ON THE CONDITIONAL DISTRIBUTIONS OF LOW-DIMENSIONAL PROJECTIONS FROM HIGH-DIMENSIONAL DATA

BY HANNES LEEB

University of Vienna

We study the conditional distribution of low-dimensional projections from high-dimensional data, where the conditioning is on other low-dimensional projections. To fix ideas, consider a random d-vector Z that has a Lebesgue density and that is standardized so that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. Moreover, consider two projections defined by unit-vectors α and β, namely a response $y = \alpha'Z$ and an explanatory variable $x = \beta'Z$. It has long been known that the conditional mean of y given x is approximately linear in x, under some regularity conditions; cf. Hall and Li [*Ann. Statist.* **21** (1993) 867–889]. However, a corresponding result for the conditional variance has not been available so far. We here show that the conditional variance of y given x is approximately constant in x (again, under some regularity conditions). These results hold uniformly in α and for most β's, provided only that the dimension of Z is large. In that sense, we see that most linear submodels of a high-dimensional overall model are approximately correct. Our findings provide new insights in a variety of modeling scenarios. We discuss several examples, including sliced inverse regression, sliced average variance estimation, generalized linear models under potential link violation, and sparse linear modeling.

1. Introduction.

1.1. *Informal summary.* We analyze a situation where a simple model is used when the true model is, in fact, much more complex. This situation is particularly common with many contemporary datasets where the number of potentially important covariates or the number of parameters exceeds the sample size; examples in, say, genomics or economics abound. When facing a large number of potentially important covariates or parameters, and a small sample size, the search for simple models is typically motivated by either one of two types of assumptions: Parametric assumptions, which postulate that the true data-generating process is given by a simple finite-dimensional model; and nonparametric assumptions, which postulate that the true data-generating process can be approximated, with arbitrary accuracy, by comparatively simple finite-dimensional models. In either case, the underlying postulates can be difficult to verify in practice. And this difficulty is often further

compounded by the relatively small sample size. The results that we obtain here provide an alternative justification for the use of simple models. We analyze a scenario where most simple submodels are approximately correct, provided only that the overall model is sufficiently complex, irrespective of whether or not the true data-generating process is given by, or can be closely approximated by, a finite-dimensional simple model. This is the main conceptual contribution of this paper.

On a technical level, we extend and refine a method pioneered by Hall and Li [11]. In that reference, the authors propose a novel approach for studying conditional means of linear projections under weak distributional assumptions; cf. Theorem 3.2 of [11]. Our technical core contribution is an extension of Hall and Li's approach to also cover conditional variances and higher conditional moments, and a more explicit control of error terms that allows us to prove strong statements like (1.5) and (1.6), which follow. Note that we deal here with the largest singular value of conditional covariance matrices of increasing dimension, which turns out to be considerably more challenging than handling the norm of the conditional mean vectors that are treated in [11].

The paper is organized as follows: We continue this section with a more detailed overview of our findings, and with a discussion of some interesting consequences. In Section 2, we present our main result, namely Theorem 2.1, and give an outline of its proof. The proof is based on five basic steps that correspond to five propositions that are also given in Section 2. The (more technical) proofs of these propositions are relegated to the supplementary material [13].

1.2. *Overview of results.*   Consider a random d-vector Z that has a Lebesgue density and that is standardized so that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. Throughout, we will study projections of Z of the form $\alpha'Z$ and $\beta'Z$ for unit d-vectors α and β. The conditional mean of $\alpha'Z$ given $\beta'Z = x$ will be denoted by $\mathbb{E}[\alpha'Z \| \beta'Z = x]$; other conditional expectations are defined similarly.[1] Our main results are concerned with the conditional mean and with the conditional variance of $\alpha'Z$ given that $\beta'Z = x$. To introduce these, consider the following two conditions: The vector β is such that:

(i) for each α, the conditional mean of $\alpha'Z$ given $\beta'Z = x$ is linear in $x \in \mathbb{R}$;
(ii) for each α, the conditional variance of $\alpha'Z$ given $\beta'Z = x$ is constant in $x \in \mathbb{R}$.

Suppose that α is an unknown parameter and that one can observe y and x given by $y = \alpha'Z$ and $x = \beta'Z$, respectively. If β is such that both (i) and (ii) hold,

---

[1]There is a measurable function $g : \mathbb{R} \to \mathbb{R}$ so that $\mathbb{E}[\alpha'Z \| \beta'Z] = g(\beta'Z)$ holds, and we write $\mathbb{E}[\alpha'Z \| \beta'Z = x]$ for $g(x)$; the existence of g is guaranteed by, say, Theorem 4.2.8 of [7]. When we say that $\mathbb{E}[\alpha'Z \| \beta'Z = x]$ is linear in x [as in condition (i), which follows], we mean that $g(x)$ can be chosen to be linear. Similar considerations apply, mutatis mutandis, to expressions like $\mathbb{E}[Z \| \beta'Z = x]$, $\mathbb{E}[(\alpha'Z)^2 \| \beta'Z = x]$ or $\mathbb{E}[ZZ' \| \beta'Z = x]$.

then y can be decomposed into the sum of a linear function of x and a remainder-term, or error-term, whose conditional mean given x is zero and whose conditional variance given x is constant; in other words, the model

$$y = \gamma x + u$$

applies, where $\mathbb{E}[u\|x] = 0$ and $\text{Var}[u\|x] = \sigma^2$, and where $\gamma \in \mathbb{R}$ and $\sigma^2 \geq 0$ are unknown parameters that are given by $\gamma = \alpha'\beta$ and $\sigma^2 = 1 - (\alpha'\beta)^2$, respectively. (Indeed, (i) implies that $\mathbb{E}[y\|x] = \mathbb{E}[\alpha'Z\|\beta'Z] = \mu + \gamma(\beta'Z)$ for some real constants μ and γ. It now follows from $\mathbb{E}Z = 0$ that $\mu = 0$, and $\mathbb{E}ZZ' = I_d$ implies that $\gamma = \alpha'\beta$. Moreover, (ii) entails that $\text{Var}[\alpha'Z\|\beta'Z] = \sigma^2$ for some constant σ^2; hence $1 = \text{Var}(y) = \text{Var}(\alpha'Z) = \mathbb{E}[\text{Var}[\alpha'Z\|\beta'Z]] + \mathbb{E}[(\mathbb{E}[\alpha'Z\|\beta'Z])^2] = \sigma^2 + (\alpha'\beta)^2$, so that σ^2 is given by $\sigma^2 = 1 - (\alpha'\beta)^2$.) These observations continue to hold also if the vectors α and β are not normalized to unit length, mutatis mutandis.

Conditions (i) and (ii) are satisfied for each β if Z is normally distributed, that is, $Z \sim N(0, I_d)$. But besides the Gaussian law, the class of distributions that satisfy (i) and (ii) for each β appears to be quite small: Indeed, if Z satisfies (i) for each β, then the law of Z is spherically symmetric [9]. And if, in addition, also (ii) holds for some β, then Z is Gaussian [1], Theorem 4.1.4.

Under comparatively mild conditions on the distribution of Z, we here show that both conditions (i) and (ii) are *approximately* satisfied for *most* unit-vectors β, namely for a set of unit-vectors β whose size, as measured by the uniform distribution on the unit-sphere in $\mathbb{R}^d$, goes to one as $d \to \infty$. To state this more formally, we first describe two preliminary results, namely (1.3) and (1.4), which follow, and then extend these to our main results in (1.5) and (1.6) below.

