Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Linear Hypothesis Testing in Dense High-Dimensional Linear Models

Yinchu Zhu & Jelena Bradic

To cite this article: Yinchu Zhu & Jelena Bradic (2018) Linear Hypothesis Testing in Dense High-Dimensional Linear Models, Journal of the American Statistical Association, 113:524, 1583-1600, DOI: 10.1080/01621459.2017.1356319

To link to this article: https://doi.org/10.1080/01621459.2017.1356319

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 524, 1583–1600: Theory and Methods
https://doi.org/10.1080/01621459.2017.1356319

Linear Hypothesis Testing in Dense High-Dimensional Linear Models

Yinchu Zhu and Jelena Bradic

Department of Mathematics, Rady School of Management, University of California at San Diego, La Jolla, CA

ABSTRACT

We propose a methodology for testing linear hypothesis in high-dimensional linear models. The proposed test does not impose any restriction on the size of the model, that is, model sparsity or the loading vector representing the hypothesis. Providing asymptotically valid methods for testing general linear functions of the regression parameters in high-dimensions is extremely challenging—especially without making restrictive or unverifiable assumptions on the number of nonzero elements. We propose to test the moment conditions related to the newly designed restructured regression, where the inputs are transformed and augmented features. These new features incorporate the structure of the null hypothesis directly. The test statistics are constructed in such a way that lack of sparsity in the original model parameter does not present a problem for the theoretical justification of our procedures. We establish asymptotically exact control on Type I error without imposing any sparsity assumptions on model parameter or the vector representing the linear hypothesis. Our method is also shown to achieve certain optimality in detecting deviations from the null hypothesis. We demonstrate the favorable finite-sample performance of the proposed methods, via a number of numerical and a real data example. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received June 2016
Revised June 2017

KEYWORDS
Dantzig; High-dimensional linear models; Inference; Lasso; Linear Testing; Nonsparse models

1. Introduction

A high-dimensional inference is a fundamental topic of interest in modern scientific problems that are widely recognized to be of high-dimensional nature, that is, that require estimation of parameters with dimensionality exceeding the number of observations. Applications span a wide variety of scientific fields, such as biology, medicine, genetics, neuroscience, economics, and finance. Minimizing a suitably regularized (quasi-)likelihood function was developed (Tibshirani 1996; Fan and Li 2001) as a suitable approach for the estimation in such models. In particular, high-dimensional linear models have been studied extensively in recent years and take the following form

$$y_i = x_i^\top \beta_* + \varepsilon_i, \qquad i = 1, 2, \ldots, n \tag{1}$$

for a response $y_i \in \mathbb{R}$, a feature vector $x_i \in \mathbb{R}^p$, and the noise $\varepsilon_i \in \mathbb{R}$, such that $E[\varepsilon_i] = 0$ and $E[\varepsilon_i^2] = \sigma_\varepsilon^2$ with $0 < \sigma_\varepsilon^2 < \infty$. The vector $\beta_* \in \mathbb{R}^p$ is the unknown model parameter and we allow for $p \gg n$. We consider a random design setting with the feature vectors satisfying $Ex_i = 0$ and $E[x_i x_i^\top] = \Sigma_X$. Under certain regularity conditions on the design matrix $X = (x_1, x_2, \ldots, x_n)^\top$, regularized methods with a suitable choice of the tuning parameter have been shown to achieve the optimal rate of estimation as long as the vector β_* is sparse in that $\|\beta_*\|_0 = o(n/\log p)$.

The goal of the present article is to address the testing problem for linear hypotheses of the form

$$H_0 : a^\top \beta_* = g_0, \tag{2}$$

where the loading vector $a \in \mathbb{R}^p$ is prespecified and $g_0 \in \mathbb{R}$ is given, and design an asymptotically valid test statistic that does not rely on sparsity assumptions. Some central challenges have hindered the systematic development of tools for statistical inference in such settings. The nonsparse nature of the model parameter β_* poses serious challenges to consistent estimation; moreover, the size and structure of the loading vector a introduce additional difficulty for the inference. However, in this article we consider potentially dense vectors β_* with $0 \le \|\beta_*\|_0 \le p$. We also allow for the nonsparse loadings with $1 \le \|a\|_0 \le p$. The inference problem for the mean of the response y_i conditional on $x_i = a$ is a prototypical case for the general functional $a^\top \beta_*$ and is a representative case for dense loading a.

We develop the principles of *restructured regression*, where a hypothesis-driven *feature synthesization* is introduced. The feature augmentation is done in such a way to separate useful inferential information from the useless one, by "projecting" the original feature space to the space spanned by the vector a and the space orthogonal to a. This orthogonal projection is introduced to achieve the above separation and avoid the curse of dimensionality. Then, an appropriate moment condition is invoked on the restricted regression and a suitable test statistic constructed. The structure of the moment condition and its test depend on whether or not the covariance of the features Σ_X is known. When prior knowledge of Σ_X is available, the synthesized features can be created in such a way that the resulting moment condition and testing procedure do not depend on β_*; thus, estimation of β_* is completely avoided. As a result, no assumption on the sparsity of β_* is required. We establish theoretical guarantees for Type I error control and show that the test can

CONTACT Jelena Bradic jbradic@ucsd.edu Department of Mathematics, Rady School of Management, University of California at San Diego, La Jolla, CA 92093.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

detect the deviation from the null hypothesis of the order $O(\|a\|_2/\sqrt{n})$. To the best of our knowledge, our approach provides the first result on testing general linear hypothesis (2) in high-dimensional linear models with potentially nonsparse (dense) parameters.

When prior knowledge of Σ_X is unavailable, the orthogonalization and perfect separation is not achievable due to the unknown projection matrix. We design an estimator of the projection matrix and further condition the new and augmented features in such a way that their correlations are estimable and yet the format of the restructured regression remains unchanged. The developed hypothesis-driven feature separation diminishes the impact of the inaccuracy of an estimator of a transformation of β_*. Consequently, we can establish asymptotically exact control of Type I error. We believe there is currently no result on testing $a^\top\beta_*$ in the case where Σ_X is unknown, and both β_* and a are allowed to be dense. Moreover, when sparsity assumptions hold, our procedure is shown to achieve optimality guarantees; hence, it does not loose efficiency.

Since we do not assume sparsity in β_*, our work does not directly compare to the existing results, which are only valid for sparse β_*. However, in some cases, our work generalizes existing results to the nonsparse models. For example, Cai and Guo (2017) showed that when Σ_X is known, the minimax length of the confidence interval for $a^\top\beta_*$ is of the order $O(\|a\|_2/\sqrt{n})$ if $\|\beta_*\|_0 = O(n/\log p)$. As confidence sets for $a^\top\beta_*$ can be easily constructed by inverting the proposed tests, our results indicate that their conclusion continues to hold for nonsparse models, where $\|\beta_*\|_0$ can be as large as p. For the case of dense a, we do not impose any constraint on a. However, existing work, such as Cai and Guo (2017), imposes a lower bound (in terms of $\|a\|_\infty$) on the minimal nonzero coordinate of a—a condition that is seldom satisfied for inference of conditional mean, when a is typically drawn from a continuous distribution (e.g., a is drawn from the same distribution as the distribution of the x_i's).

1.1. Relation to Existing Literature

Confidence intervals and hypothesis testing play a fundamental role in statistical theory and applications. However, compared to the point estimation there is still much work to be done for statistical inference of high-dimensional models. Existing work on the inference problems predominantly focuses on individual coordinates of β_*. Early work typically imposes conditions that guarantee consistent variable selection (see Fan and Li 2001; Zou 2006; Zhao and Yu 2006) or develops methods that lead to conservative inferential guarantees (e.g., Bühlmann 2013). However, recent work focusses on asymptotically accurate inference without relying on the variable selection consistency. Current advances in this domain are, however, restricted to the ultra-sparse case, where $\|\beta_*\|_0 = o(\sqrt{n}/\log p)$; see Zhang and Zhang (2014), Belloni, Chernozhukov, and Hansen (2014), Van de Geer et al. (2014), Javanmard and Montanari (2014a), Ning and Liu (2017), Javanmard and Montanari (2015), Mitra and Zhang (2016), Bühlmann and van de Geer (2015), Belloni, Chernozhukov, and Kato (2015), and Chernozhukov, Hansen, and Spindler (2015). Under such sparsity condition, the expected length of the confidence intervals for individual coordinates is of the order $O(1/\sqrt{n})$ (van de Geer and Jankova 2016). Cai and Guo (2017) studied the length of the confidence intervals allowing for $\|\beta_*\|_0 = o(n/\log p)$ and discovered that lack of explicit knowledge of $\|\beta_*\|_0$ can fundamentally limit the efficiency of confidence intervals.

However, there is little reason to believe that the sparsity of β_* needs to hold in practice (Pritchard 2001; Ward 2009; Jin and Ke 2014; Hall, Jin, and Miller 2014). Unfortunately, there is almost no work on estimating or testing the true sparsity level of the underlying parameter. Hence, the theory of hypothesis testing under general sparsity structures is still a very challenging and important open problem. In particular, progress is very much required when $\|\beta_*\|_0$ is allowed to grow faster than $n/\log p$ and perhaps even larger than the sample size n. There are several articles showing that the regularized procedures have nonvanishing estimation errors in such settings (Donoho and Johnstone 1994; Raskutti, Wainwright, and Yu 2011; Cai and Guo 2018). However, is it still possible to develop a general methodology for testing β_* in this case? Can one construct valid inference procedures that do not require knowledge of $\|\beta_*\|_0$?

In the proposed inference procedure, we handle the high-dimensional, possibly nonsparse model parameters and/or nonsparse loadings, by developing a new methodology for testing. The proposed methodology is centered around a construction of augmented and synthesized features that are driven by a specific form of the null hypothesis. Compared with the previous approaches of de-biasing (Zhang and Zhang 2014; Javanmard and Montanari 2014a; Van de Geer et al. 2014; Mitra and Zhang 2014), scoring (Ning and Liu 2017; Chernozhukov, Hansen, and Spindler 2015), and double-selection (Belloni, Chernozhukov, and Hansen 2014; Belloni, Chernozhukov, and Kato 2015), our new approach has two major distinctive features:

- We do not rely on an $l_1$norm consistent estimation of the unknown model parameters. In high-dimensional models with the lack of sparsity in the parameters, this may no longer be possible. Instead, we propose to reformulate the original parametric null hypothesis into a moment condition that can be successfully estimated even without sparsity in the model. This moment condition is different from the score equations employed for estimation as those are not estimable in nonsparse high-dimensional models.
- We advocate for a study and exploration of the correlation between feature vectors (and not the model parameters); this proves to be a valuable tool that overcomes the limit of estimation. Namely, we propose that the features be split and projected onto the loading vector a of the hypothesis (2), thereby fully using the null hypothesis structure.
  This "decoupling" scheme allows for a successful estimation of the moment condition even without sparsity assumption. As a result the developed method provides a rich alternative to the classical Wald or Score principles.

1.2. Notation and Organization of the Article

We briefly describe notations used in the article. We use $\to^d$ to denote convergence in distribution and $\mathcal{N}(0,1)$ to denote the standard normal distribution with its cumulative distribution function denoted by $\Phi(\cdot)$. The (multivariate) normal

distribution with mean (vector) μ and variance (matrix) Σ is denoted by $\mathcal{N}(\mu, \Sigma)$. We use $^\top$ to denote the transpose of (a vector or matrix) and denote by I_p the $p \times p$ identity matrix. For a vector $a = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$, its l_0 norm is the cardinality of $\text{supp}(a) = \{i \mid a_i \neq 0\}$ and $\|a\|_\infty = \max\{|a_1|, \ldots, |a_p|\}$; $\|a\|_1$ and $\|a\|_2$ denote the l_1 and l_2 norm of a, respectively. In this case, a_{-i} denotes the vector a with its ith coordinate removed. For two sequences of positive constants a_n and b_n, we use $a_n \asymp b_n$ to denote that $a_n/b_n = O(1)$ and $b_n/a_n = O(1)$. For two real numbers a_1 and a_2, $a_1 \vee a_2$ and $a_1 \wedge a_2$ denote $\max\{a_1, a_2\}$ and $\min\{a_1, a_2\}$, respectively.

The rest of this article is organized as follows. Section 2 introduces the main methodology under known Σ_X and establishes theoretical properties of the proposed test. Section 3 extends the proposed methodology to the case of the unknown Σ_X and provides theoretical results. Section 4 contains examples illustrating new methods that the proposed methodology brings to the literature on high-dimensional inference. Section 5 contains detailed numerical experiments on a number of dense high-dimensional linear models, including sparse and dense loadings a. In Section 5.1, we demonstrate the excellent finite-sample performance of the proposed methods through Monte Carlo simulations; in Section 5.2, we illustrate our method via a real data study. Appendix A contains complete details of the theoretical derivations.

2. Testing $H_0: a^\top \beta_* = g_0$ with Prior Knowledge of Σ_X

In this section, we promote a unified approach to a wide class of decision problems. Our main building block (which we believe is important in its own right) is a construction, named *restructured regression* allowing, under weak assumptions, to build tests for hypotheses on $a^\top \beta_*$, where β_* and/or a can be nonsparse. Considering the potential failure of sparsity in many practical problems, we strongly believe that our approach permits a diverse spectrum of applications. In this section, our focus is to introduce the method with known Σ_X (an assumption relaxed in the next section).

Throughout the article, we denote $\Omega_X = \Sigma_X^{-1}$. In the sequel, given the feature vector $x_i \in \mathbb{R}^p$ and loading vector $a \in \mathbb{R}^p$, we consider the following decomposition:

$$x_i = a z_i + w_i, \tag{3}$$

with a scalar

$$z_i = \left(\frac{\Omega_X a}{a^\top \Omega_X a} \right)^\top x_i$$

and a p-dimensional vector

$$w_i = \left[I_p - \frac{a a^\top \Omega_X}{a^\top \Omega_X a} \right] x_i.$$

Observe that az_i can be viewed as the projection of x_i onto the vector a—taking into account Ω_X, hence extracting information in x_i regarding the null hypothesis. Notice that the model (1) and decomposition (3) imply

$$y_i = z_i \cdot (a^\top \beta_*) + w_i^\top \beta_* + \varepsilon_i, \tag{4}$$

referred to as *restructured regression*. The proposed construction gives rise to the method of *feature customization*. Given covariate vector x_i and the loading vector a representing the structure of the null hypothesis, we create the synthesized features $\tilde{x}_i := (z_i, w_i^\top)^\top$ so that the regression coefficient for z_i in the restructured regression (4) is the quantity under testing.

