Fisher Lecture*: Dimension Reduction in Regression

R. Dennis Cook†
School of Statistics
University of Minnesota

November 7, 2005

Abstract

Beginning with a discussion of R. A. Fisher's early written remarks that relate to dimension reduction, this article revisits the method of principal components as a reductive method in regression, develops several model-based extensions and ends with descriptions of general approaches to model-based and model-free dimension reduction in regression. It is argued that the role for principal components and related methodology may be broader than previously seen and that the common practice of conditioning on observed values of the predictors may unnecessarily limit the choice of regression methodology.

Key words and phrases: Central subspace, Grassman manifolds, inverse regression, principal components, principal fitted components, sliced inverse regression, sufficient dimension reduction.

1 Introduction

R. A. Fisher is responsible for the context and mathematical foundations of a substantial portion of contemporary theoretical and applied statistics. One purpose of this article is to consider insights into the long-standing and currently prominent problem of "dimension reduction" that may be available from his writings. Two papers are discussed in this regard, Fisher's pathbreaking 1922 paper on the theoretical foundations of statistics (Section 1.1), and a later applications paper on the yield of wheat at Rothamsted (Section 1.2). The discussion of these articles makes liberal use of quoted material, in an effort to preserve historical flavor and reflect Fisher's style.

*This article is the written version of the 2005 Fisher lecture delivered at the Joint Statistical Meetings in Minneapolis, MN.

†313 Ford Hall, 224 Church Street SE, Minneapolis, MN 55455. email: dennis@stat.umn.edu. Research for this article was supported in part by National Science Foundation Grant DMS-0405360.

Principal component analysis is one of the oldest and best known methods for reducing dimensionality in multivariate problems. Another purpose of this article is to provide an exposition on principal components as a reductive method in regression, with emphasis on connections to known reductive methods and on the development of a new method – *Principal Fitted Components (PFC)* – that may outperform principal components. The discussion will be related to and guided by Fisher's writings, as much as the nature of the case permits. In particular, the philosophical spirit of the methods discussed in this article derives largely from Fisher's notion of sufficiency.

We briefly review principal components in Section 2, and starting in Section 3 we focus on principal components in regression. Principal fitted components are introduced in Section 4. In Sections 5–7, we expand the themes of Sections 3 and 4, gradually increasing the scope of dimension reduction methodology. A general model-based paradigm for dimension reduction in regression is described in Section 8.1. In keeping with the Fisherian theme of this article, the development is model-based, but in Section 8.2 we describe its relation to recent ideas for model-free reductions. The appendix contains justification for propositions and some of the other results. Emphasis is placed on ideas and methodological directions, rather than on the presentation of fully-developed comprehensive methods.

Reduction by principal components has been proposed as adjunct methodology for linear regression. It does not arise as a particular consequence of the model itself, but is used to mitigate the variance inflation that often accompanies collinearities among the predictors. Indeed, while collinearity is the main and often the only motivation for use of principal components in regression, it will play no role in the evolution of the methods in this article. It is argued in Sections 5.1, 5.2, 5.3 and 8 that the utility of principal component reduction is broader than previously seen and need not be tied to the presence of collinearity. This conclusion is a consequence of postulating inverse regression models that lead to principal components and principal fitted components as maximum likelihood estimators of reductive subspaces.

Ordinary least squares (OLS) is widely recognized as a reasonable first method of regression when the response and predictors follow a nonsingular multivariate normal distribution. Nevertheless, examples are given in Sections 5 and 6.4 to demonstrate that in this context reduction by principal components and principal fitted components may dominate OLS in estimation and prediction without invoking collinearity. Sliced inverse regression (SIR, Li 1991) is a relatively recent reductive method for regression that has received notable attention in the literature. New drawbacks of SIR will emerge from its relations to principal fitted components described in Sections 6.3 and 7.5. It is demonstrated by example that the method of principal fitted components can dominate SIR as well as OLS in the multivariate normal setting. The notions of principal components and principal fitted components are presented here as reductive frames, not necessarily tied to normality. Their construction in the context of exponential families

is sketched in Section 3.3 and elsewhere.

Conditioning on the observed values of the predictors is a well-established practice in regression, even when the response and the predictors have a joint distribution. However, as a consequence of the exposition on principal components and the development of related reductive methods, we argue in Section 8 that this practice may unnecessarily restrict our choice of regression methodology and that it may be advantageous in some regressions to make explicit use of the variability in the predictors through the multivariate inverse regression of the predictors on the response.

1.1 Fisher, 1922

Much of contemporary statistical thought began with Fisher's 1922 article "On the mathematical foundations of theoretical statistics", which set forth a new conceptual framework for statistics, including definitions of Consistency, Efficiency, Likelihood, Specification, and Sufficiency. Fisher also introduced the now familiar terms "statistic", "maximum likelihood estimate" and "parameter". The origins of this remarkable work, which in many ways is responsible for the ambient texture of statistics today, were traced by Stigler (1973, 2005), who emphasized that in his 1922 article Fisher was the first to use the word "parameter" in its modern context, 57 times (Stigler 1976). More than any other single notion, this word reflects the starting point – parametric families – for Fisher's constructions. Many of his ideas are now common knowledge, to the point that he is no longer credited when they are first introduced in some statistics texts. This paper, perhaps more than any other of Fisher's, should be required reading in every statistics curriculum.

Fisher provided a focal point for statistical methods at the outset of his 1922 article:

> ... the objective of statistical methods is the reduction of data. A quantity of data,..., is to be replaced by relatively few quantities which shall adequately represent ... the relevant information contained in the original data.
>
> Since the number of independent facts supplied in the data is usually far greater than the number of facts sought, much of the information supplied by an actual sample is irrelevant. It is the object of the statistical process employed in the reduction of data to exclude this irrelevant information, and to isolate the whole of the relevant information contained in the data.

In these statements Fisher signified the goal of statistical methods as a type of dimension reduction, the reduced data containing the relevant and only the relevant information. The fundamental idea that statistical methods deal with the reduction of data did not originate with Fisher, and was probably widely understood well before

his birth in 1890 (See, for example, Edgeworth 1884). However, Fisher's approach is unique because it changed the course of statistical history.

Fisher identified three distinct problems in his reductive process: 1. "Problems of Specification", selecting a parametric family, 2. "Problems of Estimation", selecting statistics for parameter estimation and 3. "Problems of Distribution", deriving the sampling distribution of the selected statistics or functions thereof. Problems of specification and estimation are both reductive in nature. Surely, selecting a finitely and parsimoniously parameterized family from the infinity of possible choices is a crucial first reductive step that can overshadow any subsequent reduction of the data for the purpose of parameter estimation.

Regarding estimation, Fisher said that once the model is specified

> ... the statistic chosen should summarize the whole of the relevant information supplied by the sample.

Any operational version of this idea must include some way of parsing data into the relevant and irrelevant. For Fisher, the reductive process started with a parametric family, targeted information on its parameters θ and was guided by sufficiency: If D represents the data, then a statistic $t(D)$ is sufficient if

$$D|(\theta, t) \sim D|t \tag{1}$$

so that t contains all of the relevant information about θ. Subsequent commentaries on sufficiency by Fisher and others address existence, minimal sufficiency, specializations and variations, relation to the method of maximum likelihood, and the factorization theorem.

Fisher was quite specific on how to approach the second reductive step, estimation, but was less so regarding the overarching first reductive step, specification. In stating that problems of specification "... are entirely a matter for the practical statistician," Fisher positioned them as a nexus between applied and theoretical statistics. A model is required before proceedings to problems of estimation, and model specification falls under the purview of the practical statistician. He also offered the following helpful but non-prescriptive advice:

> ... we may know by experience what forms are likely to be suitable, and the adequacy of our choices may be tested *a posteriori*. We must confine ourselves to those forms which we know how to handle, ..."

In these statements Fisher acknowledged a role for statisticians as members of scientific teams, anticipated the development of diagnostic methods for model criticism and recognized a place for off-the-shelf models. Interest in diagnostic methods for regression was particularly high from a period starting near the time of Fisher's death in 1962

and ending in the late 1980's (Anscombe 1961, Anscombe and Tukey 1963, Box 1980, Cook and Weisberg 1982, Cook 1986). Fisher also linked model complexity with the amount of data, and evidently didn't require that models be "true",

> More or less elaborate forms will be suitable according to the volume of the data. Evidently these are considerations the nature of which may change greatly during the course of a single generation.

Fisher's first reductive step is the most challenging and elusive. Indeed, Fisher's views launched a debate over modeling that continues today. Writing on Fisher's discovery of sufficiency, Stigler's (1973) introduction began with the sentence

> Because Fisher's concept of sufficiency depends so strongly on the assumed form of the population distribution, its importance to applied statistics has been questioned in recent years.

Box's (1979) memorable statement that "All models are wrong, but some are useful" has been taken to imply, perhaps from a position of devil's advocate, that D is the only sufficient statistic. At least two Fisher lectures, Lehmann in 1988 (Lehmann 1990) and Cox in 1989 (Cox 1990), were on the issue of model specification. And then there are the modeling cultures of Breiman (2001), and McCullagh's (2002) rather esoteric answer to the question "What is a statistical model?" Fourteen years after his 1922 article, Fisher suggested that it might be possible to develop inductive arguments for model specification, a promise that has yet to be realized:

> Clearly, there can be no operation properly termed 'estimation' until the parameter to be estimated has been well defined, and this requires that the mathematical form of the distribution shall be given. Nevertheless, we need not close our eyes to the possibility that an even wider type of inductive argument may some day be developed, which shall discuss methods of assigning from the data the functional form of the population. – *Uncertain Inference*, Fisher 1936.

1.2 Fisher, 1924

In Fisher's classic 1922 article we see him thinking primarily as a theoretical statistician, while in his 1924 paper "III. The influence of rainfall on the yield of wheat at Rothamsted" we see him as an applied statistician. Having recently developed a new framework for theoretical statistics, he might have quickly integrated those ideas into his applied work, but that does not appear to be the case. He did (Fisher 1924, p. 103) rely on is 1922 article to justify a claim that a skewness statistics is the "most efficient

statistic" for a test of normality, but otherwise I found no clear formal links between the two works.

The opening issue that Fisher addressed in 1924 involves sparsity of data in regression, possibly "$n < p$". According to Fisher, large sample regression methods are appropriate when the sample size n is much larger than the number of predictors p, preferably $n > 1,000$ but at least in the 100's. However, the number p of meteorological variables that might plausibly affect yield could have easily exceed the length n of the longest run of available crop records. Fisher then faced a dimension reduction problem rather like those encountered in the analysis of contemporary genomics data (e.g. microarray data). He didn't provide a general solution, but did give a clear opinion on what not to do. It was common practice at the time to preprocess the potential meteorological predictors by plotting them individually against yield to select the predictors for the subsequent regression. Fisher was critical of this practice, concluding that

> The meteorological variables to be employed must be chosen without reference to the actual crop record.

