The Annals of Statistics
2012, Vol. 40, No. 1, 353–384
DOI: 10.1214/11-AOS962

ESTIMATING SUFFICIENT REDUCTIONS OF THE PREDICTORS IN ABUNDANT HIGH-DIMENSIONAL REGRESSIONS

BY R. DENNIS COOK[1], LILIANA FORZANI AND ADAM J. ROTHMAN[2]

University of Minnesota, Instituto de Matemática Aplicada del Litoral and University of Minnesota

We study the asymptotic behavior of a class of methods for sufficient dimension reduction in high-dimension regressions, as the sample size and number of predictors grow in various alignments. It is demonstrated that these methods are consistent in a variety of settings, particularly in abundant regressions where most predictors contribute some information on the response, and oracle rates are possible. Simulation results are presented to support the theoretical conclusion.

1. Introduction. There are many facets to the analysis of data in high dimensions, depending on the type of application, some relying on dimension reduction, others relying on variable selection and a few employing both tactics. There has been considerable interest in dimension-reduction methods for the regression of a real response Y on a random vector of predictors $\mathbf{X} \in \mathbb{R}^p$ since the introduction of sliced inverse regression [SIR; Li (1991)] and sliced average variance estimation [SAVE; Cook and Weisberg (1991)]. A common goal of these and many other methods is to reduce the dimension of the predictor vector without loss of information about the response. The aim is to estimate a reduction $\mathbf{R}: \mathbb{R}^p \to \mathbb{R}^d$, $d \le p$, with the property that $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{R}(\mathbf{X})$ or, equivalently, $\mathbf{X}|(Y, \mathbf{R}(\mathbf{X})) \sim \mathbf{X}|\mathbf{R}(\mathbf{X})$ [Cook (2007)]. In this way $\mathbf{R}$ is sufficient because it captures all the information about Y that is available from $\mathbf{X}$. Sufficient reductions are not determined uniquely by this definition because any bijective transformation of $\mathbf{R}$ is also sufficient.

Nearly all methods for sufficient dimension reduction (SDR) restrict attention to the class of linear reductions, which arise naturally in many contexts. Linear reduction can be represented conveniently in terms of the projection $\mathbf{P}_{\mathcal{S}}\mathbf{X}$ of $\mathbf{X}$ onto a subspace $\mathcal{S} \subseteq \mathbb{R}^p$. If $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}$, then $\mathcal{S}$ is called a dimension-reduction subspace. Under mild conditions the intersection of any two dimension-reduction subspaces is again a dimension-reduction subspace and that being so the central subspace $\mathcal{S}_{Y|\mathbf{X}}$, defined as the intersection of all dimension-reduction subspaces,

Received February 2011; revised August 2011.
[1]Supported by NSF Grants DMS-10-07547 and DMS-11-05650.
[2]Supported by NSF Grant DMS-11-05650.
MSC2010 subject classifications. Primary 62H20; secondary 62J07.
Key words and phrases. Central subspace, oracle property, SPICE, sparsity, sufficient dimension reduction, principal fitted components.

is taken as the inferential target [Cook (1994, 1998)]. A minimal sufficient linear reduction is then of the form $\mathbf{R}(\mathbf{X}) = \boldsymbol{\eta}^T\mathbf{X}$, where $\boldsymbol{\eta}$ is any basis for $\mathcal{S}_{Y|\mathbf{X}}$.

SDR has a long history of successful application and is still an active research area. Recent novel SDR methods include likelihood-based sufficient dimension reduction [Cook and Forzani (2009)], kernel dimension reduction [Fukumizu, Bach and Jordan (2009)], shrinkage inverse regression estimation [Bondell and Li (2009)], dimension reduction for nonelliptically distributed predictors [Li and Dong (2009), Dong and Li (2010)], cumulative slicing estimation [Zhu, Zhu and Feng (2010)], dimension reduction for survival models [Xia, Zhang and Xu (2010)] and dimension reduction for spatial point processes [Guan and Wang (2010)]. This body of work reflects three different but related frontiers in SDR: extensions that require progressively fewer assumptions, development of likelihood-based methods and adaptations for specific areas of application. Almost all SDR methods rely on traditional asymptotic reasoning for support, letting the sample size $n \to \infty$ with p fixed. They nearly all require the inverse of a $p \times p$ sample covariance matrix and thus application is problematic when $n < p$. Since accurate estimation of a general $p \times p$ covariance matrix can require $n \gg p$ observations, it has seemed inevitable that SDR methods would encounter estimation problems when n is not sufficiently large. Chiaromonte and Martinelli (2002), Li and Li (2004) and others circumvented these issues by performing reduction in two stages, first replacing the p predictors with $p^* \ll n$ principal components and then applying an SDR method to the regression of the response on the selected p^* components. However, recent results on the eigenvectors of sample covariance matrices in high-dimensional settings raise questions on the value of such two-stage methods [see, e.g., Johnstone and Lu (2009)]. Cook, Li and Chiaromonte (2007) proposed an SDR method that avoids computation of inverses and reduces to partial least squares in a special case. Chun and Keleş (2010) showed recently that the partial least squares estimator of the coefficient vector in the linear regression of Y on $\mathbf{X}$ is inconsistent unless $p/n \to 0$ and this raises questions about the behavior of the Cook et al. SDR estimator when n is not large relative to p. Li and Yin (2008) used the least squares formulation of sliced inverse regression originated by Cook (2004) to develop a regularized version that allows $n < p$ and achieves simultaneous predictor selection and dimension reduction. This seems to be a promising method, but its asymptotic properties are unknown and it may not work well when the regression is not sparse. Wu and Li (2011) studied the asymptotic properties of a family of SDR estimators, using a SCAD-type penalty for variable selection in sparse regressions where the number of relevant variables is fixed as $p \to \infty$. Their method requires that $p/n \to 0$ for consistency.

While sparsity is an important concept in high-dimensional regression, not all high-dimensional regressions are sparse. For example, near-infrared reflectance is often measured at many wavelengths to predict the composition of matter, like the protein content of a grain. There is not normally an expectation that only a few wavelengths are needed to predict content. While some wavelengths may be

better predictors than others, it is the cumulative information provided by many wavelengths that is often relevant. Partial least squares has been the dimension-reduction method of choice in this type of regression. The regressions implied by this and other nonsparse applications share similar characteristics: (1) the predictor vectors are high-dimensional and typical area-specific analyses have employed some type of dimension reduction; (2) while assessing the relative importance of the predictors may be of interest, prediction is the ultimate goal; (3) information on the response is thought to accumulate, albeit perhaps slowly, as predictors are added, and (4) sparsity is not a driving notion.

In this article we introduce a family of SDR methods for studying high-dimensional regressions that differs from past approaches in at least three important ways. First, we do not require sparsity but rather we emphasize *abundant regressions* where most of the predictors contribute some information about the response. In the logic of Friedman et al. (2004), the bet-on-sparsity principle arose because, to continue the metaphor, there is otherwise little chance of a reasonable payoff. We show in contrast that reasonable payoffs can be obtained in abundant regressions with prediction as the ultimate goal, leading to a contrasting *bet-on-abundance* principle. Second, SDR studies have largely focused on properties of estimators of $\mathcal{S}_{Y|\mathbf{X}}$. We bypass this step and instead consider the limiting behavior of estimators of the sufficient reduction $\mathbf{R}(\mathbf{X})$ itself, assuming that the dimension d of $\mathbf{R}$ is fixed. More specifically, letting $\widehat{\mathbf{R}}$ denote an estimated reduction, we establish rates of convergence in the following sense. Let $\mathbf{X}_N$ denote a new observation on $\mathbf{X}$. If $\widehat{\mathbf{R}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(r(n, p))$ and if $r(n, p) \to 0$ as $n, p \to \infty$, then $\widehat{\mathbf{R}}(\mathbf{X}_N)$ is consistent for $\mathbf{R}(\mathbf{X}_N)$ and its convergence rate is at least r^{-1}. Third, we integrate recent work on the estimation of high-dimensional covariance matrices into our approach. In particular, we estimate a critical matrix of weights by using sparse permutation invariant covariance estimation (SPICE) as developed by Rothman et al. (2008).

In sum, by considering the reduction $\mathbf{R}$ itself rather than the central subspace $\mathcal{S}_{Y|\mathbf{X}}$, we both introduce a novel viewpoint for addressing dimension reduction and develop theoretically grounded SDR methodology for $n < p$ regressions where other methods either have no asymptotic support or must necessarily fail.

We describe the model for our study and its sufficient reduction in Section 2. The class of estimators that we use is described in Section 3, and stabilizing restrictions are presented in Section 4. Sections 5 and 6 contain theoretical conclusions for selected estimators from the class described in Section 3. Simulation results are presented in Sections 5.2 and 7. We turn to a spectroscopy application in Section 8 and a concluding discussion is given in Section 9. All proofs and additional simulation results are available in a supplemental article [Cook, Forzani and Rothman (2012)].

The following notational conventions will be used in our exposition. We use $\mathbb{R}^{p\times q}$ to denote the collection of all real $p \times q$ matrices. We use $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_F$ to denote the spectral and Frobenius norms of $\mathbf{A}$. The largest and smallest eigenvalues

of $\mathbf{A} \in \mathbb{R}^{p\times p}$ are denoted $\varphi_{\max}(\mathbf{A})$ and $\varphi_{\min}(\mathbf{A})$. If $\mathbf{A} \in \mathbb{R}^{p\times p}$, then $\operatorname{diag}(\mathbf{A}) \in \mathbb{R}^{p\times p}$ is the diagonal matrix with diagonal elements equal to those of $\mathbf{A}$. $\operatorname{vec}(\mathbf{A})$ is the operator that maps $\mathbf{A} \in \mathbb{R}^{p\times q}$ to $\mathbb{R}^{pq}$ by stacking its columns. If $\mathbf{B} \in \mathbb{R}^{p\times q}$ and $\mathbf{A} \in \mathbb{R}^{p\times p}$ is symmetric and positive definite, then the operator that projects in the $\mathbf{A}$ inner product onto $\operatorname{span}(\mathbf{B})$, the subspace spanned by the columns of $\mathbf{B}$, has the matrix representation $\mathbf{P}_{\mathbf{B}(\mathbf{A})} = \mathbf{B}(\mathbf{B}^T\mathbf{A}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{A}$, and $\mathbf{Q}_{\mathbf{B}(\mathbf{A})} = \mathbf{I}_p - \mathbf{P}_{\mathbf{B}(\mathbf{A})}$. $\mathbf{P}_{\mathbf{B}}$ indicates the projection onto $\operatorname{span}(\mathbf{B})$ in the usual inner product. A basis matrix for a subspace $\mathcal{S} \subseteq \mathbb{R}^p$ of dimension d is any matrix $\mathbf{B} \in \mathbb{R}^{p\times d}$ whose columns form a basis for $\mathcal{S}$. For nonstochastic sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp b_n$ if there are constants m, M and N such that $0 < m < |a_n/b_n| < M < \infty$ for all $n > N$. Similarly, for stochastic sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \asymp_p b_n$ if $a_n = O_p(b_n)$ and $b_n = O_p(a_n)$. For $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$, $\mathbf{A} > \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive definite. $\otimes$ denotes the Kronecker product, and $X \sim Y$ means X and Y are equal in distribution. Deviating slightly from convention, we do not index quantities by n and p, preferring instead to avoid notation proliferation by giving reminders from time to time.

2. Model. We assume throughout that the data consist of independent observations $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, and that p is increased by taking additional measurements on each of n sampling units. We treat the process of adding predictors as conditional on the response and so our approach is based on inverse reduction, $\mathbf{X}|(Y, \mathbf{R}(\mathbf{X})) \sim \mathbf{X}|\mathbf{R}(\mathbf{X})$, which is a standard SDR paradigm. Limits as $n, p \to \infty$ are thus conditional on the responses unless specifically indicated otherwise.

2.1. *Inverse regression model.* The model that we employ is engendered by standard SDR assumptions. We first review those assumptions briefly and then give a statement of the model.

Classical methods like SIR require two predominant and well-known conditions for estimation of $\mathcal{S}_{Y|\mathbf{X}}$. The first, called the *linearity condition*, insists that $\mathrm{E}(\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$ be a linear function of $\boldsymbol{\eta}^T\mathbf{X}$, where $\boldsymbol{\eta}$ is a basis matrix for $\mathcal{S}_{Y|\mathbf{X}}$. It must be valid only at a true basis and not for all $\boldsymbol{\eta} \in \mathbb{R}^{p\times d}$. This condition is generally regarded as mild since it holds to a good approximation when p is large [Hall and Li (1993)]. Let $\mathcal{S}_{E(\mathbf{X}|Y)}$ denote the subspace spanned by $\mathrm{E}(\mathbf{X}|Y = y) - \mathrm{E}(\mathbf{X})$ as y varies in the sample space of Y. Then the linearity condition implies that $\mathcal{S}_{E(\mathbf{X}|Y)} \subseteq \operatorname{var}(\mathbf{X})\mathcal{S}_{Y|\mathbf{X}}$, which is used as a basis for estimating a subspace of $\mathcal{S}_{Y|\mathbf{X}}$. The second, called the *coverage condition*, requires that $\mathcal{S}_{E(\mathbf{X}|Y)} = \operatorname{var}(\mathbf{X})\mathcal{S}_{Y|\mathbf{X}}$ and it too is generally regarded as mild in many applications. While the linearity and coverage conditions are standard requirements for root-n consistent methods based on the inverse mean function $\mathrm{E}(\mathbf{X}|Y)$, the actual performance of those methods depends also on the inverse variance function $\operatorname{var}(\mathbf{X}|Y)$. For instance, Bura and Cook (2001) concluded based on simulations that SIR works best when $\operatorname{var}(\mathbf{X}|Y)$ is nonstochastic, and Cook and Ni (2005) demonstrated analytically that

SIR can be quite inefficient when $\operatorname{var}(\mathbf{X}|Y)$ varies. We refer to the requirement that $\operatorname{var}(\mathbf{X}|Y)$ be nonstochastic as the *covariance condition*. If the linearity and covariance conditions hold, then $\mathrm{E}(\operatorname{var}(\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})|Y)$ must be nonstochastic as well. This is related to the usual covariance condition—$\operatorname{var}(\mathbf{X}|\boldsymbol{\eta}^T\mathbf{X})$ is constant—used by SAVE and other second-order methods. See Li (1991), Cook and Ni (2005) and Li and Dong (2009) for further discussion of these conditions.

We assume the linearity, coverage and covariance conditions as a basis for our study. Under these conditions it can be shown straightforwardly that $\operatorname{var}(\mathbf{X})\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Delta}\mathcal{S}_{Y|\mathbf{X}}$, where $\boldsymbol{\Delta} = \operatorname{var}(\mathbf{X}|Y)$. Consequently, letting $\tilde{\boldsymbol{\Gamma}} \in \mathbb{R}^{p\times d}$ be a basis matrix for $\boldsymbol{\Delta}\mathcal{S}_{Y|\mathbf{X}}$, we are led to the model

$$\mathbf{X}_i = \boldsymbol{\mu} + \tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\xi}}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$, the error vectors $\boldsymbol{\varepsilon}_i$ are independent copies of a random vector $\boldsymbol{\varepsilon} \in \mathbb{R}^p$ with mean 0 and variance $\operatorname{var}(\boldsymbol{\varepsilon}) = \boldsymbol{\Delta}$, and $\tilde{\boldsymbol{\xi}}_i = \tilde{\boldsymbol{\xi}}(Y_i)$ is the ith instance of an unknown vector-valued function $\tilde{\boldsymbol{\xi}}(Y) \in \mathbb{R}^d$ with $\mathrm{E}(\tilde{\boldsymbol{\xi}}) = 0$ that gives the coordinates of $\mathrm{E}(\mathbf{X}|Y) - \mathrm{E}(\mathbf{X})$ in terms of $\tilde{\boldsymbol{\Gamma}}$ and is independent of $\boldsymbol{\varepsilon}$. In this representation $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Delta}^{-1}\operatorname{span}(\tilde{\boldsymbol{\Gamma}})$.

Neither $\tilde{\boldsymbol{\Gamma}}$ nor $\tilde{\boldsymbol{\xi}}$ is identified in model (2.1), because for any conforming nonsingular matrix $\mathbf{A}$, $\tilde{\boldsymbol{\Gamma}}\tilde{\boldsymbol{\xi}} = (\tilde{\boldsymbol{\Gamma}}\mathbf{A})(\mathbf{A}^{-1}\tilde{\boldsymbol{\xi}})$, leading to a different parameterization. This nonuniqueness has been mitigated in past studies with p fixed by requiring $\tilde{\boldsymbol{\Gamma}}^T\tilde{\boldsymbol{\Gamma}} = \mathbf{I}_d$. However, that parameterization is not workable when allowing $p \to \infty$, and for this reason we adopt different restrictions.