Remark 1. The synthesized features are not only an artifact of our new methodology but also admit intuitive interpretations. Consider the case where Σ_X is known to be I_p. The synthesized features z_i and w_i represent the relevant and the irrelevant information with respect to the null, respectively. To see this, suppose that the true distribution of the data is known. With the population expectations, we can identify the parameters in the restructured regression (4): $E(z_i y_i) = E z_i^2 (a^\top \beta_*)$ and $E w_i y_i = E w_i w_i^\top \beta_*$. Notice that the latter equation contains no information regarding $a^\top \beta_*$ because it can be shown that a is orthogonal to columns in $E w_i w_i^\top$. In other words, knowing $E w_i w_i^\top \beta_*$ does not lead to knowing $a^\top \beta_*$. Therefore, $a^\top \beta_*$ is identified with the distribution of (y_i, z_i) and w_i does not contain information about the null hypothesis.

It is not hard to verify that, by the construction of the transformed features, $E[w_i z_i] = 0$. It follows that $E[z_i(y_i - z_i g_0)] = E[z_i(\varepsilon_i + w_i^\top \beta_* + z_i(a^\top \beta_* - g_0))] = E[z_i^2 (a^\top \beta_* - g_0)]$. Observe that the last expression is 0 if and only if the null hypothesis (2) holds. As a result, testing H_0 in (2) is equivalent to testing the following moment condition:

$$H_0 : E\left[z_1 (y_1 - z_1 g_0) \right] = 0. \tag{5}$$

To test the above condition, we propose a studentized test statistic, $T_n(g_0)$, taking the form

$$T_n(g_0) := \frac{n^{-1/2} \sum_{i=1}^n l_i(g_0)}{\sqrt{n^{-1} \sum_{i=1}^n l_i(g_0)^2}}, \tag{6}$$

with $l_i(g_0) = z_i(y_i - z_i g_0)$. For a test of H_0 with nominal size $\alpha \in (0, 1)$, we reject H_0 if

$$|T_n(g_0)| > \Phi(1 - \alpha/2).$$

The methodology proposed above is novel in a number of aspects. Unlike Wald or Score or Likelihood principles, centered around a consistent estimator of β_*, our methodology allows for extremely fast implementation and does not estimate the unknown parameter β_*. The novel methodology consists of two stages. At the first stage, our procedure establishes a data-driven feature decomposition based on the structure of the null hypothesis directly. At the second stage, only "a moment condition" of the restructured regression is tested. It is critical to observe that restructured regression by itself is not sufficient to guarantee valid inference. The novel properties of the proposed method are based on the built-in, that is, designed orthogonality of the synthesized features z_i and w_i. As such it enables us to construct a test statistic that does not contain the unknown parameter β_*, thereby allowing our methodology to handle dense (and thus possibly nonestimable) β_*. Moreover, no assumption is imposed on the loadings a either. As we will see in the next section, these properties under known Σ_X propagate to the case of the unknown Σ_X and underline all further developments.

Assumption 1. Let the following hold: (i) there exists a positive constant C such that $E|z_i\sigma_z^{-1}|^8 \le C$, $E\varepsilon_i^8 \le C$, and $E|w_i^\top \beta_*|^8 < C$ with $C < \infty$. Moreover, (ii) there exists a constant $c \in (0, \infty)$, such that $\sigma_\varepsilon \ge c$. Lastly, (iii) there exist constants $D_1, D_2 > 0$ such that the eigenvalues of Σ_X lie in $[D_1, D_2]$.

The stated conditions in Assumption 1 are very weak and intuitive. Assumption 1(i) requires components in the restructured regression (4) to have bounded eighth moments. Assumption 1(ii) rules out the noiseless regression setting in the original model (1). Assumption 1(iii) is very weak in that it only imposes well-designed covariance matrix of the features x_i (see Bickel, Ritov, and Tsybakov 2009).

Notice that Assumption 1 does not require any condition regarding the sparsity of β_*. Even in the case of sparse a, existing work, such as the debiasing method, heavily relies on the sparsity of β_*. Results regarding dense a are very limited even for sparse β_*. Cai and Guo (2017) imposed the condition of $\max_{j\in\text{supp}(a)} |a_j| / \min_{j\in\text{supp}(a)} |a_j| = O(1)$; however, such a condition is quite hard to satisfy if a is drawn from a continuous distribution whose support contains zero. In contrast, our results do not require any condition on a and, hence, bridge the gap in the existing literature on high-dimensional inference.

Theorem 1. Consider the model in (1) and the definition of z_i and w_i as in (3). Suppose that Assumption 1 holds. Under H_0 in (2), we have that (1) the test statistic T_n, (6), satisfies $T_n(g_0) \to^d \mathcal{N}(0, 1)$ as $n, p \to \infty$ and that (2)

$$\lim_{n,p\to\infty} P\Big(|T_n(g_0)| > \Phi^{-1}(1-\alpha/2)\Big) = \alpha.$$

Theorem 1 gives an asymptotic approximation for the null distribution of the test statistic $T_n(g_0)$ under general sparsity structure. The result of Theorem 1 has two striking features. The first is that it holds, no matter the size or sparsity of the loading vector a. The second is that the proposed test guarantees Type I error control when $p \ge n$ and $p, n \to \infty$ no matter of the sparsity of β_* and without the knowledge of the noise level σ_ε; in particular, it allows $\|\beta_*\|_0 = p$. Therefore, our test is fully adaptive, in the sense that its validity does not depend on in the sparse/dense level of either the model parameter β_* or the hypothesis loading a. We also show that our test can detect deviations from the null that are larger than $O(\|a\|_2/\sqrt{n})$ while allowing β_* to be nonsparse and $p \ge n$.

Theorem 2. Under the conditions of Theorem 1, suppose that $a^\top\beta_* = g_0 + h_n$ and $\sqrt{n}|h_n|/\|a\|_2 \to \infty$. Then, for any $\alpha \in (0, 1)$.

$$\lim_{n,p\to\infty} P\Big(|T_n(g_0)| > \Phi^{-1}(1-\alpha/2)\Big) = 1.$$

Remark 2. Theorem 2 also suggests that we can expect the length of the confidence interval for $a^\top\beta_*$ (obtained by inverting the proposed test) to be of the order of $O(\|a\|_2/\sqrt{n})$ regardless of the sparsity of β_* or a. To the best of our knowledge, it is the first result to explicitly allow nonsparse and simultaneously high-dimensional parameters β^* or vector loadings a. It is also closely connected with the existing results for the case of sparse parameters β^*. Cai and Guo (2017) stated that under Gaussianity and sparsity in both β_* and a together with known Σ_X and σ_ε, the optimal expected length of confidence intervals for $a^\top\beta_*$ is of the order $O(\|a\|_2/\sqrt{n})$ (see Theorem 7 therein). Observe that our procedure achieves the same optimality without the knowledge of σ_ε and allowing dense vectors β_*.

We do not formally claim that this is the optimal rate for dense β_*, but we can consider an obvious benchmark. Let $\bar\beta$ be an estimator that attains an efficiency similar to (ordinary least square) OLS in low dimensions, that is, $\bar\beta$ is distributed as $\mathcal{N}(\beta_*, \Omega_X\sigma_\varepsilon^2/n)$. Then $a^\top\bar\beta$ follows $\mathcal{N}(a^\top\beta_*, a^\top\Omega_X a\sigma_\varepsilon^2/n)$ distribution. Since Ω_X has eigenvalues bounded away from infinity, the standard deviation of $a^\top\bar\beta$ is of the order $\|a\|_2/\sqrt{n}$. Such an estimator might not be feasible in practice, but could serve as a benchmark for dense β_*. A rigorous study of the efficiency issue is likely to yield results that are quite different from current literature since existing results, for example, Cai and Guo (2017), do not naturally extend to dense problems. For example, consider the case of $\|a\|_0 = \|\beta_*\|_0 = p$, naively extending Theorem 8 of Cai and Guo (2017) would conclude that the minimax expected length of a confidence interval for $a^\top\beta_*$ is of the order $\|a\|_\infty p\sqrt{(\log p)/n}$; however, this rate is larger than the rate $\|a\|_2/\sqrt{n}$, which is bounded above by $\|a\|_\infty\sqrt{p/n}$. Lastly, according to Theorem 2 our proposed test achieves the same rate at the benchmark $\bar\beta$.

3. Testing $H_0: a^\top\beta_* = g_0$ Without Prior Knowledge of Σ_X

The approach proposed in this section tackles the high-dimensional inference problem in a very general setting. The focus is the more realistic scenario in which the covariance matrix Σ_X and the variance of the model (1) are both unknown. We synthesize new features, create a new reference model, and explore the correlations therein to design a suitable inferential procedure that is stable without sparsity assumption.

3.1. Feature Synthesization and Restructured Regression

To design inference when Σ_X unknown, we take on a new perspective and build upon the methodology of Section 2. Consider feature synthesization of Section 2 where Σ_X is naively treated as I_p,

$$z_i = \Big(\frac{a}{a^\top a}\Big)^\top x_i \in \mathbb{R} \quad \text{and} \quad w_i = \big(I_p - aa^\top/(a^\top a)\big)\, x_i \in \mathbb{R}^p. \tag{7}$$

Although the decomposition $x_i = az_i + w_i$ still holds, features z_i and w_i might be correlated (because $\Sigma_X \ne I_p$). If such correlation is estimated successfully, we can use certain decoupling method to eliminate the impact of dense parameters while allowing exponentially growing dimensions.

The first challenge is that directly estimating the correlation between z_i and w_i (as defined) is not achievable (as the restricted eigenvalue (RE) condition (Bickel, Ritov, and Tsybakov 2009) on $W = (w_1, \ldots, w_n)^\top$ is violated). To address this problem, we propose to *stabilize* the feature vector w_i and define *stabilized* features $\tilde w_i$. We stabilize the features in such a way that the RE condition on the stabilized design $\tilde W = (\tilde w_1, \ldots, \tilde w_n)^\top$ is satisfied with high probability. Since $I_p - aa^\top/(a^\top a)$ is a projection

matrix, we can find $U_a \in \mathbb{R}^{p\times(p-1)}$ an orthogonal matrix such that

$$U_a^\top U_a = I_{p-1} \qquad \text{and} \qquad I_p - aa^\top/(a^\top a) = U_a U_a^\top.$$

Then

$$W\beta_* = X(I_p - aa^\top/(a^\top a))\beta_* = XU_aU_a^\top\beta_* = \tilde{W}\pi_*,$$

where

$$\tilde{W} = WU_a \qquad \text{and} \qquad \pi_* = U_a^\top\beta_*.$$

Since $y_i = z_i\cdot(a^\top\beta_*) + w_i^\top\beta_* + \varepsilon_i$, we have the *stabilized model*

$$y_i = z_i\cdot(a^\top\beta_*) + \tilde{w}_i^\top\pi_* + \varepsilon_i. \tag{8}$$

The model is balanced in the sense that $E\tilde{W}^\top\tilde{W}/n = U_a^\top\Sigma_X U_a \in \mathbb{R}^{(p-1)\times(p-1)}$ with eigenvalues bounded away from zero and infinity. Therefore, RE condition on $\tilde{W}$ holds under weak conditions; see Rudelson and Zhou (2013).

Remark 3. The synthesized feature $w_i \in \mathbb{R}^p$ is consolidated into $\tilde{w}_i \in \mathbb{R}^{p-1}$, in that $\tilde{w}_i$ has a smaller dimensionality and can be used to recover w_i via $w_i = U_a\tilde{w}_i$. In this sense, $\tilde{w}_i$ contains all the information in w_i. As an example, consider the case with a being the first column of I_p. In this case, it is not hard to verify that $z_i = x_{i,1}$, $w_i = (0, x_{i,2}, \ldots, x_{i,p})^\top \in \mathbb{R}^p$, $U_a = (0, I_{p-1})^\top \in \mathbb{R}^{p\times(p-1)}$ and thus $\tilde{w}_i = U_a^\top w_i = (x_{i,2}, \ldots, x_{i,p})^\top \in \mathbb{R}^{p-1}$.

We now introduce an additional model to account for the dependence between the *synthesized feature* z_i and the *stabilized feature* $\tilde{w}_i$:

$$z_i = \tilde{w}_i^\top\gamma_* + u_i, \tag{9}$$

where $\gamma_* \in \mathbb{R}^{p-1}$ is an unknown parameter and u_i is independent of $\tilde{w}_i$ with $Eu_i = 0$ and $Eu_i^2 = \sigma_u^2$.

In this article, we will assume that γ_* is sparse, to decouple the dependence between z_i and $\tilde{w}_i$ with the unknown Σ_X. In fact, sparse γ_* is a generalization of the sparsity condition on the precision matrix Ω_X, a regularity condition typically imposed in the literature; see Van de Geer et al. (2014), Belloni, Chernozhukov, and Hansen (2014), Belloni, Chernozhukov, and Kato (2015), and Ning and Liu (2017). Recall the example in Remark 3. Since $x_{i,1} = z_i = \tilde{w}_i^\top\gamma_* + u_i = x_{i,-1}^\top\gamma_* + u_i$, it is not hard to show that the first row of Ω_X is $(\sigma_u^{-2}, -\sigma_u^{-2}\gamma_*^\top)$. Hence, the sparsity of γ_* is equivalent to the sparsity in the first row of Ω_X. The sparsity of γ_* can be justified for dense a as well. Consider the case of $\Sigma_X = cI_p$ for some $c > 0$; a prototypical model in compressive sensing corresponds to $c = 1$ (Nickl and van de Geer 2013). In this case, one can easily show that z_i and $\tilde{w}_i$ are uncorrelated, meaning that $\gamma_* = 0$ for any a. The synthesized features also admit intuitive interpretations in this case: the feature z_i contains useful information in testing the null hypothesis $a^\top\beta_* = g_0$, while the consolidated $\tilde{w}_i$ contains information not useful for inference.

Now, we are ready to construct the moment condition of interest. Observe that under H_0 in (2), $y_i - z_ig_0 - \tilde{w}_i^\top\pi_* = \varepsilon_i$ is uncorrelated with $z_i - \tilde{w}_i^\top\gamma_* = u_i$. If H_0 is false, then $y_i - z_ig_0 - \tilde{w}_i^\top\pi_* = \varepsilon_i + z_i(\theta_* - g_0) = \varepsilon_i + \tilde{w}_i^\top\gamma_*(\theta_* - g_0) + u_i(\theta_* - g_0)$ has nonzero correlation with $u_i = z_i - \tilde{w}_i^\top\gamma_*$. Hence, the initial null hypothesis, (2) is equivalent to the following null hypothesis:

$$H_0 : E\Big[\big(z_1 - \tilde{w}_1^\top\gamma_*\big)\big(y_1 - z_1g_0 - \tilde{w}_1^\top\pi_*\big)\Big] = 0. \tag{10}$$

Directly testing this moment condition is not feasible, due to the unknown values of parameters γ_* and π_*. As a result, we first provide estimates for these unknown parameters and consider the test statistic given by the studentized statistics.

