larger than $\theta' \Sigma \theta$, then the scaled true model is essentially given by $y_i/\sigma \approx \epsilon_i/\sigma$. Since the F-statistic is scale-invariant and ϵ is independent of X, we then have $\hat{F}(X, Y) = \hat{F}(X, Y/\sigma) \approx \hat{F}(X, (\epsilon_i)_{i=1}^n/\sigma) = \hat{F}(X, (\epsilon_i)_{i=1}^n)$. In that case, the F-statistic will essentially follow the null-distribution and we expect a rejection probability close to the nominal level, irrespective of θ and R.

[2] The spiked covariance model corresponds to a factor model where the identity matrix is perturbed by a low rank matrix. It has received much attention in the literature on high dimensional random matrices (e.g., Johnstone, 2001; Baik and Silverstein, 2006; Cai, Ma, and Wu, 2013; Donoho, Gavish, and Johnstone, 2018). We have repeated the simulations also with covariance matrices of an AR(1) process and obtained essentially the same results.

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

Table 1. Average absolute differences $\bar{D} = \frac{1}{100}\sum_{r=1}^{100} |\bar{p}_r - \alpha|$ of simulated rejection probabilities $\bar{p}_r = \frac{1}{1000}\sum_{j=1}^{1000} \mathbf{1}\{F_{j,r} > F^{-1}_{p,n-p-1,0}(1-\alpha)\}$ and nominal significance level $\alpha = 0.05$ of the F-test for $H_0 : \beta = 0$.

$d \backslash p$	1	2	5	25	1	2	5	25
		$t(5)$				Exp(1)		
2	0.077				0.141			
4	0.056	0.076			0.093	0.140		
10	0.032	0.047	0.066		0.052	0.071	0.109	
50	0.009	0.013	0.017	0.019	0.014	0.015	0.020	0.033
100	0.007	0.008	0.009	0.010	0.009	0.009	0.012	0.015
200	0.006	0.007	0.006	0.008	0.007	0.007	0.006	0.009
		$t(3)$				Unif$[-1,1]$		
2	0.188				0.025			
4	0.158	0.225			0.020	0.023		
10	0.122	0.167	0.238		0.011	0.014	0.016	
50	0.062	0.084	0.116	0.123	0.006	0.006	0.007	0.007
100	0.048	0.061	0.081	0.082	0.005	0.006	0.006	0.005
200	0.033	0.044	0.057	0.055	0.005	0.005	0.005	0.006
		$t(2)$				Gauss		
2	0.335				0.005			
4	0.332	0.458			0.006	0.005		
10	0.301	0.411	0.563		0.005	0.005	0.006	
50	0.250	0.335	0.456	0.518	0.005	0.006	0.005	0.005
100	0.228	0.314	0.412	0.457	0.005	0.005	0.006	0.005
200	0.212	0.286	0.383	0.407	0.005	0.005	0.006	0.006

to the $d \times p$ matrix whose kth column is the kth standard basis vector in $\mathbb{R}^d$, $1 \leq k \leq p$. In other words, we consider a submodel that includes only the first p regressors (out of d). For the parameter $\theta \in \mathbb{R}^d$, we need to ensure that the null-hypothesis is satisfied, that is, that $\beta = (M'\Sigma M)^{-1}M'\Sigma\theta = 0$. By construction of Σ, $M'\Sigma M$ is regular, and we choose $\theta = (I_d - P_{\Sigma M})V/\|(I_d - P_{\Sigma M})V\|$, for one realization of $V \sim N(0, I_d)$, to guarantee that $M'\Sigma\theta = 0$. Notice that, in particular, this construction leads to a true model that contains variables which are not included in the working model (first p regressors).

The results of the simulations are summarized in Table 1 and Figures 1 and 2. From Table 1, the overall picture we get is consistent with what was predicted by our theory. For all distributions except the Gaussian, the average absolute difference between the true (simulated) rejection probabilities and the nominal level decreases as d increases. This phenomenon is most pronounced for the exponential distribution, which has a finite moment generating function around the origin, and is weakest for the $t(2)$-distribution, which does not even have

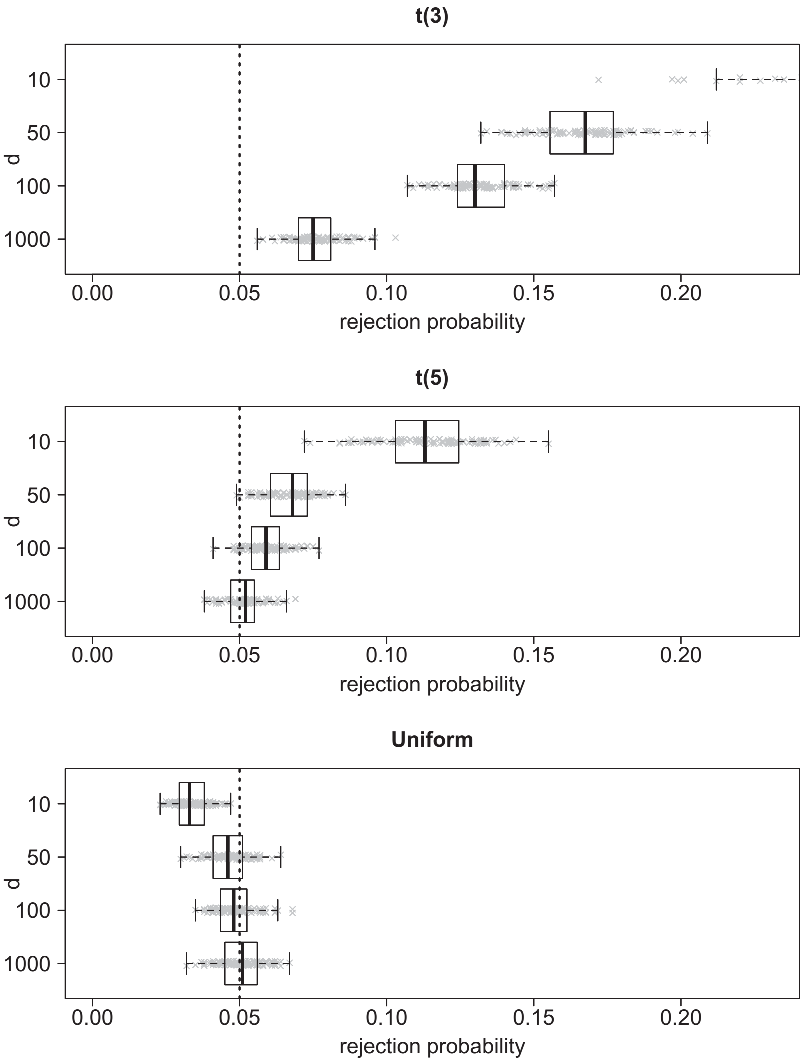

Figure 1. Box-plots of simulated rejection probabilities $(\bar{p}_r)_{r=1}^{100}$ (gray crosses) of the F-test with $n = 50$, $p = 5$ and $d = 10, 50, 100, 1{,}000$, for different design distributions. Every $r \in \{1, \ldots, 100\}$ corresponds to a different R_r applied to the standardized design $\tilde{z}$.