To introduce the two preliminary results mentioned earlier, we note that (i) and (ii) together are equivalent to the requirement that $\mathbb{E}[Z\|\beta'Z = x] = \beta x$ and $\mathbb{E}[ZZ'\|\beta'Z = x] = I_d + (x^2 - 1)\beta\beta'$ hold for each $x \in \mathbb{R}$. (In other words, the first two moments of the conditional distribution coincide with what they would be in the Gaussian case.) From this, it is easy to see that (i) and (ii) are also equivalent to the requirement that both

(1.1) $$\left\|\mathbb{E}[Z\|\beta'Z = x]\right\|^2 - x^2 = 0 \quad \text{and}$$

(1.2) $$\left\|\mathbb{E}[ZZ'\|\beta'Z = x] - (I_d + (x^2 - 1)\beta\beta')\right\| = 0$$

hold for each $x \in \mathbb{R}$. Note that we use the notation $\|\cdot\|$ to denote both the Euclidean norm of vectors as in (1.1) and the operator norm of matrices as in (1.2). The left-hand side of (1.1) can be written as

$$\left\|\mathbb{E}[Z\|\beta'Z = x]\right\|^2 - x^2 = \sup_{\alpha}\left|\mathbb{E}[\alpha'Z\|\beta'Z = x] - \alpha'\beta x\right|^2,$$

with the supremum taken over all unit-vectors $\alpha \in \mathbb{R}^d$, as is elementary to verify.[2] Hence, the left-hand side of (1.1) is always nonnegative and can be interpreted as the worst-case deviation of the regression function $\mathbb{E}[\alpha'Z \| \beta'Z = x]$ from a linear function at x. The left-hand side of (1.2) can be interpreted in a similar fashion. For fixed $x \in \mathbb{R}$, the condition (1.1) is approximately satisfied for most β's if d is large, under the assumptions of Theorem 3.2 in [11] [see also equation (1.5) in that reference], in the sense that

$$(1.3) \qquad \upsilon\{\beta \in \mathbb{R}^d : \|\mathbb{E}[Z\|\beta'Z = x]\|^2 - x^2 > \varepsilon\} \overset{d \to \infty}{\longrightarrow} 0$$

for each fixed $x \in \mathbb{R}$ and for each $\varepsilon > 0$, where υ denotes the uniform distribution on the unit-sphere in $\mathbb{R}^d$. We here show that condition (1.2) is similarly approximately satisfied for most β's if d is large, in the sense that

$$(1.4) \qquad \upsilon\{\beta \in \mathbb{R}^d : \|\mathbb{E}[ZZ'\|\beta'Z = x] - (I_d + (x^2 - 1)\beta\beta')\| > \varepsilon\} \overset{d \to \infty}{\longrightarrow} 0$$

for each $x \in \mathbb{R}$ and for each $\varepsilon > 0$ under the assumptions of Theorem 2.1 in Section 2.

So far, we have seen for *fixed* $x \in \mathbb{R}$ and for most β's that (1.1) and (1.2) are approximately satisfied, in the sense that (1.3) and (1.4) hold under some conditions. Our main result is that (1.1) and (1.2) are approximately satisfied for most β's and for most x's: Under the assumptions of Theorem 2.1, there are Borel subsets B_d of $\mathbb{R}^d$ satisfying $\upsilon(B_d) \overset{d \to \infty}{\longrightarrow} 1$ so that

$$(1.5) \qquad \sup_{\beta \in B_d} \mathbb{P}\big(\|\mathbb{E}[Z\|\beta'Z]\|^2 - (\beta'Z)^2 > \varepsilon\big) \overset{d \to \infty}{\longrightarrow} 0 \quad \text{and}$$

$$(1.6) \quad \sup_{\beta \in B_d} \mathbb{P}\big(\|\mathbb{E}[ZZ'\|\beta'Z] - (I_d + ((\beta'Z)^2 - 1)\beta\beta')\| > \varepsilon\big) \overset{d \to \infty}{\longrightarrow} 0$$

hold for each $\varepsilon > 0$.

Following a referee's suggestion, we now compare our findings to the work of Diaconis and Freedman [5], which is an important precursor to the results of [11] and hence, a fortiori, also to the results in this paper; see also the discussion surrounding the displays (1.7)–(1.8) in [11]. (Moreover, the recent work of Dümbgen and Zerial [8] should be mentioned here, where several extensions and generalizations of the results of [5] are provided.) Under the assumptions of Theorem 2.1, Proposition 5.2 of [5] entails, for large d, that the (bivariate) joint distribution of $\alpha'Z$ and $\beta'Z$ is approximately normal, with zero means, unit variances, and covariance $\alpha'\beta$, for most pairs of unit-vectors α and β (in the sense of weak convergence in probability with respect to the product measure $\upsilon \otimes \upsilon$ as $d \to \infty$). Because

---

[2] This easily follows from the fact that $\mathbb{E}[Z\|\beta'Z]$ (resp., $\beta\beta'Z$) is the orthogonal projection of Z into the space of all measurable (resp., linear) functions of $\beta'Z$ in $L_2(\mathbb{P})$, and from the relation between the unconditional and the conditional variance.

the normal distribution has linear conditional means and constant conditional variances, this suggests, but does not prove, that

$$(1.7) \qquad \mathbb{P}\big(\big|\mathbb{E}[\alpha' Z \| \beta' Z] - \alpha'\beta\beta' Z\big|^2 > \varepsilon\big) \overset{d\to\infty}{\longrightarrow} 0$$

for each $\varepsilon > 0$ and for most pairs of unit-vectors α and β in $\mathbb{R}^d$ (in the sense of convergence in probability as a function of α and β with respect to $\upsilon \otimes \upsilon$). If α is treated as an unknown parameter, and if the observations $\alpha' Z$ and $\beta' Z$ are treated as response and explanatory variable, respectively, then (1.7) entails, for *most α's and β's*, that the response can be approximated by a linear function of the explanatory variable plus an error term with zero mean conditional on $\beta' Z$, provided that d is large. The approximating linear function is $\alpha'\beta(\beta' Z)$. But for large d, we also have $\alpha'\beta \approx 0$ for most α's and β's (with respect to $\upsilon \otimes \upsilon$). The statement in (1.5), on the other hand, is equivalent to

$$(1.8) \qquad \mathbb{P}\Big(\sup_\alpha \big|\mathbb{E}[\alpha' Z \| \beta' Z] - \alpha'\beta\beta' Z\big|^2 > \varepsilon\Big) \overset{d\to\infty}{\longrightarrow} 0$$

for each $\varepsilon > 0$ (in probability as a function of β with respect to υ); cf. the discussion following (1.2). The statement in (1.8) is obviously much stronger than that in (1.7). And it guarantees that the conditional mean of $\alpha' Z$ is approximately linear in $\beta' Z$, *for all α's and for most β's*; this includes, in particular, the statistically interesting case where α is parallel, or close to parallel, to β. Finally, as already observed in [11], "Diaconis and Freedman's result does not provide clues as to whether [statements like (1.8)] might be true or false." Similar observations also apply to conditional variances, mutatis mutandis.

1.3. *Discussion.* If the left-hand sides of (1.5) and (1.6) are both small, and if $\beta \in B_d$, then the simple linear model, where the response $\alpha' Z$ is explained by a linear function of the explanatory variable $\beta' Z$ plus an error that has zero mean and constant variance given $\beta' Z$, is approximately correct, irrespective of the unit-vector α. Here, "approximately correct" means that the expressions on the left-hand sides of (1.1) and (1.2) are at most ε for a range of values x that contains the explanatory variable $\beta' Z$ with high probability. Under the conditions of Theorem 2.1, a sufficiently large dimension is enough to guarantee that B_d is large and that the left-hand sides of (1.5) and (1.6) are small.[3]

The statistical impact of our results is most pronounced in situations where the sample size is small and the dimension is large. Assume that Theorem 2.1 applies, and consider a collection of n independent copies of the pair $(\alpha' Z, \beta' Z)$ that we

---

[3]Note that this disentangles the issue of (approximate) model validity and the issue of model performance: The model is approximately valid if $\beta \in B_d$, irrespective of α; the performance of this model, on the other hand, that is, the performance of $\beta' Z$ as a predictor for $\alpha' Z$, depends on both β and α. Under classical parametric or nonparametric assumptions, a simple model that is (approximately) correct typically also performs well.

denote by $(\alpha' Z_i, \beta' Z_i)$, $i = 1, \ldots, n$, with $\beta \in B_d$. If d is large and n is comparatively small, so that the left-hand sides of both (1.5) and (1.6) are still small even when multiplied by n, then the simple linear model discussed in the preceding paragraph can also be used to approximately describe the relation between $\alpha' Z_i$ and $\beta' Z_i$ for each $i = 1, \ldots, n$, irrespective of α.