We make a few remarks about the above proposed methodology. As mentioned above, the existing literature on high-dimensional inference adopts the approach of relying on an (almost) unbiased estimate of the model parameter to distinguish the null and alternative hypotheses. The existing methods largely differ by the means of constructing the unbiased estimate and/or its asymptotic variance. Many use an approximation of a one-step Newton method (Zhang and Zhang 2014; Van de Geer et al. 2014; Javanmard and Montanari 2014a) to achieve consistency in estimation of possibly all p parameters. To test $a^\top\beta_*$ in this framework, one need to show that the debiased estimator for β_* can be used to construct an asymptotically unbiased and normal estimator for $a^\top\beta_*$; to the best of our knowledge, a formal theoretical justification is yet to be established even under sparse β_*. Other than the debiasing technique, some proposals center around Neyman's score orthogonalization ideas (Belloni, Chernozhukov, and Hansen 2014; Ning and Liu 2017; Belloni, Chernozhukov, and Kato 2015; Chernozhukov, Hansen, and Spindler 2015). It is worth pointing out that such a method requires a clear separation of parameter under testing and the nuisance parameter. In the original problem, the model parameter is β_* and the quantity under testing is $a^\top\beta_*$; hence, it is not clear how to define the nuisance parameter since the $a^\top\beta_*$ is not just one entry (or a subset) of the parameter vector β_*. Lastly, the work of Cai and Guo (2017) proposes a minimax optimal test that allows for dense loadings vector a, however in the dense case it provides a conservative error bounds and requires the knowledge of the sparsity size s.

However, our proposal deviates from the above methodologies in a few aspects. First, we design a test statistic irrespective of a consistency of high-dimensional estimators for the model parameter; hence, any refitting or one-step approximations are unnecessary. Second, we aim to orthogonalize design features (rather than model parameters) by directly taking into account the structure of the null hypothesis (represented by a and g_0). In this way, we achieve full adaptivity to the hypothesis testing problem of interest. Third, we reformulate the original parametric hypothesis into a moment condition of which we provide adaptive estimators. The moment condition itself is not a simple first-order optimality identification (related to Z-estimators), but rather a moment that uses the special feature of orthogonalization and fusion. Hence, even in setting where the existing work applies, our proposed method provides an alternative. However, apart from existing work, our proposed method applies much more broadly.

3.2. Adaptive Estimation of the Unknown Quantities

In this subsection, we start with a brief introduction of the Dantzig selector, which is the basis of our estimators. Then we

introduce the intuition and steps of our estimator as well as implementation details.

3.2.1. Dantzig Selector Review

Numerous studies have been conducted in regards to the consistent estimation of high-dimensional parameters in linear models. The canonical examples of successful estimators represent Lasso and Dantzig selector, defined as $\hat{\beta}_l$ and $\hat{\beta}_d$ below,

$$\hat{\beta}_l = \arg\min_{\beta\in\mathbb{R}^p} \left\{\|Y - X\beta\|_2^2 + \lambda_l\|\beta\|_1\right\},$$
$$\hat{\beta}_d = \arg\min_{\beta\in\mathbb{R}^p} \|\beta\|_1 \quad s.t \quad \left\|n^{-1}X^\top(Y - X\beta)\right\|_\infty \le \lambda_d. \tag{11}$$

Although Lasso and Dantzig selector are defined in different times, Bickel, Ritov, and Tsybakov (2009) established equivalence between the two estimators under the conditions of moderate design correlations and model sparsity, $\|\beta_*\|_0 \ll n$. Between these two estimator, the Dantzig selector, $\hat{\beta}_d$, offers easy implementation through linear programming techniques. Moreover, the constraint in the Dantzig selector can be interpreted as a relaxation of the least-square normal equations, $X^\top Y = X^\top X\beta$. However, the performance of both estimators is tightly connected to the choice of their respective tuning parameters λ_l and λ_d, that is, the size of such relaxation. Several empirical and theoretical studies emphasized that tuning parameters should be chosen proportionally to the noise standard deviation σ_ε, that is, $\lambda_d = \lambda_d(\sigma_\varepsilon) = \sigma_\varepsilon\sqrt{(\log p)/n}$. In such settings, one can guarantee $\|\hat{\beta}_l - \beta_*\|_1 = O(\|\beta_*\|_0\sqrt{(\log p)/n})$. Unfortunately, in most applications, the variance of the noise is unavailable. It is therefore vital to design statistical procedures that estimate unknown parameters together with the size of model variance in a joint fashion. This topic received special attention, see Giraud, Huet, and Verzelen (2012) and the references therein. Most popular σ-adaptive procedures, the square-root Lasso (Belloni, Chernozhukov, and Wang 2011), the scaled Lasso (Sun and Zhang 2012), and the self-tuned Dantzig selector (Gautier and Tsybakov 2013; Belloni, Chernozhukov, and Hansen 2016) can be seen as maximum a posteriori estimators with a particular choice of prior distribution. However, they do not provide estimates that are reasonable in nonsparse and high-dimensional models—after all in such settings it is impossible to consistently estimate the model parameters (see for more details Raskutti, Wainwright, and Yu 2011; Cai and Guo 2018). The aim of the present section is to present an alternative to these methods, which are closely related, but presents some advantages in terms of implementation and a more transparent theoretical analysis in not necessarily sparse models; the main benefit is that our estimates are well controlled in certain sense.

3.2.2. Modified Dantzig Selector: Adaptive to Signal-to-Noise Ratio

We start with the estimator for π_*, a parameter that is high-dimensional and yet not necessarily sparse. We extend the Dantzig selector above to conform to the testing problem that we have to perform. We begin by splitting the tuning parameter into a constant independent of the variance of the noise and introduce a parameter ρ, a square root of the noise to response ratio as an unknown in the optimization problem. At the population level, ρ is intended to represent $\sigma_\varepsilon/\sqrt{E(y_1 - z_1 g_0)^2}$ and ρ_0 is a lower bound for this ratio. One might attempt to use scaled Lasso by Sun and Zhang (2012) or self-tuning Dantzig selector proposed by Gautier and Tsybakov (2013), but for nonsparse π_*, these methods cannot ensure that the estimated noise variance is bounded away from zero whenever the vector π_* is a dense vector (a case of special interest here).

For $Z = (z_1, \ldots, z_n)^\top$ and $Y = (y_1, \ldots, y_n)^\top$ defined in (7), we introduce the following version of Dantzig selector of π_*:

$$\begin{aligned}(\hat{\pi}, \hat{\rho}) = &\arg\min_{(\pi,\rho)\in\mathbb{R}^{p-1}\times\mathbb{R}} \|\pi\|_1 \\ \text{s.t} \quad &\|\tilde{W}^\top(Y - Zg_0 - \tilde{W}\pi)\|_\infty \le \eta\,\rho\,\sqrt{n}\|Y - Zg_0\|_2 \\ &(Y - Zg_0)^\top(Y - Zg_0 - \tilde{W}\pi) \ge \rho_0\,\rho\,\|Y - Zg_0\|_2^2/2 \\ &\rho \in [\rho_0, 1],\end{aligned} \tag{12}$$

where $\eta \asymp \sqrt{n^{-1}\log p}$ and $\rho_0 \in (0, 1)$ are scale-free tuning parameters.

The estimator (12) is different from (11) in two ways. First, the estimator (12) simultaneously estimates π_* and ρ. We introduce a ρ_0 the lower bound for ρ as a tuning parameter. Second, the estimator (12) has an additional constraint, which essentially serves as an upper bound for ρ. The intuition of this bound is the following. When π is replaced by the true π_* and the null hypothesis holds, this constraint (scaled by $1/n$) becomes $\pi_*^\top\tilde{W}^\top\varepsilon/n + \varepsilon^\top\varepsilon/n \ge \rho_0\rho\|\tilde{W}\pi_* + \varepsilon\|_2^2/n$. By the law of large numbers, this means that $o_P(1) + \sigma_\varepsilon^2 \ge \rho_0\rho E(y_1 - z_1 g_0)^2$, which is satisfied if $\rho = \sigma_\varepsilon/\sqrt{E(y_1 - z_1 g_0)^2}$ and $\rho > \rho_0$.

The vector $\varepsilon = Y - Zg_0 - \tilde{W}\pi_*$ is a residual vector of the stabilized model (8) under the null hypothesis H_0. The first constraint on the residual vector imposes that for each i, much like the Dantzig selector, $\hat{\beta}_l$, maximal correlation $\|\tilde{W}^\top\varepsilon/n\|_\infty$ is not larger than the noise level $\eta\sigma_\varepsilon$. Yet, in contrast to $\hat{\beta}_l$, our estimator treats ρ as an unknown quantity and estimates it simultaneously with π_*. Moreover, we introduce the second constraint to stabilize estimation of the moment of interest (10) in the presence of nonsparse vectors π_*. Under the null hypothesis, this constraint prevents choice of ρ that is too large, namely, it constraints $\rho \le C(Y - Zg_0)^\top\varepsilon/\left\|Y - Zg_0\right\|_2^2$ for a finite constant $C > 0$. In sparse settings, this additional constraint is redundant, so we remove it from our estimator of γ_* defined below (a vector that is assumed to be sparse). Hence, we consider the following estimator, $\hat{\gamma}$

$$\hat{\gamma} = \arg\min_{\gamma\in\mathbb{R}^{p-1}} \|\gamma\|_1 \quad s.t \quad \left\|n^{-1}\tilde{W}^\top(Z - \tilde{W}\gamma)\right\|_\infty \le \lambda n^{-1/2}\|Z\|_2, \tag{13}$$

where $\lambda \asymp \sqrt{n^{-1}\log p}$ is a scale-free tuning parameter and $n^{-1/2}\|Z\|_2$ serves as an upper bound of the unknown σ_u in the model (9). It is worth pointing out that the defined estimators change with a change in the hypothesis testing problem (2) through the new, synthesized, and stabilized feature vectors $\tilde{W}$ and Z together with g_0. We present a few examples in Section 4.

3.2.3. Implementation

The optimization problem in (12), a generalization of the Dantzig selector (Candes and Tao 2007), can be recast as a linear program; the computational burden of our method is comparable to the Dantzig selector. Define scalars $d_1 = \rho_0\|Y - Zg_0\|_2^2/2$, $d_2 = \|Y - Zg_0\|_2^2$, vectors $D_1 = \tilde{W}^\top(Y - Zg_0) \in \mathbb{R}^{p-1}$ and $D_2 = \sqrt{n}\eta\|Y - Zg_0\|_2\mathbf{1}_{p-1}$ and matrix $D_3 = \tilde{W}^\top\tilde{W} \in \mathbb{R}^{(p-1)\times(p-1)}$.

Then, (12) is equivalent to the following linear program:

$$\begin{array}{lccc} \min_{(c,\pi,\rho)\in\mathbb{R}^{p-1}\times\mathbb{R}^{p-1}\times\mathbb{R}} & \mathbf{1}_{p-1}^\top c & & \\ s.t. & -c \le & \pi & \le c \\ & \rho_0 \le & \rho & \le 1 \\ & & d_1\rho + D_1^\top\pi & \le d_2 \\ & -D_2\rho \le & D_1 - D_3\pi & \le D_2\rho, \end{array} \tag{14}$$

where the optimization variables are $c \in \mathbb{R}^{p-1}$, $\pi \in \mathbb{R}^{p-1}$ and $\rho \in \mathbb{R}$. For application purposes, we propose to choose the following choices of the tuning parameters: $\rho_0 = 0.01$ and $\eta = \sqrt{2\log(p)/n}$. They are universal choices and we show in simulations that they provide good results.

3.3. Test Statistic

With defined estimators of γ_* and π_*, we are ready to define a sample analog of the moment condition 10. Under our proposed method, a test of nominal size $\alpha \in (0, 1)$ rejects H_0 in (2) if $|S_n| > \Phi^{-1}(1 - \alpha/2)$, where

$$S_n = \sqrt{n}\frac{(Z - \tilde{W}\hat{\gamma})^\top(Y - Zg_0 - \tilde{W}\hat{\pi})}{\|Z - \tilde{W}\hat{\gamma}\|_2\|Y - Zg_0 - \tilde{W}\hat{\pi}\|_2}. \tag{15}$$

Other estimators of the first moment (10) are certainly possible, however we focus and analyze the natural case above; we leave future efficiency studies for future work since it is not apparent that any other choice is preferred. Moreover, the self-normalizing statistic above is directly dependent on the hypothesis of interest and is a function of synthesized features. Compared with the existing approaches where the normalization factor is a consistent estimator of the asymptotic variance, our self-normalized approach adopts an inconsistent estimator as the normalization factor, which in a sense corresponds to "inefficient Studentizing" (see Shao 2010). However, we establish that the asymptotic distribution of the resulting statistic is pivotal and its percentiles can be obtained from the normal distribution.

In constructing estimates of γ_* and π_*, we do not impose any assumption regarding the sparsity of π_* or β_*. Notice that, except for the case of sparse a, it is in general unreasonable to expect sparsity in π_*, even if β_* is sparse. Although we use estimates for both γ_* and π_* denoted by $\hat{\gamma}$ and $\hat{\pi}$, respectively, we only require l_1 consistency properties for $\hat{\gamma}$; in fact, $\hat{\pi}$ only serves to satisfy our decoupling argument in the proof and does not need to be consistent. We now briefly explain this point. The constraints imposed in the estimator (12) guarantee that for the test statistic S_n, the term $n^{-1/2}(Z - \tilde{W}\hat{\gamma})^\top(Y - Zg_0 - \tilde{W}\hat{\pi})$ can be approximated by a product of two independent terms, that is, $n^{-1/2}(Z - \tilde{W}\gamma_*)^\top(Y - Zg_0 - \tilde{W}\hat{\pi})$. Then, the only requirement needed is to guarantee that the second term in the last expression does not grow to fast (it does not need to converge to zero) which in turn is provided by the constraints of the optimization problem (12).

3.4. Theoretical Properties

In deriving the theoretical properties of our test, we impose the following assumption.

Assumption 2. Let (i) x_i and ε_i have Gaussian distributions, $\mathcal{N}(0, \Sigma_X)$ and $\mathcal{N}(0, \sigma_\varepsilon^2)$, respectively. Moreover, assume (ii) that there exist constants $c_1, c_2 > 0$, such that σ_ε and the eigenvalues of Σ_X lie in $[c_1, c_2]$. Lastly, let (iii) there exist constants $c_3, c_4 \in (0, 1)$, such that $\sigma_u^2/\sigma_z^2 \ge c_3$ and $\sigma_\varepsilon^2/\sigma_y^2 \ge c_4$.

Assumption 2(i) is only imposed to simplify the proof. In high-dimensional literature, Gaussian design is a very common assumption (e.g., Javanmard and Montanari 2014b; Cai and Guo 2017). The same results, at the expense of more complicated proofs, can be derived for sub-Gaussian designs and errors. Assumption 2(ii) is very standard in high-dimensional literature (see Bickel, Ritov, and Tsybakov 2009; Ning and Liu 2017; Van de Geer et al. 2014 for more details).