finite variance. For uniformly distributed design, which is bounded, the effect of misspecification on the size of the F-test is relatively mild already for small dimensions. In the Gaussian case, all sub-models of the form (4) are correct in the sense that the error e is Gaussian with mean zero and independent of x, so that theoretically the corresponding panel in Table 1 should contain only zeros. The numbers therefore represent only the simulation error and serve as a benchmark

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

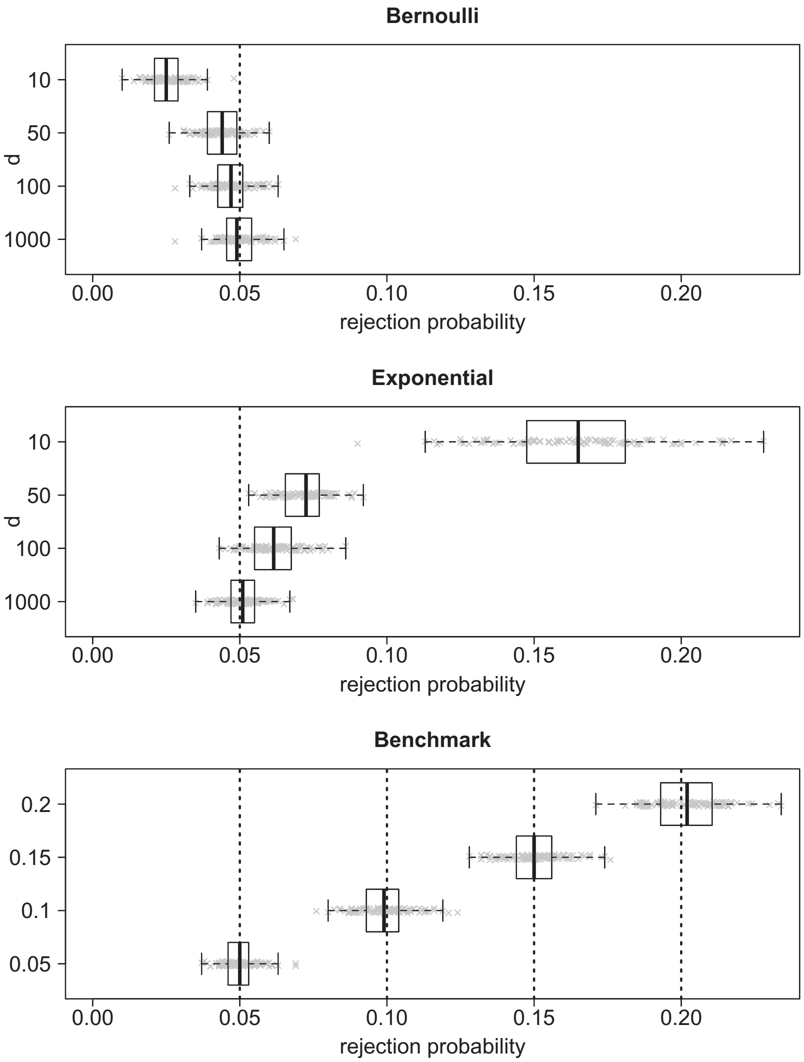

FIGURE 2. Box-plots of simulated rejection probabilities $(\bar{p}_r)_{r=1}^{100}$ (gray crosses) of the F-test with $n = 50, p = 5$ and $d = 10, 50, 100, 1,000$, for Bernoulli$\{-1,1\}$ and exponential design distributions and a benchmark panel of Binomial samples with different success probabilities.

for the other panels. We also see a monotonic increase, in the deviation of the size of the F-test from the nominal level, as the dimension p of the submodel increases, which was also suggested by our theory. However, if we fix the ratio $p/d = 1/2$, that is, if we move along the staircase pattern in each of the panels, except for the heavy tailed distributions $t(3)$ and $t(2)$, we still see the effect of misspecification

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

decrease as d increases. This suggests that convergence of $n^2/\log(d) \sim p^2/\log(d)$ to zero, as required in Theorem 4.1, may not be necessary, at least in the scenarios considered here.

In Table 1, the effect of the orthogonal matrix R on the actual significance level of the F-test was compressed into one summary statistic, namely the mean absolute deviation from the nominal significance level. To get a more comprehensive picture, Figures 1 and 2 show plots of the sample $(\bar{p}_r)_{r=1}^{100}$ (gray crosses) and superimposed box-plots for different design distributions. Due to limited space we present only the results for submodels of dimension $p = 5$. In view of Theorem 4.1, we expect that the size of $\mathbb{U}$, that is, the family of matrices R for which (7) and (8) get small, grows with d. Consequently, we expect that many of the $\bar{p}_r$ should be close to $\alpha = 0.05$. On the other hand, if d is not large then many matrices R will lead to a biased rejection probability due to misspecification of the working model. In other words, if d is not large, then for many matrices R the approximations in (10) will be poor, which invalidates the use of the conventional F-test. This is exactly what we observe in Figures 1 and 2. For small values of d, the rejection probabilities $\bar{p}_r$ are systematically biased and we see an increased variability of their values due to the variation in the choice of R_r (compare benchmark panel in Figure 2). Both the bias and the variability in $\bar{p}_r$ reduce when d increases, which is what we expected, as for large d, most R_r will be favorable and we obtain small misspecification errors uniformly over these favorable R_r. What is remarkable is the systematic over-rejection in case of the t- and exponential distribution and the under-rejection for Bernoulli and uniformly distributed designs. We currently cannot explain the mechanism that is responsible for this pattern. Finally, the benchmark panel shows i.i.d. samples $(\tilde{p}_r)_{r=1}^{100}$ with $\tilde{p}_r \sim \text{Binomial}(1,000,\alpha)/1,000$ and success probabilities $\alpha = 0.05, 0.1, 0.15, 0.2$. This provides some idea what portion of the variability observed in the other panels is due to random simulation error. Clearly, the results in the benchmark panel could have been equivalently obtained by repeating the previous simulation for the F-test with Gaussian design at significance levels $\alpha = 0.05, 0.1, 0.15, 0.2$.