He also reiterating a theme of his 1922 paper, "Relationships of a complicated character should be sought only when long series of crop data are available."

The bulk of Fisher's article is devoted to the development of models for the regression of yield on rainfall, models that were painstakingly tailored to the substantive characteristics of the problem, as one would expect in careful application. Fisher's study seems true to his general point that model specification is "entirely a matter for the practical statistician," and the corollary that it depends strongly on context. Regarding dimension reduction, there is at least one general theme in Fisher's analysis: An "$n < p$" regression might be usefully transformed into an "$n > p^*$" regression, but the methodology for doing so should not depend on the response.

Sufficiency aside (and that's a lot to set aside), I found little transportable methodology in Fisher's writing that might be applied to contemporary dimension reduction problems in regression. Nevertheless, dimension reduction was an issue during Fisher's era and before. In the next section we turn to a widely recognized dimension reduction method in statistics – principal components – whose beginnings occurred well before Fisher's time.

2 Introduction to Principal Components

It is well-established that principal components are a useful and important foundation for reducing the dimension of a multivariate random sample, represented here by the vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ in $\mathbb{R}^p$. Letting $\lambda_1 \geq \ldots \geq \lambda_p$ and $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$ denote the

eigenvalues and corresponding vectors of $\mathbf{\Sigma} = \mathrm{Var}(\mathbf{X})$, the population principal components are defined as the linearly transformed variables $\{\boldsymbol{\gamma}_1^T\mathbf{X}, \ldots, \boldsymbol{\gamma}_p^T\mathbf{X}\}$. We call the $\boldsymbol{\gamma}_j$s the *principal component (PC) directions*. The sample principal components are $\{\hat{\boldsymbol{\gamma}}_1^T\mathbf{X}, \ldots, \hat{\boldsymbol{\gamma}}_p^T\mathbf{X}\}$, where $\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_p$ are the eigenvectors (sample PC directions) corresponding to eigenvalues $\hat{\lambda}_1 > \ldots > \hat{\lambda}_p$ of the usual sample covariance matrix $\widehat{\mathbf{\Sigma}}$.

The history of principal components goes back at least to Adcock (1878) who wished to

> Find the most probable position of the straight line determined by the measured coordinates, each measure being equally good or of equal weight, $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ of n points, ...

Adcock identified his solution as the "principal axis", or first sample PC direction. Subsequent milestones were given in articles by Pearson (1901), Spearman (1904) and Hotelling (1933), but I found no direct reference to principal components in Fisher's writings. It has been discovered over time that the leading principal components, say $\{\boldsymbol{\gamma}_1^T\mathbf{X}, \ldots, \boldsymbol{\gamma}_k^T\mathbf{X}\}$, $k << p$, have a number of properties that may be helpful, depending on application-specific requirements. Reviews of principal components were given by Seber (1984), Christensen (2001) and Jolliffe (2002). Gould (1981, ch. 6) provided an interesting and illuminating historical account on the use of principal components in the social sciences. New properties and methods seem to be communicated often in contemporary statistical literature (see, for example, Jong and Kotz 1999; Tipping and Bishop 1999; Maronna 2005).

Principal components have also been studied in the context of regression, where $\mathbf{X}$ is now the vector of predictors that we would like to reduce prior to performing a regression with response Y. There are several reasons why such reduction may be useful in practice, including the possibilities of mitigating the effects of collinearity, facilitating model specification by allowing visualization of the regression in low dimensions, and providing a relatively small set of predictors on which to base prediction or interpretation. Collinearity in particular has been the main motivation for using principal components as a reductive method in regression. One persistent idea is that perhaps we can use the leading principal components in place of $\mathbf{X}$ with little loss of information; That is, the first few principal components should contain essentially the same information about Y as the original predictors $\mathbf{X}$, which is in the spirit of Fisher's idea of sufficiency. Kendall (1957, p. 75), Hocking (1976) and others suggested using principal components in this way. Scott (1992) suggested that the predictors might be "compressed" by using their principal components prior to additive nonparametric modeling. Such procedures have been questioned because principal components are computed from the marginal distribution of $\mathbf{X}$ and consequently the leading components may have little necessary relation with the response. It does seem optimistic to

think that the marginal of $\mathbf{X}$ would necessarily be structured so that the leading principal components contain the essential information about the response. Nevertheless, in support of the leading principal components, Mosteller and Tukey (1977, p. 397) turned this cause for concern into a desirable goal by posing the question

> ... how can we find linear combinations of the [predictors] that will be likely, or unlikely, to pick up regression from some as yet unspecified y?

It might seem unusual that Mosteller and Tukey would begin by asking how to perform linear reduction without reference to the response, but their approach is in agreement with Fisher's (1924) point that predictors "... be chosen without reference to the actual crop record." Mosteller and Tukey answered their question with the philosophical point that

> A malicious person who knew our x's and our plan for them could always invent a y to make our choices look horrible. But we don't believe nature works that way – more nearly that nature is, as Einstein put it (in German), "tricky, but not downright mean."

On the other hand, Cox (1968, p. 272) wrote in reference to reducing $\mathbf{X}$ by using the leading principal components

> A difficulty seems to be that there is no logical reason why the dependent variable should not be closely tied to the least important principal component.

A similar sentiment was expressed by Hotelling (1957), Hawkins and Fatti (1984) and others. Evidently, many authors did not trust nature in the same way as Mosteller and Tukey. Some gave examples of regressions where Cox's prediction apparently holds, with a few trailing principal components contributing important information about the response (Jolliffe 1982, Hadi and Ling 1998). Are there limits to the maliciousness of nature? If nature can produce regressions where there is useful information about the response in the last principal component $\boldsymbol{\gamma}_p^T\mathbf{X}$, can it also produce setting where the regression information is concentrated in a few of the original predictors, but is spread evenly across many of the principal components, making analysis on the principal component scale harder than analysis in the original scale? This possibility is relevant in recent proposals to ignore the hierarchical structure of the principal components and use instead a general subset M of them when developing linear regression models (Hwang and Nettleton 2003). In this regard, Jolliffe (2002, p. 177) commented that "... the choice of M for PC regression remains an open question."

On balance, the role for principal components in regression seems less clear-cut than their role in reducing $\mathbf{X}$ marginally. The advantages that principal components

enjoy in the multivariate setting, where the marginal distribution of $\mathbf{X}$ is of primary interest, may be of limited relevance in the regression context, where the conditional distribution of $Y|\mathbf{X}$ is of interest. It seems particularly appropriate to question the usefulness of principal components when $\mathbf{X}$ is fixed and controlled by the experimenter. In some experimental designs $\widehat{\mathbf{\Sigma}}$ is proportional to the identity, with the consequence that there is clearly no useful relation between its eigenstructure and the regression. And yet interest in using principal components for dimension reduction in regression has persisted, notably in the analysis of microarrays where principal components have been called "eigengenes" (Alter, et al. 2000). Chiaromonte and Martinelli (2002) and L. Li and H. Li (2004) used principal components for preprocessing microarray data, allowing subsequent application of other dimension reduction methodology that requires $n > p$. Acknowledging that preprocessing by principal components might be a viable alternative, Bura and Pfeiffer (2003) used marginal t-tests from the regression of the response on each predictor for prior selection of genes, a method to which Fisher would likely have objected.

3 Principal Components in Regression

In the rest of this article, we focus on regressions where $\mathbf{X}$ is random, so Y and $\mathbf{X}$ have a joint distribution, since a different structure seems necessary when $\mathbf{X}$ is fixed and subject to experimental control. In this context we may pursue dimension reduction through the conditional distribution of $Y|\mathbf{X}$ (forward regression), the conditional distribution of $\mathbf{X}|Y$ (inverse regression), or the joint distribution of $(Y, \mathbf{X})$. All three settings have been used in arguments for the relevance of principal components and related methodology. Forward regression with fixed predictors is perhaps the most common, particularly in early articles. Helland (1992) and Helland and Almoy (1994) assumed that $(Y, \mathbf{X})$ follows a multivariate normal in their development of "relevant components." Oman (1991) used an inverse model for $\mathbf{X}|Y$ in a heuristic argument that the coefficient vector $\boldsymbol{\alpha}$ in a linear regression model for $Y|\mathbf{X}$ should fall in or close to the space spanned by the first few principal components. Similarly, in their development of multiple-shrinkage principal component regression, George and Oman (1996) used a model for the joint distribution of $(Y, \mathbf{X})$ to reach the same conclusion.

In this article, we concentrate on $\mathbf{X}|Y$, although the goal is still to reduce the dimension of $\mathbf{X}$ with little or no loss of information on $Y|\mathbf{X}$. Model-based regression analyses traditionally condition on the observed values of the predictors, a characteristically Fisherian operation, even if $\mathbf{X}$ is random (See Savage 1976, p. 468, for further discussion of this point). Nevertheless, the conditional distribution of $\mathbf{X}|Y$ may provide a better handle on reductive information, and I found nothing in Fisher's writings that would compel consideration of only forward regressions. To facilitate the expo-

sition, let $\mathbf{X}_y$ denote a random variable distributed as $\mathbf{X}|(Y = y)$. Subspaces are indicated as $\mathcal{S}_{(\cdot)}$ where the argument is a spanning matrix, and $\mathbf{U} \perp\!\!\!\perp \mathbf{V}|\mathbf{W}$ means that the random vectors $\mathbf{U}$ and $\mathbf{V}$ are conditionally independent given any value for the random vector $\mathbf{W}$.

3.1 A first regression model implicating principal components

Consider the following multivariate model for the inverse regression of $\mathbf{X}$ on Y,

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \sigma\boldsymbol{\varepsilon}. \tag{2}$$

Here $\boldsymbol{\mu} = \mathrm{E}(\mathbf{X})$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$, $\boldsymbol{\Gamma}^T\boldsymbol{\Gamma} = I_d$, $\sigma \geq 0$ and d is assumed to be known. The coordinate vector $\boldsymbol{\nu}_y \in \mathbb{R}^d$ is an unknown function of y centered to have mean 0, $\sum_y \boldsymbol{\nu}_y = 0$, and $\mathrm{Var}(\boldsymbol{\nu}_Y) > 0$, but is otherwise unconstrained. The error vector $\boldsymbol{\varepsilon} \in \mathbb{R}^p$ is assumed to be independent of Y, and to be normally distributed with mean 0 and identity covariance matrix. This model specifies that, after translation by the intercept $\boldsymbol{\mu}$, the conditional means fall in the d-dimensional subspace $\mathcal{S}_{\boldsymbol{\Gamma}}$ spanned by the columns of $\boldsymbol{\Gamma}$. The vector $\boldsymbol{\nu}_y$ contains the coordinates of the translated conditional mean $\mathrm{E}(\mathbf{X}_y) - \boldsymbol{\mu}$ relative to the basis $\boldsymbol{\Gamma}$. The mean function is quite flexible, even when d is small, say at most 3 or 4. Aside from the subscript y, nothing on the right side of this model is observable.