Assumption 2(iii) imposes nondegeneracy of signal-to-noise ratios for models (1) and (9). Since $\|a\|_2$ is allowed to tend to infinity, $\sigma_z^2 = a^\top\Sigma_X a/(a^\top a)^2$ can tend to zero and thus it is too restrictive to assume that σ_u is bounded away from zero. Hence, Assumption 2(iii) is a relaxation, as it only rules out the uninteresting case of asymptotic noiselessness.

Remark 4. The sparsity condition is imposed on neither a nor β_*. Theorem 3 below says that we can conduct valid inference of a nonsparse linear combination of a nonsparse high-dimensional parameter without knowing Σ_X. To the best of our knowledge, this is the first result that allows for such generality.

Theorem 3. Let Assumption 2 hold. Consider estimators (12) and (19) with suitable choice of tuning parameters: $\eta, \lambda \asymp \sqrt{n^{-1}\log p}$, $\rho_0^{-1} = O(1)$ and $\rho_0 \le [1 + c_2c_1^{-1}(c_3^{-1} - 1)]^{-1/2}$. Suppose that $\|\gamma_*\|_0 = o(\sqrt{n}/\log p)$. Then, under H_0 in (2), optimization problems (12) and (19) are feasible with probability approaching one and

$$\lim_{n,p\to\infty} P\left(|S_n| > \Phi^{-1}(1 - \alpha/2)\right) = \alpha \qquad \forall\alpha \in (0, 1),$$

where S_n is defined in Equation (15).

Theorem 3 establishes that the proposed test is asymptotically exact regardless of how sparse the model parameter or the loading vector are. In that sense, the result is unique in the existing literature as it covers cases of β sparse and a sparse (SS), β sparse and a dense (SD), β dense and a sparse (DS), and especially β dense and a dense (DD). The (SS) case appears in a number of existing works (see Belloni, Chernozhukov, and Hansen 2014; Van de Geer et al. 2014; Javanmard and Montanari 2014b; Ning and Liu 2017), case (SD) appears in Cai and Guo (2017). Whenever (SS) case holds, our result above matches the above-mentioned work, see Theorem 4. In the special setting of (SD), our result generalizes the one of Cai and Guo (2017) as Theorem 3 does not impose any restriction on the size of the loading vector a. The last two cases of (DS) and (DD) present extremely challenging cases in which inference based

on estimation (much like Wald or Rao or Likelihood principles) fails due to the inherit limit of detection—work of Cai and Guo (2016) provides details of impossibility of estimation in such settings. However, despite these challenges our method is able to provide asymptotically valid inference as we have developed inference based on a specifically designed moment condition (and not a parameter estimation alone).

The result in Theorem 3 is based on the assumption that $\hat{\pi}_*$ is a possibly inconsistent estimator of the parameter vector π_*, that is, the full model is dense with all nonzero entries. In the following, we will show that if the model is a sparse model, the proposed test (15) maintains strong power properties. To facilitate the mathematical derivations, we consider the local alternatives of the form

$$H_{1,n}: \ a^\top \beta_* = g_0 + n^{-1/2}(a^\top \Omega_X a)^{1/2}\sigma_\varepsilon d, \tag{16}$$

where $d \in \mathbb{R}$ is a fixed constant. The following result shows that the proposed test achieves certain optimality in detecting alternatives $H_{1,n}$.

Theorem 4. Consider z_i and w_i defined in (7). Let Assumption 2 hold and consider the choice of tuning parameters, as in Theorem 3. Suppose that $\|\gamma_*\|_0 \vee \|\beta_*\|_0 \vee \|a\|_0 = o(\sqrt{n}/\log p)$. Then, under $H_{1,n}$ in (16), optimization problems (12) and (19) are feasible with probability approaching one and

$$\lim_{n,p\to\infty} P\left(|S_n| > \Phi^{-1}(1-\alpha/2)\right) = \Psi_\alpha(d) \qquad \forall \alpha \in (0,1),$$

where $\Psi_\alpha(d) := \Phi(-\Phi^{-1}(1-\alpha/2)+d) + \Phi(-\Phi^{-1}(1-\alpha/2)-d)$.

To better understand the optimality of the result above, consider the estimator (possibly infeasible) discussed at the end of Section 2: let $\bar{\beta}$ denote an estimator satisfying $\sqrt{n}(\bar{\beta}-\beta_*) \sim \mathcal{N}(0, \Omega_X\sigma_\varepsilon^2)$. Notice that, for the low-dimensional components of β_*, $\bar{\beta}$ achieves semiparametric efficiency; see Robinson (1988). Therefore, for sparse a, $a^\top\bar{\beta}$ is a semiparametrically efficient estimator for $a^\top\beta_*$. Notice that $\sqrt{n}(a^\top\bar{\beta} - a^\top\beta_*) \sim \mathcal{N}(0, a^\top\Omega_X a\sigma_\varepsilon^2)$. Based on such efficient estimator, one might consider an "oracle" test: for a test of nominal size α, reject the null $H_0: \ a^\top\beta_* = g_0$ if and only if

$$\frac{\sqrt{n}|a^\top\bar{\beta} - g_0|}{(a^\top\Omega_X a)^{1/2}\sigma_\varepsilon} > \Phi^{-1}(1-\alpha/2). \tag{17}$$

It is easy to verify that the power of this "oracle" test of nominal size α against the local alternatives $H_{1,n}$ (16) is asymptotically equal to $\Psi_\alpha(d)$. Therefore, Theorem 4 says that our test asymptotically achieves the same power as the "oracle" test under sparse a and β_*, that is, it is as efficient as the "oracle" test.

Moreover, in light of recent inferential results in the high-dimensional sparse models, the rate of Theorem 4 can also be shown to be optimal. We discuss in particular the case of Van de Geer et al. (2014) (VBRD) and Belloni, Chernozhukov, and Hansen (2014) (BCH). VBRD uses one-step correction technique to correct the bias of the lasso estimator and then proceeds to construct a Wald test around such unbiased estimator. BCH differs from VBRD in that it uses two different variable selection steps followed by a final estimation step that unifies the results of the two selected sets and corrects the bias by a subsequent refitting step. Both methods are shown to be optimal for $a = e_j$ where e_j is a coordinate vector, $1 \le j \le p$. We discuss the relations of our work in this specific setting. We note that the tests based on VBRD and BCH are asymptotically equivalent to the "oracle" test (17) and hence have the same asymptotic local power; the power of Wald or Score inferential methods (see Theorem 2.2 in Van de Geer et al. (2014), Theorem 1 in Belloni, Chernozhukov, and Hansen (2014) or Theorem 4.7 in Ning and Liu (2017)) and that of Javanmard and Montanari (2014b) (see Theorem 2.3 therein) is asymptotically equal to and converges to $\Psi_\alpha(d)$, respectively. This, in turn, implies that the proposed method is semiparametrically efficient and asymptotically minimax. For vectors a that have more than one nonzero coordinate, we can only compare our work with that of Cai and Guo (2017), where we observe that the result of Theorems 1 and 3 therein matches those of Theorem 4 covering the case of extremely sparse beta and potentially dense vectors a. However, observe that the confidence intervals developed therein require specific knowledge of the sparsity of the parameter β_*, $\|\beta_*\|_0$, a quantity rarely known in practice. Unlike their method, our method can be directly implemented without the knowledge of the sparsity of β_* and yet achieves the same optimality guarantees.

4. Applications to Nonsparse High-Dimensional Models

This section is devoted to three concrete applications of the general methodological results developed in Sections 2 and 3—hence, showcasing the wide impact of the developed theories.

4.1. Testing Pairwise Homogeneity

The previous section deals with situations in which each coordinate of the parameters is allowed to vary independently and any subset of the coordinates can be nonzero simultaneously. This condition will not be satisfied if we are interested in testing pairwise homogeneity in the linear model (group effect), that is, if we are interested in testing the hypothesis

$$H_0: \beta_{*,k} = \beta_{*,j}$$

for $k, j \in \{1, 2, \ldots, p\}$ while also allowing β to be a dense and high-dimensional vector. To the best of our knowledge, such tests were not designed in the existing literature. The proposed methodology easily extends to this case, where the loading vector a takes the form $a = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^\top$, with the location of the 1's at the jth and kth coordinate, respectively. Without loss of generality, we assume that $k = 1$ and $j = 2$. Then it is not hard to show that $z_i = (x_{i,1} - x_{i,2})/2$ and $\tilde{w}_i = ((x_{i,1} + x_{i,2})/\sqrt{2}, x_{i,3}, \ldots, x_{i,p})^\top \in \mathbb{R}^{p-1}$. The proposed methodology for this problem simplifies, then, to finding $\hat{\pi}$ and $\hat{\rho}$ that satisfy

$$\begin{aligned}(\hat{\pi}, \hat{\rho}) \quad = \quad & \underset{(\pi,\rho)\in\mathbb{R}^{p-1}\times\mathbb{R}_+}{\arg\min} \|\pi\|_1 \\ s.t \quad & \tilde{W} = [(X_1 + X_2)/\sqrt{2}, X_3, \ldots, X_p] \\ & \|\tilde{W}^\top(Y - \tilde{W}\pi)\|_\infty \le \eta\rho\sqrt{n}\|Y\|_2 \\ & Y^\top\left(Y - \tilde{W}\pi\right) \ge \rho_0\,\rho\,\|Y\|_2^2/2 \\ & \rho \in [\rho_0, 1]\end{aligned} \tag{18}$$

and $\hat{\gamma}$ that satisfies

$$\begin{aligned}\hat{\gamma} \quad &= \quad \underset{\gamma\in\mathbb{R}^{p-1}}{\arg\min}\,\|\gamma\|_1 \\ \text{s.t} \quad &\tilde{W} = [(X_1+X_2)/\sqrt{2}, X_3, \ldots, X_p] \\ &\|\tilde{W}^\top(X_1 - X_2 - 2\tilde{W}\gamma)\|_\infty \le \lambda\sqrt{n}\|X_1 - X_2\|_2, \quad (19)\end{aligned}$$

for $\lambda, \eta \asymp \sqrt{n^{-1}\log p}$.

Consequently, we reject $H_0 : \beta_{*,1} = \beta_{*,2}$ if $|S_n| > \Phi^{-1}(1-\alpha/2)$, where

$$S_n = \sqrt{n}\frac{(X_1 - X_2 - 2\tilde{W}\hat{\gamma})^\top(Y - \tilde{W}\hat{\pi})}{\|X_1 - X_2 - 2\tilde{W}\hat{\gamma}\|_2\|Y - \tilde{W}\hat{\pi}\|_2}. \quad (20)$$

4.2. Inference of Conditional Mean

Our methodology can also be used for the inference regarding the average value of the response, that is, regarding the conditional mean of the regression model. Suppose that the object of interest is $E(y_i \mid \zeta_i)$, where $y_i \in \mathbb{R}$ and $\zeta_i \in \mathbb{R}^k$. For a given value $d \in \mathbb{R}^k$ and $g_0 \in \mathbb{R}$, the focus is to test

$$H_0 : E(y_i \mid \zeta_i = d) = g_0.$$

Assuming that for some given dictionary of transformations of $\{\phi_j(\cdot)\}_{j=1}^p$, the conditional mean function admits the representation: $E(y_i \mid \zeta_i) = \sum_{j=1}^p \beta_{*,j}\phi_j(\zeta_i)$ for some vector $\beta_* = (\beta_{*,1}, \ldots, \beta_{*,k})^\top \in \mathbb{R}^p$. Then the conditional mean model can be written as

$$y_i = x_i^\top\beta_* + \varepsilon_i, \quad (21)$$

where $x_i = (\phi_1(\zeta_i), \ldots, \phi_p(\zeta_i))^\top \in \mathbb{R}^p$ and $E(\varepsilon_i \mid x_i) = 0$. In turn, the confidence intervals for the regression mean can be designed simply by inverting the test statistics

$$S_n = \sqrt{n}\frac{(Z - \tilde{W}\hat{\gamma})^\top(Y - Zg_0 - \tilde{W}\hat{\pi})}{\|Z - \tilde{W}\hat{\gamma}\|_2\|Y - Zg_0 - \tilde{W}\hat{\pi}\|_2}$$

designed for the inference problem

$$H_0 : a^\top\beta_* = g_0,$$

where $a = (\phi_1(d), \ldots, \phi_p(d))^\top \in \mathbb{R}^p$ and $U_aU_a^\top = (I_p - aa^\top/\sum_{j=1}^p\phi_j^2(d))$ with

$$z_i = \frac{\sum_{j=1}^p\phi_j(d)\phi_j(\zeta_i)}{\sum_{j=1}^p\phi_j^2(d)},$$

$$\text{and} \qquad \tilde{w}_{ij} = \sum_{l=1}^p\{U_a\}_{lj}\phi_l(\zeta_i), \qquad 1 \le j \le p-1.$$

Notice that we do not assume that the vector β_* is sparse and we allow for $p \gg n$. Therefore, representing the conditional mean function in terms of a large number of transformations of ζ_i, while simultaneously allowing all to be nonzero, does not lose much in generality. Additionally, it is worth mentioning that inference for such models has not been addressed in the existing literature: most of the existing work is strictly focused around sparse or sparse additive models. With the general model considered here, one can consider tests regarding treatment effects (when viewed as the conditional mean) and allow for fully dense models and loading vectors, that is, the treatment being a dense combination of many variables. Existing work, such as Belloni, Chernozhukov, and Hansen (2014), only allows the treatment to be a single variable.

4.3. Decomposition of Conditional Mean

In practice, the researcher might be interested in how much a certain group of features contribute to the conditional mean. Let $\mathcal{G} \subseteq \{1, \ldots, p\}$. The goal is to conduct inference on linear functionals of $\{\beta_{*,j}\}_{j\in\mathcal{G}}$, that is, $\sum_{j\in\mathcal{G}} c_j\beta_{*,j}$ for some known $\{c_j\}_{j\in\mathcal{G}}$.

For example, consider the notations from Section 4.2. Let $\zeta_i = (\zeta_{i,1}, \ldots, \zeta_{i,k})^\top$ and suppose that one is interested in the impact of $\zeta_{i,1}$ on the conditional mean for $\zeta = d$. This is equivalent to quantifying $\sum_{j\in\mathcal{G}_1}\phi_j(d)\beta_{*,j}$, where the set contains all the indexes j such that the first entry of ζ_i has nonzero effect on $\phi_j(\zeta_i)$, that is, $\mathcal{G}_1 = \{j : \phi_j(\zeta) \text{ is not constant in } \zeta_1\}$. If $\phi_j(\cdot)$'s are transformations of individual entries of $\{\zeta_{i,j}\}_{j=1}^k$, then $\mathcal{G}_1$ corresponds to transformations of $\zeta_{i,1}$. For another example, suppose that all the p features are genes. The domain scientist (biologist, doctor, geneticist, etc.) might be interested in how much a group of genes contributes to the expected value of the response variable.