Remark 6.1. Finally, we want to point out once more that, as already described in Section 5, there are two possible sources of error that could lead to the observed deviations of the true rejection probabilities from the nominal ones for small values of d. First, the data are non-Gaussian and therefore the F-statistic will not be exactly F-distributed in finite samples (recall that here $n = 50$) even if the working model is correctly specified. Secondly, the approximations in (10) may be poor due to the omitted variable bias, which will also lead to a bias in the distribution of the F-statistic. In light of simulation results in Steinberger (2016, Figure 3), however, the small sample non-Gaussianity effect appears to be rather negligible already for $n = 50$ even for heavy tailed and asymmetric error and design distributions. Thus, most of the deviations from the nominal rejection probabilities we observe seem to be attributable to violations of (10). Our theory predicts that theses violations diminish as d increases relative to p and n.

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

APPENDIX

A. Proofs

We begin with some preliminary considerations that connect this paper with the results of Steinberger and Leeb (2018). In particular, we use Theorem 2.1, parts (ii) and (iii), in that reference with $Z = \tilde{z}$ and $\tau = 1/2$: If $f_{\tilde{z}} \in \mathcal{F}_{d,20}(D,E)$, then the assumptions of that result are satisfied in view of Example 3.1 in Steinberger and Leeb (2018). The theorem guarantees existence of a Borel subset $\mathbb{G} = \mathbb{G}(f_{\tilde{z}}) \subseteq \mathcal{V}_{d,p}$ 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 $t > 0$ both

$$\sup_{B \in \mathbb{G}} \mathbb{P}\left(\left\| \mathbb{E}[\tilde{z} \| B'\tilde{z}] - BB'\tilde{z} \right\| > t\right)$$

and

$$\sup_{B \in \mathbb{G}} \mathbb{P}\left(\left\| \mathbb{E}[\tilde{z}\tilde{z}' \| B'\tilde{z}] - (I_d - BB' + BB'\tilde{z}\tilde{z}'BB') \right\| > t\right)$$

are bounded from above by

$$\frac{1}{t} d^{-1/20} + 4\gamma \frac{p}{\log d}, \tag{11}$$

such that

$$\nu_{d,p}(\mathbb{G}^c) \ \le \ \kappa d^{-(1-20\gamma\frac{p}{\log d})/20}, \tag{12}$$

where $\nu_{d,p}$ denotes the uniform distribution on the Stiefel manifold, and such that the set $\mathbb{G}$ is right-invariant under the action of $\mathcal{O}_p$, that is, $\mathbb{G}R = \mathbb{G}$ whenever $R \in \mathcal{O}_d$. Here, the constant $\gamma = \gamma(D)$ depends only on D, and the constant $\kappa = \kappa(E)$ 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

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

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 $\mathbb{G}$,

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

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. It follows that $\nu_d(\mathbb{U}^c)$ is bounded by the expression on the right-hand side of (12) whenever $f_{\tilde{z}} \in \mathcal{F}_{d,20}(D,E)$, which establishes the first claim of Theorem 4.1. The proof of the second claim is more elaborate.

The results in the preceding paragraph also show that the error e in the working model (4) is such that $\mathbb{E}[e\|x]$ is approximately zero and $\mathrm{Var}[e\|x]$ is approximately constant, provided that $R \in \mathbb{U}$: we first re-write the error e in a convenient form. Set $\tilde{\theta} = R'\Sigma^{1/2}\theta$ and $\tilde{M} = R'\Sigma^{1/2}M$. Then it is easy to see that $e = \tilde{\theta}'(I_d - P_{\tilde{M}})\tilde{z} + \epsilon$ and hence

$$\begin{aligned} \mathbb{E}[e\|x] &= \tilde{\theta}'(I_d - P_{\tilde{M}})\left\{\mathbb{E}[\tilde{z}\|P_{\tilde{M}}\tilde{z}] - P_{\tilde{M}}\tilde{z}\right\} \quad \text{and} \\ \mathbb{E}[e^2\|x] - s^2 &= \tilde{\theta}'(I_d - P_{\tilde{M}})\left\{\mathbb{E}[\tilde{z}\tilde{z}'\|P_{\tilde{M}}\tilde{z}] - ((I_d - P_{\tilde{M}}) + P_{\tilde{M}}\tilde{z}\tilde{z}'P_{\tilde{M}})\right\}(I_d - P_{\tilde{M}})\tilde{\theta}; \end{aligned} \tag{13}$$

see also (4) and (5). Our goal is to show that the expressions in the preceding two displays are approximately zero. To this end, we focus on the expressions in curly brackets and use Cauchy–Schwarz: For each $t > 0$, we have

$$\mathbb{P}(|\mathbb{E}[e\|x]| > t) \leq \mathbb{P}\left(\left\|\mathbb{E}[\tilde{z}\|P_{\tilde{M}}\tilde{z}] - P_{\tilde{M}}\tilde{z}\right\| > t/\|(I_d - P_{\tilde{M}})\tilde{\theta}\|\right) \quad \text{and}$$

$$\mathbb{P}(|\mathbb{E}[e^2\|x] - s^2| > t) \leq P\left(\left\|\mathbb{E}[\tilde{z}\tilde{z}'\|P_{\tilde{M}}\tilde{z}] - ((I_d - P_{\tilde{M}}) + P_{\tilde{M}}\tilde{z}\tilde{z}'P_{\tilde{M}})\right\| > t/\|(I_d - P_{\tilde{M}})\tilde{\theta}\|^2\right).$$

Now if $R \in \mathbb{U}(M, \Sigma, f_{\tilde{z}})$, then by definition of $\mathbb{U}$, we have $\tilde{M}(\tilde{M}'\tilde{M})^{-1/2} \in \mathbb{G}(f_{\tilde{z}})$. Because conditioning on $P_{\tilde{M}}\tilde{z}$ is equivalent to conditioning on $(\tilde{M}'\tilde{M})^{-1/2}\tilde{M}'\tilde{z}$, it follows that $\mathbb{P}(|\mathbb{E}[e\|x]| > t)$ is bounded from above by (11) with t replaced by $t/\|(I_d - P_{\tilde{M}})\tilde{\theta}\|$ and that $\mathbb{P}(|\mathbb{E}[e^2\|x] - s^2| > t)$ is bounded by (11) with t replaced by $t/\|(I_d - P_{\tilde{M}})\tilde{\theta}\|^2$.