The next step is to reparameterize model (2.1) to satisfy constraints that will facilitate our development. Specifically, we construct an equivalent model

$$\mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i, \qquad i = 1, \ldots, n, \tag{2.2}$$

where $\operatorname{span}(\boldsymbol{\Gamma}) = \operatorname{span}(\tilde{\boldsymbol{\Gamma}})$, $\boldsymbol{\xi}_i = \boldsymbol{\xi}(Y_i)$ is the coordinate vector relative to the new basis matrix $\boldsymbol{\Gamma}$, and we center the $\boldsymbol{\xi}_i$'s in the sample so that $\bar{\boldsymbol{\xi}} = 0$. Let the rows of $\tilde{\boldsymbol{\Xi}} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{\Xi} \in \mathbb{R}^{n\times d}$ be $\tilde{\boldsymbol{\xi}}_i^T$ and $\boldsymbol{\xi}_i^T$, $i = 1, \ldots, n$. Without loss of generality, we impose on model (2.2) the constraints that (1) $n^{-1}\boldsymbol{\Xi}^T\mathbf{M}_n\boldsymbol{\Xi} = \mathbf{I}_d$, where $\mathbf{M}_n$ is a *scaling matrix* that is defined in Section 4.1, and (2) $\boldsymbol{\Gamma}^T\mathbf{W}\boldsymbol{\Gamma}$ is a diagonal matrix where $\mathbf{W} \geq 0$ is a symmetric $p \times p$ population *weight matrix* that is discussed in Section 3. To see how this is done starting from model (2.1), let $\mathbf{T} = (n^{-1}\tilde{\boldsymbol{\Xi}}^T\mathbf{M}_n\tilde{\boldsymbol{\Xi}})^{-1/2}$ and let $\mathbf{O} \in \mathbb{R}^{d\times d}$ be an orthogonal matrix so that $\mathbf{O}^T\mathbf{T}^{-1}\tilde{\boldsymbol{\Gamma}}^T\mathbf{W}\tilde{\boldsymbol{\Gamma}}\mathbf{T}^{-1}\mathbf{O}$ is a diagonal matrix. Then $\tilde{\boldsymbol{\Xi}}\tilde{\boldsymbol{\Gamma}}^T = \tilde{\boldsymbol{\Xi}}\mathbf{T}\mathbf{O}\mathbf{O}^T\mathbf{T}^{-1}\tilde{\boldsymbol{\Gamma}}^T = \boldsymbol{\Xi}\boldsymbol{\Gamma}^T$, where $\boldsymbol{\Xi} = \tilde{\boldsymbol{\Xi}}\mathbf{T}\mathbf{O}$ and $\boldsymbol{\Gamma}^T = \mathbf{O}^T\mathbf{T}^{-1}\tilde{\boldsymbol{\Gamma}}^T$ satisfy the constraints by construction. Model (2.2) can be represented also in matrix form as

$$\mathbb{X} = \mathbf{1}_n\boldsymbol{\mu}^T + \boldsymbol{\Xi}\boldsymbol{\Gamma}^T + \mathbf{e}, \tag{2.3}$$

where $\mathbb{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{e} \in \mathbb{R}^{n\times p}$ have rows $\mathbf{X}_i^T$ and $\boldsymbol{\varepsilon}_i^T$, and $\mathbf{e}$ has mean 0 and variance $\operatorname{var}(\operatorname{vec}(\mathbf{e}^T)) = \mathbf{I}_n \otimes \boldsymbol{\Delta}$.

Since bijective transformations of a sufficient reduction are themselves sufficient, we define $\mathbf{R}$ to be the coordinates of the projection of $\mathbf{X} - \boldsymbol{\mu}$ onto $\operatorname{span}(\boldsymbol{\Gamma})$ in the $\boldsymbol{\Delta}^{-1}$ inner product:

$$\mathbf{R}(\mathbf{X}) = (\boldsymbol{\Gamma}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \boldsymbol{\Delta}^{-1} (\mathbf{X} - \boldsymbol{\mu}), \tag{2.4}$$

where the first factor $(\boldsymbol{\Gamma}^T \boldsymbol{\Delta}^{-1} \boldsymbol{\Gamma})^{-1}$ stabilizes $\mathrm{E}(\mathbf{R})$ as $p \to \infty$.

3. Estimation. Without further structure it is not possible to use model (2.2) to estimate $\mathbf{R}(\mathbf{X})$ since the coordinate vectors $\boldsymbol{\xi}_i$ are unknown. However, estimation is possible by approximating the coordinate vectors as $\boldsymbol{\xi}_i \approx \mathbf{b}\mathbf{f}_i$, where $\mathbf{b} \in \mathbb{R}^{d \times r}$, $d \le r$, and $\mathbf{f}_i = \mathbf{f}(Y_i)$ is the ith realization of a known user-selected vector-valued function $\mathbf{f}(Y) \in \mathbb{R}^r$. Without loss of generality we center the sample $\sum_{i=1}^n \mathbf{f}_i = 0$. Let $\mathbb{F} \in \mathbb{R}^{n \times r}$ be the matrix with rows $\mathbf{f}_i^T$, and assume that $\boldsymbol{\Phi}_n = \mathbb{F}^T \mathbb{F}/n > 0$ and $\boldsymbol{\Phi} = \lim_n \boldsymbol{\Phi}_n > 0$, unless $r = 0$ and then of course $\mathbb{F}$ is nil. We next describe the class of estimators we use, postponing discussion of possible choices for $\mathbf{f}$ until Section 3.2.

3.1. *Estimators.* The class of reduction estimators $\widehat{\mathbf{R}}$ that we study is based on using estimators of $(\boldsymbol{\mu}, \mathbf{b}, \boldsymbol{\Gamma})$ from a subclass of the family of inverse regression estimators proposed by Cook and Ni (2005):

$$(\widehat{\boldsymbol{\mu}}, \widehat{\mathbf{b}}, \widehat{\boldsymbol{\Gamma}}) = \arg\min \operatorname{tr}\{(\mathbb{X} - \mathbf{1}_n \boldsymbol{\mu}^T - \mathbb{F}\mathbf{b}^T \boldsymbol{\Gamma}^T) \widehat{\mathbf{W}} (\mathbb{X} - \mathbf{1}_n \boldsymbol{\mu}^T - \mathbb{F}\mathbf{b}^T \boldsymbol{\Gamma}^T)^T\},$$

where the minimization is over $\boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times d}$ and $\mathbf{b} \in \mathbb{R}^{d \times r}$ subject to the constraints that $\mathbf{b}\boldsymbol{\Phi}_n \mathbf{b}^T = \mathbf{I}_d$ and that $\boldsymbol{\Gamma}^T \widehat{\mathbf{W}} \boldsymbol{\Gamma}$ is a diagonal matrix. A particular estimator is determined by the choice of the sample weight matrix $\widehat{\mathbf{W}} \in \mathbb{R}^{p \times p}$ with $\mathbf{W}$ being a population version. Some instances of $\widehat{\mathbf{W}}$ that we introduce later correspond to $\mathbf{W} = \mathbf{I}_p$, $\boldsymbol{\Delta}^{-1}$ and $\operatorname{diag}^{-1}(\boldsymbol{\Delta})$. We next report the estimators from this family. A sketch of the derivation is available in the supplemental article [Cook, Forzani and Rothman (2012)].

It is easily seen that $\widehat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$. Let $\mathbb{Z} = \mathbb{X} - \mathbf{1}_n \widehat{\boldsymbol{\mu}}^T$, and let $\widehat{\mathbf{B}} = \mathbb{Z}^T \mathbb{F} (\mathbb{F}^T \mathbb{F})^{-1} \in \mathbb{R}^{p \times r}$ denote the matrix of regression coefficients from the least squares fit of $\mathbf{X}$ on $\mathbf{f}$, assuming that $n > r + 1$. Also, let the columns of $\widehat{\mathbf{V}}_d \in \mathbb{R}^{r \times d}$ be the first d eigenvectors of

$$\widehat{\mathbf{K}} = \boldsymbol{\Phi}_n^{1/2} \widehat{\mathbf{B}}^T \widehat{\mathbf{W}} \widehat{\mathbf{B}} \boldsymbol{\Phi}_n^{1/2} \in \mathbb{R}^{r \times r}, \tag{3.1}$$

assuming that the dth eigenvalue of $\widehat{\mathbf{K}}$ is strictly larger than the $(d+1)$st eigenvalue. This assumption will be true with probability 1 for the choices of $\widehat{\mathbf{W}}$ considered here. Then the estimators are $\widehat{\mathbf{b}} = \widehat{\mathbf{V}}_d^T \boldsymbol{\Phi}_n^{-1/2}$, $\widehat{\boldsymbol{\Gamma}} = \widehat{\mathbf{B}} \boldsymbol{\Phi}_n^{1/2} \widehat{\mathbf{V}}_d$, $\widehat{\boldsymbol{\Gamma}}^T \widehat{\mathbf{W}} \widehat{\boldsymbol{\Gamma}} = \widehat{\mathbf{V}}_d^T \widehat{\mathbf{K}} \widehat{\mathbf{V}}_d$ and

$$\widehat{\mathbf{R}}(\mathbf{X}) = (\widehat{\boldsymbol{\Gamma}}^T \widehat{\mathbf{W}} \widehat{\boldsymbol{\Gamma}})^{-1} \widehat{\boldsymbol{\Gamma}}^T \widehat{\mathbf{W}} (\mathbf{X} - \bar{\mathbf{X}}). \tag{3.2}$$

The diagonal elements of the diagonal matrix $\widehat{\boldsymbol{\Gamma}}^T \widehat{\mathbf{W}} \widehat{\boldsymbol{\Gamma}}$ consist of the first (largest) d eigenvalues of $\widehat{\mathbf{K}}$, and $\widehat{\mathbf{b}}$ is the coefficient matrix from the OLS fit of $\widehat{\mathbf{R}}(\mathbf{X}_i)$ on $\mathbf{f}_i$.

3.2. *Choice of* $\mathbf{f}$. Clearly, we should strive to choose an $\mathbf{f}(Y)$ so that, for some $\mathbf{b} \in \mathbb{R}^{d\times r}$, $\boldsymbol{\xi}_i$ is well approximated by $\mathbf{b}\mathbf{f}_i$ for $i = 1, \ldots, n$. This intuition is manifested in various calculations via the requirement that $\operatorname{rank}(n^{-1}\boldsymbol{\Xi}^T\mathbb{F}) = d$ for all n. As an extreme first instance, suppose that $\boldsymbol{\Xi}^T\mathbb{F} = 0$. Then $\widehat{\mathbf{B}} = \mathbb{Z}^T\mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1} = (\boldsymbol{\Gamma}\boldsymbol{\Xi}^T\mathbb{F} + \mathbf{e}^T\mathbb{F})(\mathbb{F}^T\mathbb{F})^{-1} = \mathbf{e}^T(\mathbb{F}^T\mathbb{F})^{-1}$ and consequently $\widehat{\mathbf{B}}$ provides no information on $\operatorname{span}(\boldsymbol{\Gamma})$. Since $n^{-1}\boldsymbol{\Xi}^T\mathbb{F} = \widehat{\operatorname{cov}}(\boldsymbol{\xi}(Y), \mathbf{f}(Y))$, this requirement is essentially that the true coordinate vector $\boldsymbol{\xi}$ is sufficiently correlated with its approximation, and it is equivalent to the condition derived by Cook and Forzani (2008) under a PFC model with normal errors and p fixed. In the remainder of this article we assume that, for all n,

$$\operatorname{rank}(n^{-1}\boldsymbol{\Xi}^T\mathbb{F}) = d \quad \text{and} \quad n^{-1}\boldsymbol{\Xi}^T\mathbb{F} = O(1) \qquad \text{as } n \to \infty, \tag{3.3}$$

where the order is nonstochastic because we condition on the observed values of Y in our formal asymptotic calculations.

There are many ways to choose an appropriate $\mathbf{f}$ in practice. Under model (2.2), each coordinate X_j, $j = 1, \ldots, p$, of $\mathbf{X}$ follows a univariate linear model with predictor vector $\mathbf{f}(Y)$. When Y is quantitative we can use inverse response plots [Cook (1998), Chapter 10] of X_j versus Y, $j = 1, \ldots, p$, to gain information about suitable choices for $\mathbf{f}$. When p is too large for a thorough graphical investigation, which is the case in the context of this article, we can consider $\mathbf{f}$'s that contain a reasonably flexible set of basis functions, like polynomial terms in Y. In some regressions there may be a natural choice for $\mathbf{f}$. For example, suppose that Y is categorical, taking values in one of m categories C_k, $k = 1, \ldots, m$. We can then set $r = m - 1$ and specify the kth coordinate of $\mathbf{f}$ to be the indicator function $J(y \in C_k)$. Another option consists of "slicing" the observed range of a continuous Y into m bins (categories) C_k, $k = 1, \ldots, m$. A rudimentary set of basis functions can then be constructed by specifying the kth coordinate of $\mathbf{f}$ as for the case of a categorical Y. This has the effect of approximating each conditional mean $\mathrm{E}(X_j|Y)$ as a step function of Y with m steps. This option is of particular interest because it exactly reproduces SIRs estimate of $\mathcal{S}_{Y|\mathbf{X}}$ in the traditional large-n setting [Cook and Forzani (2008)]. Cubic splines with the endpoints of the bins as the knots are a more responsive option that is less prone to loss of intra-slice information.

4. Universal context. Our goal is to study the limiting behavior of $\widehat{\mathbf{R}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N)$, where $\mathbf{R}$ and $\widehat{\mathbf{R}}$ are given by (2.4) and (3.2). This difference depends on the choice of $\widehat{\mathbf{W}}$ and the behavior of the true reduction $\mathbf{R}$ as $p \to \infty$. In this section we measure and constrain the interaction between $\mathbf{W}$ and the model. We also place weak constraints on $\mathbf{R}$ to help ensure well-behaved limits. The context described in this section will be assumed to hold throughout this article, without necessarily being declared in formal statements. We will revisit these constraints occasionally during the discussion, particularly when some of them are implied by other structure.

4.1. *Scaling matrix* $\mathbf{M}_n$. Since bijective transformations of sufficient reductions are also sufficient, we need to have $\widehat{\mathbf{R}}(\mathbf{X}_N)$ and $\mathbf{R}(\mathbf{X}_N)$ in the same scale to ensure that their difference $\widehat{\mathbf{R}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N)$ is a useful measure of agreement. This can be accomplished by choosing the scaling matrix $\mathbf{M}_n = \mathbf{P}_{\mathbb{F}}$ so that the first model constraint stated following (2.2) becomes $n^{-1}\boldsymbol{\Xi}^T\mathbf{M}_n\boldsymbol{\Xi} = n^{-1}\boldsymbol{\Xi}^T\mathbf{P}_{\mathbb{F}}\boldsymbol{\Xi} \in \mathbb{R}^{d\times d}$, which is a rank-$d$ matrix by (3.3). This choice ensures that $\boldsymbol{\Gamma}$ is the lead term in the asymptotic expansion of $\widehat{\boldsymbol{\Gamma}}$, which places $\widehat{\mathbf{R}}(\mathbf{X}_N)$ and $\mathbf{R}(\mathbf{X}_N)$ on the same scale.

4.2. *Signal rate*. We define the *signal rate* to be a positive monotonically increasing function $h(p) = O(p)$ that measures the rate of increase in the population signal: let $\mathbf{G}_h = \boldsymbol{\Gamma}^T\mathbf{W}\boldsymbol{\Gamma}/h(p)$, which is a diagonal matrix because $\boldsymbol{\Gamma}^T\mathbf{W}\boldsymbol{\Gamma}$ is diagonal by construction. Then $h(p)$ is chosen to meet the requirement

$$\lim_{p\to\infty} \mathbf{G}_h = \mathbf{G} \in \mathbb{R}^{d\times d}, \tag{4.1}$$

where $\mathbf{G}$ is a diagonal matrix that is assumed to be positive definite with finite elements. The signal rate is not needed for computation of $\widehat{\mathbf{R}}$ but it will play a key role in later developments. It depends via $\mathbf{W}$ on the specific estimator selected, although this is not indicated notationally. When $\mathbf{W} > 0$ the bounds $\varphi_{\min}(\mathbf{W})\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} \le \boldsymbol{\Gamma}^T\mathbf{W}\boldsymbol{\Gamma} \le \varphi_{\max}(\mathbf{W})\boldsymbol{\Gamma}^T\boldsymbol{\Gamma}$ can be used to aid intuition on the magnitude of $h(p)$ by presuming properties of $\boldsymbol{\Gamma}$ and $\mathbf{W}$. For example, consider regressions in which $\varphi_{\min}(\mathbf{W})$ and $\varphi_{\max}(\mathbf{W})$ are bounded away from 0 and ∞ as $p \to \infty$ and a positive fraction g, $0 < a \le g \le 1$, of the rows of $\boldsymbol{\Gamma}$ is sampled from a multivariate density with finite second moments and the other rows of $\boldsymbol{\Gamma}$ are all zero vectors. Then the number of nonzero rows $pg \asymp p$, $h(p) \asymp p$, and we say that the regression has an *abundant signal*. Regressions with *near abundant signals* like $h(p) \asymp p^{2/3}$ may often perform similarly in practice to regressions with abundant signals. It is technically possible to have regressions in which $p = o(h(p))$, but we would not normally expect this in practice. On the other extreme, sparse regressions are those in which $h(p) \asymp 1$, so only finitely many predictors are relevant or the signal accumulates very slowly as $p \to \infty$. Regressions with *near sparse signals* have, say, $h(p) = o(p^{1/3})$. Typically we will use $h(p)$ in definitions and formal statements, but use the abbreviated form h otherwise. We assume in the rest of this article that (4.1) holds.