Without loss of generality, we assume that $\mathcal{G} = \{1, \ldots, H\}$ and $c = (c_1, \ldots, c_H)^\top \in \mathbb{R}^H$. Let $U_c \in \mathbb{R}^{H\times(H-1)}$ satisfy $I_H - cc^\top/(c^\top c) = U_cU_c^\top$ and $U_c^\top U_c = I_{H-1}$. Then the synthesized features can be constructed by $z_i = \|c\|_2^{-2}\sum_{j=1}^H c_jx_{i,j}$ and $\tilde{w}_i = (\sum_{l=1}^H(U_c)_{l,1}x_{i,l}, \ldots, \sum_{l=1}^H(U_c)_{l,H-1}x_{i,l}, x_{i,H}, \ldots, x_{i,p})^\top \in \mathbb{R}^{p-1}$, where $(U_c)_{l,j}$ denotes the (l,j) entry of the matrix U_c. For example, whenever $H = 3$ and $c_j = 1$ for all $j = 1, 2, 3$, then

$$U_c = \begin{pmatrix} -\sqrt{3/2} & -1/\sqrt{2} \\ 0 & \sqrt{2} \\ \sqrt{3/2} & -1/\sqrt{2}\end{pmatrix}$$

and the procedure for testing $\beta_{*,1} + \beta_{*,2} + \beta_{*,3} = g_0$ would be as follows. We define

$$\begin{aligned}(\hat{\pi}, \hat{\rho}) \quad &= \quad \underset{(\pi,\rho)\in\mathbb{R}^{p-1}\times\mathbb{R}_+}{\arg\min}\,\|\pi\|_1 \\ s.t \quad &\tilde{W} = \left[\sqrt{\frac{3}{2}}(X_3 - X_1),\right. \\ &\left. -\frac{1}{\sqrt{2}}(X_1 - 2X_2 + X_3),\ X_4, \ldots, X_p\right] \\ &\|\tilde{W}^\top[Y - (X_1+X_2+X_3)g_0/3 - \tilde{W}\pi]\|_\infty \\ &\le \eta\rho\sqrt{n}\|Y - (X_1+X_2+X_3)g_0/3\|_2 \\ &(Y - (X_1+X_2+X_3)g_0/3)^\top \\ &\quad\times(Y - (X_1+X_2+X_3)g_0/3 - \tilde{W}\pi) \\ &\quad\ge \rho_0\,\rho\,\|Y - (X_1+X_2+X_3)g_0/3\|_2^2/2 \\ &\rho \in [\rho_0, 1] \quad (22)\end{aligned}$$

and $\hat{\gamma}$ that satisfies

$$\begin{aligned}\hat{\gamma} \quad &= \quad \underset{\gamma\in\mathbb{R}^{p-1}}{\arg\min}\,\|\gamma\|_1 \\ \text{s.t} \quad &\tilde{W} = \Big[\sqrt{\frac{3}{2}}(X_3 - X_1), \\ &\quad -\frac{1}{\sqrt{2}}(X_1 - 2X_2 + X_3),\ X_4, \ldots, X_p\Big] \\ &\|\tilde{W}^\top\Big((X_1 + X_2 + X_3)g_0 - 3\tilde{W}\gamma\Big)\|_\infty \\ &\le \lambda\sqrt{n}g_0\|X_1 + X_2 + X_3\|_2, \end{aligned} \tag{23}$$

for $\lambda, \eta \asymp \sqrt{n^{-1}\log p}$.

For a test of nominal size α, we reject $H_0 : \beta_{*,1} + \beta_{*,2} + \beta_{*,3} = g_0$ if $|S_n| > \Phi^{-1}(1-\alpha/2)$, where

$$S_n = \sqrt{n}\frac{((X_1 + X_2 + X_3)g_0 - 3\tilde{W}\hat{\gamma})^\top(Y - (X_1 + X_2 + X_3)g_0/3 - \tilde{W}\hat{\pi})}{\|(X_1 + X_2 + X_3)g_0 - 3\tilde{W}\hat{\gamma}\|_2\|Y - (X_1 + X_2 + X_3)g_0/3 - \tilde{W}\hat{\pi}\|_2}. \tag{24}$$

5. Numerical Results

In this section, we study the finite sample performance of the proposed methodology for both known Σ_X and unknown Σ_X. We explicitly consider dense loadings a and dense parameter vectors β_* as well as more common sparse settings. We have made our code publicly available at *http://www.jelenabradic.net/linear-testing-code.html*.

5.1. Monte Carlo Experiments

Consider the model (1) with the model error following standard normal distribution. In all the simulations, we set $n = 100$ and $p = 500$ and the nominal size of all the tests is 5%. The rejection probabilities are based on 500 repetitions. The null hypothesis we test is $H_0 : a^\top\beta_* = g_0$, where $g_0 = a^\top\beta_* + h$ and h is allowed to vary to capture both Type I and Type II error rates.

5.1.1. Setup

We consider in total four regimes on the structure of the model and the null hypothesis—sparse and dense regimes for β_* as well as sparse and dense regimes for the loading vector a.

(i) In the *Sparse parameter regime,* we consider the parameter structure with $\beta_* = (0.8, 0.8, 0, \ldots, 0)^\top$.
(ii) In the *Dense parameter regime,* we consider the parameter structure with $\beta_* = \frac{3}{\sqrt{p}}(1, 1, \ldots, 1)^\top$.
(iii) In the *Sparse loading regime,* we consider the loading vector $a = (0, 1, 0, \ldots, 0)^\top$.
(iv) In the *Dense loading regime,* we consider the loading vector $a = (1, 1, \ldots, 1)^\top$.

Observe that (iii) is an extreme sparse-loading case. We consider this special case to compare existing inferential methods,

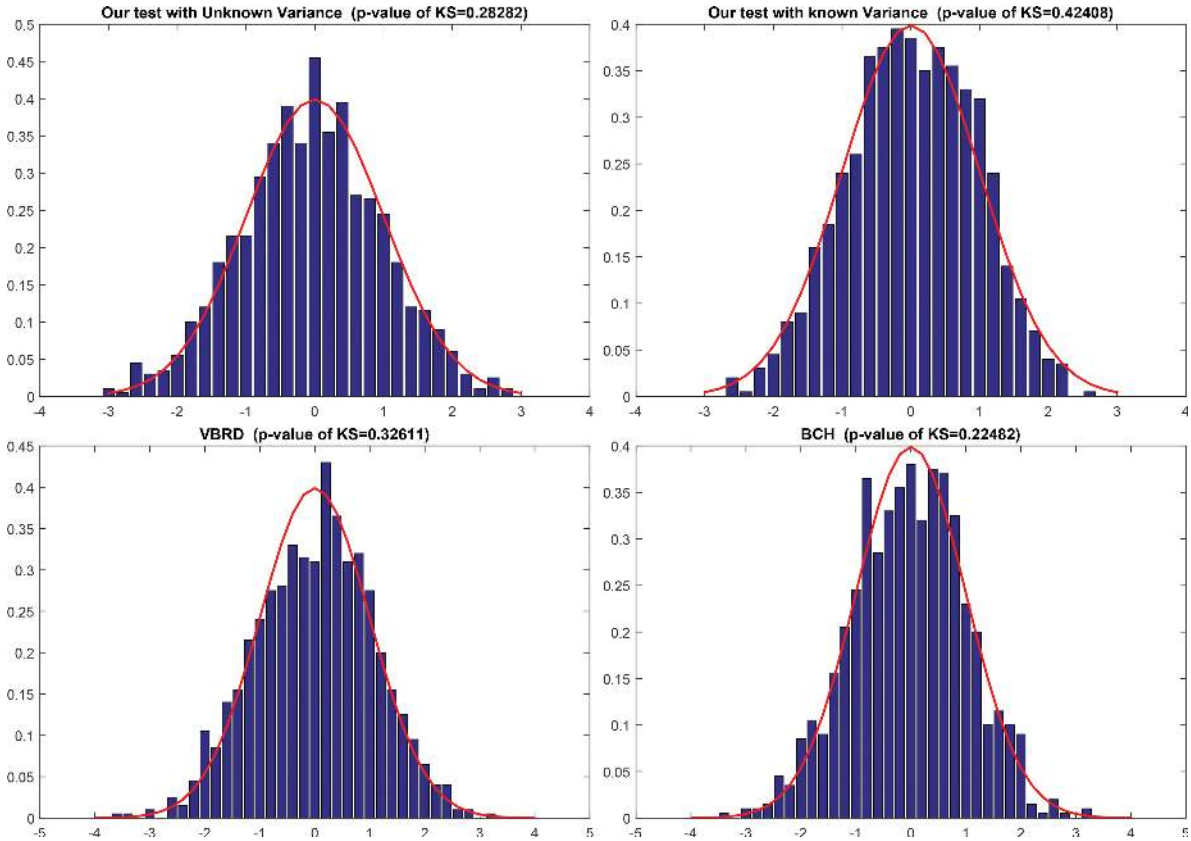

Figure 1. Distribution of the test statistics under the null hypothesis $H_0 : \beta_{*,2} = 0.8$ (in blue) and the standard normal distribution $\mathcal{N}(0,1)$ (in red) with $n = 100$ and $p = 500$. In this example, we consider sparse β and sparse a setting and compare the distribution under the null of our tests (with and without known variance) in the top row and two competing methods VBRD and BCH in the bottom row. We report p-values of the Kolmogorov–Smirnov test statistics in the subtitles. Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. Error and design are normally distributed with Toeplitz correlation structure with $\rho = 0.4$. The histograms are computed based on 500 simulation runs.

like VBRD and BCH. However, our method can be implemented for various number of nonzero elements, whereas the existing one cannot.

We present results for three different designs settings including sparse, dense, Gaussian, and non-Gaussian settings.

Example 1. Here we consider the standard Toeplitz design where the rows of X are drawn as an iid random draws from a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma_X)$, with covariance matrix $(\Sigma_X)_{i,j} = 0.4^{|i-j|}$.

Example 2. In this case, we consider a nonsparse design matrix with equal correlations among the features. Namely, rows of X are iid draws from the multivariate Gaussian distribution $\mathcal{N}(0, \Sigma_X)$, where $(\Sigma_X)_{i,j}$ is 1 for $i = j$ and is 0.4 for $i \neq j$. Observe that this case is particularly hard for most inferential methods as all features are interdependent and Ω_X is not sparse.

Example 3. In this example, we consider a highly non-Gaussian design that also has strong dependence structure. We consider the setting of Fan and Song (2010). We repeat the details here for the convenience of the reader. Let x be a typical row of X. For $j \in \{1, \ldots, 15\}$, $x_j = (\xi + c\xi_j)/\sqrt{1+c^2}$, where ξ and $\{\xi_j\}_{j=1}^{15}$ are iid $\mathcal{N}(0, 1)$ and c is chosen such that $\text{corr}(x_1, x_2) = 0.4$. For $j \in \{16, \ldots, [p/3]\}$, x_j is iid $\mathcal{N}(0, 1)$. For $j \in \{[p/3] + 1, \ldots, [2p/3]\}$, x_j is iid from a double exponential distributions with location parameter zero and scale parameter one. For $j \in \{[2p/3] + 1, \ldots, p\}$, x_j is iid from the half-half mixture of $\mathcal{N}(-1, 1)$ and $\mathcal{N}(1, 0.5)$. Observe that in this case 2/3 of the features follow non-Gaussian distributions. Thus, in this case it is extremely difficult to even obtain consistent estimation of the model parameters.

5.1.2. Implementation Details

We compare the proposed tests with VBRD and BCH; methods proposed in Cai and Guo (2017) contain constants whose values could be very conservative in finite samples. Our tests with known and unknown Σ_X are implemented as discussed in Sections 2 and 3, respectively.

The VBRD method is implemented for both dense and sparse loadings as follows. We first compute the debiased estimator $\hat{\beta}_{\text{debias}}$ and the nodewise Lasso estimator $\hat{\Omega}_{\text{Lasso}}$ for the precision matrix Σ_X as in VBRD. Then test is to reject H_0 if and only if

$$\sqrt{n}|a^\top \hat{\beta}_{\text{debias}} - g_0|/\sqrt{a^\top \hat{\Omega}_{\text{Lasso}} \hat{\Sigma}_X \hat{\Omega}_{\text{Lasso}}^\top a \sigma_\varepsilon^2} > \Phi^{-1}(1 - 0.05/2).$$

The BCH method is only implemented for the sparse loadings. We compute the generic post-double-selection estimator for the second entry of β as in eq. (2.8) of BCH and compute the standard error as in Theorem 2 therein. Then a usual t-test is conducted. It is not clear how BCH can be extended to handle any loading vector a different from an extremely sparse case (see

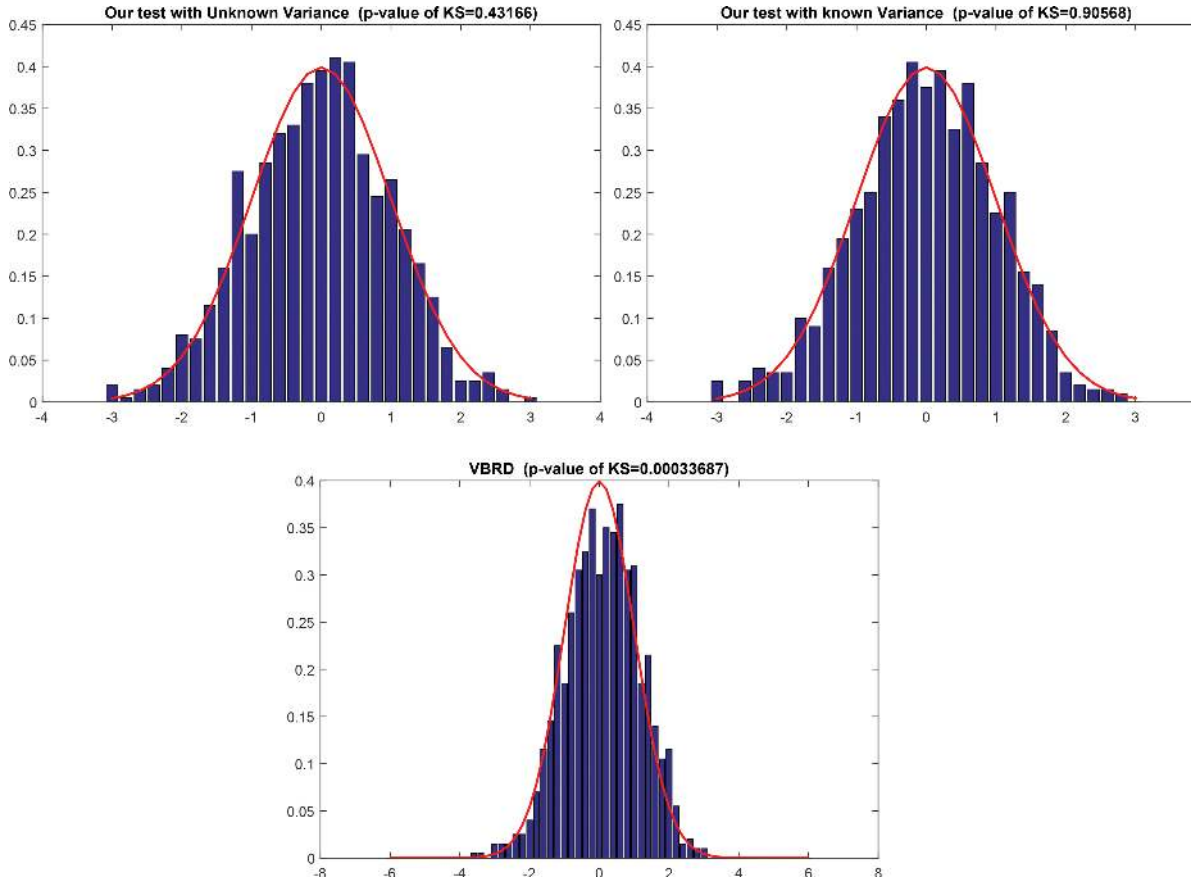

Figure 2. Distribution of the test statistics under the null hypothesis $H_0 : \sum_{j=1}^{p} a_j \beta_{*,j} = 1.6$ (in blue) and the standard normal distribution $\mathcal{N}(0, 1)$ (in red) with $n = 100$ and $p = 500$. In this example, we consider sparse β and dense a setting and compare the distribution under the null of our tests (with and without known variance) in the top row and two competing methods VBRD and BCH in the bottom row. We report p-values of the Kolmogorov–Smirnov test statistics in the subtitles. Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. Error and design are normally distributed with Toeplitz correlation structure with $\rho = 0.4$. The histograms are computed based on 500 simulation runs.