While the mean function for model (2) is permissive, the variance function is restrictive, requiring essentially that the predictors be in the same scale and that, conditional on Y, they be independent with the same variance. Nevertheless, this model provides a useful starting point for our consideration of principal components, and may be appropriate for some applications dealing with measurement error, image recognition, microarray data, and calibration (Oman 1991). And it may serve as a useful first model when the data are not plentiful, following the spirit of Fisher's recommendations.

Model (2) is similar in form to the functional model for multivariate problems studied by Anderson (1984), but there are important conceptual differences. Model (2) is a formal statement about the regression of $\mathbf{X}$ on Y, with the $\boldsymbol{\nu}_y$'s being fixed because we condition on Y, in the same way forward regression models are conditional on $\mathbf{X}$. Functional models for multivariate data postulate latent fixed effects without conditioning. This model will lead in directions that are not available when considering functional models for multivariate data.

The following proposition connects the inverse regression model (2) with the forward regression of Y on $\mathbf{X}$:

Proposition 1 *Under the inverse model (2), the distribution of $Y|\mathbf{X}$ is the same as the distribution of $Y|\boldsymbol{\Gamma}^T\mathbf{X}$ for all values of $\mathbf{X}$.*

According to this proposition, $\mathbf{X}$ can be replaced by the *sufficient reduction* $\mathbf{\Gamma}^T\mathbf{X}$ without loss of information on the regression of Y on $\mathbf{X}$, and without specifying the marginal distribution of Y or the conditional distribution of $Y|\mathbf{X}$. Here "sufficient reduction" is used in the same spirit as Fisher's "sufficient statistic". One difference is that sufficient statistics are observable, while sufficient reductions may contain unknown parameters, $\mathbf{\Gamma}$ in this instance, and thus need to be estimated. A reasonable estimate of $\mathbf{\Gamma}$ is needed for the sufficient reduction $\mathbf{\Gamma}^T\mathbf{X}$ to be useful in practice.

3.2 Estimation via the method of maximum likelihood

Even with the previously imposed constraint that $\mathbf{\Gamma}^T\mathbf{\Gamma} = I_d$, $\mathbf{\Gamma}$ and $\boldsymbol{\nu}_y$ are not simultaneously estimable under model (2) since, for any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{d\times d}$, we can always rewrite $\mathbf{\Gamma}\boldsymbol{\nu}_y = (\mathbf{\Gamma}\mathbf{O})(\mathbf{O}^T\boldsymbol{\nu}_y)$, leading to a different factorization. However, the *reductive subspace* $\mathcal{S}_{\mathbf{\Gamma}} = \mathrm{Span}(\mathbf{\Gamma})$ is estimable, and that is the focus of our inquiry. The matrix $\mathbf{\Gamma}$ is of interest only by virtue of the subspace generated by its columns, the condition $\mathbf{\Gamma}^T\mathbf{\Gamma} = I_d$ being imposed for convenience. This implies that two reductions $\mathbf{\Gamma}^T\mathbf{X}$ and $\mathbf{A}\mathbf{\Gamma}^T\mathbf{X}$ that are connected by a full-rank linear transformation $\mathbf{A}$ are regarded as equivalent for present purposes. To emphasize this distinction we will refer to the parameter space for $\mathcal{S}_{\mathbf{\Gamma}}$ as a Grassman manifold $\mathcal{G}_{p\times d}$ of dimension d in $\mathbb{R}^p$.

Consider a function $G(\mathbf{A})$ that is defined on the set of $p \times d$ matrices with $d < p$ and $\mathbf{A}^T\mathbf{A} = I_d$ and has the property that $G(\mathbf{A}\mathbf{O}) = G(\mathbf{A})$ for any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{d\times d}$. The function $G(\mathbf{A})$ depends only on the span of its argument: If $\mathrm{Span}(\mathbf{A}) = \mathrm{Span}(\mathbf{B})$ then $G(\mathbf{A}) = G(\mathbf{B})$. The set of d dimensional subspaces of $\mathbb{R}^p$ is called a Grassman manifold; a single point in a Grassman manifold being a subspace. A Grassman manifold is the natural parameter space for the $\mathbf{\Gamma}$ parameterization in this article. While no technical use will be made of Grassman manifolds, the terminology is proper in this context and it may serve as a reminder that only the subspace generated by the columns of $\mathbf{\Gamma}$ is of interest. For background on Grassman manifolds, see Edelman, et al. (1998).

There are several ways to approach estimation of a sufficient reduction, but staying with Fisher we will use the method of maximum likelihood. Assuming that the data consist of a random sample of size n from $(Y, \mathbf{X})$, the maximum likelihood estimator of $\mathcal{S}_{\mathbf{\Gamma}}$ can be constructed by holding $\mathbf{\Gamma}$ and σ^2 fixed at possible values $\mathbf{G}$ and s^2 and then maximizing the log likelihood over $\boldsymbol{\mu}$ and $\boldsymbol{\nu}_y$. This yields $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$ and, in the absence of replicate y's, a separate value for each of the n d-dimensional vectors $\boldsymbol{\nu}_y$ as a function of $(\mathbf{G}, s^2)$: $\boldsymbol{\nu}_y(\mathbf{G}, s^2) = \mathbf{G}^T(\mathbf{X}_y - \bar{\mathbf{X}})$. Substituting back, the partially

maximized log likelihood $M_{\rm PC}$ is then apart from constants

$$\begin{aligned} M_{\rm PC}(\mathbf{G}, s^2) &= -(np/2)\log(s^2) - (1/2s^2)\sum_y \|\mathbf{X}_y - \bar{\mathbf{X}} - P_{\mathbf{G}}(\mathbf{X}_y - \bar{\mathbf{X}})\|^2 \\ &= -(np/2)\log(s^2) - (n/2s^2)\,\mathrm{trace}(\widehat{\mathbf{\Sigma}} Q_{\mathbf{G}}), \end{aligned}$$

where $P_{\mathbf{G}} = \mathbf{G}\mathbf{G}^T$ is the projection onto $\mathcal{S}_{\mathbf{G}}$ and $Q_{\mathbf{G}} = I - P_{\mathbf{G}}$. Although in this and later likelihood functions we use $\mathbf{G}$ as an argument, the function itself depends only on $\mathcal{S}_{\mathbf{G}}$ and thus maximization is over the Grassman manifold $\mathcal{G}_{p\times d}$. Recalling that $\{\hat{\lambda}_j\}$ and $\{\hat{\gamma}_j\}$ are the eigenvalues and eigenvectors of $\widehat{\mathbf{\Sigma}}$, it follows that the maximum likelihood estimator $\widehat{\mathcal{S}}_{\mathbf{\Gamma}}$ of $\mathcal{S}_{\mathbf{\Gamma}}$ is

$$\widehat{\mathcal{S}}_{\mathbf{\Gamma}} = \mathrm{Span}\{\hat{\gamma}_1, \ldots, \hat{\gamma}_d\},$$

and that the estimator of σ^2 is $\hat{\sigma}^2 = \sum_{j=d+1}^{p} \hat{\lambda}_j/p$. A sufficient reduction is thus estimated by the first d sample principal components, and for this reason we will refer to model (2) as a *PC regression model*. As mentioned previously, for a given d, this approach does not distinguish between the first d principal components and any full rank linear transformation of them.

In the absence of replication in the y's, observed responses play no role other than acting collectively as a conditioning argument. The same reduction would be obtained with a continuous multivariate response. In fact, it is not necessary for the response to be even observed and thus model (2) is one possible route to satisfying Mosteller and Tukey's desire to reduce $\mathbf{X}$ for an "... as yet unspecified y," and is in accord with Fisher's requirement that "... the variables to be employed must be chosen without reference to the actual crop record." However, model (2) does not hold for all potential responses, but requires that the predictors be conditionally independent with common variance.

The fact that all full-rank linear transformations $\mathbf{A}\widehat{\mathbf{\Gamma}}^T\mathbf{X}$ of the first d sample principal components $\widehat{\mathbf{\Gamma}}^T\mathbf{X}$ are equivalent reductions in the context of model (2) may cast doubt on the usefulness of *post hoc* reification of principal components. The safest interpretations are perhaps the ones that suggest regression mechanisms that can be verified independently of the principal components themselves.

As mentioned previously, the PC model (2) may be appropriate for some applications, but there is also value in the paradigm it suggests for extension to other settings. In the next section we sketch at the conceptual level how the ideas behind (2) can be applied to exponential families, returning to normal models in Section 4.

3.3 Extensions to exponential families

As in the PC model, we assume that the predictors are independent given Y, but instead of requiring $\mathbf{X}_y$ to be normally distributed we assume that the j-th conditional predic-

tor X_{yj} is distributed according to a one-parameter exponential family with density or mass function of the form

$$f_j(x|\eta_{yj}, Y = y) = a_j(\eta_{yj})b_j(x)\exp(x\eta_{yj}). \tag{3}$$

We also assume that the natural parameter η_{yj} follows the model for the conditional mean $\mathrm{E}(\mathbf{X}_y)$ in the normal case:

$$\eta_{yj} = \mu_j + \boldsymbol{\gamma}_j^T \boldsymbol{\nu}_y, \;\; j = 1, \ldots, p,$$

where μ_j is the j-th coordinate of $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}_j^T$ is the j-th row of $\boldsymbol{\Gamma}$. Under this setup $\boldsymbol{\Gamma}^T\mathbf{X}$ is again a sufficient reduction:

Proposition 2 *Under the inverse exponential model (3), the distribution of* $Y|\mathbf{X}$ *is the same as the distribution of* $Y|\boldsymbol{\Gamma}^T\mathbf{X}$ *for all values of* $\mathbf{X}$.

Given y, $\boldsymbol{\mu}$ and $\boldsymbol{\Gamma}$ the likelihood for $\boldsymbol{\nu}_y$ is constructed from the products of (3) for $j = 1, \ldots, p$. This corresponds to fitting a generalized linear model with offsets $\boldsymbol{\mu}_j$, "predictors" $\boldsymbol{\gamma}_j$ and regression coefficient $\boldsymbol{\nu}_y$. The remaining parameters $\boldsymbol{\mu}$ and $\mathcal{S}_{\boldsymbol{\Gamma}} \in \mathcal{G}_{p\times d}$ can then be estimated by combining these intermediate partially maximized likelihoods over y. The essential point here is that *generalized PC models* can be constructed from the normal PC model (2) in the way that generalized linear models are constructed from linear models. Marx and Smith (1990) developed a method of principal component estimation for generalized linear models. Their approach, while recognizing issues that come with generalized linear models, seems quite different from the approach suggested here.