4.3. *Limiting reduction*. Our second requirement is that, as $p \to \infty$, the spectral norm $\|\mathrm{var}(\mathbf{R})\|$ converges to a finite constant, which may be 0. If this is not so, then the variance of some linear combinations $\mathbf{a}^T\mathbf{R}$ will diverge as $p \to \infty$, and the very notion of dimension reduction for large-p regressions becomes problematic.

DEFINITION 4.1. *A reduction* $\mathbf{R}(\mathbf{X})$ *is stable if* $\lim_{p\to\infty}\|\mathrm{var}(\mathbf{R}(\mathbf{X}))\| < \infty$.

The following lemma gives a sufficient condition for stability that incorporates the signal rate. In preparation, define

$$\boldsymbol{\rho} = \mathbf{W}^{1/2}\boldsymbol{\Delta}\mathbf{W}^{1/2} \in \mathbb{R}^{p\times p}, \tag{4.2}$$

which is manifested in various asymptotic expansions as a measure of the agreement between $\mathbf{W}$ and $\boldsymbol{\Delta}^{-1}$.

LEMMA 4.1. *If* $\|\boldsymbol{\rho}\| = O(h(p))$, *then reduction* (2.4) *is stable.*

According to Lemma 4.1, if $\mathbf{W} = \boldsymbol{\Delta}^{-1}$, then the reduction is stable regardless of h. All regressions in which $\|\boldsymbol{\rho}\|$ is bounded yield stable reductions. Bounded eigenvalues have been required in various studies to avoid ill-conditioned covariance matrices [see, e.g., Bickel and Levina (2008a) and Rothman et al. (2008)]. More generally, the requirement for a stable reduction is that $h(p)$ must increase at a rate that is no less than the rate at which $\|\boldsymbol{\rho}\|$ increases. In particular, if $h = O(1)$, then the sufficient condition of Lemma 4.1 requires that $\|\boldsymbol{\rho}\|$ is bounded.

The rates that we develop depend on the functions $\kappa(n, p) = [p/\{h(p)n\}]^{1/2}$ and $\psi(n, p, \boldsymbol{\rho}) = \|\boldsymbol{\rho}\|_F/\{h(p)\sqrt{n}\}$. The interpretation and roles of these functions will be discussed later; for now we state across-the-board requirements that, as $n, p \to \infty$,

$$\kappa(n, p) = O(1) \quad \text{and} \quad \psi(n, p, \boldsymbol{\rho}) = O(1). \tag{4.3}$$

These functions will nearly always be written without their arguments. We assume in the rest of this article that all regressions are stable and that the orders given in (4.3) hold.

4.4. *Joint constraints on the weights* $\mathbf{W}$ *and errors* $\boldsymbol{\varepsilon}$. So far we have assumed only that the errors $\boldsymbol{\varepsilon}_i$ are independent copies of the random vector $\boldsymbol{\varepsilon}$ which has mean 0 and variance $\boldsymbol{\Delta}$. We impose two additional constraints to ensure well-behaved limits: as $p \to \infty$,

(W.1) $\mathrm{E}(\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon}) = O(p)$, and
(W.2) $\mathrm{var}(\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon}) = O(p^2)$.

Condition (W.1), which is equivalent to $\mathrm{tr}(\boldsymbol{\rho})/p = O(1)$, seems quite mild. For example, if we use unweighted fits with $\widehat{\mathbf{W}} = \mathbf{W} = \mathbf{I}_p$, then this condition is simply that the average error variance $\mathrm{tr}(\boldsymbol{\Delta})/p$ is bounded. Condition (W.2) can be seen as placing constraints on $\mathbf{W}$ and on the fourth moments of $\boldsymbol{\varepsilon}$. It too seems mild, although it is more constraining than the first condition. The following lemma describes a few settings in which condition (W.2) holds.

LEMMA 4.2. *Let* $\boldsymbol{\epsilon} = (\epsilon_i) = \boldsymbol{\Delta}^{-1/2}\boldsymbol{\varepsilon}$ *so that* $\mathrm{E}(\boldsymbol{\epsilon}) = 0$ *and* $\mathrm{var}(\boldsymbol{\epsilon}) = \mathbf{I}_p$. *Let* $\phi_{ij} = \mathrm{E}(\epsilon_i^2\epsilon_j^2)$, $i, j = 1, \ldots, p$. *Then*:

(i) *If* $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$, *then condition* (W.1) *implies condition* (W.2).
(ii) *If* $\mathbf{W} = \boldsymbol{\Delta}^{-1}$ *and if* $\phi_{ij} \le \phi < \infty$ *as* $p \to \infty$, *then condition* (W.2) *holds*.
(iii) *If* (a) *the elements* ϵ_i *of* $\boldsymbol{\epsilon}$ *have symmetric distributions or are independent and if* (b) $\phi_{ij} \le \phi < \infty$ *as* $p \to \infty$, *then condition* (W.1) *implies condition* (W.2).

We assume in the rest of this article that conditions (W.1) and (W.2) hold.

5. $\widehat{\mathbf{W}}$ converging in spectral norm. In this section we describe our results for the limiting reduction when $\widehat{\mathbf{W}}$ converges in the spectral norm. Specifically, we require the following two conditions:

(S.1) There exists a population weight matrix $\mathbf{W} \in \mathbb{R}^{p\times p}$ so that the spectral norm of $\mathbf{S} \equiv \mathbf{W}^{-1/2}(\widehat{\mathbf{W}} - \mathbf{W})\mathbf{W}^{-1/2}$ converges to 0 in probability at rate at most $\omega^{-1}(n, p)$ as $n, p \to \infty$; equivalently, $\|\mathbf{S}\| = O_p(\omega(n, p))$ as $\omega \to 0$.
(S.2) $\|\mathrm{E}(\mathbf{S}^2)\| = O(\omega^2(n, p))$ as $\omega \to 0$.

These conditions are implied by the stronger condition that $\mathrm{E}(\|\mathbf{S}^2\|) = O(\omega^2)$, (S.1) following from Chebyshev's inequality and (S.2) arising since $\|\mathrm{E}(\mathbf{S}^2)\| \le \mathrm{E}(\|\mathbf{S}^2\|)$. All of the weight matrices discussed in this article can satisfy these conditions, as well as other weight matrices that we have considered, so these conditions do not seem burdensome. For ease of reference we denote the corresponding estimator as $\widehat{\mathbf{R}}_{\widehat{\mathbf{W}}}$.

5.1. *Theoretical results.* We state and discuss one of our main results in this section. In preparation, let

$$\mathbf{K} = n^{-2}\boldsymbol{\Phi}_n^{-1/2}\mathbb{F}^T\,\boldsymbol{\Xi}\mathbf{G}_h\boldsymbol{\Xi}^T\mathbb{F}\boldsymbol{\Phi}_n^{-1/2} \in \mathbb{R}^{r\times r}, \tag{5.1}$$

where $\mathbf{G}_h$ is as defined in (4.1), let $\bar{\rho} = \mathrm{tr}(\boldsymbol{\rho})/p$ and define the random vector $\boldsymbol{\nu} = \mathbf{R}_{\mathbf{W}}(\boldsymbol{\varepsilon}_N) - \mathbf{R}(\boldsymbol{\varepsilon}_N) \in \mathbb{R}^d$, where $\mathbf{R}_{\mathbf{W}}(\boldsymbol{\varepsilon}_N) = (\boldsymbol{\Gamma}^T\mathbf{W}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\mathbf{W}\boldsymbol{\varepsilon}_N$ is the population reduction using $\mathbf{W}$ applied to the error $\boldsymbol{\varepsilon}_N$ of a new observation and $\mathbf{R}(\boldsymbol{\varepsilon}_N)$ is the targeted population reduction (2.4) applied to the same error. The vector $\boldsymbol{\nu} \in \mathbb{R}^d$ measures the population-level agreement between the user-selected reduction $\mathbf{R}_{\mathbf{W}}$ and the ideal reduction. It has mean $\mathrm{E}(\boldsymbol{\nu}) = 0$ and variance

$$\mathrm{var}(\boldsymbol{\nu}) = \mathbf{G}_h^{-1/2}\left\{\frac{\boldsymbol{\gamma}^T\boldsymbol{\rho}\boldsymbol{\gamma} - (\boldsymbol{\gamma}^T\boldsymbol{\rho}^{-1}\boldsymbol{\gamma})^{-1}}{h(p)}\right\}\mathbf{G}_h^{-1/2} \in \mathbb{R}^{d\times d}, \tag{5.2}$$

where the semi-orthogonal matrix $\boldsymbol{\gamma} = \mathbf{W}^{1/2}\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\mathbf{W}\boldsymbol{\Gamma})^{-1/2}$.

The next proposition describes asymptotic properties of $\widehat{\mathbf{K}} \in \mathbb{R}^{r\times r}$ and $\widehat{\mathbf{R}}_{\widehat{\mathbf{W}}}$, where $\widehat{\mathbf{K}}$ was defined in (3.1).

PROPOSITION 5.1. *Assume conditions* (S.1) *and* (S.2). *Then*:

(i) *If* $\mathbf{S} = 0$, *then* $h^{-1}(p)\mathrm{E}(\widehat{\mathbf{K}}) = \mathbf{K} + \kappa^2\bar{\rho}\mathbf{I}_r$.
(ii) $h^{-1}(p)(\widehat{\mathbf{K}} - \mathbf{K}) = O_p(\kappa) + O_p(\omega)$.

(iii) *If, in addition,* $\|\boldsymbol{\rho}\| = O(h(p))$, *then*

$$\widehat{\mathbf{R}}_{\widehat{\mathbf{W}}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = \boldsymbol{\nu} + O_p(\kappa) + O_p(\psi) + O_p(\omega),$$

where κ *and* ψ *were defined in* (4.3).

This proposition shows that the asymptotic properties of $\widehat{\mathbf{K}}$ and $\widehat{\mathbf{R}}_{\widehat{\mathbf{W}}}$ depend on four quantities, each with a different role. We defer discussion of the order $O_p(\omega)$ that measures the asymptotic behavior of $\widehat{\mathbf{W}}$ to later sections. In the rest of this section we concentrate on the remaining terms—$\boldsymbol{\nu}$, $O_p(\kappa)$ and $O_p(\psi)$—by assuming that $\widehat{\mathbf{W}}$ is nonstochastic, so $\widehat{\mathbf{W}} = \mathbf{W}$, $\mathbf{S} = 0$ and $\widehat{\mathbf{R}}_{\widehat{\mathbf{W}}} = \widehat{\mathbf{R}}_{\mathbf{W}}$. This involves no loss since none of these terms depends on $\mathbf{S}$.

Terms involving κ affect both estimation $\widehat{\mathbf{K}}$ and prediction $\widehat{\mathbf{R}}_{\mathbf{W}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N)$. If κ were to diverge, then, from Proposition 5.1(i), eventually $\mathrm{E}(\widehat{\mathbf{K}})/h$ will look like a diagonal matrix and there will be little signal left, which is why we imposed the universal condition (4.3) that $\kappa = O(1)$. If $\bar{\rho}$ diverges, then again $\mathrm{E}(\widehat{\mathbf{K}})/h$ will resemble a diagonal matrix, but this is prohibited by universal condition (W.1). Here we also see the role of the rank condition (3.3) introduced at the beginning of Section 3.2. If $\operatorname{rank}(n^{-1}\boldsymbol{\Xi}^T\mathbb{F}) < d$, then $\operatorname{rank}(\mathbf{K}) < d$ and again some signal will lost. In the extreme, if $\boldsymbol{\Xi}^T\mathbb{F} = 0$, then $h^{-1}\mathrm{E}(\widehat{\mathbf{K}}) = \kappa^2\bar{\rho}\mathbf{I}_r$ and all information on $\operatorname{span}(\boldsymbol{\Gamma})$ is lost.

If $\kappa \to 0$, then hn increases faster than p increases. In effect there is a synergy between the signal rate and the sample size. If the regression is abundant, so $h \asymp p$, then $\kappa \asymp n^{-1/2}$. Useful results can also be obtained when $h \asymp p^{2/3}$, because then $\kappa \asymp p^{1/6}n^{-1/2}$, which may be small in some regressions. In sparse regressions where $h \asymp 1$, $\kappa \asymp (p/n)^{1/2}$, and we are back to the usual requirement that $p/n \to 0$ for consistency of $\widehat{\mathbf{K}}$. Equally important, since κ does not depend on the weight matrix, the results indicate that it may not be possible to develop from the family of estimators considered rates of convergence that are faster than κ^{-1}.

The $\boldsymbol{\nu} \in \mathbb{R}^d$ and $O_p(\psi)$ terms arise from prediction. The first term $\boldsymbol{\nu}$ has a characteristic that is different from the others because it does not depend on n. Since $\operatorname{var}(\boldsymbol{\nu}) \le \mathbf{G}_h^{-1}\|\boldsymbol{\rho}\|/h$, the sufficient stability requirement $\|\boldsymbol{\rho}\| = O(h)$ of Lemma 4.1 guarantees that $\operatorname{var}(\boldsymbol{\nu})$ is bounded as $p \to \infty$. While the contribution of $\boldsymbol{\nu}$ will be negligible if $\operatorname{var}(\boldsymbol{\nu})$ is sufficiently small, for the best results we should have $\operatorname{var}(\boldsymbol{\nu}) \to 0$ as $p \to \infty$. Let $(\boldsymbol{\gamma}, \boldsymbol{\gamma}_0)$ be an orthogonal matrix. Using a result from Cook and Forzani [(2009), eq. (A.4)], $\boldsymbol{\gamma}^T\boldsymbol{\rho}\boldsymbol{\gamma} - (\boldsymbol{\gamma}^T\boldsymbol{\rho}^{-1}\boldsymbol{\gamma})^{-1} = \boldsymbol{\gamma}^T\boldsymbol{\rho}\boldsymbol{\gamma}_0(\boldsymbol{\gamma}_0^T\boldsymbol{\rho}\boldsymbol{\gamma}_0)^{-1}\boldsymbol{\gamma}_0^T\boldsymbol{\rho}\boldsymbol{\gamma}_0$. Consequently, $\operatorname{var}(\boldsymbol{\nu}) = 0$ when $\operatorname{span}(\boldsymbol{\gamma})$ is a reducing subspace of $\boldsymbol{\rho}$, even if $h \asymp 1$. This result is similar to Zyskind's (1967) classical findings about conditions for equality of the best and simple least squares linear estimators in linear models. If $\mathbf{W} = \boldsymbol{\Delta}^{-1}$, then $\boldsymbol{\rho} = \mathbf{I}_p$ and $\operatorname{span}(\boldsymbol{\gamma})$ is trivially a reducing subspace of $\boldsymbol{\rho}$. If $\boldsymbol{\Delta}$ is a generalized inverse of $\mathbf{W}$, $\mathbf{W}\boldsymbol{\Delta}\mathbf{W} = \mathbf{W}$, and if $\boldsymbol{\Gamma}^T\mathbf{W}\boldsymbol{\Gamma} > 0$, then again $\operatorname{span}(\boldsymbol{\gamma})$ reduces $\boldsymbol{\rho}$ and $\operatorname{var}(\boldsymbol{\nu}) = 0$.

Turning to $O_p(\psi)$, since $\|\boldsymbol{\rho}\|_F \le \sqrt{p}\|\boldsymbol{\rho}\|$, it follows that $\psi \le \kappa\|\boldsymbol{\rho}\|/\sqrt{h}$. Consequently, if $\|\boldsymbol{\rho}\| = O(\sqrt{h})$, then $O_p(\kappa) + O_p(\psi) = O_p(\kappa)$ and $\widehat{\mathbf{R}}_{\mathbf{W}}(\mathbf{X}_N) -$

$\mathbf{R}(\mathbf{X}_N) = \boldsymbol{\nu} + O_p(\kappa)$. In a worst-case scenario, if $\psi = \kappa\|\boldsymbol{\rho}\|/\sqrt{h}$ and $\|\boldsymbol{\rho}\|/\sqrt{h}$ diverges, then $O_p(\kappa) + O_p(\psi) = O_p(\sqrt{p}/\sqrt{n})$ because of the requirement $\|\boldsymbol{\rho}\| = O(h)$ in Proposition 5.1(iii), and these terms reduce to the usual condition that $p/n \to 0$.