We stress that additional data may give reason to dismiss the simple linear model considered in the preceding paragraphs in favor of a more complex one, because the error suffered from using a model that is only approximately correct will typically become apparent if n increases to a value that is no longer sufficiently small relative to d. This is in line with R. A. Fisher's 1922 observation that "more or less elaborate forms [of models] will be suitable according to the volume of data;" cf. [10]. And we stress that our results cannot guarantee that a given simple model, like that discussed in the preceding paragraphs, is correct. But we can guarantee, under the assumptions of Theorem 2.1, that most simple models are approximately correct, in the sense that $\upsilon(B_d)$ is large and in the sense that the left-hand sides of (1.5) and (1.6) are small, provided only that d is sufficiently large. This also underscores the need for critical examination of the data and of the model fit, irrespective of whether or not d is large. To this end, a very useful diagnostic tool is introduced by Li in [18], namely a method to estimate, for a given unit-vector β, that unit-vector α for which the conditional mean of $\alpha' Z$ given $\beta' Z$ is most nonlinear in $\beta' Z$; see also Section 6.1 in that reference.

The results obtained in this paper do not suggest that one should abandon the search for complex and potentially nonlinear relations in the data. But after such complex and/or nonlinear relations have been accounted for, or in the case where none such can be found, our results show how the use of simple linear models can be justified without imposing strong regularity conditions on the true data-generating process.

The discussion so far prompts for two extensions of our results that are beyond the scope of this paper. The first one is to extend our findings to the case of more than one explanatory variable, that is, the case where the conditioning is not on $\beta' Z$ but on $(\beta'_1 Z, \beta'_2 Z, \ldots, \beta'_p Z)$ for a collection of p mutually orthogonal unit-vectors $\beta_1, \ldots, \beta_p$. In fact, Hall and Li [11] sketch an extension of their results to that situation, so that an appropriate generalization of (1.3) holds. We will consider a corresponding generalization of (1.4) and also of the main results in (1.5) and (1.6) elsewhere. The second extension is to provide explicit upper bounds for the expressions on the left-hand sides of (1.5) and (1.6) that converge to zero as $d \to \infty$ at a fast rate; and to provide an explicit lower bound for $\upsilon(B_d)$ that converges to one as $d \to \infty$ also at a fast rate.

1.4. *Some consequences*.

1.4.1. *SIR, SAVE and related methods*.   Many modern dimension reduction methods, like those based on inverse conditional moments, rely on conditions

like (i) and (ii) in Section 1.2. In particular, first-moment-based methods like Sliced Inverse Regression [16] are based on a linear conditional mean requirement as in (i). (Besides, this requirement is also used in several important results on generalized least squares under possible link misspecification; see, for example, [19] and the references cited therein.) And second-moment-based techniques like the Sliced Average Variance Estimator [4], Principal Hessian Directions [17], or Directional Regression [15], are based on both a linear conditional mean requirement as in (i), and on a constant conditional variance requirement as in (ii). Both conditions (i) and (ii) are also used in recent works such as [2, 3, 6, 14].

Given observations from a potentially rather complex data-generating process, the dimension-reduction methods mentioned in the preceding paragraph aim at finding a simpler model that also describes the data. To justify the dimension reduction, these methods make assumptions to the effect that requirements like (i) or (ii) are satisfied, for *one particular projection*, namely for the projection on the so-called central subspace. Under such assumptions, the central subspace or, equivalently, the projection onto it, can be recovered from the data with good accuracy. But as outlined in the Introduction, verifying such assumptions in practice can be hard, particularly in situations where the sample size is comparatively small.

Our results provide an alternative justification for requirements like (i) and (ii). In particular, in the setting of Theorem 2.1, we see that both (i) and (ii) are approximately satisfied for *most projections* $\beta'Z$ in the sense of (1.5) and (1.6), provided only that the underlying dimension is large. For the linear conditional mean condition, we stress that the relation (1.3) has been derived much earlier in [11].

1.4.2. *Sparse linear modeling.*   Consider the linear model with univariate response y and a d-vector of explanatory variables w, that is,

$$(1.9) \qquad\qquad y = \theta'w + \varepsilon,$$

where $\theta \in \mathbb{R}^d$ is unknown, and where the error ε has zero mean and constant variance conditional on w. We also assume that y and w are square integrable and centered so that $\mathbb{E}w = 0$. The leading case we have in mind is a situation where d, that is, the number of available regressors, is as large as, or even much larger than, the sample size. To deal with such situations, it is common to assume that the true model (1.9) is equal to, or can be closely approximated by, a "sparse" submodel that uses only a few explanatory variables, and to use the available data to select and fit a sparse submodel. Such sparsity assumptions are clearly restrictive. In the following, we argue that the results in this paper provide weaker, and hence less restrictive, assumptions that also justify the fitting of sparse submodels.

For illustration, consider now an extremely sparse model where y is regressed on just one explanatory variable, say, w_1, that is,

$$(1.10) \qquad\qquad y = cw_1 + e,$$

where $c \in \mathbb{R}$ is unknown, and where e has zero mean and constant variance given w_1. To consider various possible justifications of the sparse submodel (1.10), we first rewrite the overall model (1.9) as

$$y = (\theta_1 w_1 + \mathbb{E}[\theta'_{\neg 1} w_{\neg 1} \| w_1]) + (\theta'_{\neg 1} w_{\neg 1} - \mathbb{E}[\theta'_{\neg 1} w_{\neg 1} \| w_1] + \varepsilon),$$

where $\theta_{\neg 1}$ and $w_{\neg 1}$ are obtained from θ and w, respectively, by deleting the first component.

One possibility to justify the sparse model (1.10) is to impose the extreme sparsity assumption that all coefficients of $\theta_{\neg 1}$ are zero, so that $\theta'_{\neg 1} w_{\neg 1} = 0$. Then the relation in the preceding display obviously reduces to $y = \theta_1 w_1 + \varepsilon$, and (1.10) applies with $c = \theta_1$ and $e = \varepsilon$. Under this extreme sparsity assumption we obtain, in particular, that the sparse model (1.10) is equivalent to the overall model (1.9) in terms of prediction, because $\mathbb{E}[y \| w_1] = \mathbb{E}[y \| w]$. A slightly relaxed sparsity condition is to assume that the coefficients of $\theta_{\neg 1}$ are possibly nonzero but otherwise negligible, that is, $\theta'_{\neg 1} w_{\neg 1} \approx 0$, so that the sparse model (1.10) is approximately valid with $c \approx \theta_1$ and $e \approx \varepsilon$.

An alternative justification of (1.10) is to impose the assumption that, given w_1, the conditional mean of $\theta'_{\neg 1} w_{\neg 1}$ is linear in w_1 and the conditional variance of $\theta'_{\neg 1} w_{\neg 1}$ is constant in w_1. In that case, the relation in the preceding display also reduces to (1.10), but now with $c = \text{Cov}[\theta' w, w_1]/\text{Var}[w_1] = \theta_1 + \sum_{i=2}^{d} \theta_i \text{Cov}[w_1, w_i]/\text{Var}[w_1]$, and $e = \theta'_{\neg 1} w_{\neg 1} - \mathbb{E}[\theta'_{\neg 1} w_{\neg 1} \| w_1] + \varepsilon$. Under this alternative assumption, the model (1.10) is valid but typically less accurate in terms of prediction than the overall model (1.9), because, typically, $\mathbb{E}[y \| w_1] \neq \mathbb{E}[y \| w]$ and hence $\text{Var}[y \| w_1] > \text{Var}[y \| w]$. As before, these assumptions can be relaxed by requiring that the conditional mean is approximately linear and the conditional variance is approximately constant.

In the preceding two paragraphs, we have considered two types of justifications for fitting the submodel (1.10). Type (a): Exact or approximate sparsity assumptions. And type (b): Exact or approximate linear conditional mean and constant conditional variance assumptions. In practice, verifying either of these assumptions for a given submodel can be difficult. This raises the question as to which set of assumptions, that is, (a) or (b), is more restrictive. To this end, we first note that (a) obviously implies (b). For the more detailed comparison that we give in the following, we assume that the law of w is nondegenerate so that w can be written as $w = MZ$ for a d-vector Z satisfying $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. The $d \times d$ matrix M is a square root of the variance/covariance matrix of w, which is nondegenerate by assumption; which can be assumed to be symmetric; and which need not be known in practice. Then $\theta'_{\neg 1} w_{\neg 1}$ and w_1 can be written as $\theta'_{\neg 1} w_{\neg 1} = a'_\circ Z$ and $w_1 = b'_\circ Z$ with $a'_\circ = (0, \theta_2, \ldots, \theta_d)M$ and $b'_\circ = (1, 0, \ldots, 0)M$.