The consideration in the preceding paragraph suggests that the effect of misspecification in (4), where $\mathbb{E}[e\|x]$ may be nonzero and $\mathrm{Var}[e\|x]$ may be nonconstant, may be negligible in an asymptotic setting where $p/\log d$ becomes small, provided that $f_{\tilde{z}} \in \mathcal{F}_{d,20}$ and that $R \in \mathbb{U}(M, \Sigma, f_{\tilde{z}})$. This idea is formalized in the following two results, which show that the distribution of certain statistics is unaffected asymptotically if the error e is replaced by a substitute error e^* that has mean zero and constant variance conditional on x. The following results are stated for sequences where the data-generating model (1) and (2) as well as the working model (4) are allowed to depend on n, that is, a "triangular array" setting where all parameters depend on n.

LEMMA A.1. *Fix finite positive constants D and E. For every $n \in \mathbb{N}$, let $p_n \leq d_n$ be positive integers so that $np_n/\log d_n \to 0$ as $n \to \infty$. For each n, consider (y,z,x) as in (1)–(3) but with d_n and p_n replacing d and p, respectively, with $f_{\tilde{z}} \in \mathcal{F}_{d_n,20}(D,E)$ and with $R \in \mathbb{U}(M, \Sigma, f_{\tilde{z}})$. And for each n, consider a sample of n i.i.d. observations (y_i, z_i, x_i), $1 \leq i \leq n$, of (y,z,x), stack the values of the individual variables into a vector Y and matrices Z and X, respectively, and write $\varepsilon = Y - \alpha\iota - X\beta = (e_1, \ldots, e_n)'$ for the vector of errors from (4). Finally, define a vector $\varepsilon^* = (e_1^*, \ldots, e_n^*)'$ of substitute errors through $e_i^* = s(\mathrm{Var}[e_i\|x_i])^{-1/2}(e_i - \mathbb{E}[e_i\|x_i])$. Then, for every $k \in \mathbb{R}$ and (possibly random) symmetric idempotent $n \times n$ matrices P_n,*

$$n^k\|\varepsilon - \varepsilon^*\|/s \xrightarrow{p} 0 \quad \textit{and} \tag{14}$$

$$n^k|\varepsilon' P_n \varepsilon - \varepsilon^{*\prime} P_n \varepsilon^*|/s^2 \xrightarrow{p} 0, \tag{15}$$

as $n \to \infty$. As a by product, we also obtain that

$$\max_{i=1,\ldots,n} |\mathrm{Var}[e_i\|x_i]/s^2 - 1| \xrightarrow{p} 0.$$

Proof. First, note that $\mathrm{Var}[e_i\|x_i] = \mathrm{Var}[y_i\|x_i] = \mathrm{Var}[\theta' z_i\|x_i] + \sigma^2 > 0$, so that e_i^* is well defined (almost surely). For the claim in (14), fix $k \in \mathbb{R}$ and $t > 0$, and consider $\mathbb{P}(n^k\|\varepsilon - \varepsilon^*\|/s > t) \leq n\mathbb{P}(n^{2k+1}|e_1 - e_1^*|^2/s^2 > t^2)$. Now, using the simple observation $|\sqrt{\mathrm{Var}[e_1\|x_1]} - s| = |\mathrm{Var}[e_1\|x_1] - s^2|/|\sqrt{\mathrm{Var}[e_1\|x_1]} + s| \leq |\mathrm{Var}[e_1\|x_1] - s^2|/s$, we get

$$|e_1 - e_1^*|/s = (s^2\,\mathrm{Var}[e_1\|x_1])^{-1/2}\left|e_1(\sqrt{\mathrm{Var}[e_1\|x_1]} - s) + s\mathbb{E}[e_1\|x_1]\right|$$

$$\leq \frac{s}{\sqrt{\mathrm{Var}[e_1\|x_1]}}\left(\frac{|e_1|}{s}\frac{|\mathrm{Var}[e_1\|x_1] - s^2|}{s^2} + \frac{|\mathbb{E}[e_1\|x_1]|}{s}\right),$$

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

and furthermore

$$
\begin{aligned}
&\mathbb{P}(n^{2k+1}|e_1 - e_1^*|^2/s^2 > t^2)\\
&\leq \mathbb{P}\left(n^{k+1/2}\left|\frac{|e_1|}{s}\frac{|\operatorname{Var}[e_1\|x_1]-s^2|}{s^2}+\frac{|\mathbb{E}[e_1\|x_1]|}{s}\right| > t/\sqrt{2}\right)\\
&\qquad + \mathbb{P}\left(\frac{s^2}{\operatorname{Var}[e_1\|x_1]} > 2\right)\\
&\leq \mathbb{P}\left(\left|\frac{\operatorname{Var}[e_1\|x_1]}{s^2}-1\right| > \frac{1}{2}\right)+\mathbb{P}\left(n^{k+1/2}\frac{|e_1|}{s}\frac{|\operatorname{Var}[e_1\|x_1]-s^2|}{s^2} > t/2^{3/2}\right)\\
&\qquad + \mathbb{P}\left(n^{k+1/2}\frac{|\mathbb{E}[e_1\|x_1]|}{s} > t/2^{3/2}\right)\\
&\leq \mathbb{P}\left(\frac{|\operatorname{Var}[e_1\|x_1]-s^2|}{s^2} > \frac{1}{2}\right)+\mathbb{P}\left(n^{k+3/2}\frac{|\operatorname{Var}[e_1\|x_1]-s^2|}{s^2} > t/2^{3/2}\right)\\
&\qquad + \mathbb{P}\left(\frac{|e_1|}{s} > n\right)+\mathbb{P}\left(n^{k+1/2}\frac{|\mathbb{E}[e_1\|x_1]|}{s} > t/2^{3/2}\right).
\end{aligned}
\tag{16}
$$

The claim (14) will follow if each of the four terms in (16) is of the order $o(1/n)$. Because $f_{\tilde{z}} \in \mathcal{F}_{d_n,20}(D,E)$ and $R \in \mathbb{U}(M,\Sigma,f_{\tilde{z}})$, the considerations leading up to Lemma A.1 apply. Also note that $\|(I_d - P_{\tilde{M}})\tilde{\theta}\|^2 \leq s^2$. For the last term in (16), we obtain, for every $t > 0$, that

$$
\mathbb{P}\left(n^{k+1/2}\frac{|\mathbb{E}[e_1\|x_1]|}{s} > t\right) \leq t^{-1}n^{k+1/2}d_n^{-1/20} + 4\gamma\frac{p_n}{\log d_n},
$$

and the upper bound goes to zero as $o(1/n)$ in view of the assumption that $np_n/\log d_n \to 0$. For the second-to-last term in (16), we have $\mathbb{P}(|e_1|/s > n) \leq n^{-2}\mathbb{E}[e_1^2/s^2] = 1/n^2$. For the second term in (16), we proceed like for the last term in (16). In particular, we obtain, for any $t > 0$, that

$$
\mathbb{P}\left(n^{k+3/2}\frac{|\operatorname{Var}[e_1\|x_1]-s^2|}{s^2} > t\right) \tag{17}
$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(n^{k+3/2}\frac{|\mathbb{E}[e_1^2\|x_1]-s^2|}{s^2} > t/2\right)+\mathbb{P}\left(n^{k+3/2}\frac{|\mathbb{E}[e_1\|x_1]|^2}{s^2} > t/2\right)\\
&\leq \frac{2}{t}n^{k+3/2}d^{-1/20} + \left(\frac{2}{t}n^{k+3/2}\right)^{1/2}d^{-1/20} + 8\gamma\frac{p_n}{\log d_n}.
\end{aligned}
\tag{18}
$$