In sparse regressions, the condition $\kappa \to 0$ reduces to $p/n \to 0$ and $\|\boldsymbol{\rho}\| = O(h)$ means that $\|\boldsymbol{\rho}\|$ must be bounded. These two conditions imply that $\psi \to 0$. Consequently, the best results in sparse regressions will be achieved when (a) $\operatorname{span}(\boldsymbol{\gamma})$ is a reducing subspace of $\boldsymbol{\rho}$, (b) $p/n \to 0$ and (c) $\|\boldsymbol{\rho}\|$ is bounded. In nonsparse settings $\widehat{\mathbf{R}}_{\mathbf{W}}$ will yield the best results when $\|\boldsymbol{\rho}\|$ is bounded and the regression is abundant or near abundant. We summarize some implications of these conditions in the following corollary.

COROLLARY 5.1. *Assume that* $\|\boldsymbol{\rho}\| = O(1)$ *and that* $\mathbf{S} = 0$. *Then*:

(i) $\widehat{\mathbf{R}}_{\mathbf{W}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(h^{-1/2}) + O_p(\kappa)$. *If the regression is abundant, then* $\widehat{\mathbf{R}}_{\mathbf{W}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(n^{-1/2})$.

(ii) *If* $\operatorname{span}(\boldsymbol{\gamma})$ *is a reducing subspace of* $\boldsymbol{\rho}$, *then* $\widehat{\mathbf{R}}_{\mathbf{W}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(\kappa)$. *If, in addition, the regression is abundant, then* $\widehat{\mathbf{R}}_{\mathbf{W}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(n^{-1/2})$.

The conclusions of Proposition 5.1 and its Corollary 5.1 hold as $n \to \infty$ and $p \to \infty$ in the required relationships and as $n \to \infty$ with p fixed. In the latter case the results simplify to those given in the next corollary.

COROLLARY 5.2. *If* $p = O(1)$ *and* $\mathbf{S} = 0$, *then*:

(i) $\widehat{\mathbf{R}}_{\mathbf{W}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = \operatorname{var}(\boldsymbol{\nu}) + O_p(n^{-1/2})$.

(ii) *If, in addition,* $\operatorname{span}(\boldsymbol{\gamma})$ *is a reducing subspace of* $\boldsymbol{\rho}$, *then* $\widehat{\mathbf{R}}_{\mathbf{W}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(n^{-1/2})$.

It may be clear from the previous discussion that the best possible rate is achieved when $\widehat{\mathbf{W}} = \mathbf{W} = \boldsymbol{\Delta}^{-1}$ and then $\widehat{\mathbf{R}}_{\mathbf{W}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(\kappa)$. Consequently, we refer to κ^{-1} as the *oracle rate*.

5.2. *Simulations*. We conducted a simulation study with $\widehat{\mathbf{W}} = \mathbf{W} = \mathbf{I}_p$ to show the importance of $\boldsymbol{\nu}$, to demonstrate that Proposition 5.1 and Corollaries 5.1 and 5.2 give good qualitative characterizations of the limiting behavior of $\widehat{\mathbf{R}}_{\mathbf{I}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N)$, and to provide some intuition into the nonasymptotic behavior of $\widehat{\mathbf{R}}_{\mathbf{I}}$. The simulated data were generated using a simple version of model (2.2): $\mathbf{X} = \boldsymbol{\Gamma} Y + \boldsymbol{\varepsilon}$, $Y \sim N(0, 1)$, $\boldsymbol{\Gamma} \in \mathbb{R}^p$, was constructed as a vector of standard normal random variables and $\boldsymbol{\Delta}$ is a diagonal matrix. Specific scenarios may differ on n, p and the choice of $\boldsymbol{\Delta}$. In any case, the true model then has $d = 1$. We confined attention to regressions with $d = 1$ because we found that there is nothing in principle to be gained from settings with $d > 1$. For each sample size n we used the

estimators described at the end of Section 3 with $\widehat{\mathbf{W}} = \mathbf{I}_p$ and Y^j, $j = 1, 2, 3, 4$, as the elements of $\mathbf{f}$, so $\boldsymbol{\rho} = \boldsymbol{\Delta}$ and $r = 4$.

For each data set constructed in this way, we determined $\widehat{\mathbf{R}}_{\mathbf{I}}(\mathbf{X}_{N,j})$ and $\mathbf{R}(\mathbf{X}_{N,j})$ at $j = 1, \ldots, 100$ new data points generated from the original model. We summarized each data set by computing the absolute sample correlation between $\widehat{\mathbf{R}}_{\mathbf{I}}$ and $\mathbf{R}$ and the sample standard deviation of the difference $\widehat{\mathbf{R}}_{\mathbf{I}} - \mathbf{R}$ over the 100 new data points. This process was repeated at least 400 times for $p \leq 300$, 200 times for $300 < p < 800$ and 50 times for $p > 800$. The average absolute correlation and average standard deviation were used as overall summary statistics. Our decision to use a diagonal matrix for $\boldsymbol{\Delta}$ was based on the observation that the asymptotic results depend on $\boldsymbol{\Delta}$ largely via $\|\boldsymbol{\rho}\|$ and with $\widehat{\mathbf{W}} = \mathbf{I}_p$, $\|\boldsymbol{\rho}\| = \|\boldsymbol{\Delta}\|$.

Simulation I: *Illustrations of Proposition* 5.1. In the first set of simulations we fixed $p = 50$ and generated $\boldsymbol{\Gamma}$ once at the outset, giving $\|\boldsymbol{\Gamma}\| = 8.2$. The diagonal elements of $\boldsymbol{\Delta}$ were 50 regularly spaced values between 1 and 101. According to Corollary 5.2(i), as $n \to \infty$ the variance of $\widehat{\mathbf{R}}_{\mathbf{I}} - \mathbf{R}$ should converge to $\operatorname{var}(\boldsymbol{\nu}) = 0.595^2$, where the value was determined by substituting the simulation parameters into (5.2). This conclusion is supported by the results in Figure 1(a), where we see that the standard deviation curve A converges nicely to the predicted asymptotic value marked by line segment B. Curve C in Figure 1(a) was generated by setting $\boldsymbol{\Gamma} = (8.2, 0, \ldots, 0)^T$, so $\operatorname{span}(\boldsymbol{\gamma})$ is now a reducing subspace of $\boldsymbol{\rho} = \boldsymbol{\Delta}$. In this case Corollary 5.2(ii) predicts that $\widehat{\mathbf{R}}_{\mathbf{I}} - \mathbf{R}$ will converge to 0 at the usual root-n rate. That conclusion is supported by curve C. Curves A and C in Figure 1(b)

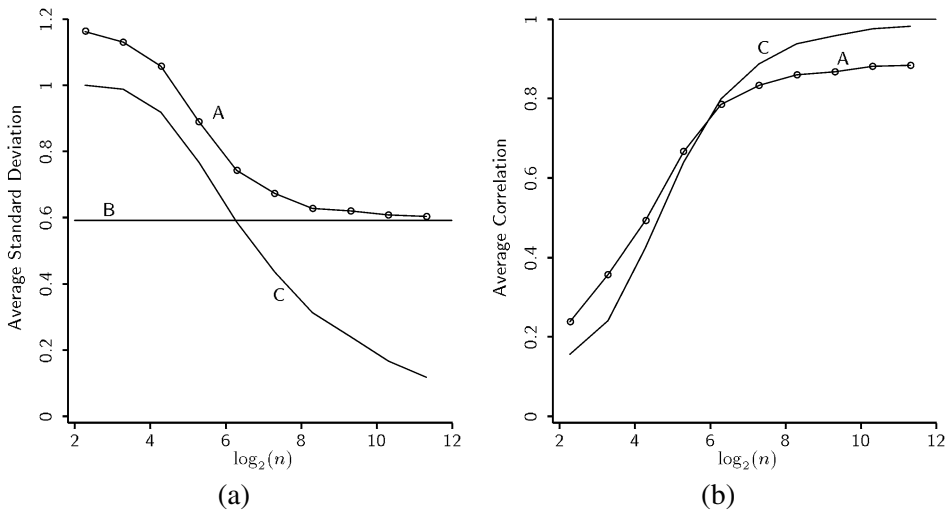

FIG. 1. *Results of the first simulation described in Section 5.2 with fixed* $p = 50$, $\widehat{\mathbf{W}} = \mathbf{I}_p$, $Y \sim N(0, 1)$ *and* $\operatorname{diag}(\boldsymbol{\Delta}) \sim \operatorname{Uniform}(1, 101)$. *A*: $\boldsymbol{\Gamma}$ *generated as a vector of* $N(0, 1)$ *random variables. B*: *Theoretical lower bound on the standard deviation of* $\widehat{\mathbf{R}}_{\mathbf{I}} - \mathbf{R}$ *for curve A. C*: $\boldsymbol{\Gamma} = (8.2, 0, \ldots, 0)^T$. (a) *Standard deviation of* $\widehat{\mathbf{R}}_{\mathbf{I}} - \mathbf{R}$; (b) *correlation* $(\widehat{\mathbf{R}}_{\mathbf{I}}, \mathbf{R})$.

correspond to those in Figure 1(a), except the average correlation is plotted on the vertical axis. Evidently the correlations for curve A are converging to a value close to 0.9, while those for curve C are converging to 1. Curve A suggests that when $\operatorname{var}(\boldsymbol{\nu}) > 0$ we might still gain useful results in practice if the correlation is sufficiently large.

Simulation II: *Bounded and unbounded error variances.* The structure of the second set of simulations was like the first set, except we varied $n = p/2$ and, at the start of each replication, $\boldsymbol{\Gamma}$ and Y were generated anew and the diagonal elements of $\boldsymbol{\Delta}$ were sampled from the uniform distribution on the interval $(1, u)$, where $u = 101$ or $u = p + 1$. The regeneration at the start of each replication was used to avoid tying results to a particular parameter configuration. For this simulation the ratio of the systematic variation in $\mathbf{X}$ to its total variation is $\operatorname{var}^{-1/2}(\mathbf{X}) \operatorname{var}(\mathrm{E}(\mathbf{X}|\boldsymbol{\Gamma}, Y, \boldsymbol{\Delta})) \operatorname{var}^{-1/2}(\mathbf{X}) = (2/(u+3))\mathbf{I}_p$, where the moments were computed over the simulation distributions of $\mathbf{X}$, Y, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Delta}$. Consequently, it seems fair to characterize the signal as weak when $u \geq 101$, since then the systematic variation accounts for less than 2% of the total variation.

The regression is abundant and $\kappa = o(1)$ in this simulation. Additionally, when $u = 101$, $\|\boldsymbol{\rho}\|$ is bounded and the results of Corollary 5.1(i) apply, $\widehat{\mathbf{R}}_{\mathbf{I}} - \mathbf{R} = O_p(n^{-1/2})$. The simulation results for this case, which are shown by the curves labeled "$u = 101$" in Figure 2, support the claim that standard deviation of $\widehat{\mathbf{R}}_{\mathbf{I}} - \mathbf{R}$ is converging to 0 and the correlation between $\widehat{\mathbf{R}}_{\mathbf{I}}$ and $\mathbf{R}$ is approaching 1. On the other hand, when $u = p + 1$, $\|\boldsymbol{\rho}\|$ is unbounded and the results of Corollary 5.1 do not apply. However, we still might expect the standard deviation and correlation to

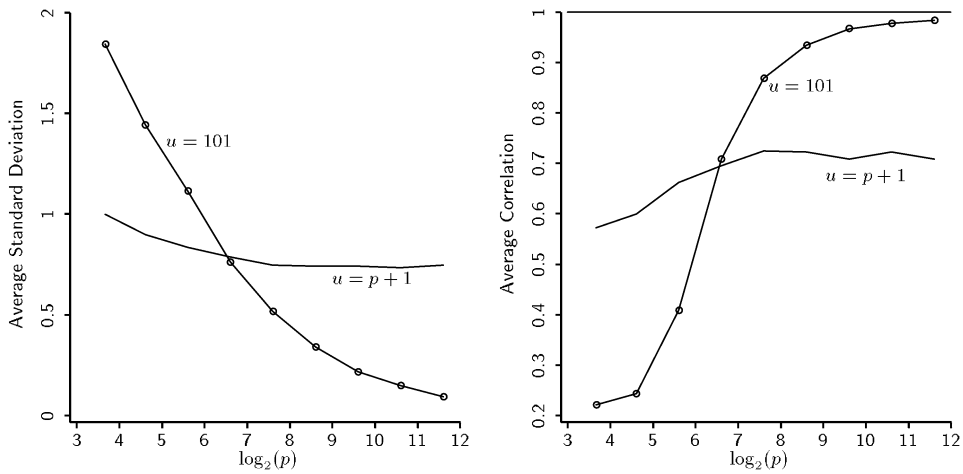

FIG. 2. *Results from simulation II described in Section 5.2 with* $n = p/2$, $\mathbf{W} = \mathbf{I}_p$, $Y \sim N(0, 1)$, $\boldsymbol{\Gamma} \sim N(0, \mathbf{I}_p)$ *and* $\operatorname{diag}(\boldsymbol{\Delta}) \sim \operatorname{Uniform}(1, u)$. (a) *Standard deviation of* $\widehat{\mathbf{R}}_{\mathbf{I}} - \mathbf{R}$; (b) *correlation* $(\widehat{\mathbf{R}}_{\mathbf{I}}, \mathbf{R})$.

converge to values away from 0 and 1. The simulation results shown by the curves labeled $u = p + 1$ in Figure 2 sustain this expectation.

5.3. *SPICE.* Beginning with a little background, we turn in this section to weight matrices $\widehat{\mathbf{W}} \in \mathbb{R}^{p\times p}$ selected by using SPICE.

Restricting attention to population weight matrices that are equal to the inverse of the error covariance matrix, $\mathbf{W} = \mathbf{\Delta}^{-1}$, allows for the application of some modern regularized inverse covariance estimators with reasonable convergence rates of $\widehat{\mathbf{W}} \in \mathbb{R}^{p\times p}$ as $n, p \to \infty$. For example, rates of convergence have been established for banding or tapering [Bickel and Levina (2008a)] and element-wise thresholding [Bickel and Levina (2008b)] of the sample covariance matrix. However, using these approaches to estimate the population weight matrix is problematic. Banding/tapering the sample covariance matrix is not invariant under permutations of variable labels, and in finite samples, both banding/tapering and thresholding the sample covariance matrix may produce an estimator of the covariance matrix that is not positive definite. To avoid these problems, we further restrict our attention to covariance estimators that are positive definite in finite samples and are invariant to permutations of variable labels. Several authors have recently analyzed the high-dimensional inverse covariance estimator formed through L_1-penalized likelihood optimization. We will use a particular version to estimate the population weight matrix (equivalently the inverse error covariance matrix) which was named SPICE by Rothman et al. (2008) for "sparse permutation invariant covariance estimator."

Let $\tilde{n} = n - r - 1$ and let

$$\widehat{\mathbf{\Delta}} = \tilde{n}^{-1}\mathbb{Z}^T\mathbf{Q}_{\mathbb{F}}\mathbb{Z} \in \mathbb{R}^{p\times p} \tag{5.3}$$

denote the residual sample covariance matrix from the linear regression of $\mathbf{X}$ on $\mathbf{f}$. We construct $\widehat{\mathbf{W}}_\lambda = \widehat{\mathbf{\Delta}}_\lambda^{-1}$ using

$$\begin{aligned}\widehat{\mathbf{\Omega}}_\lambda &= \arg\min_{\mathbf{\Omega}>0}\Bigg[\operatorname{tr}\{\operatorname{diag}^{-1/2}(\widehat{\mathbf{\Delta}})\widehat{\mathbf{\Delta}}\operatorname{diag}^{-1/2}(\widehat{\mathbf{\Delta}})\mathbf{\Omega}\} - \log|\mathbf{\Omega}| + \lambda\sum_{i\neq j}|\Omega_{ij}|\Bigg],\\ \widehat{\mathbf{\Delta}}_\lambda^{-1} &= \operatorname{diag}^{-1/2}(\widehat{\mathbf{\Delta}})\widehat{\mathbf{\Omega}}_\lambda\operatorname{diag}^{-1/2}(\widehat{\mathbf{\Delta}}),\end{aligned} \tag{5.4}$$

where $\lambda \geq 0$ is a tuning parameter and Ω_{ij} is the (i, j) element of $\mathbf{\Omega}$. The optimization in (5.4) produces a sparse estimator $\widehat{\mathbf{\Omega}}_\lambda$ of the inverse error correlation matrix $\mathbf{\Omega}$, which is then rescaled by the residual sample standard deviations to give the inverse error covariance estimator $\widehat{\mathbf{\Delta}}_\lambda^{-1}$. The use of L_1-penalized likelihood optimization to estimate a sparse inverse covariance matrix has been studied extensively in the literature [d'Aspremont, Banerjee and El Ghaoui (2008), Yuan and Lin (2007), Friedman, Hastie and Tibshirani (2008), Rothman et al. (2008), Lam and Fan (2009)]. Although we impose sparsity on the off-diagonal entries of the inverse error covariance matrix $\mathbf{\Delta}^{-1}$, we expect that in many situations $\mathbf{\Delta}^{-1}$ may not be sparse. We selected this estimator since it is able to adapt, with its tuning parameter, to both sparse and less sparse estimates, and can lead to a substantial

reduction in variability of our reduction estimator when p is large; however, using small values of the tuning parameter λ to give less-sparse inverse error covariance estimates leads to slow convergence of algorithms to solve (5.4) when p is large.