For example, if the coordinates of $\mathbf{X}_y$ are independent Bernoulli random variables, then we can express the model in terms of a multivariate logit defined coordinate-wise as

$$\mathrm{multlogit}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y,$$

where the right hand side is the same as the mean function for the PC model (2). By analogy with the normal case, this leads to *principal components for binary variables*, but they are not computed from the eigenvectors of a covariance matrix and they seem distinct from the recent proposal by de Leeuw (2006) for marginal reduction of $\mathbf{X}$.

4 Principal Fitted Components

In the PC model (2) no direct use is made of the response, which plays the role of an implicit conditioning argument. In principle, once the response is known, we should be able to tailor our reduction to that response and thereby obtain a supervised reduction.

One way to adapt the reduction for a specific response is by modeling $\boldsymbol{\nu}_y$. This can be facilitated by graphical analyses when the response is bivariate or univariate. Recalling that X_j denotes the j-th predictor in $\mathbf{X}$, we can gain information from the data on the mean function $\mathrm{E}(X_j|Y)$ by investigating the p two or three-dimensional inverse response plots of X_j versus Y, $j = 1, \ldots, p$, as described by Cook and Weisberg (1994, Ch. 8) and Cook (1998, Ch. 10). While such a graphical investigation might not be practical if p is in the hundreds or more, it might be doable when p is in the tens, and certainly if p is less that, say, 25.

Assume then that $\boldsymbol{\nu}_y = \boldsymbol{\beta}\mathbf{f}_y$, where $\boldsymbol{\beta} \in \mathbb{R}^{d\times r}$, $d \leq r$, has rank d and $\mathbf{f}_y \in \mathbb{R}^r$ is a known vector-valued function of the response with $\sum_y \mathbf{f}_y = 0$. For example, if it is decided that each inverse mean function $\mathrm{E}(X_j|Y = y)$ can be modeled adequately by a cubic polynomial in y, then $\mathbf{f}_y$ equals $(y, y^2, y^3)^T$ minus its sample average. When Y is univariate and graphical guidance is not available, $\mathbf{f}_y$ could be constructed by first partitioning the range of Y into r "slices" or bins H_k, and then setting the kth coordinate f_{yk} of $\mathbf{f}_y$ to

$$f_{yk} = J(y \in H_k) - n_k/n,\ k = 1, \ldots, h, \tag{4}$$

where J is the indicator function and n_k is the number of observations falling in H_k. The jth coordinate ν_{yj} of $\boldsymbol{\nu}_y$ is then modeled as a constant in each slice H_k:

$$\nu_{yj} = \sum_{k=1}^{r} \beta_{jk} f_{yk} = \sum_{k=1}^{r} \beta_{jk}(J(y \in H_k) - n_k/n),$$

where β_{jk} is the jk-th element of $\boldsymbol{\beta}$. Each coordinate of the vector $\boldsymbol{\nu}_y = \boldsymbol{\beta}\mathbf{f}_y$ is now a step function that is constant within slices. Many other possibilities for basis functions are available in the literature. For example, we might adapt a classical Fourier series form (see, for example, Eubank 1988, p. 82) and set $\mathbf{f}_y = \mathbf{g}_y - \bar{\mathbf{g}}$, where

$$\mathbf{g}_y = (\cos(2\pi y), \sin(2\pi y), \ldots, \cos(2\pi k y), \sin(2\pi k y))^T$$

with $r = 2k$. We will use the slice basis function (4) later in this article because it leads to a connection with inverse regression (SIR, Li 1991). However, no claim is made that this is the "best" or even a generally reasonable nonparametric choice for $\mathbf{f}_y$. The parameterization $\boldsymbol{\nu}_y = \boldsymbol{\beta}\mathbf{f}_y$ can be used with exponential families as described in Section 3.3 and the PC model (2). In particular, the slice basis function can be used to allow for replication in model (2), with the slices corresponding to the different values of y.

Substituting $\boldsymbol{\nu}_y = \boldsymbol{\beta}\mathbf{f}_y$ into model (2) we obtain the new model

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \sigma\boldsymbol{\varepsilon} \tag{5}$$

for which $\mathbf{\Gamma}^T\mathbf{X}$ is still a sufficient reduction. Let $\mathbb{X}$ denote the $n \times p$ matrix with rows $(\mathbf{X}_y - \bar{\mathbf{X}})^T$, let $\mathbf{F}$ denote the $n \times r$ matrix with rows $\mathbf{f}_y^T$, and let $\widehat{\mathbb{X}} = P_{\mathbf{F}}\mathbb{X}$ denote the $n \times p$ matrix of centered fitted values from the multivariate linear regression of $\mathbf{X}$ on $\mathbf{f}_y$, including an intercept. Holding $\mathbf{\Gamma}$ and σ fixed at $\mathbf{G}$ and s, and maximizing the likelihood over $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$, we obtain $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$, $\hat{\boldsymbol{\beta}} = \mathbf{G}^T\mathbb{X}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^-$, and the partially maximized log likelihood (see Appendix C for details)

$$M_{\mathrm{PFC}}(\mathbf{G}, s^2) = (-np/2)\log(s^2) - (n/2s^2)\{\,\mathrm{trace}[\widehat{\mathbf{\Sigma}}] - \,\mathrm{trace}[P_{\mathbf{G}}\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}P_{\mathbf{G}}]\} \quad (6)$$

where $\widehat{\mathbf{\Sigma}}_{\mathrm{fit}} = \widehat{\mathbb{X}}^T\widehat{\mathbb{X}}/n$ is the sample covariance matrix of the fitted values $\widehat{\mathbb{X}}$. The likelihood again depends only on $\mathcal{S}_{\mathbf{G}}$. The rank of $\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}$ is at most r and typically $\mathrm{rank}(\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}) = r$. In any event, we assume that $\mathrm{rank}(\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}) \geq d$. The likelihood is then maximized by setting $\widehat{\mathcal{S}}_{\mathbf{\Gamma}}$ equal to the span of eigenvectors $\hat{\boldsymbol{\phi}}_1, \ldots, \hat{\boldsymbol{\phi}}_d$ corresponding to largest d eigenvalues $\hat{\lambda}_i^{\mathrm{fit}}, i = 1, \ldots, d$, of $\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}$.

We call the corresponding components $\hat{\boldsymbol{\phi}}_1^T\mathbf{X}, \ldots, \hat{\boldsymbol{\phi}}_p^T\mathbf{X}$ *principal fitted components (PFC)*, and call the associated eigenvectors $\hat{\boldsymbol{\phi}}_1, \ldots, \hat{\boldsymbol{\phi}}_d$ *PFC directions*. The corresponding estimate of scale is $\hat{\sigma}^2 = (\sum_{i=1}^p \hat{\lambda}_i - \sum_{i=1}^d \hat{\lambda}_i^{\mathrm{fit}})/p$. A sufficient reduction under model (5) is then estimated by the first d PFC's. We call model (5) a *PFC model* to distinguish it from the PC model (2).

5 Comparing PC's and PFC's

In the PC model (2) there are $(n-1)d$ ν-parameters to be estimated, while in the PFC model (5) the corresponding number is just dr, which does not increase with n. Consequently, assuming that (5) is reasonable, we might expect it to yield more accurate estimates that the unsupervised model (2).

5.1 Simulating $\widehat{\mathcal{S}}_{\mathbf{\Gamma}}$

A small simulation using multivariate normal $(Y, \mathbf{X})$ was conducted to obtain first insights into the operating characteristics of principal components and principal fitted components. First, Y was generated as a normal random variable with mean 0 and variance σ_Y^2, and then $\mathbf{X}_y$ was generated according to the inverse model

$$\mathbf{X}_y = \mathbf{\Gamma}y + \sigma\boldsymbol{\varepsilon}, \quad (7)$$

with $\mathbf{\Gamma}^T = (1, 0, \ldots, 0)$, $p = 10$ and, to avoid singularities in later discussion, $\sigma > 0$. The forward regression $Y|\mathbf{X}$ follows a textbook normal linear regression model where the mean function depends on $\mathbf{X}$ only via $\mathbf{\Gamma}^T\mathbf{X}$, which is equal to X_1 for this simulation. Each data set was summarized by computing the angle between $\mathbf{\Gamma}$ and each

of $\hat{\gamma}_1$ (PC), $\hat{\phi}_1$ (PFC) from the fit of model (5) with $\mathbf{f}_y = y - \bar{y}$, and the estimated coefficient vector from the forward ordinary least squares (OLS) fit of Y on $\mathbf{X}$. Three aspects of the simulation model were varied, the sample size n, the conditional error standard deviation σ and the marginal standard deviation of Y, σ_Y. Simulation models with multivariate normal $(Y, \mathbf{X})$ are useful because they allow straightforward comparisons with other methods like forward OLS, and they limit the flexibility for tuning to emphasize the advantages of a particular method.

Shown in Figure 1a are average angles taken over 500 replications versus n, with $\sigma = \sigma_Y = 1$. On the average, the OLS vector was observed to be a bit closer to $\mathcal{S}_{\mathbf{\Gamma}}$ than the PC vector $\hat{\gamma}_1$, except for small samples when $\hat{\gamma}_1$ was the better of the two. More importantly, the PFC vector $\hat{\phi}_1$ was observed to be more accurate than the other two estimators at all sample sizes. The difference between the PFC and OLS estimators can be increased by varying σ_Y, as is apparent from Figure 1b.

Figure 1b shows the results of a second series of simulations at various values of σ_Y with $n = 40$ and $\sigma = 1$. The method of principal fitted components is seen to be superior for small values of σ_Y, while it is essentially equivalent to principal components for large values. Perhaps surprisingly, the OLS estimator is clearly the worst method over most of the range of σ_Y. Figure 1c shows average angles as σ varies with $n = 40$ and $\sigma_Y = 1$. Again, PFC is seen to be the best method, with the relative performance of PC and PFC depending on the value of σ.