The type (a) condition that $\theta_{\neg 1} = 0$ entails that $a_\circ$ is equal to zero; the collection of θ's that satisfy this condition is the 1-dimensional subset of the parameter space $\mathbb{R}^d$ that is spanned by $M^{-1} b_\circ = (1, 0, \ldots, 0)'$ (this collection depends

on $b_\circ$). More generally, for y as in (1.9) and for each vector b, the simple model with response y and with explanatory variable $b'Z$, that is, $y = c(b'Z) + e$, can be justified by the type (a) condition that θ is parallel to $M^{-1}b$. And, for each vector b, the approximate type (a) condition, that θ is approximately parallel to $M^{-1}b$, is satisfied if θ belongs to an appropriately small neighborhood of the span of $M^{-1}b$.

To study type (b) conditions on the conditional moments of $\theta'_{\neg 1}w_{\neg 1}$ given w_1 or, equivalently, on the conditional moments of $a'_\circ Z$ given $b'_\circ Z$, we may replace the vectors $a_\circ$ and $b_\circ$ by standardized versions $\alpha_\circ$ and $\beta_\circ$ that have length one, for example, $\alpha_\circ = a_\circ/\|a_\circ\|$. (Indeed, the conditional mean of $a'_\circ Z$ given $b'_\circ Z$ is linear, or approximately linear, in $b'_\circ Z$, if and only if the same is true for the conditional mean of $\alpha'_\circ Z$ given $\beta'_\circ Z$; and a similar statement applies for the conditional variances, mutatis mutandis.) If Theorem 2.1 applies and if d is large, then *for most β's and uniformly in α*, the conditional mean of $\alpha'Z$ given $\beta'Z$ is approximately linear and the conditional variance of $\alpha'Z$ given $\beta'Z$ is approximately constant, in the sense of (1.5) and (1.6). In terms of the original parameter θ, we note that uniformity in α corresponds to uniformity in $\theta \in \mathbb{R}^d \setminus \{0\}$.

2. Main result and outline of proof. Our main result is that (1.3)–(1.4) and also (1.5)–(1.6) hold, for sets B_d with $\lim_{d\to\infty} \upsilon(B_d) = 1$. We will establish this under the basic condition that Z has a Lebesgue density and is standardized so that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$ for each d. For the method of proof that we employ, we also rely on two technical conditions, which follow.

CONDITION (t1). *For fixed $k \in \mathbb{N}$ and for each d, set $S_k = (Z'_i Z_j/d)^k_{i,j=1}$, where the Z_i's are i.i.d. copies of Z.*

(a) *We have $\mathbb{E}[(\sqrt{d}\|S_k - I_k\|)^{2k+1}] = O(1)$ as $d \to \infty$. Moreover, let H be a monomial in the elements of $S_k - I_k$ of degree $h \leq 2k$. If H has a linear factor, then $d^{h/2}\mathbb{E}H = o(1)$. And if H consists only of quadratic factors in the elements of $S_k - I_k$ above the diagonal, then $d^{h/2}\mathbb{E}H = 1 + o(1)$.*

(b) *Consider two monomials G and H of degree g and h, respectively, in the elements of $S_k - I_k$. If G is given by $Z'_1 Z_2 Z'_2 Z_3 \cdots Z'_{g-1} Z_g Z'_g Z_1/d^g$, if H depends at least on those Z_i's with $i \leq g$, and if $2 \leq h < g \leq k$, then $d^g \mathbb{E}GH = o(1)$.*

CONDITION (t2). *For fixed $k \in \mathbb{N}$, for each d, and for any orthogonal $d \times d$ matrix R, the marginal density of the last $d - k + 1$ components of RZ is bounded by $\binom{d}{k-1}^{1/2} B^{d-k+1}$ for some constant B that does not depend on d or R.*

Conditions (t1) and (t2) are always satisfied, for any fixed k, if the components of Z are independent, with bounded marginal densities and bounded marginal moments of sufficiently high order; cf. Example A.1 in the supplementary material [13]. Also, if $\mathbb{E}[(\sqrt{d}\|S_k - I_k\|)^{2k+1}] = O(1)$, then condition (t1)(a) is satisfied

if the elements of $\sqrt{d}(S_k - I_k)$ jointly converge to a Gaussian; cf. Example A.2 in the supplementary material [13]. However, these conditions are more general than that and allow, in particular, for situations where the components of Z are dependent and/or where the elements of $\sqrt{d}(S_k - I_k)$ do not converge in distribution. Also note that both conditions are orthogonally invariant: If Z satisfies any one of them, then the same is true for any orthogonal transformation of Z.

The first requirement of condition (t1)(a) entails that $d^{h/2}\mathbb{E}H = O(1)$ for any monomial H in the elements of $S_k - I_k$ of degree up to $2k + 1$. Condition (t1)(b) strengthens parts of condition (t1)(a) in the following sense: Consider monomials G and H as in condition (t1)(b). If condition (t1)(a) is satisfied, then $\mathbb{E}GH = o(d^{-(g+h)/2})$ (because GH is a monomial of degree $g + h$ that has a linear factor). Condition (t1)(b) requires that $\mathbb{E}GH$ converges to zero at the faster rate $o(d^{-g})$. Condition (t2) ensures that the distribution of Z is not "too concentrated" in certain directions, and is used together with (t1)(a) to guarantee uniform integrability of $d/Z'Z$ and related quantities (see Proposition E.1 in the supplementary material [13]). Also, our conditions should be compared to those used in [11].[4]

We can now state the main result of this paper.

THEOREM 2.1. *For each d, consider a random d-vector Z that has a Lebesgue density and that is standardized such that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$. If conditions (t1)(a) and (t2) are satisfied with $k = 2$, then there are Borel sets $B_d \subseteq \mathbb{R}^d$ satisfying $\lim_{d\to\infty} \upsilon(B_d) = 1$, such that (1.5) holds for each $\varepsilon > 0$. If condition (t1) and condition (t2) are satisfied with $k = 4$, then the sets B_d can be chosen so that also (1.6) holds for each $\varepsilon > 0$. [Moreover, for each $x \in \mathbb{R}$ and each $\varepsilon > 0$, the relation (1.3) obtains under conditions (t1)(a) and (t2) with $k = 2$, and the relation (1.4) holds under conditions (t1) and (t2) with $k = 4$.]*

In the remainder of this section, we give an outline of the proof of Theorem 2.1. The proof is comprised of five main steps corresponding to the five propositions that follow.

As the first step, it will be convenient to replace the usual reference measure on $\mathbb{R}^d$, that is, Lebesgue measure, by the d-variate standard Gaussian measure, that is, $N(0, I_d)$. The effect of this change of measure on conditional densities and on conditional expectations involving Z is described by the next result.

PROPOSITION 2.2. *Fix $d \geq 1$, and consider a random d-vector Z with Lebesgue density f. Let $V \sim N(0, I_d)$, and write $\phi(\cdot)$ for the Lebesgue density of V. Moreover, for a fixed unit-vector $\beta \in \mathbb{R}^d$ and for each $x \in \mathbb{R}$, set*

---

[4] The results in [11] are stated under high-level assumptions that are less specific but harder to verify; see, for example, conditions (3.21), (3.28) and (3.29) of Theorem 3.2 in that paper. The only specific example that is actually shown in [11] to satisfy all three of these high-level conditions is the normal distribution; cf. Example 4.2 and Remark 4.2 in that paper.

$W^{x|\beta} = x\beta + (I_d - \beta\beta')V$. *Then the function $h(\cdot|\beta)$ defined by*

$$h(x|\beta) = \mathbb{E}\left[\frac{f(W^{x|\beta})}{\phi(W^{x|\beta})}\right]$$

for $x \in \mathbb{R}$ is a density of $\beta'Z$ with respect to the univariate standard Gaussian measure [i.e., $h(x|\beta)\phi_1(x)$ is a Lebesgue density of $\beta'Z$ if ϕ_1 denotes the $N(0,1)$-density]. Moreover, if $\Psi : \mathbb{R}^d \to \mathbb{R}$ is such that $\Psi(Z)$ is integrable, then a conditional expectation $\mathbb{E}[\Psi(Z)\|\beta'Z = x]$ of $\Psi(Z)$ given $\beta'Z$ satisfies

$$\mathbb{E}[\Psi(Z)\|\beta'Z = x]h(x|\beta) = \mathbb{E}\left[\Psi(W^{x|\beta})\frac{f(W^{x|\beta})}{\phi(W^{x|\beta})}\right]$$

whenever $x \in \mathbb{R}$ is such that $h(x|\beta) < \infty$.