Rothman et al. (2008) established the consistency of $\widehat{\boldsymbol{\Delta}}_\lambda^{-1}$ in a special case of model (2.2) characterized by the following four conditions: (A) $r = 0$, so $\widehat{\boldsymbol{\Delta}} = \mathbb{Z}^T\mathbb{Z}/\tilde{n}$ is the usual estimator of the marginal covariance matrix of $\mathbf{X}$. (B) The errors $\boldsymbol{\varepsilon}$ are normally distributed with mean 0 and variance $\boldsymbol{\Delta}$. (C) The largest $\varphi_{\max}(\boldsymbol{\Delta})$ and smallest $\varphi_{\min}(\boldsymbol{\Delta})$ eigenvalues of $\boldsymbol{\Delta}$ are uniformly bounded; that is, for all p

$$0 < \underline{k} \le \varphi_{\min}(\boldsymbol{\Delta}) \le \varphi_{\max}(\boldsymbol{\Delta}) \le \overline{k} < \infty, \tag{5.5}$$

where $\underline{k}$ and $\overline{k}$ are constants. (D) $\lambda \asymp \sqrt{\frac{\log p}{n}}$. Then [Rothman et al. (2008)]

$$\|\widehat{\boldsymbol{\Delta}}_\lambda^{-1} - \boldsymbol{\Delta}^{-1}\| = O_p\big(\sqrt{n^{-1}\big(s(p)+1\big)\log p}\big), \tag{5.6}$$

where $s(p)$ is the total number of nonzero off-diagonal entries of $\boldsymbol{\Delta}^{-1}$, which could grow with p.

In this application $\mathbf{S} = \mathbf{W}^{-1/2}(\widehat{\mathbf{W}} - \mathbf{W})\mathbf{W}^{-1/2} = \boldsymbol{\Delta}^{1/2}(\widehat{\boldsymbol{\Delta}}_\lambda^{-1} - \boldsymbol{\Delta}^{-1})\boldsymbol{\Delta}^{1/2}$. To find the order of $\|\mathbf{S}\|$ and thereby allow application of Proposition 5.1 with SPICE, we relax the conditions of Rothman et al. (2008) by allowing (A*) $r > 0$ and assuming that (B*) the elements ε_j of $\boldsymbol{\varepsilon}$ are *uniformly sub-Gaussian random variables*. That is, we assume there exist positive constants K_1 and K_2 such that for all $t > 0$ and all p

$$P(|\varepsilon_j| > t) \le K_1 e^{-K_2 t^2}, \qquad j = 1, \ldots, p. \tag{5.7}$$

This assumption is compatible with universal conditions (W.1) and (W.2). When the ε_j's are normal, it is equivalent to requiring that their variances are uniformly bounded as $p \to \infty$. The following lemma gives the orders of $\|\mathbf{S}\|$ and $\|\mathrm{E}(\mathbf{S}^2)\|$ required for conditions (S.1) and (S.2).

LEMMA 5.1. *Let $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Delta}}_\lambda^{-1}$, let $s \equiv s(p)$ be the total number of nonzero off-diagonal entries of $\boldsymbol{\Delta}$ and let $\omega_{\mathrm{spice}}(n, p) = \sqrt{n^{-1}(s+1)\log p}$. Under conditions* (B*), (C) *and* (D), $\|\mathbf{S}\| = O_p(\omega_{\mathrm{spice}})$ *and* $\|\mathrm{E}(\mathbf{S}^2)\| = O(\omega_{\mathrm{spice}}^2)$.

Applying Proposition 5.1(iii) in the context of SPICE, $\boldsymbol{\rho} = \mathbf{I}_p$ because $\mathbf{W} = \boldsymbol{\Delta}^{-1}$. Consequently, $\mathrm{var}(\boldsymbol{\nu}) = 0$ and $\psi = \sqrt{p}/h\sqrt{n} \le \kappa$. From this we immediately get the following convergence rate bound for $\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\Delta}}_\lambda} \equiv \widehat{\mathbf{R}}_{\mathrm{spice}}$.

PROPOSITION 5.2. *Assume that the conditions and notation of Lemma 5.1 hold. Then*

$$\widehat{\mathbf{R}}_{\mathrm{spice}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(\kappa) + O_p(\omega_{\mathrm{spice}}). \tag{5.8}$$

If the number s of nonzero off-diagonal entries of $\mathbf{\Delta}^{-1}$ is bounded as $p \to \infty$ and the regression is abundant ($\kappa \asymp n^{-1/2}$), then $\widehat{\mathbf{R}}_{\text{spice}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(n^{-1/2}\log^{1/2} p)$. This allows for the number of variables p to grow much faster than the sample size, so long as s is bounded. On the other extreme, if there is no sparsity so that $s = p(p-1)$, then the bounding rate implied in Proposition 5.2 would seem to indicate that n needs to be large relative to $\{p(p-1)+1\}\log p$ for good results. Based on our results for normal errors in Section 6, we anticipate that this rate is not sharp, most notably when $\mathbf{\Delta}^{-1}$ is not sparse, and that it can be improved, particularly when additional structure is imposed (see Section 6).

Conditions (B*), (C) and (D) required for Proposition 5.2 are in addition to the active universal constraints stated in Section 4. Of the universal constraints, we still require the signal rate property (4.1), the order $\kappa = O(1)$ and (W.2). The remaining universal constraints are implied by the conditions of the proposition: since $\boldsymbol{\rho} = \mathbf{I}_p$, the regression is stable by Lemma 4.1, (W.1) holds and $\psi \le \kappa$ so (4.3) holds.

5.4. *A diagonal weight matrix.* We include in this section a discussion of the asymptotic behavior of the reduction based on the diagonal weight matrix $\widehat{\mathbf{W}} = \text{diag}^{-1}(\widehat{\mathbf{\Delta}})$, where $\widehat{\mathbf{\Delta}}$ was defined in (5.3). This weight matrix, which corresponds to the population weight matrix $\mathbf{W} = \text{diag}^{-1}(\mathbf{\Delta})$, ignores the residual correlations and adjusts only for residual variances. However, in contrast to SPICE, there is no tuning parameter involved and so its computation is not an issue.

With $\mathbf{W} = \text{diag}^{-1}(\mathbf{\Delta})$, $\boldsymbol{\rho}$ is the residual correlation matrix and

$$\mathbf{S} = \text{diag}^{1/2}(\mathbf{\Delta})\big(\text{diag}^{-1}(\widehat{\mathbf{\Delta}}) - \text{diag}^{-1}(\mathbf{\Delta})\big)\,\text{diag}^{1/2}(\mathbf{\Delta}).$$

The following lemma gives the orders of $\|\mathbf{S}\|$ and $\|\mathrm{E}(\mathbf{S}^2)\|$ in preparation for application of Proposition 5.1.

LEMMA 5.2. *Let* $\widehat{\mathbf{W}} = \text{diag}^{-1}(\widehat{\mathbf{\Delta}})$, *let* $\omega_{\text{diag}} = n^{-1/2}\log^{1/2} p$ *and assume that the elements* ε_j *of the errors* $\boldsymbol{\varepsilon}$ *are sub-Gaussian random variables. Then* $\|\mathbf{S}\| = O_p(\omega_{\text{diag}})$ *and* $\|\mathrm{E}(\mathbf{S}^2)\| = O(\omega_{\text{diag}}^2)$.

In contrast to Lemma 5.1, here we require only that the errors are sub-Gaussian and not uniformly sub-Gaussian. This is because the diagonal weight matrix effectively standardizes the variances.

PROPOSITION 5.3. *Assume that the conditions and notation of Lemma 5.2 hold. Then*

$$\widehat{\mathbf{R}}_{\text{diag}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = \boldsymbol{\nu} + O_p(\kappa) + O_p(\psi) + O_p(\omega_{\text{diag}}). \tag{5.9}$$

The order (5.9) for the diagonal weight matrix can be smaller than, equal to, or greater than the order (5.8) for the SPICE weight matrix, depending on

the underlying structure of the regression. Using results from the discussion of Section 5.1, if $\|\boldsymbol{\rho}\| = O(\sqrt{h})$ and $\operatorname{span}(\boldsymbol{\gamma})$ is a reducing subspace of $\boldsymbol{\rho}$, then $\widehat{\mathbf{R}}_{\text{diag}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(\kappa) + O_p(\omega_{\text{diag}})$, and thus the diagonal weight matrix can produce a better rate because $\omega_{\text{diag}} \le \omega_{\text{spice}}$. If $\boldsymbol{\Delta}$ is itself a diagonal matrix, then $s = 0$, $\operatorname{var}(\boldsymbol{\nu}) = 0$, $\kappa \ge \psi$ and $\omega_{\text{diag}} = \omega_{\text{spice}}$, and consequently

$$\widehat{\mathbf{R}}_{\text{diag}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(\kappa) + O_p(\omega_{\text{diag}}). \tag{5.10}$$

In this case the two weight matrices result in the same order, so there seems to be no asymptotic loss incurred by SPICE when $\boldsymbol{\Delta}$ is diagonal.

6. Normal errors and $\boldsymbol{\xi} = \boldsymbol{\beta}\mathbf{f}$. We consider in this section the relatively ideal situation in which the errors $\boldsymbol{\varepsilon}$ are normally distributed and the user-selected basis function $\mathbf{f}$ correctly models the true coordinates, so $\boldsymbol{\xi}_i = \boldsymbol{\beta}\mathbf{f}_i$, $i = 1, \ldots, n$, for some fixed matrix $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$ of rank d. The diagonal weight matrix is revisited under these assumptions in Section 6.1. In Section 6.2 we add the sample size constraint $n > p + r + 4$ so $\widehat{\boldsymbol{\Delta}} > 0$. The importance of this setting is that we can obtain the oracle rate.

6.1. *Diagonal weight matrix.* The rates (5.10) for a diagonal $\boldsymbol{\Delta}$ can be sharpened considerably when the errors are normal and $\boldsymbol{\xi} = \boldsymbol{\beta}\mathbf{f}$:

PROPOSITION 6.1. *Assume that* $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Delta})$, $\boldsymbol{\Delta}$ *is a diagonal matrix*, $\boldsymbol{\xi} = \boldsymbol{\beta}\mathbf{f}$, $n > r + 5$ *and* $\widehat{\mathbf{W}} = \operatorname{diag}^{-1}(\widehat{\boldsymbol{\Delta}})$. *Then* $\widehat{\mathbf{R}}_{\text{diag}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(\kappa)$.

We see from this proposition that when the errors are normal, the order $O_p(\omega_{\text{diag}})$ that appears in (5.10) is no longer present and the rate for $\widehat{\mathbf{R}}_{\text{diag}}$ reduces to the oracle rate κ^{-1}. The condition $n > r + 5$ stated in Proposition 6.1 is needed to insure the existence of the variance of an inverse chi-squared random variable. Its proof is omitted from the supplement since it follows the same lines as the proof of Proposition 5.1.

6.2. $n > p + r + 4$. Recalling that $\widehat{\boldsymbol{\Delta}} \in \mathbb{R}^{p\times p}$ is the residual covariance matrix defined in (5.3), this constraint on n means that $\widehat{\boldsymbol{\Delta}} > 0$ with probability 1, allowing the straightforward use of $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Delta}}^{-1}$ as the weight matrix. The population weight matrix is $\mathbf{W} = \boldsymbol{\Delta}^{-1}$ with corresponding $\mathbf{S} = \boldsymbol{\Delta}^{1/2}\widehat{\boldsymbol{\Delta}}^{-1}\boldsymbol{\Delta}^{1/2} - \mathbf{I}_p$. The asymptotic behavior of $\|\mathbf{S}\|$ in this setting can be obtained as follows: phrased in the context of this article, suppose that $r = 0$ so $\widehat{\boldsymbol{\Delta}} = (n-1)^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$. Johnstone (2001) [see also Bai (1999), page 635] showed that $\varphi_{\max}(\boldsymbol{\Delta}^{-1/2}\widehat{\boldsymbol{\Delta}}\boldsymbol{\Delta}^{-1/2}) - 1 \asymp_p (\sqrt{p/n})$ and Paul (2005) showed that $\varphi_{\min}(\boldsymbol{\Delta}^{-1/2}\widehat{\boldsymbol{\Delta}}\boldsymbol{\Delta}^{-1/2}) - 1 \asymp_p (\sqrt{p/n})$ when $p/n \to a \in [0, 1)$. Together these results imply that $\|\mathbf{S}\| \asymp_p (\sqrt{p/n})$ and therefore $\|\mathbf{S}\| \to 0$ in probability if and only if $p/n \to 0$, which gives the order for condition (S.1). Still with $r = 0$, since $(n-1)\widehat{\boldsymbol{\Delta}}^{-1}$ is distributed as the inverse of

a Wishart $W_p(\boldsymbol{\Delta}, n-1)$ matrix, we used results from von Rosen (1988) to verify condition (S.2). Combining these results with Proposition 5.1 we have

$$\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\Delta}}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p\big(\sqrt{p/n}\big), \tag{6.1}$$

where $\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\Delta}}}$ denotes the reduction with $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Delta}}^{-1}$. This suggests that it may be reasonable to use $\widehat{\boldsymbol{\Delta}}^{-1}$ as the weight matrix only when n is large relative to p, essentially sending us back to the usual requirement. In the next proposition we show this is not the case with $r > 0$ because the order in (6.1) is in fact too large.

PROPOSITION 6.2. *Assume that* $\boldsymbol{\varepsilon} \sim N_p(0, \boldsymbol{\Delta})$, $\boldsymbol{\xi} = \boldsymbol{\beta}\mathbf{f}$, $n > p + r + 4$ *and* $p/n \to [0, 1)$. *Let* $a = (n - p - 1)^{-1}$. *Then when* $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Delta}}^{-1}$:

(i) $h^{-1}\mathrm{E}(\widehat{\mathbf{K}}) = a(n - r - 1)\{\mathbf{K} + \kappa^2\mathbf{I}_r\}$,
(ii) $h^{-1}(p)\{\widehat{\mathbf{K}} - \mathrm{E}(\widehat{\mathbf{K}})\} = O_p(1/\sqrt{n})$,
(iii) $\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\Delta}}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(\kappa)$.

Recall from Corollary 5.1 that when $\widehat{\mathbf{W}} = \boldsymbol{\Delta}^{-1}$, we obtain the oracle rate $\widehat{\mathbf{R}}_{\boldsymbol{\Delta}}(\mathbf{X}_N) - \mathbf{R}(\mathbf{X}_N) = O_p(\kappa)$. Comparing this result with the conclusion of Proposition 6.2(iii), we see that there is no impact on the rate of convergence when using $\widehat{\mathbf{W}} = \widehat{\boldsymbol{\Delta}}^{-1}$ instead of $\widehat{\mathbf{W}} = \boldsymbol{\Delta}^{-1}$, both weight matrices giving estimators that converge at the oracle rate. In contrast to the gross bound given in (6.1), Proposition 6.2 indicates that it may in fact be quite reasonable to use $\widehat{\boldsymbol{\Delta}}^{-1}$ as the weight matrix when $n > p + r + 4$, without requiring n to be large relative to p.

Properties of the normal and the accurate modeling of $\boldsymbol{\xi} = \boldsymbol{\beta}\mathbf{f}$ surely contribute to the oracle rate of Proposition 6.2. Recall from (3.1) that $\widehat{\mathbf{K}} = \boldsymbol{\Phi}_n^{1/2}\widehat{\mathbf{B}}^T\widehat{\mathbf{W}}\widehat{\mathbf{B}}\boldsymbol{\Phi}_n^{1/2}$. Under the conditions of Proposition 6.2, $\widehat{\mathbf{B}} \perp\!\!\!\perp \widehat{\mathbf{W}}$, which is not required for our most general results given in Proposition 5.1. Additionally, we were able to use exact calculations in places where bounding was otherwise employed. The inverse residual sum of squares matrix $(n - r - 1)\widehat{\boldsymbol{\Delta}}^{-1}$ is distributed as the inverse of a Wishart $W_p(\boldsymbol{\Delta}, n - r - 1)$ matrix and the moments of an inverse Wishart derived by von Rosen (1988) were used extensively in the proof of Proposition 6.2. For instance, although $\widehat{\boldsymbol{\Delta}}$ has an inverse with probability 1 when $n > p + r + 1$, the constraint on n in the hypothesis is needed to ensure the existence of the variance of the inverse Wishart.