Under the PC model (2), the accuracy of the PC or PFC estimates of $\mathcal{S}_{\mathbf{\Gamma}}$ should improve as n increases, or as σ_Y increases or as σ decreases. The results in Figure 1 agree with this expectation. To explain the relative behavior of the OLS estimates in the simulation results for model (7), write the corresponding forward regression model as

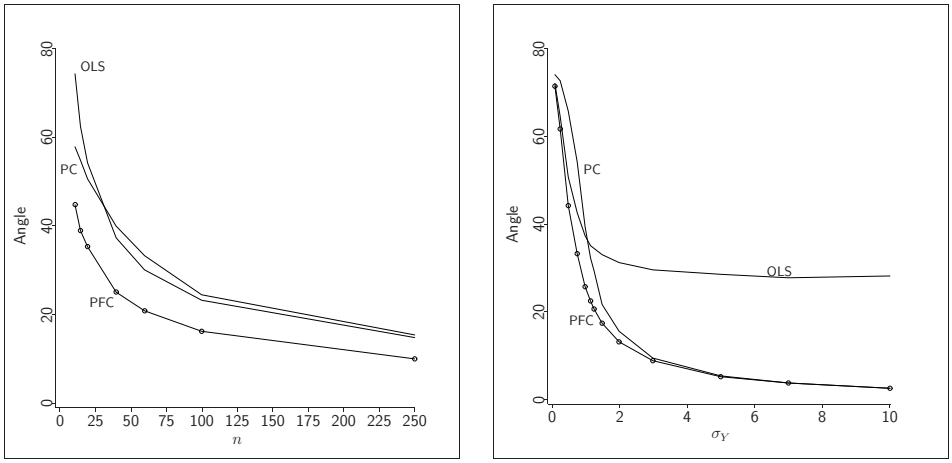
$$Y = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{X} + \sigma_{Y|\mathbf{X}} \epsilon,$$

let $R = \sigma_Y^2/(\sigma_Y^2 + \sigma^2)$, and let $\widehat{\boldsymbol{\alpha}}$ denote the OLS estimator of $\boldsymbol{\alpha}$. Then it can be shown that $\boldsymbol{\alpha} = R\mathbf{\Gamma}$, that

$$\mathrm{Var}(\widehat{\boldsymbol{\alpha}}) = RQ_{\mathbf{\Gamma}} + R(1-R)P_{\mathbf{\Gamma}}$$

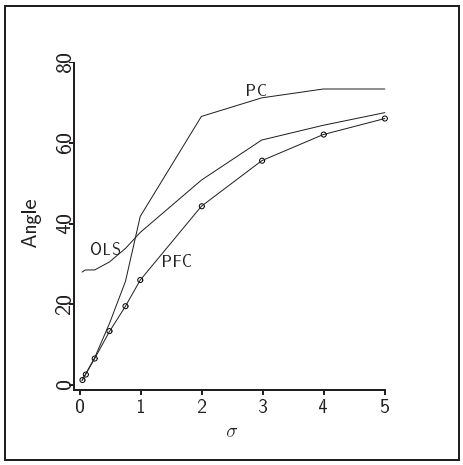
and that $\boldsymbol{\alpha}^T[\mathrm{Var}(\widehat{\boldsymbol{\alpha}})]^{-1}\boldsymbol{\alpha} = R/(1-R)$. As $\sigma_Y \to \infty$, $R \to 1$ but $\mathrm{Var}(\widehat{\boldsymbol{\alpha}}) \geq RQ_{\mathbf{\Gamma}}$. Thus, $\mathrm{Var}(\widehat{\boldsymbol{\alpha}})$ is bounded below, and this explains the behavior of the OLS estimates in Figure 1b. On the other hand, as $\sigma \to \infty$, $\boldsymbol{\alpha} \to 0$ and $\mathrm{Var}(\widehat{\boldsymbol{\alpha}}) \to 0$, but $\boldsymbol{\alpha}^T[\mathrm{Var}(\widehat{\boldsymbol{\alpha}})]^{-1}\boldsymbol{\alpha} \to 0$ as well. Consequently, $\boldsymbol{\alpha} \to 0$ faster than the "standard deviation" of its estimates and the performance of $\widehat{\boldsymbol{\alpha}}$ must deteriorate. These results can be extended straightforwardly to model (2) with $d = 1$, but then the behavior of the OLS estimator will depend on $\mathrm{Var}(\boldsymbol{\nu}_Y)$ and $\mathrm{Cov}(\boldsymbol{\nu}_Y, Y)$.

Letting Γ_k denote the kth element of the $\mathbf{\Gamma}$ in model (7), the correlation between

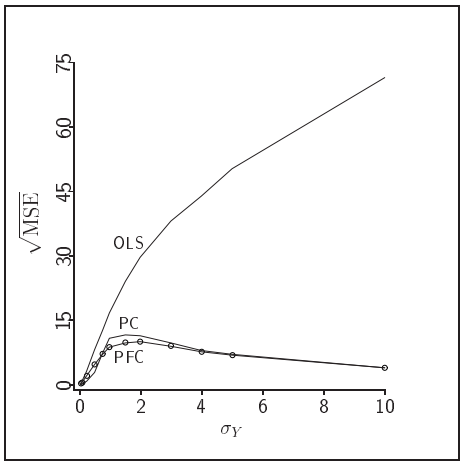


a. Angle vs. n

b. Angle vs. σ_Y



c. Angle vs. σ



d. $\sqrt{\text{MSE}}$ vs. σ_Y

Figure 1: Simulation results for model (7). Figures a-d display average simulation angles between estimated and true direction versus (a) sample size with $\sigma_Y = \sigma = 1$; (b) σ_Y with $n = 40$, $\sigma = 1$; and (c) σ with $n = 40$, $\sigma_Y = 1$. Figure d is the square root of the standardized MSE versus σ_Y with $n = 40$ and $\sigma = 1$.

the jth and kth predictors is

$$\rho_{jk} = \frac{\Gamma_k \Gamma_j \sigma_Y^2}{\sqrt{(\sigma^2 + \Gamma_j^2 \sigma_Y^2)(\sigma^2 + \Gamma_k^2 \sigma_Y^2)}}.$$

If Γ_j and Γ_k are both nonzero, then $|\rho_{jk}| \to 1$ as $\sigma_Y^2 \to \infty$ with σ^2 fixed. However, $\boldsymbol{\Gamma} = (1, 0, \ldots, 0)^T$ in the version of model (7) used to produce Figure 1, with the consequence that the predictors are both marginally and conditionally independent. Thus, the bottoming out of the OLS estimator in Figure 1b for large values of σ_Y and in Figure 1c for small values of σ has nothing to do with collinearity. To gain further insights into the impact of collinearity in this situation, we repeated the simulation leading to Figure 1b with $\boldsymbol{\Gamma} = (1, \ldots, 1)^T/\sqrt{10}$. Now, $\rho_{jk} \to 1$ as $\sigma_Y \to \infty$ for all pairs of predictors. However, within the simulation error, these new results were identical to those of Figure 1b. This suggests that the value of principal component estimators does not rest solely with the presence of collinearity.

5.2 Multivariate normal $(Y, \mathbf{X})$

In this section we provide theoretical results to support the observations from the previous simulations. Assume that $(Y, \mathbf{X})$ follows a nonsingular multivariate normal distributions, and that the predictors are conditionally independent with common variance σ^2 given the response. Let $a_{\text{ols}}^2 = (\widehat{\boldsymbol{\alpha}}^T \boldsymbol{\Gamma})^2/\|\widehat{\boldsymbol{\alpha}}\|^2$ and let $a_{\text{pfc}}^2 = (\hat{\boldsymbol{\phi}}_1^T \boldsymbol{\Gamma})^2$, where $\boldsymbol{\Gamma} \in \mathbb{R}^p$ is as defined in (5), still with length 1. Then the squared cosines a_{ols}^2 and a_{pfc}^2 are each asymptotically distributed as a constant, respectively c_{ols} and c_{pfc}, times a chi-squared random variable with $p-1$ degrees of freedom. These constants depend only on the common conditional variance σ^2 of the predictors and the marginal variance of the response: $c_{\text{ols}} = R^{-1}$ and $c_{\text{pfc}} = \sigma^2 R^{-2}/\sigma_Y^2$, where R is as defined in Section 5.1. As $\sigma_Y \to \infty$, $c_{\text{ols}} \to 1$, while $c_{\text{pfc}} \to 0$. This again accounts for the qualitative behavior of Figure 1b. Perhaps more interestingly, the asymptotic distributions of a_{ols}^2 and a_{pfc}^2 do not depend on $\boldsymbol{\Gamma}$, so in this context the relative behavior of OLS and PFC has nothing to do with collinearity in the marginal distribution of $\mathbf{X}$.

5.3 Prediction

While this article is focused on estimation of reductive subspaces, this first simulation study provides a convenient place to touch base with predictive considerations. Since the forward regression follows a normal linear model, we characterize predictive performance by using the scaled mean squared error

$$\text{MSE} = \text{E}(Y_f - \hat{a} - \hat{b}\widehat{\boldsymbol{\Gamma}}^T \mathbf{X}_f)^2/\sigma_Y^2 \tag{8}$$

where $(Y_f, \mathbf{X}_f)$ represents a future observation on $(Y, \mathbf{X})$, and the expectation is taken over $(Y_f, \mathbf{X}_f)$ and the data. The marginal variance σ_Y^2 was used as a scaling constant because the numerator of (8) is in the units of Y^2. $\widehat{\mathbf{\Gamma}}$ can be $\hat{\boldsymbol{\gamma}}_1$ (PC), $\hat{\boldsymbol{\phi}}_1$ (PFC), or the OLS estimator of $\boldsymbol{\alpha}$ scaled to have length 1, $\widehat{\boldsymbol{\alpha}}/\|\widehat{\boldsymbol{\alpha}}\|$. The intercept $\hat{a}$ and slope $\hat{b}$ were then computed from the OLS fit of Y_i on $\widehat{\mathbf{\Gamma}}^T\mathbf{X}_i$. The MSE was estimated by first calculating the expectation explicitly over the future observations and then using 500 simulated data sets to estimate the remaining expectation over the data. Figure 1d shows the resulting MSE as function of σ_Y and the three estimates. The PFC estimator performed the best, except for small values of σ_Y where the PC estimator did a bit better. The OLS estimator is dominated by the other two, and its MSE appears to be unbounded as σ_Y increases. This phenomenon is related to the bottoming out of the OLS estimator in Figure 1b since, for any fixed angle between $\widehat{\mathbf{\Gamma}}$ and $\mathbf{\Gamma}$, the magnitude of the average predictive error increases as σ_Y increases. In other words, the average angle must approach 0 as σ_Y increases for the predictive error to be bounded. As may be clear from the discussion of Section 5.2, the PC and PFC estimators have this property under model (7), while the OLS estimator does not. As before, the results shown in Figure 1d have nothing to do with collinearity.

6 Inverse Models with Structured Errors

In this section we consider expanded PC and PFC models that allow for heterogeneous and partially correlated errors, starting with the PC model in Section 6.1 and turning to the PFC model in Section 6.2.