This result allows us, for fixed $x \in \mathbb{R}$, to study the marginal density of $\beta'Z$ at x as well as conditional expectations involving Z given $\beta'Z = x$, by considering unconditional means involving the random variable $W^{x|\beta}$, which has a $N(x\beta, I_d - \beta\beta')$-distribution.

Now, in order to derive (1.3), we follow [11] and use the following argument (which can be traced back to Hoeffding [12]; see also [8]): Set $\mu_{x|\beta} = \mathbb{E}[Z\|\beta'Z = x]$, and let b be a random vector in $\mathbb{R}^d$ with distribution υ, that is, such that b is uniformly distributed on the unit-sphere, that is, on the set of unit-vectors in $\mathbb{R}^d$. Then (1.3) is equivalent to the statement that $\|\mu_{x|b}\|^2 - x^2$ converges to zero in probability as a function of b, and this will follow if

$$\mathbb{E}[(\|\mu_{x|b}\|^2 - x^2)h^2(x|b)] \quad \text{and} \quad \mathbb{E}[(h(x|b) - 1)^2]$$

both converge to zero as $d \to \infty$, where the expectations are taken with respect to b. [Note that $\mu_{x|b}$ and $h(x|b)$ are measurable in view of Corollary B.2. Also note that both integrands in the preceding display are nonnegative.] We now compute $\mathbb{E}[(h(x|b) - 1)^2]$ as

(2.1)
$$\int_\beta \left(\mathbb{E}\left[\frac{f(W^{x|\beta})}{\phi(W^{x|\beta})}\right]\right)^2 - 2\mathbb{E}\left[\frac{f(W^{x|\beta})}{\phi(W^{x|\beta})}\right] + 1\upsilon(d\beta)$$

$$= \mathbb{E}\left[\frac{f(W_1)}{\phi(W_1)}\frac{f(W_2)}{\phi(W_2)}\right] - 2\mathbb{E}\left[\frac{f(W_1)}{\phi(W_1)}\right] + 1,$$

where the W_i's are defined as $W_i = xb + (I_d - bb')V_i$, $i = 1, 2$, with V_1 and V_2 i.i.d. $N(0, I_d)$ and independent of b. Note that W_1 and W_2 are dependent because both share the same random unit-vector b. If (2.1) converges to zero or, equivalently, if $h(x|b) \to 1$ in $L^2(b)$ as $d \to \infty$, then $h(x|b) < \infty$ with probability one for sufficiently large d. For such d, we see that the second statement of Proposition 2.2 applies for υ-almost all β, such that $\mathbb{E}[(\|\mu_{x|b}\|^2 - x^2)h^2(x|b)]$ can be written as

(2.2)
$$\mathbb{E}\left[(W_1'W_2 - x^2)\frac{f(W_1)}{\phi(W_1)}\frac{f(W_2)}{\phi(W_2)}\right],$$

by arguing as in the derivation of (2.1), as is easy to see. With this, we see that (1.3) holds if both (2.1) and (2.2) go to zero as $d \to \infty$. And if (2.1) and (2.2) converge to zero uniformly in x over compact subsets of $\mathbb{R}$, that is, if the suprema of (2.1) and (2.2) over x satisfying $|x| \leq M$ converge to zero as $d \to \infty$ for each $M > 0$, then it is not difficult to also derive (1.5) by employing standard arguments; for details, see Lemma B.4(i) in the supplementary material [13].

To establish (1.4), we employ a similar strategy and write $\Delta_{x|\beta}$ as shorthand for the $d \times d$ matrix $\Delta_{x|\beta} = \mathbb{E}[ZZ' \| \beta' Z = x] - (I_d + (x^2 - 1)\beta\beta')$. With b again uniform on the unit-sphere, the goal is to show convergence of the largest singular value $\|\Delta_{x|b}\|$ to zero in probability with respect to b. But this follows if

$$\text{trace } \Delta_{x|b}^k \xrightarrow{p} 0$$

as $d \to \infty$ for some even integer k (where, again, Corollary B.2 guarantees measurability). This, and hence also (1.4), will follow if

$$\mathbb{E}\big[\text{trace } \Delta_{x|b}^k h^k(x|b)\big]$$

converges to zero as $d \to \infty$ for some even integer k and if, in addition, also (2.1) converges to zero. (We shall find, at the end of the section, that the expression in the preceding display converges to zero for $k = 4$ but typically not for $k = 2$.) To analyze the expectation in the preceding display, define the function $\Delta_{x|\beta}(z)$ as $\Delta_{x|\beta}(z) = zz' - (I_d + (x^2 - 1)\beta\beta')$ for $z \in \mathbb{R}^d$. We now argue as in the preceding paragraph: Assume that (2.1) converges to zero, and assume that d is sufficiently large so that $h(x|\beta) < \infty$ for υ-almost all β. For such d, use Proposition 2.2 to see that $\mathbb{E}[\text{trace } \Delta_{x|b}^k h^k(x|b)]$ equals

$$\mathbb{E}\bigg[\text{trace } \Delta_{x|b}(W_1) \cdots \Delta_{x|b}(W_k) \frac{f(W_1)}{\phi(W_1)} \cdots \frac{f(W_k)}{\phi(W_k)}\bigg],$$

where, similarly to before, $W_i = xb + (I_d - bb')V_i$ with the V_i, $1 \leq i \leq k$, i.i.d. $N(0, I_d)$ independent of b. Instead of computing the trace in the preceding display directly, we find it convenient to break it into smaller, more manageable, pieces. Indeed, we find that the expression in the preceding display can be written as the weighted sum of the terms

$$\sum_{j=1}^{k} \binom{k}{j} (-1)^j \mathbb{E}\bigg[\big(W_1' W_2 \cdots W_j' W_1 - d + 1 - x^{2j}\big) \frac{f(W_1)}{\phi(W_1)} \cdots \frac{f(W_k)}{\phi(W_k)}\bigg],$$

(2.3)

$$\mathbb{E}\bigg[\bigg(\prod_{i=1}^{m} W_{j_{i-1}+1}' W_{j_{i-1}+2} \cdots W_{j_i-1}' W_{j_i} - x^{2(j_m - m)}\bigg) \frac{f(W_1)}{\phi(W_1)} \cdots \frac{f(W_k)}{\phi(W_k)}\bigg]$$

for $m \geq 1$ and indices $j_0, \ldots, j_m$ satisfying $j_0 = 0$, $j_m < k$, and $j_{i-1} + 1 < j_i$ whenever $0 < i \leq m$. In addition, we find that the weights in this weighted sum depend only on k and on x, and that the weights are continuous in $x \in \mathbb{R}$; cf.

Lemma B.3 for details. [Note that, in (2.3), we write $W_1' W_2 \cdots W_j' W_1$ as shorthand for trace $\prod_{i=1}^{j} W_i W_i'$, and we also use notation like $W_1' W_2 \cdots W_{j-1}' W_j$ as shorthand for $\prod_{i=1}^{j-1} W_i' W_{i+1}$.] Hence, (1.4) holds if the expression in (2.1) and those in (2.3) go to zero as $d \to \infty$, the latter for some even integer k, and for any m and $j_0, \ldots, j_m$ as indicated. Moreover, (1.6) holds provided that the expressions in (2.1) and (2.3) all converge to zero uniformly in x over compact subsets of $\mathbb{R}$; details are given in Lemma B.4(ii) in the supplementary material [13].

To understand the large-d-behavior of the quantities in (2.1), (2.2), and (2.3), we need to understand the joint distribution of the W_j's, which is described by the next result.