Again, this upper bound goes to zero as $o(1/n)$ because $np_n/\log d_n \to 0$. Note that the considerations in the preceding display also entail that $\mathbb{P}(\max_{i=1,\dots,n}|\operatorname{Var}[e_i\|x_i]/s^2 - 1| > t) \leq n\mathbb{P}(|\operatorname{Var}[e_1\|x_1]/s^2 - 1| > t) \to 0$.

For the claim in (15), write

$$
\begin{aligned}
|\varepsilon' P_n \varepsilon - \varepsilon^{*\prime} P_n \varepsilon^*| &= |(\varepsilon-\varepsilon^*)' P_n \varepsilon + \varepsilon^{*\prime} P_n(\varepsilon - \varepsilon^*)|\\
&\leq \|\varepsilon - \varepsilon^*\|\|\varepsilon\| + \|\varepsilon - \varepsilon^*\|\|\varepsilon^*\|,
\end{aligned}
$$

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

and note that by definition of e_1^* and the variance decomposition formula, we have $\mathbb{E}[e_1^*] = \mathbb{E}[\mathbb{E}[e_1^* \| x_1]] = 0$ and $\mathrm{Var}[e_1^*] = \mathbb{E}[\mathrm{Var}[e_1^* \| x_1]] + \mathrm{Var}[\mathbb{E}[e_1^* \| x_1]] = s^2$, so that by independence $\|\varepsilon^*\|/s = O_{\mathbb{P}}(\sqrt{n})$. Premultiplying by n^k/s^2 in the previous display and applying (14) finishes the proof of the second claim. □

LEMMA A.2. *Fix* $K \in (0,\infty)$ *and an integer* $l \geq -1$. *Under the assumptions and in the notation of Lemma A.1, assume that* $\mathbb{E}[|\epsilon/\sigma|^4] \leq K$ *for each n, that* $\Delta = \mathrm{Var}(\beta' x)/\mathrm{Var}(e) = O(n^l)$ *and that* $\limsup_{n\to\infty} p_n/n < 1$. *Define substitute data* $Y^* = \iota\alpha + X\beta + \varepsilon^*$. *Then, for every* $k \in \mathbb{R}$, *we have*

$$n^k \left(\hat{F}_n(X,Y) - \hat{F}_n(X,Y^*)\right) \quad \xrightarrow{p} \quad 0$$

as $n \to \infty$.

Proof. The idea is to use Lemma A.1 to approximate $\hat{F}_n(X,Y)$ by $\hat{F}_n(X,Y^*)$. In particular, we will show that on some event C_n to be defined below, we have

$$n^k \left|\hat{F}_n(X,Y) - \hat{F}_n(X,Y^*)\right| \leq n^{k+l+1} |\delta_n^{(1)} - 1| \hat{F}_n(X,Y^*)/n^{l+1} + n^k |\delta_n^{(2)}|,$$

where $\delta_n^{(1)}$ converges to one and $\delta_n^{(2)}$ converges to zero, both at an arbitrary polynomial rate in n, and where $\hat{F}_n(X,Y^*)/n^{l+1} = O_{\mathbb{P}}(1)$. The probability of C_n will be shown to converge to one. The claim of the lemma follows from this.

Set $U = [\iota, X]$, where $\iota = (1,\dots,1)' \in \mathbb{R}^n$. With this, define the event $C_n = \{\det U'U \neq 0, \varepsilon'(I_n - P_U)\varepsilon > 0, \varepsilon^{*\prime}(I_n - P_U)\varepsilon^* > 0\}$. On C_n, by block matrix inversion, we have $[0, I_{p_n}](U'U)^{-1}U' = [X'(I_n - P_\iota)X]^{-1} X'(I_n - P_\iota)$. Using the abbreviation $V = (I_n - P_\iota)X$, we thus see that $\hat{\beta} = \beta + (V'V)^{-1}V'\varepsilon$ and that the F-statistic $\hat{F}_n(X,Y)$ can be written as

$$\begin{aligned}\hat{F}_n(X,Y) &= \frac{n-p_n-1}{p_n} \frac{\|V\hat{\beta}\|^2}{\|(I-P_U)Y\|^2} = \frac{n-p_n-1}{p_n} \frac{\varepsilon' P_V \varepsilon + 2\varepsilon' V\beta + \beta' V' V \beta}{\varepsilon'(I_n - P_U)\varepsilon} \\ &= \frac{\varepsilon^{*\prime}(I_n - P_U)\varepsilon^*}{\varepsilon'(I_n - P_U)\varepsilon} \hat{F}_n(X,Y^*) + \frac{\varepsilon' P_V \varepsilon - \varepsilon^{*\prime} P_V \varepsilon^* + 2(\varepsilon - \varepsilon^*)' V\beta}{p_n \varepsilon'(I_n - P_U)\varepsilon/(n-p_n-1)}.\end{aligned}$$

This establishes a representation $\hat{F}_n(X,Y) = \delta_n^{(1)} \hat{F}_n(X,Y^*) + \delta_n^{(2)}$ on C_n. On the complement of C_n, we set $\delta_n^{(1)} = \delta_n^{(2)} = 0$, say. We next show that for every fixed $k \in \mathbb{R}$, $n^k(\delta_n^{(1)} - 1) = o_{\mathbb{P}}(1)$ and $n^k \delta_n^{(2)} = o_{\mathbb{P}}(1)$.