6.1 An expanded PC model

Consider the following PC model with heterogeneous errors,

$$\mathbf{X}_y = \boldsymbol{\mu} + \mathbf{\Gamma}\boldsymbol{\nu}_y + \mathbf{\Gamma}_0\mathbf{\Omega}_0\boldsymbol{\varepsilon}_0 + \mathbf{\Gamma}\mathbf{\Omega}\boldsymbol{\varepsilon}. \tag{9}$$

Here $\boldsymbol{\mu}$ and $\boldsymbol{\nu}_y$ are as defined previously, and $\mathbf{\Gamma}_0 \in \mathbb{R}^{p\times(p-d)}$ is a *completion* of $\mathbf{\Gamma}$; that is, $(\mathbf{\Gamma}_0, \mathbf{\Gamma}) \in \mathbb{R}^{p\times p}$ is an orthogonal matrix. The error vectors $\boldsymbol{\varepsilon}_0 \in \mathbb{R}^{p-d}$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ are independent and normally distributed, each with mean 0 and identity covariance matrix. The full-rank matrices $\mathbf{\Omega} \in \mathbb{R}^{d\times d}$ and $\mathbf{\Omega}_0 \in \mathbb{R}^{(p-d)\times(p-d)}$ serve in part to convert the normal errors to appropriate scales and, without loss of generality, are assumed to be symmetric. Since the two error components must be independent, the variance structure in this model is still restrictive, but it is considerably more permissive than the variance structure in the first PC model (2) and the components of $\mathbf{X}$ can now be in different scales. We assume that the y's are distinct. Replication for this model can be addressed with a version of model (12), which is discussed in Section 6.2.

Linearly transforming both sides of model (9), we have $\mathbf{\Gamma}^T\mathbf{X}_y = \mathbf{\Gamma}^T\boldsymbol{\mu} + \boldsymbol{\nu}_y + \mathbf{\Omega}\boldsymbol{\varepsilon}$ and $\mathbf{\Gamma}_0^T\mathbf{X}_y = \mathbf{\Gamma}_0^T\boldsymbol{\mu} + \mathbf{\Omega}_0\boldsymbol{\varepsilon}_0$. From these representations we see that $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)^T\mathbf{X}_y$ contains two independent components. The active y-dependent component $\mathbf{\Gamma}^T\mathbf{X}_y$ consists of the coordinates of the projection of $\mathbf{X}$ onto $\mathcal{S}_{\mathbf{\Gamma}}$. It has an arbitrary mean and constant but a general covariance matrix. The other projective component lives in the orthogonal complement $\mathrm{Span}(\mathbf{\Gamma}_0) = \mathcal{S}_{\mathbf{\Gamma}}^{\perp}$ and has constant mean and variance.

We again find that $\mathbf{\Gamma}^T\mathbf{X}$ is a sufficient reduction:

Proposition 3 *Under the extended PC model (9), the distribution of $Y|\mathbf{X}$ is the same as the distribution of $Y|\mathbf{\Gamma}^T\mathbf{X}$ for all values of $\mathbf{X}$.*

Maximizing over $\boldsymbol{\mu}$, $\boldsymbol{\nu}_y$ and $\mathbf{\Omega}_0$, the partially maximized log likelihood L_{PC} is a function of possible values $\mathbf{G}$ and $\mathbf{M}$ for $\mathbf{\Gamma}$ and $\mathbf{\Omega}^2$, apart from constants:

$$L_{\mathrm{PC}}(\mathbf{G}, \mathbf{M}) = -(n/2)\log(|\mathbf{G}_0^T\widehat{\mathbf{\Sigma}}\mathbf{G}_0|) - (n/2)\log(|\mathbf{M}|),$$

where $\mathbf{G}_0$ is a completion of $\mathbf{G}$. $\mathbf{\Omega}^2$ is not estimable, but because the likelihood factors, it is maximized over $\mathbf{G}_0$ for any fixed $\mathbf{M}$ by any full-rank linear transformation of the last $p - d$ sample PC directions. Consequently, we might be tempted to take

$$\widehat{\mathcal{S}}_{\mathbf{\Gamma}} = \mathrm{Span}^{\perp}(\hat{\boldsymbol{\gamma}}_{d+1}, \ldots, \hat{\boldsymbol{\gamma}}_p). \tag{10}$$

Thus a sufficient reduction is again estimated by the first d sample principal components computed from $\widehat{\mathbf{\Sigma}}$. The likelihood for the first PC model (2) is able to tell us which principal components are *likely* to be associated with an as yet unspecified y. In contrast, the likelihood for the extended PC model (9) suggests the principal components that are *unlikely* to be associated with such y. In effect, because of the added flexibility in model (9) there is not enough information to estimate $\mathcal{S}_{\mathbf{\Gamma}}$ directly, but there is some information on its orthogonal complement. Nevertheless, additional conditions are necessary for principal components to work well under this model.

Some of the differences between the PC model (2) and the extended version (9) can be seen in the marginal covariances of the predictors. Under model (2)

$$\mathbf{\Sigma} = \sigma^2\mathbf{\Gamma}_0\mathbf{\Gamma}_0^T + \mathbf{\Gamma}\{\sigma^2 I + \mathrm{Var}(\boldsymbol{\nu}_Y)\}\mathbf{\Gamma}^T.$$

Here the smallest eigenvalue of $\mathbf{\Sigma}$ is equal to σ^2 with multiplicity $p - d$ and corresponding eigenvectors $\mathbf{\Gamma}_0$. The largest d eigenvalues, which are all larger than σ^2, are the same as the eigenvalues of $\sigma^2 I + \mathrm{Var}(\boldsymbol{\nu}_Y)$. The corresponding PC directions are $\mathbf{\Gamma}\mathbf{v}_j$, $j = 1, \ldots, d$, where $\{\mathbf{v}_j\}$ are the eigenvectors of $I\sigma^2 + \mathrm{Var}(\boldsymbol{\nu}_Y)$. In this case $\mathcal{S}_{\mathbf{\Gamma}}$ and $\mathcal{S}_{\mathbf{\Gamma}_0}$ are distinguished by the magnitudes of the corresponding eigenvalues of $\mathbf{\Sigma}$, providing the likelihood information to identify $\mathcal{S}_{\mathbf{\Gamma}}$.

Under model (9)

$$\begin{aligned}\boldsymbol{\Sigma} &= \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^2\boldsymbol{\Gamma}_0^T + \boldsymbol{\Gamma}\{\boldsymbol{\Omega}^2 + \mathrm{Var}(\boldsymbol{\nu}_Y)\}\boldsymbol{\Gamma}^T \\ &= \boldsymbol{\Gamma}_0\mathbf{V}_0\mathbf{D}_0\mathbf{V}_0^T\boldsymbol{\Gamma}_0^T + \boldsymbol{\Gamma}\mathbf{V}\mathbf{D}\mathbf{V}^T\boldsymbol{\Gamma}^T,\end{aligned}$$

where $\mathbf{V}_0\mathbf{D}_0\mathbf{V}_0^T$ and $\mathbf{V}\mathbf{D}\mathbf{V}^T$ are the spectral decompositions of $\boldsymbol{\Omega}_0^2$ and $\boldsymbol{\Omega}^2+\mathrm{Var}(\boldsymbol{\nu}_Y)$. The PC directions under model (9) can be written *unordered* as $\boldsymbol{\Gamma}_0\mathbf{V}_0$ and $\boldsymbol{\Gamma}\mathbf{V}$ with eigenvalues given by the corresponding elements of the diagonal matrices $\mathbf{D}_0$ and $\mathbf{D}$. The estimate of $\mathcal{S}_{\boldsymbol{\Gamma}}$ given in (10) will be consistent if the largest eigenvalue in $\mathbf{D}_0$ is smaller than the smallest eigenvalue in $\mathbf{D}$. In other words, the likelihood-based estimator (10) should be reasonable if the signal as represented by $\mathbf{D}$ is larger than the noise as represented by $\mathbf{D}_0$.

To help fix ideas, consider the following extension of model (7):

$$\mathbf{X}_y = \boldsymbol{\Gamma}y + \sigma_0\boldsymbol{\Gamma}_0\boldsymbol{\varepsilon}_0 + \sigma\boldsymbol{\Gamma}\boldsymbol{\varepsilon}, \tag{11}$$

where $(Y, \mathbf{X})$ is normally distributed. The forward regression $Y|\mathbf{X}$ follows a normal linear regression model where the mean function depends on $\mathbf{X}$ only via $\boldsymbol{\Gamma}^T\mathbf{X}$, and

$$\boldsymbol{\Sigma} = \sigma_0^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T + (\sigma_Y^2 + \sigma^2)\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T.$$

If $\sigma_0^2 < \sigma_Y^2+\sigma^2$ then the first population principal component yields a sufficient reduction $\boldsymbol{\Gamma}^T\mathbf{X}$. If $\sigma_0^2 > \sigma_Y^2 + \sigma^2$, then a sufficient reduction is given by the last population principal component, illustrating the possibility suggested by Cox (1972). However, if $\sigma_0^2 = \sigma_Y^2 + \sigma^2$, then principal components will fail since all of the eigenvalues of $\boldsymbol{\Sigma}$ are equal.

In short, principal components may yield reasonable reductions if the signal dominates the noise in the extended PC model (9).

6.2 An extended PFC model

In this section we consider a PFC model with heterogeneous errors by modeling the ν-parameters in (9):

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\varepsilon}_0 + \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\varepsilon}, \tag{12}$$

where all terms are as defined previously. As mentioned in Section 4, this model can be used to accommodate replication in the extended PC model (9) by using the slice basis function for $\mathbf{f}_y$ to indicate identical y's. Extensions of the exponential family model (3) to allow for dependence among the predictors may depend on the particular family involved. For instance, the quadratic exponential model described by Zhao and Prentice (1990) might be useful when the predictors are correlated binary variables.

Under model (12), $\boldsymbol{\Gamma}^T\mathbf{X}$ is again a sufficient reduction, so we are still interested in estimating the reductive subspace $\mathcal{S}_{\boldsymbol{\Gamma}}$. Maximizing the log likelihood over $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$, we find the following partially maximized log likelihood apart from unimportant constants:

$$
\begin{aligned}
&-(n/2)\log|\mathbf{M}_0| - (n/2)\,\mathrm{trace}[\mathbf{M}_0^{-1}\mathbf{G}_0^T\mathbb{X}^T\mathbb{X}\mathbf{G}_0/n] \\
&-(n/2)\log|\mathbf{M}| - (n/2)\,\mathrm{trace}[\mathbf{M}^{-1}\mathbf{G}^T\{\mathbb{X}^T\mathbb{X} - \mathbb{X}^T P_{\mathbf{F}}\mathbb{X}\}\mathbf{G}/n],
\end{aligned}
$$

where $\mathbf{G}$, $\mathbf{M}$ and $\mathbf{M}_0$ represent possible values for $\boldsymbol{\Gamma}$, $\boldsymbol{\Omega}^2$ and $\boldsymbol{\Omega}_0^2$. After maximizing over $\mathbf{M}$ and $\mathbf{M}_0$, the maximum likelihood estimate of $\mathcal{S}_{\boldsymbol{\Gamma}}$ is found by maximizing

$$
L_{\mathrm{PFC}}(\mathbf{G}) = (-n/2)\log|\mathbf{G}_0^T\widehat{\boldsymbol{\Sigma}}\mathbf{G}_0| - (n/2)\log|\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}\mathbf{G}|, \tag{13}
$$

where $\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}} = \widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}$ is the sample covariance matrix of the residuals from the fit of $\mathbf{X}_y$ on $\mathbf{f}_y$. Like previous likelihoods, (13) is invariant to orthogonal transformations of $\mathbf{G}$ and so it depends only on $\mathcal{S}_{\mathbf{G}}$. But in contrast to the previous likelihoods, here there does not seem to be a recognizable estimate for $\mathcal{S}_{\boldsymbol{\Gamma}}$.