PROPOSITION 2.3. *For d and k satisfying $1 \leq k < d$, the joint distribution of $W_1, \ldots, W_k$ has a density with respect to Lebesgue measure that we denote by $\varphi_x(w_1, \ldots, w_k)$, and this density satisfies*

$$\frac{\varphi_x(w_1, \ldots, w_k)}{\phi(w_1) \cdots \phi(w_k)} = \frac{(d/2)^{-k/2} \Gamma(d/2)}{\Gamma((d-k)/2)} \det S_k^{-1/2} \left(1 - \frac{x^2}{d} \iota' S_k^{-1} \iota\right)^{(d-k-2)/2} e^{(k/2)x^2}$$

if S_k is invertible with $x^2 \iota' S_k^{-1} \iota < d$, and $\varphi_x(w_1, \ldots, w_k) = 0$ otherwise, where $S_k = (w_i' w_j / d)_{i,j=1}^{k}$ denotes the $k \times k$ matrix of scaled inner products of the w_i's, and $\iota = (1, \ldots, 1)'$ denotes an appropriate vector of ones.

Using Proposition 2.3, we can rewrite the quantities of interest in (2.1), (2.2), and (2.3) as follows: For example, (2.2) equals

$$\iint (w_1' w_2 - x^2) \frac{f(w_1)}{\phi(w_1)} \frac{f(w_2)}{\phi(w_2)} \varphi_x(w_1, w_2) \, dw_1 \, dw_2$$

$$(2.4) \qquad = \iint (w_1' w_2 - x^2) \frac{\varphi_x(w_1, w_2)}{\phi(w_1) \phi(w_2)} f(w_1) f(w_2) \, dw_1 \, dw_2$$

$$= \mathbb{E}\left[(Z_1' Z_2 - x^2) \frac{\varphi_x(Z_1, Z_2)}{\phi(Z_1) \phi(Z_2)}\right],$$

where Z_1 and Z_2 are i.i.d. copies of Z. In a similar fashion, the quantities in (2.3) can be written as

$$\sum_{j=1}^{k} \binom{k}{j} (-1)^j \mathbb{E}\left[(Z_1' Z_2 \cdots Z_j' Z_1 - d + 1 - x^{2j}) \frac{\varphi_x(Z_1, \ldots, Z_k)}{\phi(Z_1) \cdots \phi(Z_k)}\right],$$

$$(2.5)$$

$$\mathbb{E}\left[\left(\prod_{i=1}^{m} Z_{j_{i-1}+1}' Z_{j_{i-1}+2} \cdots Z_{j_i-1}' Z_{j_i} - x^{2(j_m-m)}\right) \frac{\varphi_x(Z_1, \ldots, Z_k)}{\phi(Z_1) \cdots \phi(Z_k)}\right]$$

for Z_i, $i = 1, \ldots, k$, i.i.d. as Z. And finally (2.1) reduces to

$$(2.6) \qquad \mathbb{E}\left[\frac{\varphi_x(Z_1, Z_2)}{\phi(Z_1) \phi(Z_2)}\right] - 2\mathbb{E}\left[\frac{\varphi_x(Z_1)}{\phi(Z_1)}\right] + 1.$$

Recall that our goal is to show that the expressions in (2.4)–(2.6) converge to zero as $d \to \infty$, uniformly in x over compact subsets of $\mathbb{R}$. We show, in fact, a slightly stronger statement, motivated by the obvious conjecture that the expected value of density ratios in (2.4)–(2.6), like $\varphi_x(Z_1, Z_2)/(\phi(Z_1)\phi(Z_2))$, converges to one, and also by the observation that the expression in (2.4) is a special case of the second expression in (2.5) with k replaced by 2. For an even integer k and for each $l = 1, \ldots, k$, for each $m \geq 0$ and for any indices $j_0, \ldots, j_m$ that satisfy $j_0 = 0$, $j_m \leq l$, and $j_{i-1} + 1 < j_i$ whenever $0 < i \leq m$, consider the expressions

$$(2.7) \quad \mathbb{E}\left[\left(\prod_{i=1}^m Z'_{j_{i-1}+1} Z_{j_{i-1}+2} \cdots Z'_{j_i-1} Z_{j_i}\right) \frac{\varphi_x(Z_1, \ldots, Z_l)}{\phi(Z_1) \cdots \phi(Z_l)}\right] - x^{2(j_m - m)}$$

and also the expression

$$(2.8) \quad \sum_{j=1}^k \binom{k}{j} (-1)^j \mathbb{E}\left[(Z'_1 Z_2 \cdots Z'_j Z_1 - d) \frac{\varphi_x(Z_1, \ldots, Z_k)}{\phi(Z_1) \cdots \phi(Z_k)}\right] - (1 - x^2)^k.$$

Convergence to zero of the expressions in (2.7) corresponding to $k = 2$ (uniformly over compacts in x) entails convergence to zero of (2.4) and (2.6) (uniformly over compacts in x). And if both (2.7) and (2.8) converge to zero for some even integer k (uniformly over compacts), then the expressions in (2.5) corresponding to that same k also converge to zero (uniformly over compacts). [Convergence to zero of (2.4) follows from convergence to zero of (2.7) with $m = 1$, $j_m = 2$ and $l = 2$ together with convergence to zero of (2.7) with $m = 0$ and $l = 2$. Convergence to zero of (2.6) follows from convergence to zero of (2.7) with $m = 0$ and $l = 1$ together with convergence to zero of (2.7) with $m = 0$ and $l = 2$. For the first expression in (2.5), convergence to zero follows from convergence to zero of (2.7) with $m = 0$ and $l = k$ and from convergence to zero of (2.8), in view of the binomial theorem. Similarly, for the second expression in (2.5), convergence to zero follows from convergence to zero of (2.7) and of convergence to zero of the special case of (2.7) where $m = 0$ and $l = k$.]

The expressions in (2.7)–(2.8) both involve expected values of a polynomial in $Z'_i Z_j$ for some pairs (i, j), multiplied by $\varphi_x(Z_1, \ldots, Z_l)/(\phi(Z_1) \cdots \phi(Z_l))$, $l = 1, \ldots, k$. To proceed, we need the polynomial approximation to $\varphi_x(Z_1, \ldots, Z_l)/(\phi(Z_1) \cdots \phi(Z_l))$ that is provided by the next result.

PROPOSITION 2.4. *Fix $M > 0$ and x satisfying $|x| \leq M$. Moreover, consider integers k and d that satisfy $k \geq 1$ and $d > \max\{3k, 2(k+1)M^2\}$, and d-vectors $w_1, \ldots, w_k$ that are such that the $k \times k$ matrix $S_k = (w'_i w_j / d)_{i,j=1}^k$ satisfies $\|S_k - I_k\| < 1/(2k)$. Then $\varphi_x(w_1, \ldots, w_k)$ is such that*

$$\frac{\varphi_x(w_1, \ldots, w_k)}{\phi(w_1) \cdots \phi(w_k)} = \psi_x(S_k - I_k) + \Delta,$$

where $\psi_x(S_k - I_k)$ is a polynomial of degree up to k in the elements of $S_k - I_k$. The coefficients of the polynomial $\psi_x(\cdot)$ depend on k, x and d, and are bounded,

in absolute value and uniformly in $x \in [-M, M]$, by a constant C, and Δ satisfies $|\Delta| < D\|S_k - I_k\|^{k+1}$, for some constants $C = C(k, M)$ and $D = D(k, M)$ that depend only on k and on M. Moreover, both $\psi_x(S_k - I_k)$ and Δ are invariant under permutations of the w_i's so that, in particular, $\psi_x(S_k - I_k)$ is unchanged when S_k is replaced by the matrix $(w'_{\pi(i)} w_{\pi(j)}/d)^k_{i, j=1}$ for any permutation π of k elements. [*The coefficients of $\psi_x(\cdot)$ and the bounds C and D can be obtained explicitly upon inspection of the proof.*]

When studying the expected values in (2.7)–(2.8), Proposition 2.4 suggests that the density ratio $\varphi_x(Z_1, \ldots, Z_k)/(\phi(Z_1)\cdots\phi(Z_k))$ can be approximated by the polynomial $\psi_x(S_k - I_k)$. The resulting approximations to (2.7) and (2.8) are

$$(2.9) \qquad \mathbb{E}\left[\left(\prod_{i=1}^m Z'_{j_{i-1}+1} Z_{j_{i-1}+2} \cdots Z'_{j_i-1} Z_{j_i}\right)\psi_x(S_l - I_l)\right] - x^{2(j_m - m)}$$

and

$$(2.10) \quad \sum_{j=1}^k \binom{k}{j}(-1)^j \mathbb{E}\left[(Z'_1 Z_2 \cdots Z'_j Z_1 - d)\psi_x(S_k - I_k)\right] - (1 - x^2)^k,$$

respectively. For these approximations to be useful, we need to show that the difference of (2.7) and (2.9), and also the difference of (2.8) and (2.10), converges to zero as $d \to \infty$, uniformly in x over compact subsets of $\mathbb{R}$. The technical difficulty here is that expressions like, for example, $Z'_1 Z_2 \cdots Z'_j Z_1 - d$ in (2.10) have zero mean but do not converge to zero in probability. To deal with this, we rely on conditions (t1)(a) and (t2).