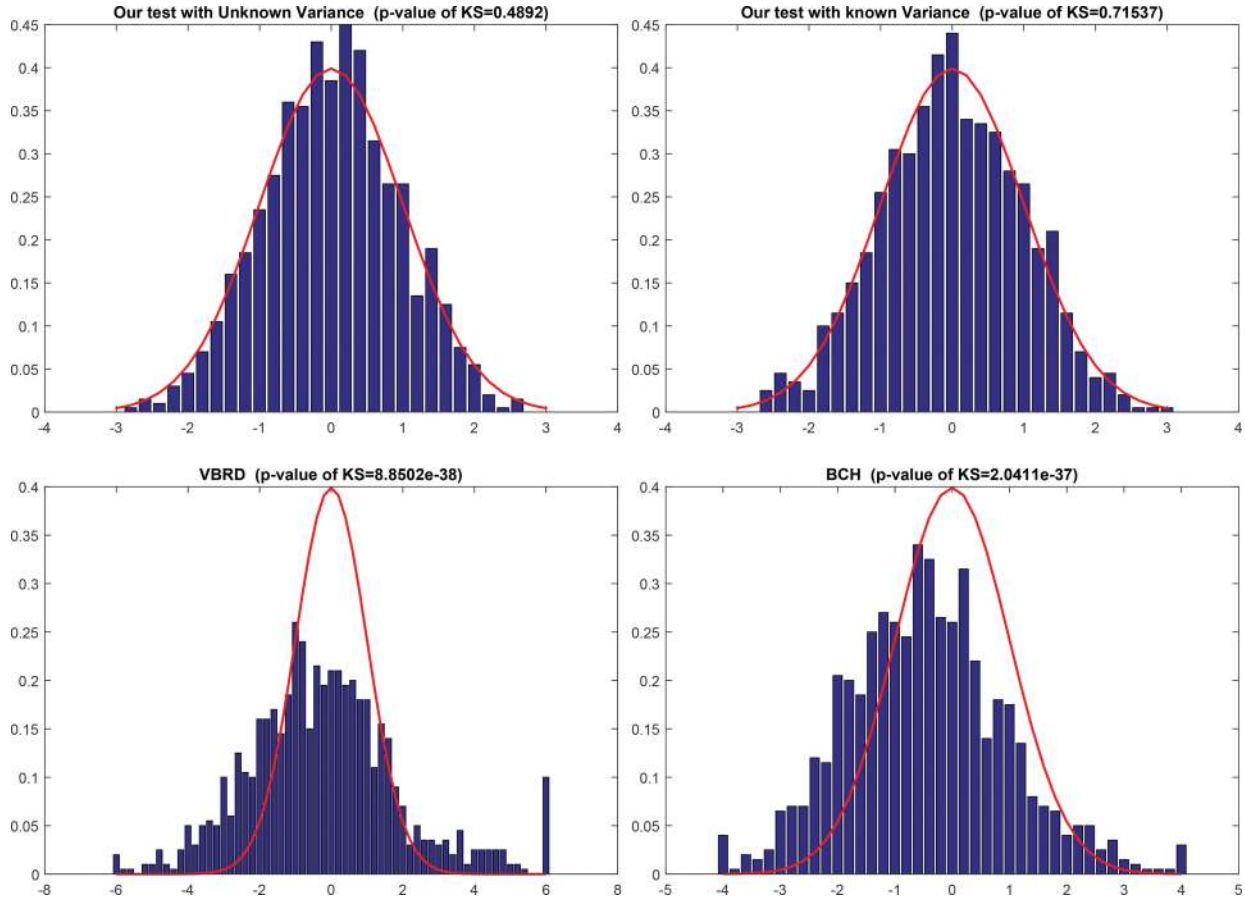

Figure 3. Distribution of the test statistics under the null hypothesis $H_0 : \beta_{*,2} = 3/\sqrt{p}$ (in blue) and the standard normal distribution $\mathcal{N}(0, 1)$ (in red) with $n = 100$ and $p = 500$. In this example, we consider dense β and sparse a setting and compare the distribution under the null of our tests (with and without known variance) in the top row and two competing methods VBRD and BCH in the bottom row. We report p-values of the Kolmogorov–Smirnov test statistics in the subtitles. Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. Error and design are normally distributed with Toeplitz correlation structure with $\rho = 0.4$. The histograms are computed based on 500 simulation runs.

(iii) above): first, for any other loading structure it is not defined how to gather selected features of what would be multiple simultaneous equations; second, naively extending the original BCH to the problem of dense a ($\|a\|_0 = p$) means running an OLS regression of the response against all the features, which is not feasible for $p > n$.

5.1.3. Results

We start with the size properties of competing tests. For this purpose, we examine the distributions of the test statistics under the null hypothesis by comparing empirical distributions of the tests with the theoretical benchmark of standard normal random variable. For simplicity of presentation, we only consider the Toeplitz design. For the testing problem with sparse β_* and sparse a, our tests, VBRD, and BCH exhibit the validity guaranteed by the theory; in Figure 1, the histograms of the test statistics are close to $\mathcal{N}(0, 1)$ with large p-values of the Kolmogorov–Smirnov (KS) tests. For all the other problems, our tests outperform existing methods. As shown in Figure 2, the histogram of VBRD test visually is still close to the standard normal distribution but the KS test suggests discernible discrepancies between the two distributions. In Figure 3, we see that lack of sparsity in β_* causes serious problems in Type I error for both VBRD and BCH. Inference under dense β_* and dense a turns out to be the most challenging problem for existing methods; in Figure 4, we see quite noticeable difference between the histogram of VBRD test and $\mathcal{N}(0, 1)$. In contrast, the distribution of the test statistics of the proposed methods closely matches $\mathcal{N}(0, 1)$ in all the scenarios, as established in Theorems 1 and 3. The Type I errors, reported in Table 1, confirm the above findings: existing methods can suffer greatly from lack of sparsity in β_* and/or a in terms of validity—observed Type I error of BCH or VBRD can easily reach 40%.

We also contrast the power properties of the proposed tests with respect to the existing methods. Results are collected in Figures 5–7, where we plot the power curves of competing methods for design Examples 1, 2, and 3 described above with hypothesis setting of (i)–(iv). The overall message is clear from

Table 1. Type I error computed over 500 repetitions of the 5% level proposed tests together with VBRD and BCH. In the table, NA symbol indicates that the method cannot be implemented "as is."

	Type I error			
Hypothesis setting	Unknown Σ_X	Known Σ_X	VBRD	BCH
Sparse β and Sparse a	7.4%	5.6%	8.2%	6.6%
Sparse β and Dense a	4.4%	4.8%	7.4%	NA
Dense β and Sparse a	3.6%	4.4%	33.4%	27.2%
Dense β and Dense a	5.6%	3.0%	67.2%	NA

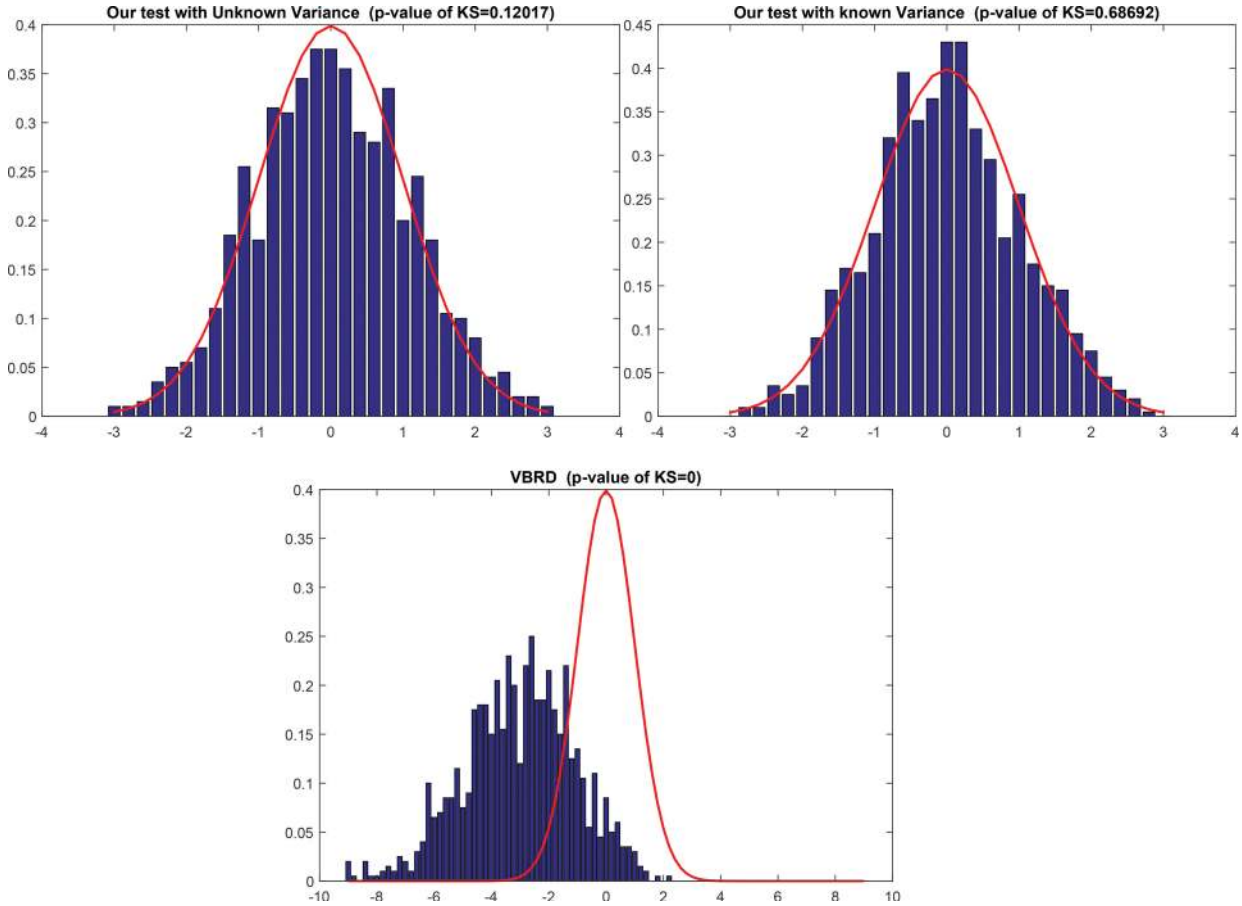

Figure 4. Distribution of the test statistics under the null hypothesis $H_0 : \sum_{j=1}^{p} \beta_{*,j} = 3\sqrt{p}$ (in blue) and the standard normal distribution $\mathcal{N}(0, 1)$ (in red) with $n = 100$ and $p = 500$. In this example, we consider dense β and dense a setting and compare the distribution under the null of our tests (with and without known variance) in the top row and two competing methods VBRD and BCH in the bottom row. We report p-values of the Kolmogorov–Smirnov test statistics in the subtitles. Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. Error and design are normally distributed with Toeplitz correlation structure with $\rho = 0.4$. The histograms are computed based on 500 simulation runs.

these figures: our tests and existing methods are quite similar for sparse β_* and sparse a, whereas our tests behave nominally for other problems with preserving both low Type I error rates and Type II error rates. The biggest advantages are seen for dense vectors β_* with other methods behaving in a manner close to random guessing. In addition to the advantages in Type I error, our methods also display certain power advantages. In the case of equal-correlation setting, we observe that our methods consistently reach faster power than BCH method even in the case of all sparse setting. Observe that the precision matrix in this setting is not sparse and our methods are still well-behaved. In the case of dense models, VBRD method completely breaks down with Type I or Type II error being close to 1. For non-Gaussian design, we see that VBRD may not be a nominal test any more regardless of the model sparsity. BCH behaves more stably in this case but fails to apply for the hypothesis settings (ii) and (iv) as described at the beginning of the section. In conclusion, we observe that our methods are stable across vastly different designs and model setting whereas existing methods fail to control either Type I error rate or Type II error rate. Hence the proposed methodology offers a robust and more widely applicable alternative to the existing inferential procedures, achieving better error control in difficult setting and not losing much in the simple cases.

5.2. Real Data Example: Equity Risk Premia

We apply the methods developed in Section 3 to inference of equity risk premia during different states of the economy. Some studies have found that the risk premia of stock market returns have different predictability, depending on whether the macroeconomy is in recession or expansion; see Rapach, Strauss, and Zhou (2010), Henkel, Martin, and Nardari (2011), and Dangl and Halling (2012). One common explanation for this is time variation in risk premia; see Henkel, Martin, and Nardari (2011). It is plausible that the stock market is riskier in recessions than in expansions and thus a higher expected return is demanded by investors, implying that the expected stock returns can be predicted by the state of the macroeconomy. In this section, we revisit this argument by directly conducting inference on the expected return of the stock market conditional on a large number of macroeconomic variables.