The following proposition gives population versions of matrices in the partially maximized likelihood function (13). It summarizes some of discussion in Section 6.1, and provides results that will be useful for studying (13).

Proposition 4 *Assume the extended PFC model (12) with uncorrelated but not necessarily normal errors,* $\mathrm{Var}((\boldsymbol{\varepsilon}_0^T, \boldsymbol{\varepsilon}^T)^T) = I_p$*. Then*

$$
\begin{aligned}
\widehat{\boldsymbol{\Sigma}} &\xrightarrow{p} \boldsymbol{\Sigma} = \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^2\boldsymbol{\Gamma}_0^T + \boldsymbol{\Gamma}\{\boldsymbol{\Omega}^2 + \boldsymbol{\beta}\mathrm{Var}(\mathbf{f}_Y)\boldsymbol{\beta}^T\}\boldsymbol{\Gamma}^T \\
\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}} &\xrightarrow{p} \boldsymbol{\Sigma}_{\mathrm{fit}} = \boldsymbol{\Gamma}\boldsymbol{\beta}\mathrm{Var}(\mathbf{f}_Y)\boldsymbol{\beta}^T\boldsymbol{\Gamma}^T \\
\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}} &\xrightarrow{p} \boldsymbol{\Sigma}_{\mathrm{res}} = \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^2\boldsymbol{\Gamma}_0^T + \boldsymbol{\Gamma}\boldsymbol{\Omega}^2\boldsymbol{\Gamma}^T
\end{aligned}
$$

To understand the behavior of the function $L_{\mathrm{PFC}}(\mathbf{G})$ (13) in a bit more detail, write it in the form, with $\mathbf{H} = P_{\mathbf{F}}\mathbb{X}\mathbf{G}/\sqrt{n}$,

$$
\begin{aligned}
-2L_{\mathrm{PFC}}(\mathbf{G})/n &= \log|\mathbf{G}_0^T\widehat{\boldsymbol{\Sigma}}\mathbf{G}_0| + \log|\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}\mathbf{G} - \mathbf{G}^T\mathbb{X}^T P_{\mathbf{F}}\mathbb{X}\mathbf{G}/n| \\
&= \log\{|\mathbf{G}_0^T\widehat{\boldsymbol{\Sigma}}\mathbf{G}_0||\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}\mathbf{G}||I_n - \mathbf{H}(\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}\mathbf{G})^{-1}\mathbf{H}^T|\} \\
&= \log\{|\mathbf{G}_0^T\widehat{\boldsymbol{\Sigma}}\mathbf{G}_0||\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}\mathbf{G}||I_n - P_{\mathbf{F}}P_{\mathbb{X}\mathbf{G}}P_{\mathbf{F}}|\}.
\end{aligned}
$$

The first product in the $\log$ is such that

$$
|\mathbf{G}_0^T\widehat{\boldsymbol{\Sigma}}\mathbf{G}_0||\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}\mathbf{G}| \geq |\widehat{\boldsymbol{\Sigma}}|, \tag{14}
$$

and it achieves its lower bound when the columns of $\mathbf{G}$ are *any* d sample PC directions and $\mathbf{G}_0$ is a completion of $\mathbf{G}$. Consequently, the function $-\log\{|\mathbf{G}_0^T\widehat{\boldsymbol{\Sigma}}\mathbf{G}_0||\mathbf{G}^T\widehat{\boldsymbol{\Sigma}}\mathbf{G}|\}$

has at least $\begin{pmatrix} p \\ d \end{pmatrix}$ local maxima of equal height. It is then up to the last term $-\log|I_n - P_{\mathbf{F}}P_{\mathbb{X}\mathbf{G}}P_{\mathbf{F}}|$ to reshape L_{PFC} into a typically multi-modal surface over $\mathcal{G}_{p\times d}$ with a single global maximum. If a $\mathbf{G}$ can be found so that $\mathrm{Span}(\mathbb{X}\mathbf{G}) \subseteq \mathrm{Span}(\mathbf{F})$ then the $|I_n - P_{\mathbf{F}}P_{\mathbb{X}\mathbf{G}}P_{\mathbf{F}}| = 0$ and the log likelihood is infinite at its maximum. If the signal is very weak, or in the extreme $\mathbf{F}^T\mathbb{X} \approx 0$, then $|I_n - P_{\mathbf{F}}P_{\mathbb{X}\mathbf{G}}P_{\mathbf{F}}| \approx 1$ for all $\mathbf{G}$, this term will contribute little to the log likelihood, and we will be left with a surface having perhaps many local maxima of similar heights.

Because the likelihood surface will typically be multi-modal, use of standard gradient optimization methods may be problematic. Consideration of a candidate set of solutions or starting values might mitigate the problems resulting from a multi-modal surface and can further illuminate the role of principal components. Candidate sets can be constructed using the following rationale. Recall from the discussion in Section 6.1 that the unordered PC directions are of the form $(\mathbf{\Gamma}_0\mathbf{V}_0, \mathbf{\Gamma}\mathbf{V})$, and they comprise one possible population candidate set. This structure means that there is a subsets of d PC directions $\boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(d)}$, such that

$$\mathcal{S}_{\mathbf{\Gamma}} = \mathrm{Span}(\mathbf{\Gamma}\mathbf{V}) = \mathrm{Span}\{\boldsymbol{\gamma}_{(1)}, \ldots, \boldsymbol{\gamma}_{(d)}\}.$$

Consequently, provided that the eigenvalues corresponding to $\mathbf{\Gamma}\mathbf{V}$ are distinct from those corresponding to $\mathbf{\Gamma}_0\mathbf{V}_0$, we can construct a consistent estimator of the sufficient reduction $\mathbf{V}^T\mathbf{\Gamma}^T\mathbf{X}$ by evaluating $L_{\mathrm{PFC}}(\mathbf{G})$ at all subsets of d sample PC directions $\hat{\boldsymbol{\gamma}}_{(1)} \ldots, \hat{\boldsymbol{\gamma}}_{(d)}$, and then choosing the subset with the highest likelihood. Call this the $\mathrm{PFC}_{\mathrm{PC}}$ method, because the PCF likelihood is being evaluated at the PC directions. This method might be awkward to implement if the number of combinations to be evaluated is large. As an alternative, we could choose PC directions sequentially:

1. Find $\hat{\boldsymbol{\gamma}}_{(1)} = \arg\max L_{\mathrm{PFC}}(\mathbf{h})$, where the maximum is taken over the $p \times 1$ vector $\mathbf{h}$ in the PC candidate set $A = \{\hat{\boldsymbol{\gamma}}_j, j = 1, \ldots, p\}$.

2. Find $\hat{\boldsymbol{\gamma}}_{(2)} = \arg\max L_{\mathrm{PFC}}(\hat{\boldsymbol{\gamma}}_{(1)}, \mathbf{h})$, where the maximum is now taken over the $p \times 1$ vector $\mathbf{h}$ in the reduced candidate set $A - \{\hat{\boldsymbol{\gamma}}_{(1)}\}$.

3. Continue until reaching the final maximization

$$\hat{\boldsymbol{\gamma}}_{(d)} = \arg\max L_{\mathrm{PFC}}(\hat{\boldsymbol{\gamma}}_{(1)}, \ldots, \hat{\boldsymbol{\gamma}}_{(d-1)}, \mathbf{h})$$

   where the maximum is now taken over $\mathbf{h}$ in $A - \{\hat{\boldsymbol{\gamma}}_{(1)}, \ldots, \hat{\boldsymbol{\gamma}}_{(d-1)}\}$.

This approach is similar in spirit to other methodology for selecting principal components (Jolliffe 2002), but here we base the selection on a likelihood.

Using Proposition 4 and following the same rationale we can construct two additional candidate sets. One consists of the PFC directions, the eigenvectors of $\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}$, and

the other contains the eigenvectors of $\widehat{\mathbf{\Sigma}}_{\rm res}$, which are called the *residual component (RC) directions*. The estimator constructed by evaluating the PFC likelihood (13) at all subsets of the candidate set consisting of the PC, PFC and RC directions will be denoted as $\mathrm{PFC}_{\rm all}$.

Before turning again to illustrative simulations, we connect principal fitted components with sliced inverse regression.

6.3 PFC and SIR

To relate SIR and PFC, $\mathbf{f}_y$ *must* be constructed using the slice basis function (4). Let $\bar{\mathbf{X}}_k = \sum_{i\in H_k}\mathbf{X}_i/n_k$ and $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$. Then it can be shown that $\widehat{\mathbf{\Sigma}}_{\rm fit}$ is the sample covariance matrix of the slice means $\bar{\mathbf{X}}_k$:

$$\widehat{\mathbf{\Sigma}}_{\rm fit} = \sum_{k=1}^{h}(n_k/n)(\bar{\mathbf{X}}_k - \bar{\mathbf{X}})(\bar{\mathbf{X}}_k - \bar{\mathbf{X}})^T = \widehat{\mathbf{\Sigma}}^{1/2}\widehat{\mathbf{\Sigma}}_{\rm sir}\widehat{\mathbf{\Sigma}}^{1/2}$$

where $\widehat{\mathbf{\Sigma}}_{\rm sir} = \widehat{\mathbf{\Sigma}}^{-1/2}\widehat{\mathbf{\Sigma}}_{\rm fit}\widehat{\mathbf{\Sigma}}^{-1/2}$ is the usual SIR kernel matrix in the standardized scale of $\mathbf{Z} = \mathbf{\Sigma}^{-1/2}(\mathbf{X} - \mathrm{E}(\mathbf{X}))$, with sample version $\hat{\mathbf{Z}} = \widehat{\mathbf{\Sigma}}^{-1/2}(\mathbf{X} - \bar{\mathbf{X}})$. Substituting this form into the partially maximized log likelihood (13), we have

$$-2L_{\rm PFC}(\mathbf{G})/n = \log|\mathbf{G}_0^T\widehat{\mathbf{\Sigma}}\mathbf{G}_0| + \log|\mathbf{G}^T\widehat{\mathbf{\Sigma}}^{1/2}\{I_p - \widehat{\mathbf{\Sigma}}_{\rm sir}\}\widehat{\mathbf{\Sigma}}^{1/2}\mathbf{G}|.$$