PROPOSITION 2.5. *For each d, consider a random d-vector Z that has a Lebesgue density, that is standardized such that $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ' = I_d$, and that satisfies conditions (t1)(a) and (t2) for some fixed integer k. Let $H(S_k - I_k)$ be a (fixed) monomial in the elements of $S_k - I_k$ whose degree, denoted by $\deg(H)$, satisfies $0 \le \deg(H) \le k$. Then*

$$\mathbb{E}\left[d^{(k+\deg(H))/2}|H(S_k - I_k)|\left|\frac{\varphi_x(Z_1, \ldots, Z_k)}{\phi(Z_1)\cdots\phi(Z_k)} - \psi_x(S_k - I_k)\right|\right]$$

converges to zero as $d \to \infty$, uniformly in x over compact subsets of $\mathbb{R}$.

If Proposition 2.5 applies, then it is not difficult to see that the difference between (2.7) and (2.9), and also the difference between (2.8) and (2.10), converges to zero, uniformly in x over compact subsets of $\mathbb{R}$, as required. For example, consider the difference of (2.8) and (2.10), which both involve k expected values indexed by $j = 1, \ldots, k$, and focus on the difference of those expected values corresponding to the index j. Also, recall that k is an even integer, so that $k > 1$. For $j = 1$, we simply use Proposition 2.5 with the monomial $(S_k - I_k)_{1,1}$, and

note that $|Z_1'Z_1 - d| = d|(S_k - I_k)_{1,1}| < d^{(k+1)/2}|(S_k - I_k)_{1,1}|$. For $j > 1$, we first write $|Z_1'Z_2 \cdots Z_j'Z_1|$ as

$$d^j|(S_k - I_k)_{1,2} \cdots (S_k - I_k)_{j,1}| \le d^{(j+k)/2}|(S_k - I_k)_{1,2} \cdots (S_k - I_k)_{j,1}|.$$

Now use Proposition 2.5 twice, first with the monomial $(S_k - I_k)_{1,2} \cdots (S_k - I_k)_{j,1}$ of degree $j \le k$, and then with the degree-zero monomial, and note that $d \le d^{(k+0)/2}$ here, to see that the difference of expected values corresponding to the index $j > 1$ also converges to zero, uniformly in x over compact subsets of $\mathbb{R}$. The difference of (2.7) and (2.9) is treated similarly.

To show that the expressions in (2.9) and (2.10) converge to zero, the following observation will be useful: If the Z_i's in (2.9) and (2.10) are replaced by independent standard normal vectors V_i (and if S_l and S_k are replaced by the corresponding Gram matrices of the V_i's), then the resulting expressions both converge to zero as $d \to \infty$, uniformly in x over compact subsets of $\mathbb{R}$. To establish convergence to zero of (2.9)–(2.10), uniformly on compacts in x, it therefore is sufficient to study the differences between the expressions in (2.9)–(2.10) and the same expressions with the Z_i's replaced by V_i's that are i.i.d. standard normal, and to show that these differences converge to zero as $d \to \infty$, uniformly in x over compacts subsets of $\mathbb{R}$. (To derive the last observation, we first note that (2.7) and (2.8), with the Z_i's replaced by the V_i's, are both equal to zero. Indeed, with this replacement, the expectation in (2.7) is equal to $\mathbb{E}[\prod_{i=1}^m W_{j_{i-1}+1}' W_{j_{i-1}+2} \cdots W_{j_i-1}' W_{j_i}]$, because $\phi(v_1) \cdots \phi(v_l)$ is the joint density of $V_1, \ldots, V_l$, and because $\varphi_x(w_1, \ldots, w_l)$ is the joint density of $W_1, \ldots, W_l$. Conditional on b, the W_i's are conditionally i.i.d., with $\mathbb{E}[W_i \| b] = xb$ and $\mathbb{E}[W_i W_i' \| b] = I_d + (x^2 - 1)bb'$. In view of this, it is elementary to verify that (2.7) with the Z_i's replaced by the V_i's is equal to zero. A similar argument applies, mutatis mutandis, to (2.8). Next, we note that Proposition 2.5 applies if Z is replaced by a standard normal vector V [that conditions (t1)(a) and (t2) hold when Z is replaced by V follows either from Example A.1 or upon a simple direct computation]. When replacing Z by V throughout, this entails that (2.9) and (2.10) converge to the same limit as (2.7) and (2.8), that is, to zero, uniformly over compacts in x.)

To put this idea to work, expand $\psi_x(S_k - I_k)$ into a weighted sum of monomials (in the elements of $S_k - I_k$), where the weight of each monomial in the sum is given by the coefficient of that monomial in $\psi_x(S_k - I_k)$; similarly, $\psi_x(S_l - I_l)$ can also be written as a weighted sum of such monomials for each $l \le k$. We see that the integrand in (2.9) for $m = 0$, that is, $\psi_x(S_l - I_l)$, can be written as the weighted sum of monomials in the elements of $S_k - I_k$. Similarly, for Z as in Theorem 2.1, the integrand in (2.9) for $m > 0$ can be written as the weighted sum of expressions of the form

$$(2.11) \qquad\qquad d^{\deg(G)}(G - \mathbb{E}[G])H$$

for two monomials G and H in the elements of $S_k - I_k$ of degree k or less, where G is given by the monomial

$$(2.12) \qquad \prod_{i=1}^{m} (S_k - I_k)_{j_{i-1}+1, j_{i-1}+2} \cdots (S_k - I_k)_{j_i-1, j_i}$$

of degree $j_m - m$, for some $m > 0$ and indices $j_0, \ldots, j_m$ that satisfy $j_0 = 0$, $j_m \leq k$, and $j_{i-1} + 1 < j_i$ whenever $0 < i \leq m$. In this weighted expansion of (2.9), note that the weight of (2.11) depends on x, on H (through its degrees) and also on d, in such a way that the weight is bounded in x and d as long as x is restricted to a compact set (cf. Proposition 2.4). Lastly, consider the integrand in (2.10), for Z as in Theorem 2.1. Arguing as before, we can write that integrand as the weighted sum of expression of the form (2.11), where here G is given by the monomial

$$(2.13) \qquad (S_k - I_k)_{1,2} \cdots (S_k - I_k)_{j-1, j} (S_k - I_k)_{j, 1}$$

of degree j for some j satisfying $1 \leq j \leq k$. And, again, in this weighted sum, the weight of each term depends on x and on H (through its degrees), and that weight is bounded in x and d over compacts in x. (Note that, under the assumptions of Theorem 2.1, it is elementary to verify that $\mathbb{E}[G] = 0$ whenever G is given by (2.12) and also whenever G is given by (2.13) with $j = 1$, and that $d^{\deg(G)} \mathbb{E}[G] = d$ if G is given by (2.13) with $j > 1$.)

PROPOSITION 2.6.   *For each $d \geq 1$, assume that Z is as in Theorem 2.1. Fix an integer $k \geq 1$, and let G and H be two (fixed) monomials in the elements of $S_k - I_k$ of degree k or less, define G^* and H^* as G and H, respectively, but with the $Z_1, \ldots, Z_k$ replaced by i.i.d. standard Gaussian d-vectors, and consider*

$$(2.14) \qquad \mathbb{E}\big[d^{\deg(G)}(G - \mathbb{E}[G])H\big] - \mathbb{E}\big[d^{\deg(G^*)}(G^* - \mathbb{E}[G^*])H^*\big].$$

(i) *Assume that condition* (t1)(a) *applies with the integer k as chosen here, and that $k \leq 4$. Then $\mathbb{E}[H] - \mathbb{E}[H^*]$ and also the expression in* (2.14) *converge to zero as $d \to \infty$ for each monomial G as in* (2.12).