The signal rate (4.1) and the order $\kappa = O(1)$ are still required for Proposition 6.2, but its hypothesis implies all the other universal conditions stated in Section 4: since $\mathbf{W} = \boldsymbol{\Delta}^{-1}$ we have $\boldsymbol{\rho} = \mathbf{I}_p$ and thus (1) the regression is stable by Lemma 4.1, (2) $\psi \leq \kappa$ so (4.3) holds, (3) $\mathrm{E}(\boldsymbol{\varepsilon}^T\mathbf{W}\boldsymbol{\varepsilon}) = p$ so condition (W.1) holds, and (4) condition (W.2) holds by Lemma 4.2(i).

7. Simulations. The effects of predictor correlations are often a concern when dealing with high-dimensional regressions. In this section, we introduce predictor

correlations into our simulations to give some intuition about their impact on the four reduction estimators.

The computation of the four reduction estimators $\widehat{\mathbf{R}}_{\mathbf{I}}$, $\widehat{\mathbf{R}}_{\text{spice}}$, $\widehat{\mathbf{R}}_{\text{diag}}$ and $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$ relies on the computation of their weight matrix estimators $\widehat{\mathbf{W}}$, which are available in closed form for $\widehat{\mathbf{R}}_{\mathbf{I}}$, $\widehat{\mathbf{R}}_{\text{diag}}$ and $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$; however, computing the weight matrix estimator for $\widehat{\mathbf{R}}_{\text{spice}}$ requires us to select its tuning parameter λ and to use an iterative algorithm.

7.1. *Computation and tuning parameter selection for* $\widehat{\mathbf{R}}_{\text{spice}}$. Several computational algorithms [d'Aspremont, Banerjee and El Ghaoui (2008), Yuan and Lin (2007), Friedman, Hastie and Tibshirani (2008), Rothman et al. (2008)] have been proposed to compute the solution to (5.4), of which we propose to use the graphical lasso (glasso) algorithm of Friedman, Hastie and Tibshirani (2008). We employ K-fold cross-validation to select the tuning parameter λ for the weight matrix estimator $\widehat{\mathbf{W}}_\lambda \equiv \widehat{\mathbf{\Delta}}_\lambda^{-1}$, where we minimize the validation negative log-likelihood. Specifically, we solve $\widehat{\lambda} = \arg\min_\lambda \sum_{k=1}^K g_k(\lambda)$ where

$$g_k(\lambda) = \operatorname{tr}\bigl[n_k^{-1}\mathbf{A}_{(k,\lambda)}^T\mathbf{A}_{(k,\lambda)}\widehat{\mathbf{W}}_\lambda^{(-k)}\bigr] - \log\bigl|\widehat{\mathbf{W}}_\lambda^{(-k)}\bigr|,$$
$$\mathbf{A}_{(k,\lambda)} = \mathbb{Z}^{(k)} - \mathbb{F}^{(k)}\widehat{\mathbf{b}}_\lambda^{(-k)T}\widehat{\mathbf{\Gamma}}_\lambda^{(-k)T}.$$

In the above expression, a superscript of (k) indicates that the quantity is based on the observations inside the kth fold, and a superscript of $(-k)$ indicates that the quantity is based on observations outside of the kth fold, and n_k is the number of observations in the kth fold.

7.2. *Simulation settings*. The simulated data were generated from $\mathbf{X}_i = \mathbf{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i$, $i = 1, \ldots, n$, where $Y_1, \ldots, Y_n$ is a sequence of independent random variables, $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are independent copies of $\boldsymbol{\varepsilon}$, a p-variate normal random vector with mean 0 and variance $\mathbf{\Delta}$, and $\boldsymbol{\xi}_i = \boldsymbol{\xi}(Y_i)$ is specified later. As in the second simulation of Section 5.2, $\mathbf{\Gamma}$ and $\mathbf{\Delta}$ were produced anew at the start of each replication, with the elements of $\mathbf{\Gamma} \in \mathbb{R}^p$ always generated as a sequence of independent standard normal variables and $\mathbf{\Delta}$ generated as $\mathbf{D}^{1/2}\mathbf{\Theta}\mathbf{D}^{1/2}$ where $\mathbf{D}$ is a diagonal matrix with diagonal elements sampled from a uniform $(1, 101)$ distribution and $\mathbf{\Theta} = (\theta_{ij})$ is a correlation matrix with exponentially decreasing correlations, $\theta_{ij} = \theta^{|i-j|}$ with parameter $0 < \theta < 1$. This structure has been used frequently to assess the impact of predictor correlations on high-dimensional methodology [see, e.g., Bickel and Levina (2008b), Li and Yin (2008)]. For all four reduction estimators, $h \asymp p$ implying $\kappa = n^{-1/2}$.

For each replication, we determined $\widehat{\mathbf{R}}(\mathbf{X}_{N,j})$ and $\mathbf{R}(\mathbf{X}_{N,j})$ at $j = 1, \ldots, 100$ new data points generated from the original model. We assess performance by computing the magnitude of the sample correlation between $\widehat{\mathbf{R}}$ and $\mathbf{R}$ over the 100 new data points. The results are based on 200 independent replications.

For a given weight matrix $\mathbf{W}$, the expected signal strength over simulation replications is determined by $\mathrm{E}_{\mathrm{sim}}(\boldsymbol{\Gamma}^T\mathbf{W}\boldsymbol{\Gamma}) = \mathrm{E}_{\mathrm{sim}}\,\mathrm{tr}(\mathbf{W})$, where $\mathrm{E}_{\mathrm{sim}}$ denotes expectation over the simulations. For $\mathbf{W} = \mathrm{diag}^{-1}(\boldsymbol{\Delta})$, $\mathrm{E}_{\mathrm{sim}}[\mathrm{tr}(\mathbf{W})] = \log(101)/100$, but for $\mathbf{W} = \boldsymbol{\Delta}^{-1}$,

$$\mathrm{E}_{\mathrm{sim}}[\mathrm{tr}(\mathbf{W})] = \mathrm{tr}(\boldsymbol{\Theta}^{-1})\log(101)/100 = \frac{2+(p-2)(1+\theta^2)}{1-\theta^2}\frac{\log 101}{100}.$$

From this we see that the expected signal strength increases with θ when using $\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\Delta}}}$ and $\widehat{\mathbf{R}}_{\mathrm{spice}}$, but is constant in θ when using $\widehat{\mathbf{R}}_{\mathrm{diag}}$ and $\widehat{\mathbf{R}}_{\mathbf{I}}$. In addition, $\|\boldsymbol{\Delta}\| = O(1)$ since $\|\boldsymbol{\Delta}\| \le \|\boldsymbol{\Delta}\|_\infty = \max_j \sum_i \delta_{ij} \le 2 \cdot 101/(1-\theta)$, where δ_{ij} is element (i, j) of $\boldsymbol{\Delta}$.

With this general setup we next report results for various combinations of n, p, $\boldsymbol{\xi}$ and $\mathbf{f}$. We extended the applicability of $\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\Delta}}}$ to regressions in which $n < p$ by using the Moore–Penrose generalized inverse of $\widehat{\boldsymbol{\Delta}}$, although we have presented no asymptotic results for this case.

7.3. *Correctly specified* $\boldsymbol{\xi} = \boldsymbol{\beta}\mathbf{f}$. In this section we consider a simple case where $\boldsymbol{\xi} = Y$, giving $d = 1$, and $\mathbf{f} = (Y, Y^2, Y^3, Y^4)$, so $r = 4$ and $\boldsymbol{\xi} = \boldsymbol{\beta}\mathbf{f}$, where $\boldsymbol{\beta} = (1, 0, 0, 0)$ and Y was generated as a standard normal variate.

7.3.1. $n = p/2$ *and* $p \to \infty$. In this setting, n and p grow with $n = p/2$. For the reduction estimators $\widehat{\mathbf{R}}_{\mathbf{I}}$ and $\widehat{\mathbf{R}}_{\mathrm{diag}}$, $\|\boldsymbol{\rho}\| = O(1)$, $\|\boldsymbol{\rho}\|_F \le \sqrt{p}\|\boldsymbol{\rho}\| = O(\sqrt{p})$ which imply that $\psi = \kappa = n^{-1/2}$. Also, $\mathrm{var}(\boldsymbol{\nu}) \le \mathbf{G}_h^{-1}\|\boldsymbol{\Delta}\|/h = O(p^{-1})$, hence $\boldsymbol{\nu} = O_p(p^{-1})$. Thus in this setting with $n \asymp p$, $\widehat{\mathbf{R}}_{\mathbf{I}}$ is $\sqrt{n}$-consistent and $\widehat{\mathbf{R}}_{\mathrm{diag}}$ is at least $\sqrt{n/\log n}$-consistent as $n, p \to \infty$.

Although $\|\boldsymbol{\Delta}\| = O(1)$ and $\boldsymbol{\Delta}^{-1}$ is a tri-diagonal matrix, our theoretical bounds for $\widehat{\mathbf{R}}_{\mathrm{spice}}$ guarantee consistency for this model only when p is bounded and $n \to \infty$; however, a result established by Ravikumar et al. (2011), which requires additional assumptions, indicates that the weight matrix estimator for $\widehat{\mathbf{R}}_{\mathrm{spice}}$ is consistent when $\boldsymbol{\Delta}^{-1}$ is tri-diagonal.

The results for p and n growing with $n = p/2$ are shown in Figure 3(a)–(c). All reduction estimators appear to be converging to the population reduction as $n, p \to \infty$, even though consistency is not guaranteed in this setting for $\widehat{\mathbf{R}}_{\mathrm{spice}}$ and $\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\Delta}}}$, which uses a Moore–Penrose generalized inverse of the residual sample covariance matrix. Interestingly, when $p > n$, $\widehat{\mathbf{R}}_{\widehat{\boldsymbol{\Delta}}}$ outperforms $\widehat{\mathbf{R}}_{\mathrm{diag}}$ and $\widehat{\mathbf{R}}_{\mathbf{I}}$ when $\theta \ge 0.9$. Results for $\widehat{\mathbf{R}}_{\mathrm{spice}}$ were computed up to $p = 400$ when $\theta \le 0.9$ and up to $p = 100$ when $\theta = 0.99$ due to intractable computation time required for the glasso algorithm. In scenarios when $\widehat{\mathbf{R}}_{\mathrm{spice}}$ was computed, it outperformed the other reduction estimators, particularly when $\theta = 0.9$. As expected, larger values of θ lead to favorable performance for the reduction estimators with population weight matrix $\mathbf{W} = \boldsymbol{\Delta}^{-1}$.

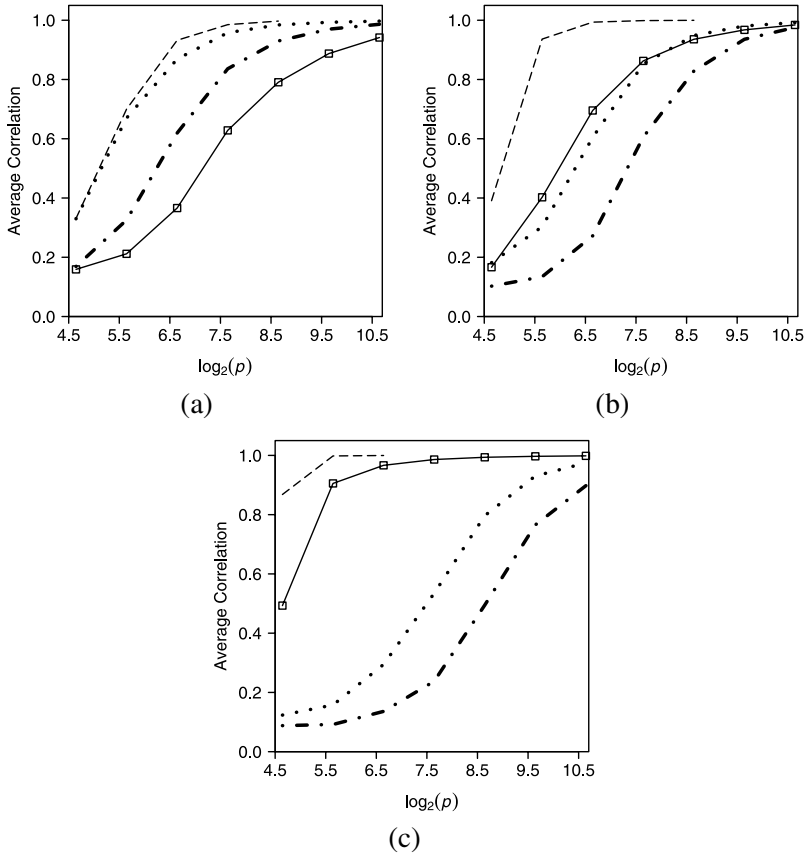

FIG. 3. *Comparison of the four estimators of* $\mathbf{R}$: $\widehat{\mathbf{R}}_{\text{spice}}$ (*dashes*), $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$ (*solid*), $\widehat{\mathbf{R}}_{\text{diag}}$ (*dots*) *and* $\widehat{\mathbf{R}}_{\mathbf{I}}$ (*dash dot*), *with exponential error correlations and* $n = p/2$. *Here* (a) $\theta_{ij} = 0.5^{|i-j|}$; (b) $\theta_{ij} = 0.9^{|i-j|}$; (c) $\theta_{ij} = 0.99^{|i-j|}$.

7.3.2. $p = 100$ *and* $n \to \infty$. In this setting we fix $p = 100$ and let n grow. Our theory guarantees that $\widehat{\mathbf{R}}_{\text{spice}}$ and $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$ are both $\sqrt{n}$-consistent. On the other hand, $\widehat{\mathbf{R}}_{\mathbf{I}}$ and $\widehat{\mathbf{R}}_{\text{diag}}$ are inconsistent since $\text{span}(\boldsymbol{\gamma})$ is not a reducing subspace of $\boldsymbol{\rho}$ and p is bounded, implying $\boldsymbol{\nu}$ fails to vanish.

The results for $p = 100$ and n growing are illustrated in Figure 4(a)–(c). As our theory suggests, the reduction estimators $\widehat{\mathbf{R}}_{\text{spice}}$ and $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$ appear to be converging to the population reduction as n increases. We see that $\widehat{\mathbf{R}}_{\text{spice}}$ outperformed the other reduction estimators, particularly for relatively small n when $\theta = 0.9$. As expected, $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$ outperforms $\widehat{\mathbf{R}}_{\text{diag}}$ and $\widehat{\mathbf{R}}_{\mathbf{I}}$ when n is much larger than p or when θ is large.

7.4. *Results for* $\boldsymbol{\xi} \neq \boldsymbol{\beta}\mathbf{f}$. In this section we present results for a misspecified $\boldsymbol{\xi}$ using $\boldsymbol{\xi} = \text{var}^{-1/2}(\exp(Y))[\exp(Y) - \text{E}(\exp(Y))]$ where $Y \sim \text{Unif}(0, 4)$. Hold-

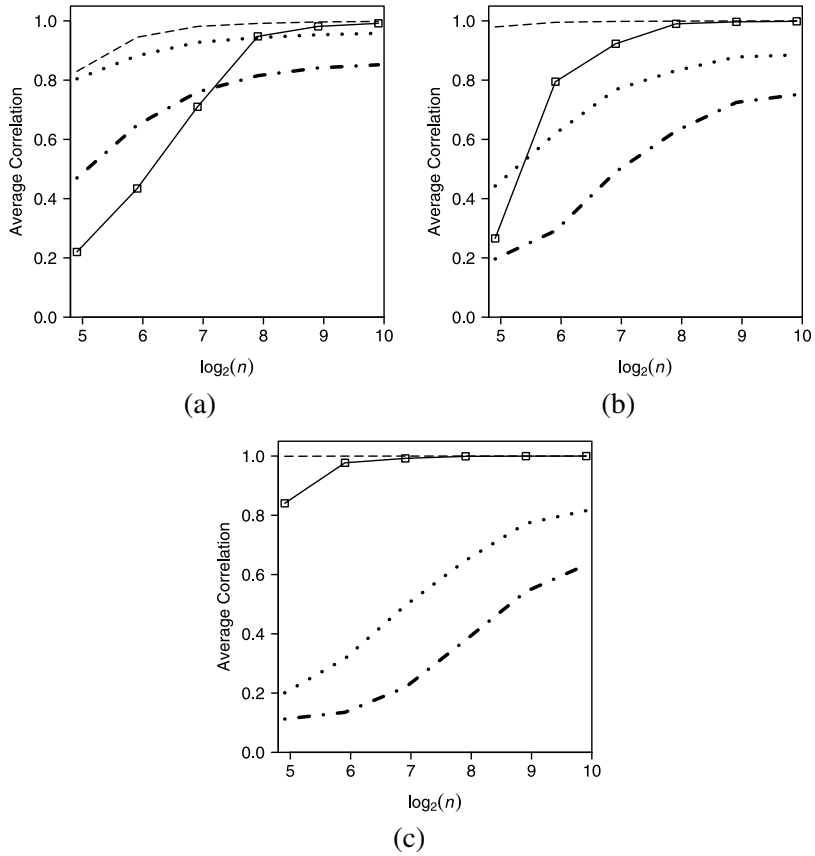

FIG. 4. *Comparison of the four estimators of* $\mathbf{R}$: $\widehat{\mathbf{R}}_{\text{spice}}$ (*dashes*), $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$ (*solid*), $\widehat{\mathbf{R}}_{\text{diag}}$ (*dots*) *and* $\widehat{\mathbf{R}}_{\mathbf{I}}$ (*dash dot*), *with exponential error correlations and* $p = 100$. *Here* (a) $\theta_{ij} = 0.5^{|i-j|}$; (b) $\theta_{ij} = 0.9^{|i-j|}$; (c) $\theta_{ij} = 0.99^{|i-j|}$.

ing $n = 50$ and $p = 100$, we varied $\mathbf{f} = (y, y^2, \ldots, y^k)^T$ for $k = 1, \ldots, 5$ and $\mathbf{f} = \exp(y)$. The results are summarized in Figure 5 as boxplots of the correlation magnitudes over the 200 replications. Most striking is how little the choice of $\mathbf{f}$ seems to affect the estimators. This is likely because there are fairly strong correlations between $\boldsymbol{\xi}$ and its approximations provided by the six $\mathbf{f}$'s used in the simulation, which satisfies condition (3.3). The relationship between the estimators are similar to those in Figure 4 at $n = 50$, regardless of the choice of $\mathbf{f}$.