Let y_t be the excess return of the U.S. stock market observed at time t and $x_{t-1} \in \mathbb{R}^p$ be a large number of macroeconomic variables observed at time $t-1$. Let $s_t \in \{0, 1\}$ denote the NBER recession indicator; $s_t = 1$ means that the economy is in recession at time t. We would like to conduct inference on $E(y_t \mid x_{t-1})$ for the two different values of s_{t-1}. Formally, we wish to construct confidence intervals for the following quantities: (a)

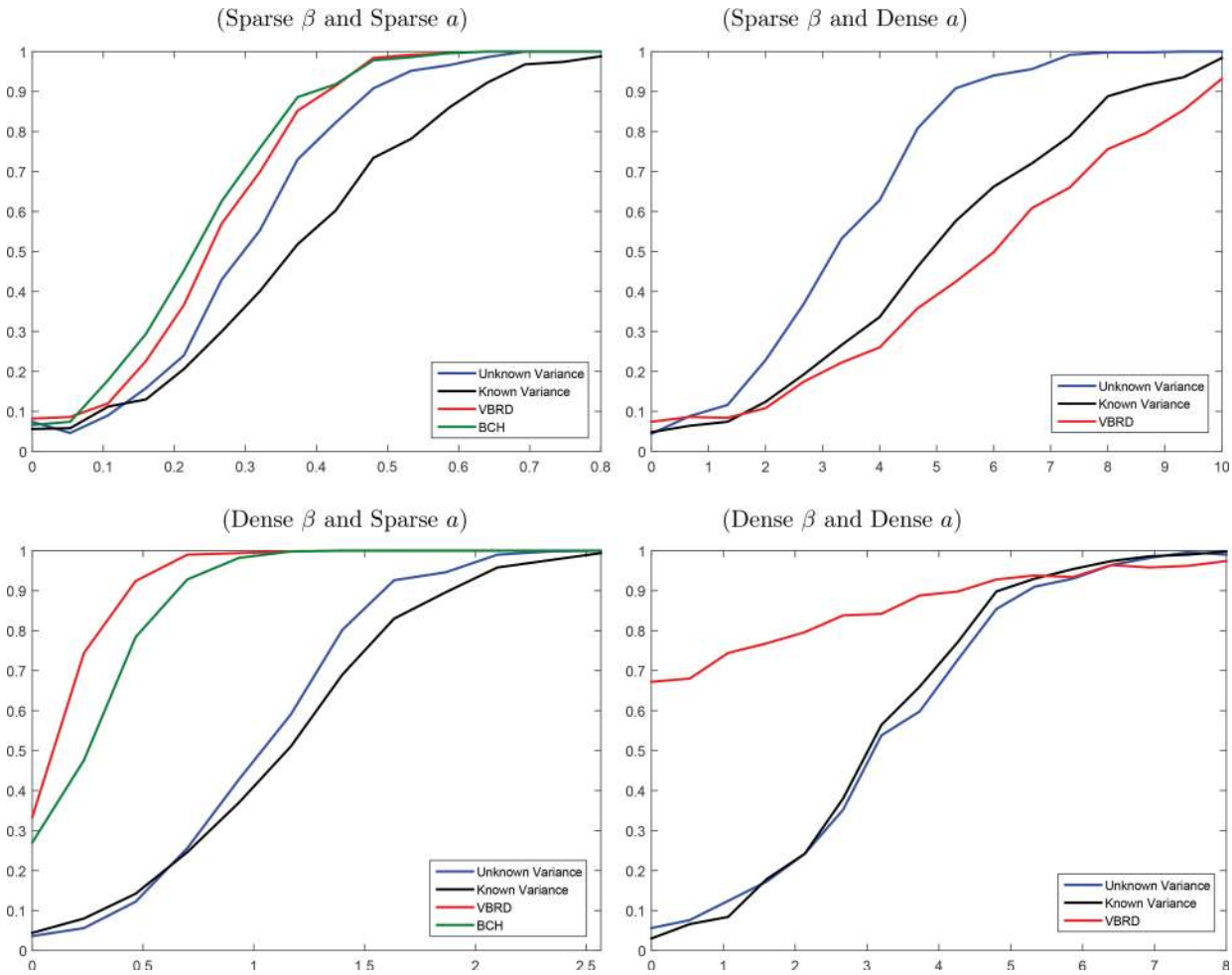

Figure 5. Power curves of competing methods across different hypothesis $a^\top \beta_* = g_0$ settings. Design settings follow Example 1 with $n = 100$ and $p = 500$. The alternative hypothesis takes the form of $a^\top \beta_* = g_0 + h$ with h presented on the x-axis. The y-axis contains the average rejection probability over 500 repetition. Therefore, $h = 0$ corresponds to Type I error and the remaining ones the Type II error. "Known variance" denotes the method as is introduced in Section 2 whereas, "unknown variance" denotes the method introduced in Section 3. VBRD and BCH refer to the methods proposed in Van de Geer et al. (2014) and Belloni, Chernozhukov, and Hansen (2014), respectively. Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. If a method could not be implemented as is proposed in its respective article, it was not included in the graph.

$E[E(y_t \mid x_{t-1}) \mid s_{t-1} = 1]$, (b) $E[E(y_t \mid x_{t-1}) \mid s_{t-1} = 0]$, and (c) $E[E(y_t \mid x_{t-1}) \mid s_{t-1} = 1] - E[E(y_t \mid x_{t-1}) \mid s_{t-1} = 0]$.

We impose a linear model on the risk premia: $E(y_t \mid x_{t-1}) = x_{t-1}^\top \beta_*$ for some unknown $\beta_* \in \mathbb{R}^p$. Hence, the quantities of interest are: $a_1^\top \beta_*$, $a_0^\top \beta_*$, and $(a_1 - a_0)^\top \beta_*$, where $a_j = E(x_{t-1} \mid s_{t-1} = j)$. The macroeconomic variables we use are from the dataset constructed by McCracken and Ng (2015). We also include the squared, cubed, and fourth power of these variables, leading to $p = 440$ (after removing variables with more than 30 missing observations). It is possible that $\beta_* \in \mathbb{R}^p$ is not a sparse vector because many macroeconomic variables might be relevant and each might only explain a tiny fraction of the equity risk premia. Therefore, the methods proposed in this article are particularly useful because they do not assume the sparsity of β_*.

Remark 5. There have been numerous attempts to include information from many macroeconomic variables in estimating the equity risk premium. Rapach, Strauss, and Zhou (2010) used the model combination approach by taking the simple average of 14 univariate linear models. Although this approach manages to reduce the variance in the predictions, it only produces a single point prediction and does not deliver a confidence interval. Moreover, under the specification of $E(y_t \mid x_{t-1}) = x_{t-1}^\top \beta_*$, we should not expect the simple average of predictions by individual components of x_{t-1} to be close to $x_{t-1}^\top \beta_*$, especially with highly correlated regressors. Another popular approach is to use factor models. This method is widely used in macroeconomics for predictions; see Stock and Watson (2002a), Stock and Watson (2002b), and McCracken and Ng (2016). The idea is to extract a few principal components (PC's) from x_t and to predict y_t using these PC's. Although the PC's account for a large variation in x_{t-1}, they are not hard-wired to have high predictive power for y_t unless we assume that the PC's capture the factors that drive y_t. In some sense, this factor approach only uses information in x_{t-1} that is relevant for predicting variations among different components of x_{t-1}; by contrast, the methodology we propose in this article allows us to use all the information in x_{t-1}.

Our dataset has 659 monthly observations starting from 1960. We use the first 20 years ($n = 240$) to train the data and the last $659 - n$ months to compute $a_j = \sum_{t=n+1}^{659} x_t \mathbf{1}\{s_t = j\} / \sum_{t=n+1}^{659} \mathbf{1}\{s_t = j\}$. In other words, we investigate the equity risk premia between 1980 and 2014. We conduct inference on the average equity risk premia in different

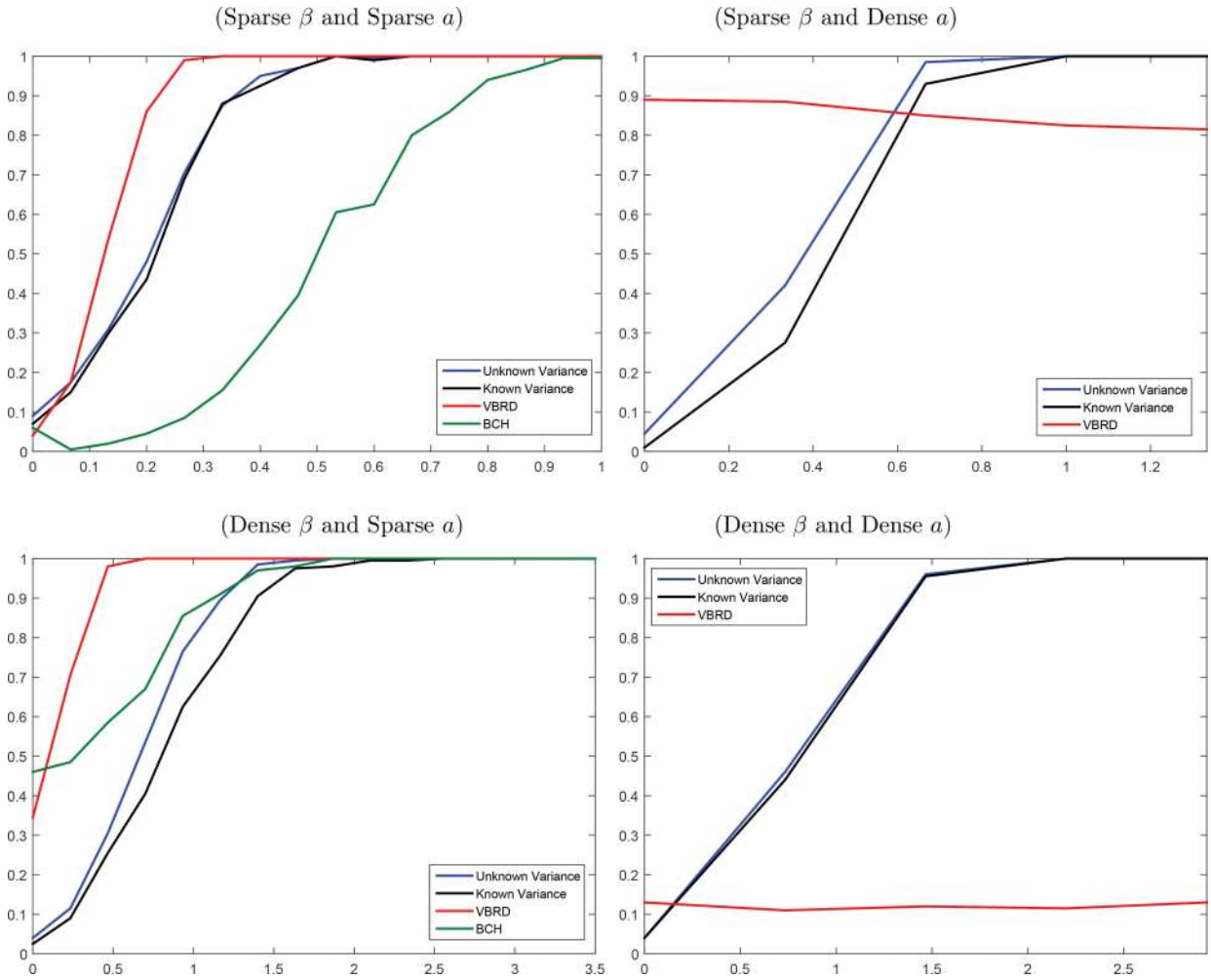

Figure 6. Power curves of competing methods across different hypothesis $a^\top \beta_* = g_0$ settings. Design settings follow Example 2 with $n = 100$ and $p = 500$. The alternative hypothesis takes the form of $a^\top \beta_* = g_0 + h$ with h presented on the x-axis. The y-axis contains the average rejection probability over 500 repetition. Therefore, $h = 0$ corresponds to Type I error and the remaining ones the Type II error. "Known variance" denotes the method as is introduced in Section 2 whereas, "unknown variance" denotes the method introduced in Section 3. VBRD and BCH refer to the methods proposed in Van de Geer et al. (2014) and Belloni, Chernozhukov, and Hansen (2014), respectively. Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. If a method could not be implemented as is proposed in its respective article, it was not included in the graph.

states of the macroeconomy. The 95% confidence intervals for $a_1^\top \beta_*$, $a_0^\top \beta_*$, and $(a_1 - a_0)^\top \beta_*$ are reported in Table 2.

The confidence intervals in Table 2 are very informative for our purpose. The results presented in Table 2 imply that the risk premia in recessions are higher than in expansions and that the magnitude of difference is economically meaningful. These results are consistent with existing literature; see Table 1 of Henkel, Martin, and Nardari (2011). Figure 8 plots the confidence intervals for $E(y_t \mid x_{t-1})$ at each t. This figure is consistent with the hypothesis that, during the Recessions (e.g., in the early 1980s or around 2008), the risk premia went up substantially.

Table 2. 95% confidence intervals for equity risk premia.

	Lower bound	Upper bound
Risk premia in expansion $a_0^\top \beta_*$:	2.79	10.94
Risk premia in recession $a_1^\top \beta_*$:	6.32	36.92
Risk premia difference $(a_1 - a_0)^\top \beta_*$:	5.13	38.30

NOTE: The values are reported in annualized percentage, that is, 2.79 means 2.79%.