To verify the claimed properties of these quantities, on C_n, consider first

$$\delta_n^{(1)} - 1 = \frac{\varepsilon^{*\prime}(I_n - P_U)\varepsilon^* - \varepsilon'(I_n - P_U)\varepsilon}{s^2(n-p_n-1)} \frac{s^2(n-p_n-1)}{\varepsilon'(I_n - P_U)\varepsilon}.$$

Using Lemma A.1, we see that the first fraction in this representation multiplied by n^k converges to zero in probability. The second fraction obviously equals $s^2/\hat{s}^2$. Define $\hat{s}^{*2}$ like $\hat{s}^2$ (see the discussion following (6)) but with Y^* replacing Y. We show that $\hat{s}^2/s^2 = \hat{s}^{*2}/s^2 + (\hat{s}^2 - \hat{s}^{*2})/s^2 \to 1$ in probability. To see this, first note that the convergence to zero of $(\hat{s}^2 - \hat{s}^{*2})/s^2$ follows again from Lemma A.1. For the ratio $\hat{s}^{*2}/s^2$, convergence to 1 in probability follows, e.g., from Lemma C.1 in Steinberger (2016), upon verifying its

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

assumptions. To this end, it remains to show that $n^{-1}\sum_{i=1}^n \mathbb{E}[(e_i^*/s)^4\|x_i] = O_{\mathbb{P}}(1)$. Using $(a+b)^4 \leq 2^3(a^4+b^4)$, for $a,b \in \mathbb{R}$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[(e_i^*/s)^4\|x_i] &\leq \max_{j=1,\dots,n}\left(\frac{s^2}{\operatorname{Var}[e_j\|x_j]}\right)^2 \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(e_i/\sigma - \mathbb{E}[e_i/s\|x_i])^4\|x_i] \\ &\leq \max_{j=1,\dots,n}\left(\frac{s^2}{\operatorname{Var}[e_j\|x_j]}\right)^2 2^4\frac{1}{n}\sum_{i=1}^n \mathbb{E}[(e_i/s)^4\|x_i].\end{aligned}$$

The maximum in the preceding display converges to one in probability if $\min_j \operatorname{Var}[e_j/s\|x_j]$ converges to one in probability, which follows from Lemma A.1. The arithmetic mean of the conditional fourth moments is $O_{\mathbb{P}}(1)$ if the unconditional mean of fourth moments is bounded in n. To this end, note that we have $e = \tilde{\theta}'(I_d - P_{\tilde{M}})\tilde{z} + \epsilon$ and $s^2 = \|(I_d - P_{\tilde{B}})\tilde{\theta}\|^2 + \sigma^2$; cf. (5) and the discussion right before (13). With this, we get

$$\begin{aligned}(e_i/s)^4 = \left(\theta'(I_d - P_{\tilde{M}})\tilde{z}_i/s + \epsilon_i/s\right)^4 &\leq 2^3[(\tilde{\theta}'(I_d - P_{\tilde{B}})\tilde{z}_i/s)^4 + (\epsilon_i/s)^4] \\ &\leq 2^3[(\tilde{\theta}'(I_d - P_{\tilde{B}})\tilde{z}_i/\|\tilde{\theta}'(I_d - P_{\tilde{B}})\|)^4 + (\epsilon_i/\sigma)^4],\end{aligned}$$

and take expectations. The claim follows now from $\mathbb{E}[(\epsilon_i/\sigma)^4] \leq K$ and the fact that the fourth spherical moment of $\tilde{z}_i$ is uniformly bounded in view of Rosenthal's inequality (Rosenthal, 1970, Theorem 3) and the assumption that $f_{\tilde{z}} \in \mathcal{F}_{d_n,20}(D,E)$. Note that this also entails $\mathbb{P}(C_n^c) \leq \mathbb{P}(\hat{s}^{*2} = 0) + \mathbb{P}(\hat{s}_n^2 = 0) \leq \mathbb{P}(|\hat{s}^{*2}/s^2 - 1| > 1/2) + \mathbb{P}(|\hat{s}^2/s^2 - 1| > 1/2) \to 0$.

To see that also $\delta_n^{(2)}$ behaves as desired, first note that on C_n,

$$n^k\delta_n^{(2)} = \frac{n^k}{p_n}\left(\frac{\varepsilon' P_V\varepsilon - \varepsilon^{*\prime}P_V\varepsilon^*}{s^2} + \frac{2(\varepsilon - \varepsilon^*)'V\beta}{s^2}\right)\frac{s^2}{\hat{s}^2}.$$

The factor n^k/p_n can be bounded by κn^{k-1} for some constant κ by assumption; the ratio $s^2/\hat{s}^2$ was shown to converge to one in probability in the preceding paragraph. The difference of quadratic forms converges to zero in probability by Lemma A.1, even when multiplied by κn^{k-1}. Noting that $\|V\beta\| = \|(I_n - P_\iota)X\beta\| \leq \|(I_n - P_\iota)X(\tilde{M}'\tilde{M})^{-1/2}\|\|(\tilde{M}'\tilde{M})^{1/2}\beta\|$, the scaled second term in parentheses, that is, $(n^k/p_n)2(\varepsilon - \varepsilon^*)'V\beta/s^2$, can be bounded by

$$2\kappa n^{k+l/2}\frac{\|\varepsilon - \varepsilon^*\|}{s}\frac{\|(\tilde{M}'\tilde{M})^{1/2}\beta\|}{sn^{l/2}}\left\|(I_n - P_\iota)X(\tilde{M}'\tilde{M})^{-1/2}\right\|/n,$$

where $n^{k+l/2}\|\varepsilon - \varepsilon^*\|/s$ converges to zero in probability by Lemma A.1 and $n^{-l}\beta'(\tilde{M}'\tilde{M})\beta/s^2 = n^{-l}\Delta = O(1)$ by assumption. It remains to show that the largest singular value of $(I_n - P_\iota)X(\tilde{M}'\tilde{M})^{-1/2}/n$ is bounded in probability. Due to the projection onto the orthogonal complement of ι, the distribution of this quantity does not depend on the parameter μ, which is why we may assume that $\mu = 0$ for this part of the argument. Abbreviate $\bar{X} = X(\tilde{M}'\tilde{M})^{-1/2}$, $\bar{x}_i = (\tilde{M}'\tilde{M})^{-1/2}x_i$ and consider $\|(I_n - P_\iota)\bar{X}/n\|^2 \leq \operatorname{trace}(\bar{X}'\bar{X}/n^2) = \sum_{i=1}^n \|\bar{x}_i\|^2/n^2$. Taking expectation, noting that $\mathbb{E}[\|\bar{x}_1\|^2] = p_n$ and $p_n/n = O(1)$, we arrive at the desired boundedness in probability.