8. Spectroscopy application. To illustrate operation of the proposed methodology, we examine data from the potential application area mentioned in the Introduction. The response variable is the percentage of fat in samples of beef or pork, and the predictors are the absorbance spectra ($\log(1/R)$) from near-infrared transmittance for fat measured at every second wavelength between 850 and 1050 nm,

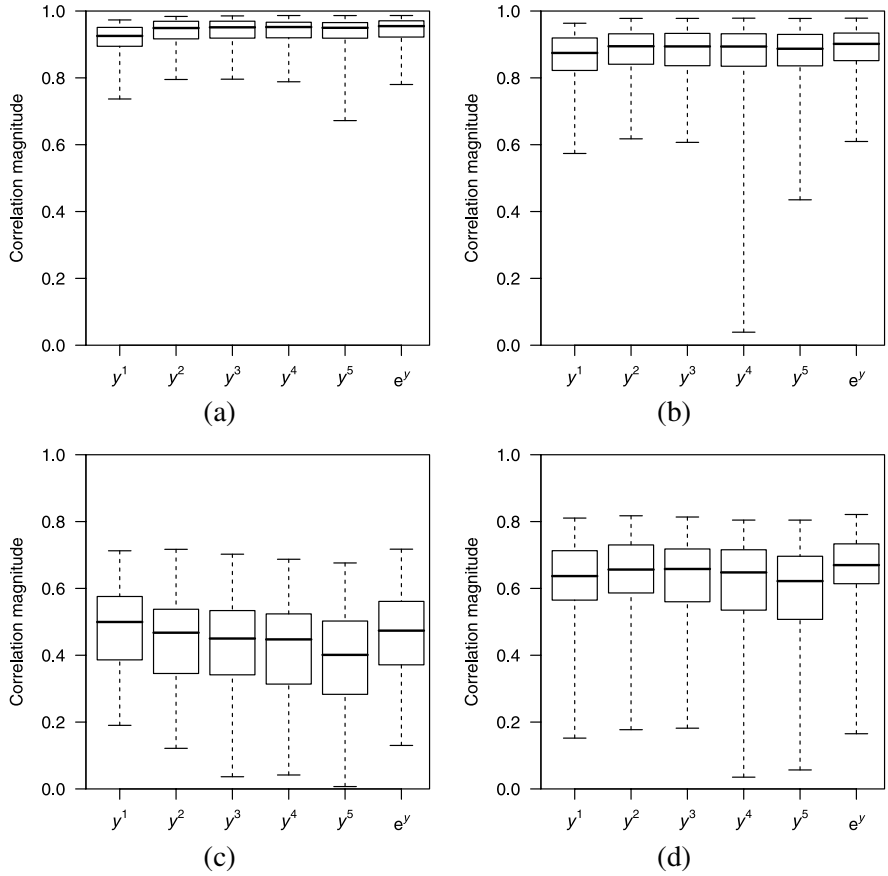

FIG. 5. *Estimators of* $\mathbf{R}$ *when* $\mathbf{f}$ *is misspecified, using exponential error correlations*: $\theta_{ij} = 0.5^{|i-j|}$, $n = 50$ *and* $p = 100$. *Boxplots are labeled by the highest order term in* $\mathbf{f}$. *Here* (a) $\widehat{\mathbf{R}}_{\text{spice}}$; (b) $\widehat{\mathbf{R}}_{\text{diag}}$; (c) $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$; (d) $\widehat{\mathbf{R}}_{\mathbf{I}}$.

giving $p = 100$ with $n = 54$ [Sæbø et al. (2007)]. The goal of the study is to predict the response from the absorbance spectra of a new sample. Our predictive framework is described in the next section. We return to the spectroscopy application in Section 8.2.

8.1. *Prediction*. We predict an unobserved response Y_N associated with a new observed vector of predictors $\mathbf{X}_N$ by using the forward regression mean function: $Y_{\text{pred}} = \mathrm{E}\{Y|\mathbf{X}_N\} = \mathrm{E}\{Y|\mathbf{R}(\mathbf{X}_N)\}$. However, the reduction $\mathbf{R}(\mathbf{X})$ was based on the inverse regression $\mathbf{X}|Y$ and the development did not produce a direct estimator of $\mathrm{E}\{Y|\mathbf{R}(\mathbf{X}_N)\}$. There are perhaps several ways of using an estimated reduction for predicting a new response. Some authors have used standard data-analytic methods to develop predictive models based on Y and the estimated reduction, and there are a variety of nonparametric regression methods that could

also be used as well. We follow Adragni and Cook (2009) and use a kernel-type estimator of $\mathrm{E}\{Y|\mathbf{R}(\mathbf{X}_N)\}$ based on the relationship $\mathrm{E}\{Y|\mathbf{X}=\mathbf{x}\} = \mathrm{E}\{Y|\mathbf{R}(\mathbf{x})\} = \mathrm{E}\{Yg(\mathbf{R}(\mathbf{x})|Y)\}/\mathrm{E}\{g(\mathbf{R}(\mathbf{x})|Y)\}$, where g is the conditional density of $\mathbf{R}|Y$. This provides a method to estimate $\mathrm{E}\{Y|\mathbf{X}\}$:

$$\begin{aligned}\widehat{\mathrm{E}}\{Y|\mathbf{X}=\mathbf{x}\} &= \sum_{i=1}^{n} w_i(\mathbf{x})Y_i,\\ w_i(\mathbf{x}) &= \frac{\widehat{g}(\widehat{\mathbf{R}}(\mathbf{x})|Y_i)}{\sum_{i=1}^{n}\widehat{g}(\widehat{\mathbf{R}}(\mathbf{x})|Y_i)},\end{aligned} \tag{8.1}$$

where $\widehat{g}$ denotes an estimate of the density and $\widehat{\mathbf{R}}$ is the estimated reduction. This estimator is reminiscent of a nonparametric kernel estimator, but there are consequential differences. The weights in a kernel estimator do not depend on the response, while the weights w_i here do. Kernel weights typically depend on the full vector of predictors $\mathbf{X}$, while the weights here depend on $\mathbf{X}$ only through the estimated reduction $\widehat{\mathbf{R}}(\mathbf{x})$. Multivariate kernels are usually taken to be the product of univariate kernels, corresponding here to treating the components of $\mathbf{R}$ as independent. Finally, there is no need for bandwidth estimation because the weights are determined entirely from $\widehat{g}$.

When d is small relative to p it may often be reasonable to assume that $\mathbf{R}(\mathbf{X})|Y$ is normally distributed, which seems appropriate for the spectroscopy data. Ignoring constants not depending on i, we have

$$g(\mathbf{R}(\mathbf{x})|Y_i) \propto \exp\{-(1/2)(\mathbf{R}(\mathbf{x})-\boldsymbol{\xi}_i)^T(\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma})(\mathbf{R}(\mathbf{x})-\boldsymbol{\xi}_i)\}.$$

Substituting the estimators $\widehat{\boldsymbol{\Gamma}}$, $\widehat{\mathbf{b}}\mathbf{f}_i$ and $\widehat{\mathbf{W}}$ for $\boldsymbol{\Gamma}$, $\boldsymbol{\xi}_i$ and $\boldsymbol{\Delta}^{-1}$ gives the weights required for (8.1). For $\widehat{\mathbf{R}}_{\mathbf{I}}$, we set $\widehat{\mathbf{W}} = (p/\operatorname{tr}(\widehat{\boldsymbol{\Delta}}))\mathbf{I}_p$.

8.2. *Spectroscopy.* Since there are only $p = 100$ predictors it was straightforward, albeit somewhat tedious, to inspect inverse response plots of X_j versus Y, $j = 1, \ldots, 100$, to gain information about the likely structure of $\mathbf{f}$. We performed two analyses; the first consisted of 54 pork samples and the second consisted of 103 meat samples of beef and pork. For a specified value of d, we assessed the fitted model by using the residual mean square $\mathrm{RMS}(d, r, \widehat{\mathbf{R}}_{(\cdot)}) = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2/n$, where the fitted values $\widehat{Y}_i = \widehat{\mathrm{E}}(Y|\mathbf{X}=\mathbf{X}_i)$ were determined using (8.1) with the indicated combination of d, r and reduction $\widehat{\mathbf{R}}_{(\cdot)}$.

We selected the value of d by adapting the permutation scenario developed by Cook and Yin (2001). The hypothesis $d = d_0$ was tested sequentially, starting at $d_0 = 0$ and estimating d as the first hypothesized value that was not rejected. The test statistic $\mathrm{RMS}(d_0+1, r, \mathbf{R}_{(\cdot)})$ was compared to the distribution of $\mathrm{RMS}(d_0+1, r, \mathbf{R}_{(\cdot)})$ induced by 1000 random permutations $\mathbf{J}$ applied to the rows of the predictor matrix $\mathbb{X}$ as follows:

$$\mathbb{X}_{\mathrm{perm}} = \mathbb{X}\mathbf{P}^T_{\widehat{\boldsymbol{\Gamma}}(\widehat{\mathbf{W}})} + \mathbf{J}\mathbb{X}\mathbf{Q}^T_{\widehat{\boldsymbol{\Gamma}}(\widehat{\mathbf{W}})},$$

where $\widehat{\mathbf{\Gamma}}$ and $\widehat{\mathbf{W}}$ were computed under the null hypothesis. This scheme leaves the signal $\mathbb{X}\mathbf{P}^T_{\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{W}})}$ intact while permuting the uninformative part of the predictors $\mathbb{X}\mathbf{Q}^T_{\widehat{\mathbf{\Gamma}}(\widehat{\mathbf{W}})}$.

The multicollinearity of the predictors made computing the weight matrix estimator for $\widehat{\mathbf{R}}_{\text{spice}}$ difficult for small values of its tuning parameter, values for which our cross-validation procedure recommended. We subsequently set its tuning parameter to $\lambda = 2^{-10}$ for both analyses, since this was the smallest value for which a numerically stable solution was available.

8.2.1. *Analysis of pork samples*. In this case we concluded that a cubic polynomial $\mathbf{f}(y) = (y, y^2, y^3)^T$ would be adequate; a representative plot is shown in Figure 6(a). Performing the permutation test to select d, the test statistic for $d = 0$ using $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$ was $\text{RMS}(1, 3, \widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}) = 0.31$ which was smaller than the smallest value 10.7 of $\text{RMS}(1, 3, \widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}})$ observed among the 1000 random permutations under the hypothesis that $d = 0$. Since the test statistic is much smaller than can be accounted for by chance under the null hypothesis, we concluded that $d \geq 1$. Similarly, to test $d = 1$, we observed $\text{RMS}(2, 3, \widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}) = 0.29$ which fell at the 80th quantile of the permutation distribution of $\text{RMS}(2, 3, \widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}})$ under the hypothesis. Consequently, we used $d = 1$ for the model. Comparisons with other values of r figured in our choice $r = 3$. Cross-validation using RMS as the criterion might also be used to select d.

Fitting each of the four estimators with $d = 1$, we found that $\text{RMS}(1, 3, \widehat{\mathbf{R}}) = 21.55$, 5.15, 0.31 and 2.15 for $\widehat{\mathbf{R}} = \widehat{\mathbf{R}}_{\mathbf{I}}$, $\widehat{\mathbf{R}}_{\text{diag}}$, $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$ and $\widehat{\mathbf{R}}_{\text{spice}}$. Plots of Y versus $\widehat{Y}$

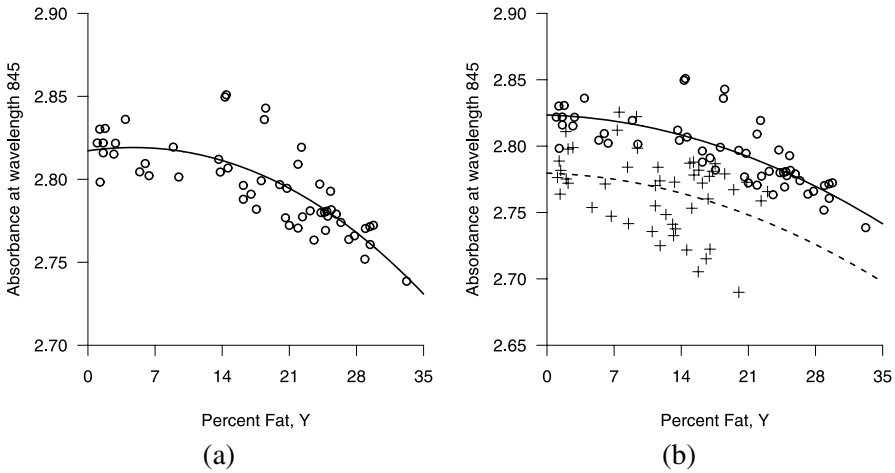

FIG. 6. *Inverse response plot of the absorbance at wavelength* 854 *versus Y for* (a) *pork samples and* (b) *pork and beef samples. The line for the pork sample is a fitted cubic polynomial. For pork and beef the lines represent a second-order polynomial fit to the data with one intercept for pork* (*solid*) *and a second intercept for beef* (*dashes*).

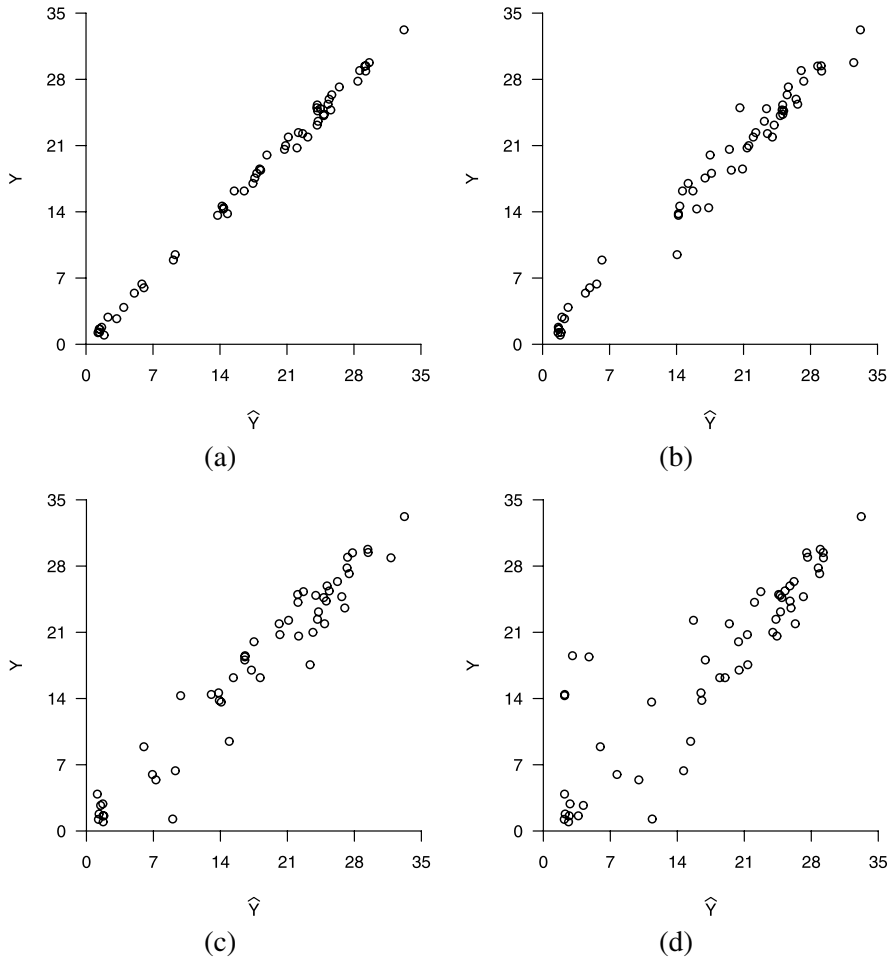

FIG. 7. *Plots of the response versus fitted values for the four estimators* (a) $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$, (b) $\widehat{\mathbf{R}}_{\text{spice}}$, (c) $\widehat{\mathbf{R}}_{\text{diag}}$ *and* (d) $\widehat{\mathbf{R}}_{\mathbf{I}}$, *using the pork samples*.

are shown in Figure 7(a)–(d). The predictors in this illustration are highly collinear, as is typical in spectral data, and the relative performance of the four estimators is qualitatively similar to that shown in previous simulations; however, numerical instability degraded the performance of $\widehat{\mathbf{R}}_{\text{spice}}$. The relative signal rates for the four estimators are reflected by the values of $\widehat{\mathbf{\Gamma}}^T\widehat{\mathbf{W}}\widehat{\mathbf{\Gamma}} = 35.3$, 128.4, 169.7 and 63.38 for $\widehat{\mathbf{W}} = (p/\operatorname{tr}(\widehat{\mathbf{\Delta}}))\mathbf{I}_d$, $\operatorname{diag}^{-1}(\widehat{\mathbf{\Delta}})$, $\widehat{\mathbf{\Delta}}^-$ and $\widehat{\mathbf{\Delta}}_{\widehat{\lambda}}^{-1}$.