Discussions

In this article, we develop new methodology for testing hypotheses on $a^\top \beta_*$, where a is given and β_* is the regression parameter of a high-dimensional linear model. Under the proposed methodology, a new restructured regression and with features that are synthesized and augmented, is constructed based on a and is used to obtain moment conditions that are equivalent to the null hypothesis. Estimators proposed are tailored to the problem at hand and solve constrained high-dimensional optimization problems. The two proposed methods deal with the scenario with known Σ_X and the scenario with unknown Σ_X, respectively. The first can be used when a prior information about correlation among the features exists; a case of independent features, whereas the second applies more broadly to many scientific examples where feature correlations need to be estimated. To solve a high-dimensional inference problem, there exists at least one competing choice. It is based on the "debiasing" principles of Zhang and Zhang (2014). However, the principles laid out therein only apply to strictly sparse linear models. Therefore, we fulfill an important gap in the

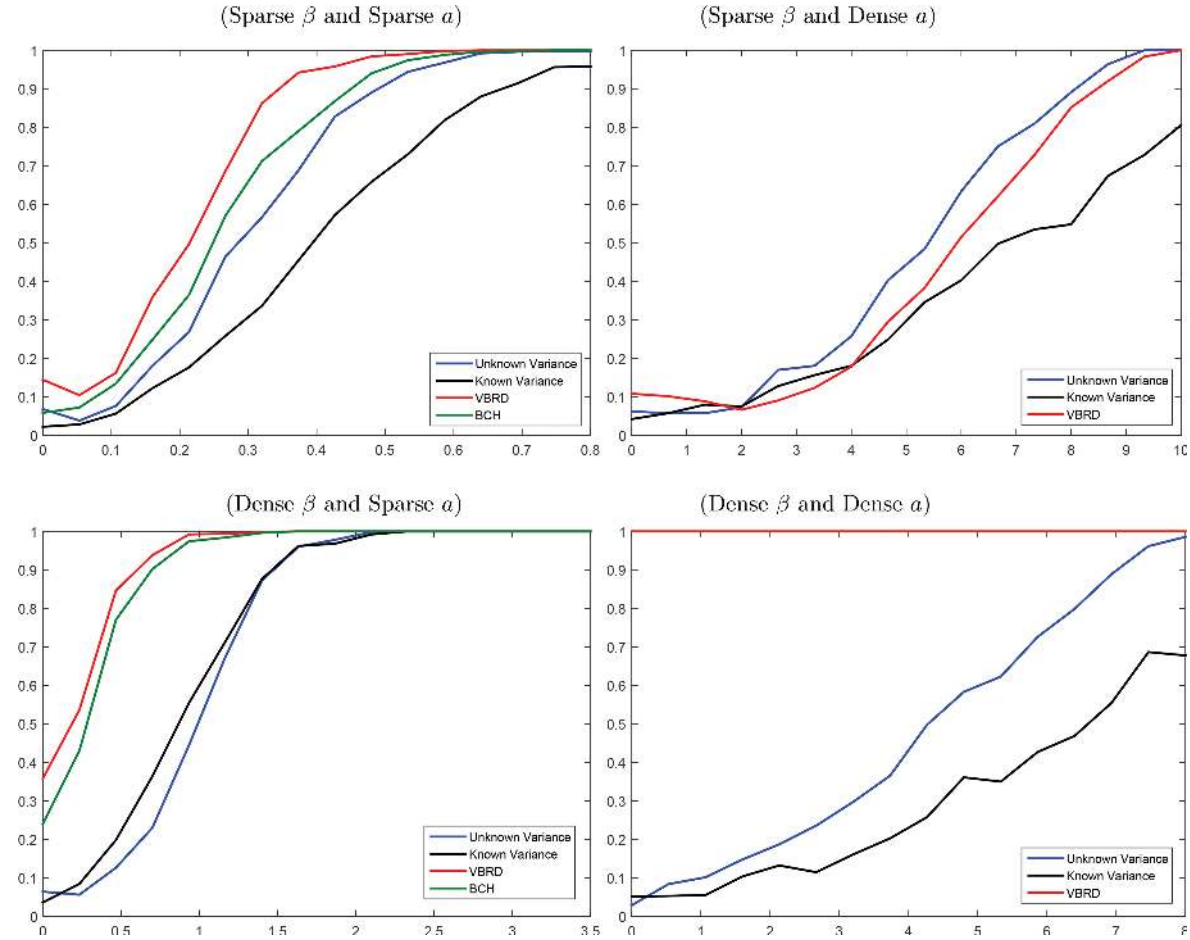

Figure 7. Power curves of competing methods across different hypothesis $a^\top \beta_* = g_0$ settings. Design settings follow Example 3 with $n = 100$ and $p = 500$. The alternative hypothesis takes the form of $a^\top \beta_* = g_0 + h$ with h presented on the x-axis. The y-axis contains the average rejection probability over 500 repetition. Therefore, $h = 0$ corresponds to Type I error and the remaining ones the Type II error. "Known variance" denotes the method as is introduced in Section 2 whereas, "unknown variance" denotes the method introduced in Section 3. VBRD and BCH refer to the methods proposed in Van de Geer et al. (2014) and Belloni, Chernozhukov, and Hansen (2014), respectively. Note that tuning parameters for all the methods are chosen according to their "oracle" theoretical values. If a method could not be implemented as is proposed in its respective article, it was not included in the graph.

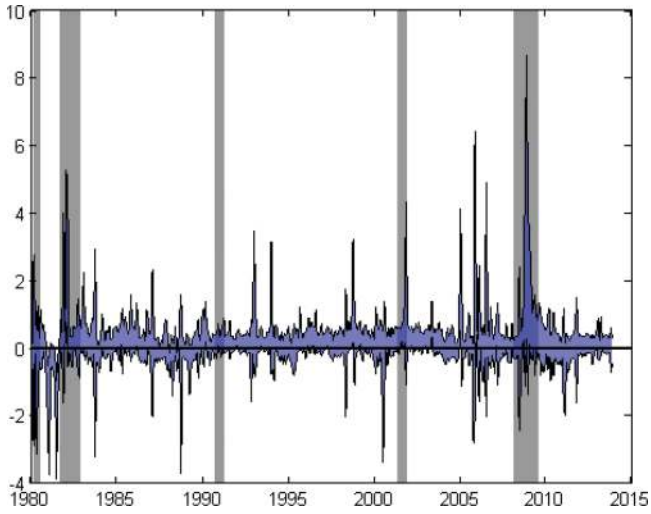

Figure 8. 95% confidence interval for the risk premia at each time period (the blue band) with the gray shades representing the NBER recession periods.

existing literature by developing methodology that allows fully nonsparse linear models.

Restructuring the model according to the hypothesis under testing allows for the high-dimensional a and β_* that are not necessarily sparse. The synthesized features are customized based on the null hypothesis and are close to being orthogonal. We note that this customization is the key, since the orthogonality per se is not useful. Techniques that only induce feature orthogonality, such as preconditioning by Jia and Rohe (2012) and DECO by Wang, Dunson, and Leng (2016), still cannot be used to test $H_0 : a^\top \beta_* = g_0$ when a and β_* are dense.

Observe that we have proposed two different inferential methods. However, it is not necessarily true that the method proposed in Section 2 dominates the one proposed in Section 3 in terms of power. The main difference between the method is in the definition of the moment condition. The method assuming knowledge of Σ_X avoids estimation of β_* and hence is extremely easy to implement; however, when β_* is sparse (and thus easy to estimate), not using information on β_* can cause some loss of power. The method proposed in Section 2 essentially treats $w_i^\top \beta_*$ as the error term. In contrast, the method proposed in Section 3 computes an estimate for $\tilde{w}_i^\top \pi_*$ (which in spirit corresponds to $w_i^\top \beta_*$); when the model turns out to be sparse, the method without knowledge of Σ_X can essentially "remove" $w_i^\top \beta_*$ from the error term, thereby achieving better power. For dense models, this reasoning does not apply and thus it is not clear which one should be more powerful.

To conclude the article, we would like to discuss here valuable topics for future research. The proposed methodology can be used to conduct inference of conditional distributions of the response, whenever the distribution function of ε, $Q(\cdot)$ is known or is consistently estimated. Specific example includes construction of prediction intervals for high-dimensional linear models—a topic of extreme importance. For $F_{Y|X}(y, x) = P(y_{n+1} \le y \mid x_{n+1} = x)$, $F_{Y|X}$ can be parameterized as $F_{Y|X}(y, x; \beta_*, Q) = Q(y - x^\top \beta_*)$. For a given x, we can obtain a confidence set for $x^\top \beta_* : \hat{I}(1-\alpha, x)$ such that $P(x^\top \beta_* \in \hat{I}(1-\alpha, x)) \to 1-\alpha$, by inverting the tests proposed in this article. This leads to a natural confidence set for the $F_{Y|X}(y, x)$: $P(F_{Y|X}(\cdot, x) \in \hat{S}(1-\alpha, x)) \to 1-\alpha$, where

$$\hat{S}(1-\alpha, x) = \{Q(\cdot - c) \mid c \in \hat{I}(1-\alpha, x)\}.$$

If we restrict the model parameters to be sparse, then we can consistently estimate ε_i (and thus $Q(\cdot)$) and consequently form valid prediction intervals—a topic of specific importance for practitioners. However, when the model is allowed to be nonsparse and high-dimensional, the question of construction of prediction intervals has not been answered and needs special considerations. Additionally, under this setup, the proposed methods also lead to an inference method for (possibly nonlinear) functionals of the conditional distribution of y_{n+1} given x_{n+1}. For example, suppose that one is interested in $H(u, x) = \inf\{y \in \mathbb{R} \mid F_{Y|X}(y, x) \ge u\}$. Following the above proposal, we can simply take

$$\hat{\mathcal{H}}(u, x, \alpha) = \{\inf\{y \in \mathbb{R} \mid Q(y - c) \ge u\} \mid c \in \hat{I}(1-\alpha, x)\}$$

as a confidence set for $H(u, x)$.

Supplementary Materials

The online supplementary materials contain additional proofs for the article.

Funding

The authors gratefully acknowledge NSF support through the grant DMS-1205296.

References

Belloni, A., Chernozhukov, V., and Hansen, C. (2014), "Inference on Treatment Effects After Selection Among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608–650. [1584,1587,1589,1590,1591]

——— (2016), "LASSO Methods for Gaussian Instrumental Variables Models," *Journal of Royal Statistical Society*, Series B. [1588]

Belloni, A., Chernozhukov, V., and Kato, K. (2015), "Uniform Post-Selection Inference for Least Absolute Deviation Regression and Other Z-Estimation Problems," *Biometrika*, 102, 77–94. [1584,1587]

Belloni, A., Chernozhukov, V., and Wang, L. (2011), "Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming," *Biometrika*, 98, 791–806. [1588]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [1586,1588,1589]

Bühlmann, P., and van de Geer, S. (2015), "High-Dimensional Inference in Misspecified Linear Models," *Electronic Journal of Statistics*, 9, 1449–1473. [1584]

Bühlmann, P. (2013), "Statistical Significance in High-Dimensional Linear Models," *Bernoulli*, 19, 1212–1242. [1584]

Cai, T. T., and Guo, Z. (2017), "Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity," *Annals of Statistics*, 45, 615–646. [1584,1586,1587,1589,1590,1593]

——— (2018), "Accuracy Assessment for High-dimensional Linear Regression," *Annals of Statistics*, 46, 1807–1836. [1584,1588,1590]

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p is Much Larger Than n," *Annals of Statistics*, 35, 2313–2351. [1589]

Chernozhukov, V., Hansen, C., and Spindler, M. (2015), "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," *Annual Review of Economics*, 7, 649–688. [1584,1587]

Dangl, T., and Halling, M. (2012), "Predictive Regressions With Time-Varying Coefficients," *Journal of Financial Economics*, 106, 157–181. [1595]

Donoho, D. L., and Johnstone, I. M. (1994), "Minimax Risk Over lp-Balls for lp-Error," *Probability Theory and Related Fields*, 99, 277–303. [1584]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1583,1584]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [1593]

Gautier, E., and Tsybakov, A. B. (2013), "Pivotal Estimation in High-Dimensional Regression via Linear Programming," in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, eds. B. Schölkopf, Z. Luo, and V. Vovk, Berlin, Heidelberg: Springer, pp. 195–204. [1588]

Giraud, C., Huet, S., and Verzelen, N. (2012), "High-Dimensional Regression with Unknown Variance," *Statistical Science*, 27, 500–518. [1588]

Hall, P., Jin, J., and Miller, H. (2014), "Feature Selection When There are Many Influential Features," *Bernoulli*, 20, 1647–1671. [1584]

Henkel, S. J., Martin, J. S., and Nardari, F. (2011), "Time-Varying Short-Horizon Predictability," *Journal of Financial Economics*, 99, 560–580. [1595,1597]

Javanmard, A., and Montanari, A. (2014a), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *The Journal of Machine Learning Research*, 15, 2869–2909. [1584,1587]

Javanmard, A., and Montanari, A. (2014b), "Hypothesis Testing in High-Dimensional Regression Under thE Gaussian Random Design Model: Asymptotic Theory," *IEEE Transactions on Information Theory*, 60, 6522–6554. [1589,1590]

——— (2015), "De-Biasing the Lasso: Optimal Sample Size for Gaussian Designs," arXiv:1508.0257v3. [1584]

Jia, J., and Rohe, K. (2012), "Preconditioning to Comply With the Irrepresentable Condition," arXiv:1208.5584. [1598]

Jin, J., and Ke, Z. T. (2014), "Rare and Weak Effects in Large-Scale Inference: Methods and Phase Diagrams," arXiv:1410.4578v1. [1584]

McCracken, M. W., and Ng, S. (2016), "FRED-MD: A Monthly Database for Macroeconomic Research," *Journal of Business & Economic Statistics*, 34, 574–589. [1596]

Mitra, R., and Zhang, C.-H. (2016), "The Benefit of Group Sparsity in Group Inference With De-Biased Scaled Group Lasso," *Electronic Journal of Statistics*, 10, 1829–1873. [1584]

Nickl, R., and van de Geer, S. (2013), "Confidence Sets in Sparse Regression," *The Annals of Statistics*, 41, 2852–2876. [1587]

Ning, Y., and Liu, H. (2017), "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models," *Annals of Statistics*, 45, 158–195. [1584,1587,1589,1590]

Pritchard, J. K. (2001), "Are Rare Variants Responsible for Susceptibility to Complex Diseases?" *The American Journal of Human Genetics*, 69, 124–137. [1584]

Rapach, D. E., Strauss, J. K., and Zhou, G. (2010), "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy," *Review of Financial Studies*, 23, 821–862. [1595,1596]

Raskutti, G., Wainwright, M. J., and Yu, B. (2011), "Minimax Rates of Estimation for High-Dimensional Linear Regression Over-Balls," *IEEE Transactions on Information Theory*, 57, 6976–6994. [1584,1588]

Robinson, P. M. (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954. [1590]

Rudelson, M., and Zhou, S. (2013), "Reconstruction From Anisotropic Random Measurements," *IEEE Transactions on Information Theory*, 59, 3434–3447. [1587]

Shao, X. (2010), "A Self-Normalized Approach to Confidence Interval Construction in Time Series," *Journal of the Royal Statistical Society*, Series B, 72, 343–366. [1589]

Stock, J. H., and Watson, M. W. (2002a), "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179. [1596]

——— (2002b), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business & Economic Statistics*, 20, 147–162. [1596]

Sun, T., and Zhang, C.-H. (2012), "Scaled Sparse Linear Regression," *Biometrika*, 99, 879–898. [1588]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1583]

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [1584,1587,1589,1590]

van de Geer, S., and Jankova, J. (2016), "Semi-Parametric Efficiency Bounds and Efficient Estimation for High-Dimensional Models," arXiv:1601.00815v4. [1584]

Wang, X., Dunson, D., and Leng, C. (2016), "DECOrrelated Feature Space Partitioning for Distributed Sparse Regression," arXiv:1602.02575. [1598]

Ward, R. A. (2009), "Compressed Sensing With Cross Validation," *IEEE Transactions on Information Theory*, 55, 5773–5782. [1584]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [1584,1587,1597]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [1584]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1584]