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

It remains to show that $\hat{F}_n(X,Y^*)/n^{l+1} = O_{\mathbb{P}}(1)$. To this end, recall that $\hat{s}^{*2}/s^2 \to 1$ in probability, and one easily verifies that

$$\mathbb{E}\left[\frac{\hat{s}^{*2}}{s^2}\hat{F}_n(X,Y^*)/n^{l+1}\right] = \mathbb{E}\left[(\varepsilon^{*\prime}P_V\varepsilon^* + 2\varepsilon^{*\prime}V\beta + \beta' V' V\beta)/(p_n s^2 n^{l+1})\right]$$
$$= \frac{1}{n^{l+1}} + \frac{n-1}{np_n}\frac{\Delta}{n^l} = O(1);$$

here, the first equality is obtained by arguing as in the first paragraph of the proof but with Y^* replacing Y, and the second equality follows upon noting that $\beta' V' V\beta = \operatorname{trace}(I_n - P_\iota)X\beta\beta' X'$ and that $X\beta$ is a vector with i.i.d. components, each of which has variance $\beta' M' \Sigma M\beta = s^2\Delta$. □

Proof of Theorem 4.1. Define $\mathbb{U} = \mathbb{U}(M,\Sigma, f_{\tilde{z}})$ as in the beginning of the appendix and note that the first statement in the theorem, concerning $\nu_d(\mathbb{U})$, has already been established there. For the second statement, concerning Ξ_n, let $p_n \leq d_n$ be positive integers so that $n^2 p_n/\log d_n \to 0$ and so that $p_n/n \to \rho \in (0,1)$ as $n \to \infty$. For each n, consider a sample of i.i.d. observations (y_i, z_i, x_i), $1 \leq i \leq n$, as in Lemma A.1, so that the underlying quantities (i.e., M, ϑ, θ, $\mathcal{L}(\epsilon)$, μ, Σ, Δ, $f_{\tilde{z}}$, and R) satisfy the restrictions in the suprema in the last display of Theorem 4.1. For given M, we stress that the restriction on Δ implicitly also restricts the parameters θ, Σ and σ^2; see the definition of Δ at the beginning of Section 4 as well as the relations in (5). We have to show that $\Xi_n \to 0$ as $n \to \infty$.

Set $a_n = 2(1/p_n + 1/(n-p_n-1))$ and $b_n = \sqrt{\frac{(1-(p_n+1)/n)(1-1/n)}{2p_n/n}}$ for each n, and define Y^* for each n as in Lemma A.2. We first show that

$$a_n^{-1/2}(\hat{F}_n(X,Y^*) - 1) - \sqrt{n}\Delta b_n \quad \xrightarrow[n\to\infty]{w} \quad N(0,1) \tag{19}$$

by verifying the assumptions of Theorem 2.1(i) in Steinberger (2016) for the sample $(y_i^*, x_i)_{i=1}^n$, with the symbols s_n, Δ_γ and R_0 in that reference equal to a_n, Δ, and $[0, I_{p_n}]$, respectively. Clearly, under the imposed conditions $\Delta < \gamma/\sqrt{n}$ or $\Delta < g(n) = o(1)$, we have $\Delta = o(p/n)$, because $p/n \to \rho \in (0,1)$. In particular, we need to verify conditions (A1).(a,b,c,d) and (A2) in that reference. The design conditions (A1).(a,c,d) are easily verified by use of Lemma A.2(i) in Steinberger (2016). And our assumptions that $f_{\tilde{z}} \in \mathcal{F}_{d_n,20}(D,E)$ and that $p_n < n-1$ imply condition (A1).(b). Assumption (A2) on the scaled errors e_i^*/s is established by an argument similar to the one also used in the third paragraph of the proof of Lemma A.2 but for the $(8+\kappa)$-th moment instead of the fourth moment: Simply decompose $e_i^* = e_i^\circ \tilde{\varepsilon}_i$, with $e_i^\circ = \sqrt{s^2/\operatorname{Var}[e_i \| x_i]}$ and $\tilde{\varepsilon}_i = e_i - \mathbb{E}[e_i \| x_i]$, and use Lemma A.1 as before to get $\max_{i=1,\ldots,n} e_i^\circ \to 1$ in probability. Then, the assumption that $\mathbb{E}[|\epsilon/\sigma|^{8+\kappa}] \leq K$ and the fact that the marginals of $\tilde{z} \in \mathcal{F}_{d_n,20}(D,E)$ have bounded 20th moment, together with Rosenthal's inequality establish the boundedness of $\mathbb{E}[|\tilde{\varepsilon}_i/s|^{8+\kappa}]$, which is sufficient for (A2), provided that p_n is of the same order as n (cf. Remark 4.3). Using Lemma A.2, noting that $a_n^{-1/2} = O(\sqrt{n})$, it follows that (19) continues to hold with $\hat{F}_n(X,Y)$ replacing $\hat{F}_n(X,Y^*)$.

Now standard arguments conclude the proof: First, note that an appropriately scaled and centered F-distributed random variable $\mathcal{F}_{p_n,n-p_n-1,n\Delta}$ with p_n and $n-p_n-1$ degrees of freedom and noncentrality parameter $n\Delta$ is also asymptotically normal, that is,

$$a_n^{-1/2}(\mathcal{F}_{p_n,n-p_n-1,n\Delta} - 1) - \sqrt{n}\Delta b_n \xrightarrow[n\to\infty]{w} N(0,1), \tag{20}$$

because $p_n/n \to \rho \in (0,1)$ implies that $p_n \to \infty$. Hence, we have

$$\begin{aligned}
&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\hat{F}_n(X,Y)\le t\right)-\mathbb{P}(\mathcal{F}_{p_n,n-p_n-1,n\Delta}\le t)\right|\\
&\quad=\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(a_n^{-1/2}(\hat{F}_n(X,Y)-1)-\sqrt{n}\Delta b_n\le t\right)\right.\\
&\qquad\left.-\,\mathbb{P}\left(a_n^{-1/2}(\mathcal{F}_{p_n,n-p_n-1,n\Delta}-1)-\sqrt{n}\Delta b_n\le t\right)\right|\\
&\quad\le\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(a_n^{-1/2}(\hat{F}_n(Y,X)-1)-\sqrt{n}\Delta b_n\le t\right)-\Phi(t)\right|\\
&\qquad+\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(a_n^{-1/2}(\mathcal{F}_{p_n,n-p_n-1,n\Delta}-1)-\sqrt{n}\Delta b_n\le t\right)-\Phi(t)\right|,
\end{aligned}$$

and the last two suprema converge to zero in view of Polya's theorem, which establishes that $\Xi_n \to 0$ in case Ξ_n equals (7). Finally, it is elementary to verify that Ξ_n also converges to zero in case Ξ_n equals (8): This follows from (19) with $\hat{F}_n(X,Y)$ replacing $\hat{F}_n(X,Y^*)$, because the quantiles of the central F-distribution satisfy $a_n^{-1/2}(F^{-1}_{p_n,n-p_n-1,0}(\alpha)-1) \to \Phi^{-1}(\alpha)$. □

Remark A.3. Inspection of the proof reveals that the assumption that σ^2 is positive is used only to guarantee that $\mathrm{Var}[e\|x] > 0$ almost surely (and hence also $s^2 = \mathrm{Var}[e] > 0$). If this assumption is dropped, we thus see that Ξ_n (defined in Theorem 4.1) converges to zero along sequences of parameters as used in the proof of Theorem 4.1, provided that $\mathrm{Var}[\theta' z\|x] > 0$ almost surely for each n (as then $\mathrm{Var}[e\|x] = \mathrm{Var}[y\|x] > 0$ a.s.).