8.2.2. *Analysis of both pork and beef samples*. In this case we concluded that a second-order polynomial and the indicator function of beef would be adequate, $\mathbf{f}(y) = (y, y^2, J(\text{beef}))^T$; a representative plot is shown in Figure 6(b). Using the permutation test approach to select d, the test statistic for $d = 0$ using $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$ was

$\mathrm{RMS}(1, 3, \widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}) = 0.55$ which fell at the 0.003 quantile of $\mathrm{RMS}(1, 3, \widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}})$ observed among the 1000 random permutations under the hypothesis that $d = 0$. Since the test statistic is much smaller than can be accounted for by chance under the null hypothesis, we concluded that $d \geq 1$. Similarly, to test $d = 1$, we observed $\mathrm{RMS}(2, 3, \widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}) = 0.01$ which fell at the 0.70 quantile of the permutation distribution of $\mathrm{RMS}(2, 3, \widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}})$ under the hypothesis. Consequently, we used $d = 1$ for the model.

Fitting each of the four estimators with $d = 1$, we found that $\mathrm{RMS}(1, 3, \widehat{\mathbf{R}}) = 41.96$, 41.65, 0.55 and 4.66 for $\widehat{\mathbf{R}} = \widehat{\mathbf{R}}_{\mathbf{I}}$, $\widehat{\mathbf{R}}_{\mathrm{diag}}$, $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$ and $\widehat{\mathbf{R}}_{\mathrm{spice}}$. Plots of Y versus $\widehat{Y}$ are shown in Figure 8(a)–(d). The relative signal rates for the four estimators are reflected by the values of $\widehat{\mathbf{\Gamma}}^T \widehat{\mathbf{W}} \widehat{\mathbf{\Gamma}} = 18.4$, 59.2, 3346.6 and 69.3 for $\widehat{\mathbf{W}} =$

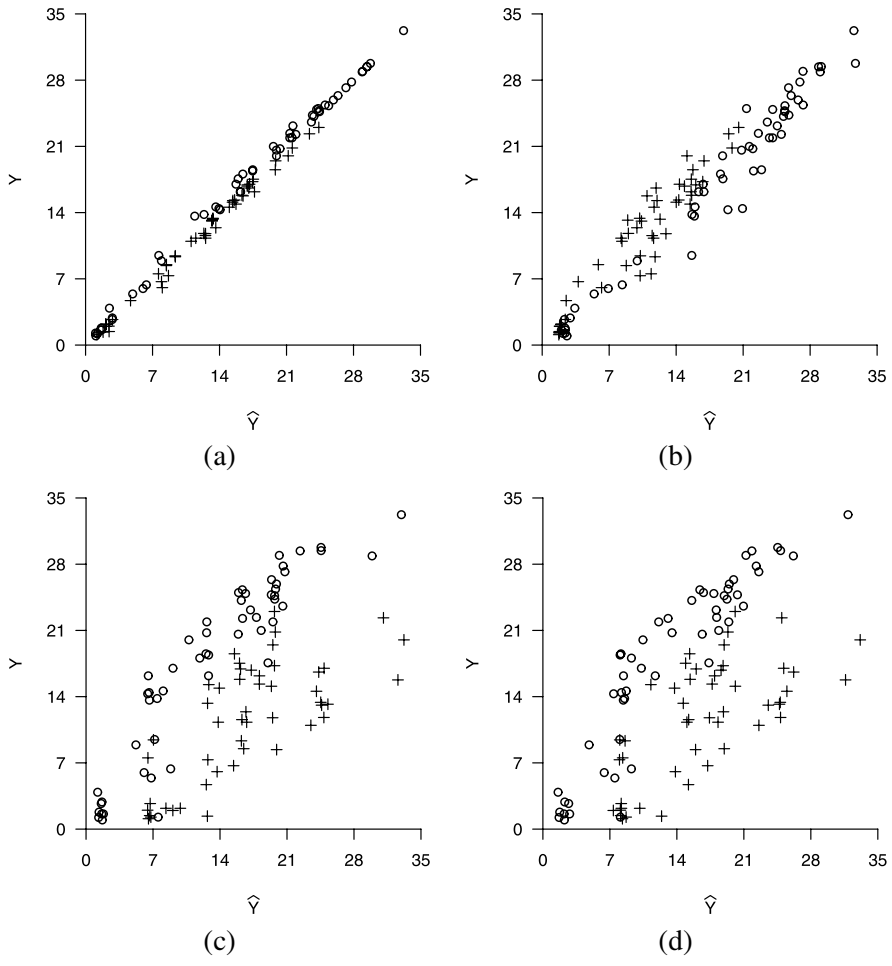

FIG. 8. *Plots of the response versus fitted values for the four estimators* (a) $\widehat{\mathbf{R}}_{\widehat{\mathbf{\Delta}}}$, (b) $\widehat{\mathbf{R}}_{\mathrm{spice}}$, (c) $\widehat{\mathbf{R}}_{\mathrm{diag}}$ *and* (d) $\widehat{\mathbf{R}}_{\mathbf{I}}$. *Circles represent pork and plus signs represent beef.*

$(p/\operatorname{tr}(\boldsymbol{\Delta}))\mathbf{I}_d$, $\operatorname{diag}^{-1}(\widehat{\boldsymbol{\Delta}})$, $\widehat{\boldsymbol{\Delta}}^{-1}$ and $\widehat{\boldsymbol{\Delta}}_{\widehat{\lambda}}^{-1}$. The relatively large signal for $\widehat{\boldsymbol{\Delta}}^{-1}$ is reflected in the plots of Figure 8.

9. Discussion. The class of estimators that we studied is limited in scope relative to the range of SDR methods presently available for $p = o(n)$ regressions. However, in the broader context of this article, which does not require $p = o(n)$, we introduced the concept of an abundant regression and presented what may be the first n, p asymptotic analysis of a class of SDR methods when the focus is on estimating a reduction $\mathbf{R}$ rather than on the underlying central subspace. These ideas can in principle be extended for other SDR methods, and we expect that the same general issues will be encountered. While each of the methods that we studied can perform usefully in the right situation, we judge the SPICE and $\boldsymbol{\Delta}^-$ weight matrices to be the best overall, although improvements for nonsparse weight matrices and for regressions with very highly correlated predictors are still needed.

Our simulation results were all based on normal errors to focus the presentation and save space. However, we have also conducted a variety of parallel simulations using Uniform $(0, 1)$, T_5 and χ_5^2 errors, each centered and appropriately scaled. The results were essentially the same as those with normal errors [Cook, Forzani and Rothman (2012)].

9.1. *Penalty functions.* Alternative penalty functions may be used for estimating the weight matrix, particularly in scenarios when the inverse error covariance matrix is not sparse. For example, the sparse-seeking penalty function in (5.4), $\lambda \sum_{i \neq j} |\Omega_{ij}|$, could be replaced with the quadratic penalty function,

$$\lambda\left(\sum_{i \neq j} \Omega_{ij}^2 + \alpha \sum_{j=1}^{p} \Omega_{jj}^2\right), \tag{9.1}$$

where $\alpha \in \{0, 1\}$ controls whether or not the diagonal of $\boldsymbol{\Omega}$, the inverse error correlation matrix, is penalized. If $\alpha = 0$, the general SPICE algorithm developed by Rothman et al. (2008) can efficiently solve (5.4) with the penalty defined in (9.1). If $\alpha = 1$, Witten and Tibshirani (2009) derived an noniterative solution to an equivalent problem to (5.4) with the penalty defined in (9.1). Recalling that $\varphi_j(\mathbf{A})$ denotes the jth eigenvalue of a matrix $\mathbf{A}$, let $\eta_j = \varphi_j(\operatorname{diag}^{-1/2}(\widehat{\boldsymbol{\Delta}})\widehat{\boldsymbol{\Delta}}\operatorname{diag}^{-1/2}(\widehat{\boldsymbol{\Delta}}))$. Witten and Tibshirani showed that the eigenvectors of $\widehat{\boldsymbol{\Omega}}_\lambda$ are equivalent to the eigenvectors of $\operatorname{diag}^{-1/2}(\widehat{\boldsymbol{\Delta}})\widehat{\boldsymbol{\Delta}}\operatorname{diag}^{-1/2}(\widehat{\boldsymbol{\Delta}})$ and that $\varphi_j(\widehat{\boldsymbol{\Omega}}_\lambda) = (4\lambda)^{-1}\{(\eta_j^2 + 8\lambda)^{1/2} - \eta_j\}$.

9.2. *Choice of* $\mathbf{f}$. The general rates given in Proposition 5.1 are not very sensitive to the choice of $\mathbf{f}$ since they hold when $\mathbf{f}$ satisfies the minimal rank condition (3.3). Nevertheless, assuming normality and a correct $\mathbf{f}$, we obtained the oracle rates of Proposition 6.2, which indicates that there are advantages to pursuing good choices. The methods sketched in Section 3.2 are often useful in practice, but it is also possible to develop semiparametric methods to estimate $\boldsymbol{\xi}$ directly

rather than passing through approximations $\boldsymbol{\beta}\mathbf{f}$. This might be accomplished iteratively: choose an initial $\mathbf{f}$ and construct the corresponding estimates $\boldsymbol{\Gamma}^1$, $\boldsymbol{\xi}^1 = \boldsymbol{\beta}^1\mathbf{f}$ and $\mathbf{R}^1_{\widehat{\mathbf{W}}}$. A new estimate of $\boldsymbol{\xi}$ can be obtained by smoothing the coordinates of $\mathbf{R}^1_{\widehat{\mathbf{W}}}$ against Y, leading to a second reduction estimate $\mathbf{R}^2_{\widehat{\mathbf{W}}}$. The process can now be continued until some convergence criterion is met.

9.3. *Variable selection*. While we did not incorporate screening or variable selection into our reduction methodology, the potential benefits of those procedures are manifested in our results. Consider, for instance, a regression in which $p^{2/3}$ of the predictors are inactive. Then the oracle rate $\kappa^{-1} = (p/hn)^{-1/2} = n^{1/2}p^{-2/3}$. However, if we remove $p^{1/3}$ of the inactive predictors, the oracle rate is increased to $\kappa^{-1} = n^{1/2}p^{-1/3}$, which should be worthwhile in most applications. Work along these lines is in progress.

Acknowledgments. The authors are grateful to the referees whose helpful comments led to significant improvements in this article.

SUPPLEMENTARY MATERIAL

Supplement to "Estimating sufficient reductions of the predictors in abundant high-dimensional regressions" (DOI: 10.1214/11-AOS962SUPP; .pdf). Owing to space constraints, we have placed the technical proofs in a supplemental article [Cook, Forzani and Rothman (2012)]. The supplement also contains several preparatory technical results that may be of interest in their own right and additional simulations. For instance, we gave in Section 7 simulation results from models with exponentially decreasing error correlations. In the supplemental article we give parallel results based on the same models but with constant error correlations.

REFERENCES

ADRAGNI, K. P. and COOK, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4385–4405. MR2546393

BAI, Z. D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statist. Sinica* **9** 611–677.

BICKEL, P. J. and LEVINA, E. (2008a). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BICKEL, P. J. and LEVINA, E. (2008b). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

BONDELL, H. D. and LI, L. (2009). Shrinkage inverse regression estimation for model-free variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 287–299. MR2655534

BURA, E. and COOK, R. D. (2001). Extending sliced inverse regression: The weighted chi-squared test. *J. Amer. Statist. Assoc.* **96** 996–1003. MR1946367

CHIAROMONTE, F. and MARTINELLI, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Math. Biosci.* **176** 123–144. MR1869195

CHUN, H. and KELEŞ, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 3–25. MR2751241

COOK, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the Section on Physical and Engineering Sciences* 18–25. Amer. Statist. Assoc., Alexandria, VA.

COOK, R. D. (1998). *Regression Graphics*: *Ideas for Studying Regressions Through Graphics*. Wiley, New York. MR1645673

COOK, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32** 1062–1092. MR2065198

COOK, R. D. (2007). Fisher lecture: Dimension reduction in regression (with discussion). *Statist. Sci.* **22** 1–26. MR2408655

COOK, R. D. and FORZANI, L. (2008). Principal fitted components for dimension reduction in regression. *Statist. Sci.* **23** 485–501. MR2530547

COOK, R. D. and FORZANI, L. (2009). Likelihood-based sufficient dimension reduction. *J. Amer. Statist. Assoc.* **104** 197–208. MR2504373

COOK, R. D., FORZANI, L. and ROTHMAN, A. J. (2012). Supplement to "Estimating sufficient reductions of the predictors in abundant high-dimensional regressions." DOI:10.1214/11-AOS962SUPP.

COOK, R. D., LI, B. and CHIAROMONTE, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika* **94** 569–584. MR2410009

COOK, R. D. and NI, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.* **100** 410–428. MR2160547

COOK, R. D. and WEISBERG, S. (1991). Discussion of "Sliced inverse regression for dimension reduction" by K.-C. Li. *J. Amer. Statist. Assoc.* **86** 382–332.

COOK, R. D. and YIN, X. (2001). Dimension reduction and visualization in discriminant analysis. *Aust. N. Z. J. Stat.* **43** 901–999.

D'ASPREMONT, A., BANERJEE, O. and EL GHAOUI, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30** 56–66. MR2399568

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

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

FRIEDMAN, J., HASTIE, T., ROSSET, R., TIBSHIRANI, R. and ZHU, J. (2004). Consistency in boosting: Discussion. *Ann. Statist.* **32** 102–107.

FUKUMIZU, K., BACH, F. R. and JORDAN, M. I. (2009). Kernel dimension reduction in regression. *Ann. Statist.* **37** 1871–1905. MR2533474

GUAN, Y. and WANG, H. (2010). Sufficient dimension reduction for spatial point processes directed by Gaussian random fields. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 367–387. MR2758117

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

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

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

LI, B. and DONG, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *Ann. Statist.* **37** 1272–1298. MR2509074

LI, L. and LI, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics* **20** 3406–3412.

LI, L. and YIN, X. (2008). Sliced inverse regression with regularizations. *Biometrics* **64** 124–131. MR2422826

PAUL, D. (2005). Nonparametric estimation of principal components. Ph.D. thesis, Dept. Statistics, Stanford Univ. MR2707156

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980.

ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391

SÆBØ, S., ALMØY, T., AARØE, J. and AASTVEIT, A. H. (2007). ST-PLS: A multi-directional nearest shrunken centroid type classifier via PLS. *Journal of Chemometrics* **20** 54–62.

VON ROSEN, D. (1988). The inverted Wishart distribution. *Scand. J. Stat.* **15** 97–109.

WITTEN, D. M. and TIBSHIRANI, R. (2009). Covariance-regularized regression and classification for high dimensional problems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 615–636. MR2749910

WU, Y. and LI, L. (2011). Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statist. Sinica* **21** 707–730. MR2829852

XIA, Y., ZHANG, D. and XU, J. (2010). Dimension reduction and semiparametric estimation of survival models. *J. Amer. Statist. Assoc.* **105** 278–290. MR2656052

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

ZHU, L.-P., ZHU, L.-X. and FENG, Z.-H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *J. Amer. Statist. Assoc.* **105** 1455–1466. MR2796563

ZYSKIND, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Ann. Math. Statist.* **38** 1092–1109. MR0214237

R. D. COOK
A. J. ROTHMAN
SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA 55455
USA
E-MAIL: dennis@stat.umn.edu
rothman@stat.umn.edu

L. FORZANI
INSTITUTO DE MATEMÁTICA APLICADA DEL LITORAL
FACULTAD DE INGENIERÍA QUÍMICA
CONICET AND UNL
GÜEMES 3450, (3000) SANTA FE
ARGENTINA
E-MAIL: liliana.forzani@gmail.com