(ii) *Assume that condition* (t1) *is satisfied with the integer k as chosen here. Let G be given by the monomial in* (2.13) *for some j, $1 \leq j \leq k$. Then the expression in* (2.14) *converges to zero as $d \to \infty$, unless either* (a) $H = (S_k - I_k)_{a,a}$ *for some a satisfying $1 \leq a \leq j$,* (b) $H = (S_k - I_k)_{a,b}$ *with $1 \leq a < b \leq j$, or* (c) $H = ((S_k - I_k)_{a,b})^2$ *with $1 \leq a < b \leq j$. In case* (a), *the expression in* (2.14) *is equal to* $\mathrm{Var}[Z_1' Z_1]/d - 2$; *in case* (b), *it is equal to* $\mathbb{E}[(Z_1' Z_2)^3]/d$; *and in case* (c), *it equals* $\mathrm{Var}[(Z_1' Z_2)^2]/d^2 - 2(1 + 3/d)$.

To complete the proof of Theorem 2.1, let Z be as in the theorem. We first assume that conditions (t1)(a) and (t2) are satisfied with $k = 2$. The relation (1.3) holds for each x and ε, if we can show that the expression in (2.9) converges

to zero for each collection of indices l, m, $j_0, \ldots, j_m$ so that $1 \le l \le 2$, $m \ge 0$, $j_0 = 0$, $j_m \le l$, and $j_{i-1} + 1 < j_i$ for each $i = 1, \ldots, m$. Moreover, (1.5) holds for each $\varepsilon > 0$, if convergence zero of (2.9) is uniform in x over compacts. To this end, consider the difference of the expression in (2.9) and of the same expression with the Z_i's replaced by V_i's that are i.i.d. standard normal d-vectors. Expanding the polynomial ψ_x into a weighted sum of monomials, the difference in question can be written as a weighted sum of expressions of the form $\mathbb{E}[H] - \mathbb{E}[H^*]$ in case $m = 0$, and as a weighted sum of expressions of the form (2.14) in case $m > 0$, where the weight is given by the coefficient of the monomial H in ψ_x, and where G is of the form (2.12) [the monomials H, G^* and H^* are as in Proposition 2.6(i)]. By Proposition 2.4, we see that the coefficients of ψ_x are bounded uniformly in d and uniformly in x over compacts. And by Proposition 2.6(i), we see that expressions of the form $\mathbb{E}[H] - \mathbb{E}[H^*]$ or of the form (2.14) with G as in (2.12) all converge to zero. Therefore, (2.9) converges to zero, uniformly in x over compacts subsets of $\mathbb{R}$.

Finally, we assume that conditions (t1) and (t2) hold with $k = 4$. To derive (1.4) for fixed x and ε, we show that the expressions in (2.9) and (2.10) converge to zero [with the indices l, m, $j_0, \ldots, j_m$ in (2.9) now so that $1 \le l \le 4$, $m \ge 0$, $j_0 = 0$, $j_m \le l$, and $j_{i-1} + 1 < j_i$ for each $i = 1, \ldots, m$]. And (1.6) holds for each $\varepsilon > 0$ if convergence in (2.9) and (2.10) is uniform in x over compact sets. Now convergence to zero of (2.9) (with $k = 4$ here), uniformly over compacts, follows by arguing as in the preceding paragraph, mutatis mutandis. To deal with (2.10), consider the difference of (2.10) and of the same expression with the Z_i's replaced by i.i.d. standard Gaussian V_i's. Again, this can be written as a weighted sum of expressions of the form (2.14), with G now as in (2.13), where the weights are bounded uniformly in d and uniformly in x over compacts in view of Proposition 2.4. And by Proposition 2.6(ii), we see that (2.14) converges to zero except for those H's that correspond to the cases (a), (b) and (c) in Proposition 2.6(ii). Write $\mathcal{H}_a$, $\mathcal{H}_b$, and $\mathcal{H}_c$ for the collection of all monomials H where the case (a), (b), or (c) of Proposition 2.6(ii) occurs, respectively. For each $H \in \mathcal{H}_a$, the value of (2.14) is given $\mathrm{Var}[Z_1' Z_1]/d - 2$ and hence does not depend on H in view of Proposition 2.6(ii). And for each $H \in \mathcal{H}_a$, the coefficient of H in the polynomial $\psi_x(S_k - I_k)$ also does not depend on H in view of Proposition 2.4, because the monomials in $\mathcal{H}_a$ can be obtained from each other by permutations (or relabelings) of Z_i's. Consider now the difference of (2.10) and the same expression with the Z_i's replaced by i.i.d. standard Gaussians. The combined contribution of the monomials in $\mathcal{H}_a$ to that difference is given by

$$\sum_{j=1}^{k} \binom{k}{j} (-1)^j j = -k \sum_{j=0}^{k-1} \binom{k-1}{j} (-1)^j$$

multiplied by a constant (namely by $\mathrm{Var}[Z_1' Z_1]/d - 2$ times the common coefficient of the monomials from $\mathcal{H}_a$ in $\psi_x(S_k - I_k)$). Similarly, the monomials in $\mathcal{H}_b$

contribute

$$\sum_{j=1}^{k} \binom{k}{j} (-1)^j \binom{j}{2} = \sum_{j=0}^{k-2} \binom{k-2}{j} (-1)^j$$

multiplied by a constant. And the combined contribution of the monomials in $\mathcal{H}_c$ is also given by the expression in the preceding display multiplied by another constant. Because we have $k = 4$ here, the expressions in the last two displays are both equal to zero. Except for the more technical arguments that we collect in the supplementary material [13], this concludes the proof of Theorem 2.1.

SUPPLEMENTARY MATERIAL

Appendix: Proofs for Section 2 (DOI: 10.1214/12-AOS1081SUPP; .pdf). The Appendix contains several more technical arguments that are used in Section 2 including, in particular, Examples A.1 and A.2, as well as the proofs of Propositions 2.2 through 2.6.

REFERENCES

[1] BRYC, W. (1995). *The Normal Distribution*: *Characterizations with Applications. Lecture Notes in Statistics* **100**. Springer, New York. MR1335228

[2] CHEN, X., ZOU, C. and COOK, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38** 3696–3723. MR2766865

[3] CHIAROMONTE, F., COOK, R. D. and LI, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *Ann. Statist.* **30** 475–497. MR1902896

[4] COOK, R. D. and WEISBERG, S. (1991). Discussion of "Sliced inverse regression for dimension reduction," by K.-C. Li. *J. Amer. Statist. Assoc.* **86** 328–332.

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

[6] DONG, Y. and LI, B. (2010). Dimension reduction for non-elliptically distributed predictors: Second-order methods. *Biometrika* **97** 279–294. MR2650738

[7] DUDLEY, R. M. (2002). *Real Analysis and Probability. Cambridge Studies in Advanced Mathematics* **74**. Cambridge Univ. Press, Cambridge. MR1932358

[8] DÜMBGEN, L. and ZERIAL, P. (2011). On low-dimensional projections of high-dimensional distributions. Available at arXiv:1107.0417v1.

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

[10] FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **222** 309–368.

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

[12] HOEFFDING, W. (1952). The large-sample power of tests based on permutations of observations. *Ann. Math. Statistics* **23** 169–192. MR0057521

[13] LEEB, H. (2013). Supplement to "On the conditional distributions of low-dimensional projections from high-dimensional data." DOI:10.1214/12-AOS1081SUPP.

[14] LI, B., ARTEMIOU, A. and LI, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Ann. Statist.* **39** 3182–3210. MR3012405

[15] LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102** 997–1008. MR2354409

[16] LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

[17] LI, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87** 1025–1039. MR1209564

[18] LI, K.-C. (1997). Nonlinear confounding in high-dimensional regression. *Ann. Statist.* **25** 577–612. MR1439315

[19] LI, K.-C. and DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052. MR1015136

DEPARTMENT OF STATISTICS
UNIVERSITY OF VIENNA
UNIVERSITÄTSSTR. 3/5
VIENNA, 1010
AUSTRIA
E-MAIL: hannes.leeb@univie.ac.at