REFERENCES

Abadie, G., G. W. Imbens & F. Zheng (2014) Inference for misspecified models with fixed regressors. *Journal of the American Statistical Association* 109, 1601–1614.

Akritas, M. & S. Arnold (2000) Asymptotics for analyis of variance when the number of levels is large. *Journal of the American Statistical Association* 95, 212–226.

Anderson, T. W. (1958) *An Introduction to Multivariate Analysis*. Wiley.

Bachoc, F., H. Leeb & B. M. Pötscher (2019) Valid confidence intervals for post-model-selection predictors. *Annals of Statistics* 47, 1475–1504.

Bai, Z. & H. Saranadasa (1996) Effect of high dimension: By an example of a two sample problem. *Statistica Sinica* 6, 311–329.

Baik, J. & J. W. Silverstein (2006) Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* 97, 1382–1408.

Bathke, A. & D. Lankowski (2005) Rank procedures for a large number of treatments. *Journal of Statistical Planning and Inference* 133, 223–238.

Boos, D. D. & C. Brownie (1995) ANOVA and rank tests when the number of treatments is large. *Statistics and Probability Letters* 23, 183–191.

Boos, D.D. & L.A. Stefanski (2013) *Essential Statistical Inference*, Springer Texts in Statistics. Springer.

Brannath, W. & M. Scharpenberg (2014) Interpretation of linear regression coefficients under mean model miss-specification. arXiv:1409.8544.

Buja, A.R., L. D. Brown, E. George, E. Pitkin, M. Traskin, K. Zhan, & L. Zhao (2014) A conspiracy of random predictors and model violations against classical inference in regression. arXiv:1404.1578.

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

Cai, T., Z. Ma & Y. Wu (2013) Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields* 161, 1–35.

Cattaneo, M. D., M. Jansson & W. K. Newey (2018) Inference in linear regression models with many covariates and heteroscedasticity. *Journal of the American Statistical Association* 113, 1350–1361.

Chen, S.-X. & Y.-L. Qin (2010) A two-sample test for high-dimensional data with applications to gene-set testing. *Annals of Statistics* 38, 808–835.

Choi, H. S. & N. M. Kiefer (2011) Geometry of the log-likelihood ratio statistic in misspecified models. *Journal of Statistical Planning and Inference* 141, 2091–2099.

Dobriban, E. & W. Su (2018) Robust inference under heteroskedasticity via the Hadamard estimator. arXiv:1807.00347.

Donoho, D. L., M. Gavish & I. M. Johnstone (2018) Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of Statistics* 46, 1742–1778.

Eicker, F. (1967) Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 59–82. University of California Press.

Fomby, T.B. & R.C. Hill (2003) *Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later*, Advances in Econometrics, 17. Elsevier.

Hall, P. & K.-C. Li (1993) On almost linearity of low dimensional projections from high dimensional data. *Annals of Statistics* 21, 867–889.

Harrar, S. & A. C. Bathke (2008) Nonparametric methods for unbalanced multivariate data and many factor levels. *Journal of Multivariate Analysis* 99, 1635–1664.

Huber, P. J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In Lucien M. Le Cam, Jerzy Neyman (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 221–233. University of California Press.

Jensen, D. R. & D. E. Ramirez (1991) Misspecified t^2 tests. I. Location and scale. *Communications in Statistics. Theory and Methods* 20, 249–259.

Jochmans, K. Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association*, first published online 19 November 2020. https://doi.org/10.1080/01621459.2020.1831924.

Johnstone, I. M. (2001) On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics* 29, 295–327.

Leeb, H. (2013) On the conditional distributions of low-dimensional projections from high-dimensional data. *Annals of Statistics* 41, 464–483.

Li, Z. & J. Yao (2019) Testing for heteroscedasticity in high-dimensional regressions. *Econometrics and Statistics* 9, 122–139.

Portnoy, S. (1984) Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large. I. Consistency. *Annals of Statististics* 12, 1298–1309.

Portnoy, S. (1985) Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Annals of Statistics* 13, 1403–1417.

Preinerstorfer, D. & B. M. Pötscher (2016) On size and power of heteroskedasticity and autocorrelation robust tests. *Econometric Theory* 32, 261–358.

Ramirez, D. E. & D. R. Jensen (1991) Misspecified t^2 tests. II. Series expansions. *Communications in Statistics. Theory and Methods* 20, 97–108.

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

Souders, T. M. & G. N. Stenbakken (1991) Cutting the high cost of testing. *IEEE Spectrum* 28, 48–51.

Steinberger, L. (2015) *Statistical inference in high-dimensional linear regression based on simple working models*. PhD thesis, University of Vienna.

Steinberger, L. (2016) The relative effects of dimensionality and multiplicity of hypotheses on the F-test in linear regression. *Electronic Journal of Statistics* 10, 2584–2640.

Steinberger, L. & H. Leeb (2018) On conditional moments of high-dimensional random vectors given lower-dimensional projections. *Bernoulli* 24, 565–591.

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press

Steinberger, L. & H. Leeb (2019) Prediction when fitting simple models to high-dimensional data. *Annals of Statistics* 47, 1408–1442.

Stock, J. H. & M. W. Watson (2002) Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* 97, 1167–1179.

van't Veer, L. J., H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards & S. H. Friend (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536.

Wang, S. & H. Cui (2013) Generalized F test for high dimensional linear regression coefficients. *Journal of Multivariate Analysis* 117, 134–149.

White, H. (1980a) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 84, 817–838.

White, H. (1980b) Using least squares to approximate unknown regression functions. *International Economic Review* 21, 149–170.

Zhong, P. S. & S. X. Chen (2011) Tests for high-dimensional regression coefficients with factorial designs. *Journal of the American Statistical Association* 106, 260–274.

https://doi.org/10.1017/S026646662100044X Published online by Cambridge University Press