Suppose now that we redefine $(\mathbf{G}_0, \mathbf{G})$ to be an orthogonal matrix in the $\widehat{\mathbf{\Sigma}}$ inner product: $(\mathbf{G}_0, \mathbf{G})^T\widehat{\mathbf{\Sigma}}(\mathbf{G}_0, \mathbf{G}) = I_p$. Then $\mathbf{G}_0^T\widehat{\mathbf{\Sigma}}\mathbf{G}_0 = I_{p-d}$ and, letting $\mathbf{G}^* = \widehat{\mathbf{\Sigma}}^{1/2}\mathbf{G}$, (13) can be written as a function of $\mathbf{G}^*$ with $\mathbf{G}^{*T}\mathbf{G}^* = I_d$:

$$L_{\rm PFC}(\mathbf{G}^*) = -(n/2)\log|\mathbf{G}^{*T}\{I_p - \widehat{\mathbf{\Sigma}}_{\rm sir}\}\mathbf{G}^*|.$$

This objective function results in the standard SIR estimates since it is maximized by setting the columns of $\mathbf{G}^*$ to be the matrix $\widehat{\mathbf{G}}^*$ whose columns are the first d eigenvectors of $\widehat{\mathbf{\Sigma}}_{\rm sir}$. This solution is then back-transformed to the original $\mathbf{X}$ scale so that $\widehat{\mathcal{S}}_{\mathbf{\Gamma}} = \widehat{\mathbf{\Sigma}}^{-1/2}\mathrm{Span}(\widehat{\mathbf{G}}^*)$. From this we see that, starting from the normal likelihood for (12), the SIR solution is found by using a sample-based inner product. Relative to the likelihood, this can have a costly effect of neglecting the information in $\widehat{\mathbf{\Sigma}}$ when estimating the sufficient reduction.

While SIR does not require normality or a particular structure for $\mathrm{Var}(\mathbf{X}_y)$, it is know that its operating characteristics can vary widely depending on these features. Bura and Cook (2001) argue that generally SIR performs the best under normality and that its performance can degrade when $\mathrm{Var}(\mathbf{X}_y)$ is not constant. Cook and Ni (2005)

show that SIR can be very inefficient when $\mathrm{Var}(\mathbf{X}_y)$ is not constant and they provide a new method called inverse regression estimation (IRE) that can dominate SIR in application. From these and other articles, we would expect SIR to be at its best under the models considered here, which all have both normality and constant $\mathrm{Var}(\mathbf{X}_y)$.

Like SIR, the fundamental population characteristics of the PC and PFC methods considered here do not hinge on normality. From (13) and Proposition 4, the normalized partially maximized log likelihood $L_{\mathrm{PFC}}(\mathbf{G})/n$ converges in probability to

$$\widetilde{L}_{\mathrm{PFC}}(\mathbf{G}) = -(1/2)\log|\mathbf{G}_0^T\boldsymbol{\Sigma}\mathbf{G}_0| - (1/2)\log|\mathbf{G}^T\boldsymbol{\Sigma}_{\mathrm{res}}\mathbf{G}|.$$

We then have

Proposition 5 *Assume the conditions of Proposition 4. Then* $\boldsymbol{\Gamma} = \arg\max_{\mathbf{G}} \widetilde{L}_{\mathrm{PFC}}(\mathbf{G})$.

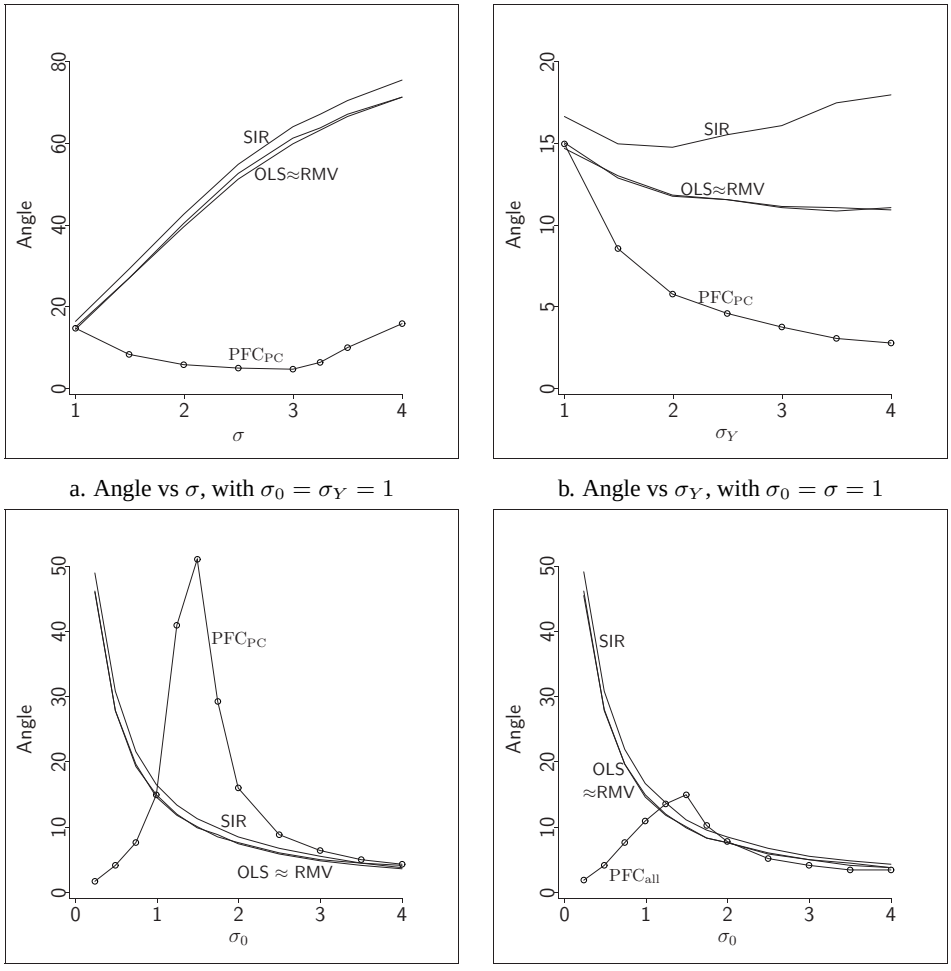
This proposition says that the likelihood objective function arising from the extended PFC model (12) produces Fisher consistent estimates when the errors are uncorrelated, but not necessarily normal, suggesting that normality per se is not crucial for the type of analysis suggested in this article.

6.4 Illustration via simulation

A simulation study was conducted to illustrate some of the results to this point. We generated Y as a $N(0, \sigma_Y^2)$ random variable, and then generated $\mathbf{X}_y$ according to model (11) with $p = 10$, sample size $n = 250$ and $\boldsymbol{\Gamma} = (1, 0, \ldots, 0)^T$. In reference to model (12), the data were generated with $\boldsymbol{\Omega}_0 = \sigma_0 I_{p-1}$ and $\boldsymbol{\Omega} = \sigma$. Four estimators were applied to each dataset: OLS, SIR with 8 slices, and $\mathrm{PFC}_{\mathrm{PC}}$ based on the likelihood for the extended PFC model (12) using the slicing construction (4) with $r = 8$ slices for $\mathbf{f}_y$. To expand the comparisons, we also included a recent semiparametric estimator RMAVE (Xia, et al. 2002), which is based on local linear smoothing of the forward regression with adaptive weights and, like SIR, is expected to do well in the context of this simulation. The results were summarized by computing the angle between each of the four estimates and $\mathcal{S}_{\boldsymbol{\Gamma}}$.

The angles plotted in Figure 2 are averages over 500 replications. Figure 2a is a plot of the average angle versus the error standard deviation σ for the signal, with $\sigma_Y = \sigma_0 = 1$. Clearly, the likelihood-based estimator $\mathrm{PFC}_{\mathrm{PC}}$ dominates SIR, OLS and RMAVE, except when σ is close to 1, so the three variances in the simulation are close. The RMAVE estimator is indistinguishable from OLS in all of the simulations of Figure 2. In Figure 2b σ_Y was varied while holding $\sigma = \sigma_0 = 1$. Again we see that PC dominates over most of the plot.

The error standard deviation σ_0 for the inactive predictors was varied for the construction of Figure 2c. Here the results are of a fundamentally different character.



a. Angle vs σ, with $\sigma_0 = \sigma_Y = 1$

b. Angle vs σ_Y, with $\sigma_0 = \sigma = 1$

c. Angle vs σ_0, $\sigma = \sigma_Y = 1$

d. Angle vs σ_0, $\sigma = \sigma_Y = 1$

Figure 2: Simulation results based on model (11). In Figures a-c the likelihood-based estimator was computed from the PC candidate set. In Figure d the estimator was computed from the full candidate set containing the PC, PFC and RC directions. RMV is short for RMAVE.

$\mathrm{PFC}_{\mathrm{PC}}$ performed the best for $\sigma_0 < 1$, and the three methods are roughly equivalent for the larger values of σ_0. But in the middle region, $\mathrm{PFC}_{\mathrm{PC}}$ performed the worst. This poor performance arises because the estimate was computed using the PC candidate set, and when $\sigma_0 = \sqrt{2}$ the population eigenvalues of $\mathbf{\Sigma} = 2I_p$ are equal. When $\mathbf{\Sigma}$ is spherical, the principal components are arbitrary and cannot be expected to convey any useful information.

Figure 2d was constructed as Figure 2c, except the estimator $\mathrm{PFC}_{\mathrm{PC}}$ was replaced by the estimator $\mathrm{PFC}_{\mathrm{all}}$ based on the candidate set of PC, PFC and RC directions. The $\mathrm{PFC}_{\mathrm{all}}$ estimate of $\mathcal{S}_{\mathbf{\Gamma}}$ always performed as well as or better than SIR, OLS and RMAVE, except in a neighborhood around $\sigma_0 = \sqrt{2}$. The vector that maximized the likelihood always came from the PC candidate set for $\sigma_0 \leq 0.75$, from the residual candidate set for $1 \leq \sigma_0 \leq 1.25$, and from the PFC candidate set for $\sigma_0 \geq 2$. At $\sigma_0 = 1.5$ this vector came about equally from the residual and PFC candidate sets. These results may provide some intuition into operational characteristics of the likelihood as reflected through the three matrices given in Proposition 4. For this simulation example those matrices are

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{\Gamma}_0\mathbf{\Gamma}_0^T\sigma_0^2 + \mathbf{\Gamma}\mathbf{\Gamma}^T(\sigma^2 + \sigma_Y^2) \\ \mathbf{\Sigma}_{\mathrm{fit}} &= \mathbf{\Gamma}\mathbf{\Gamma}^T\sigma_Y^2 \\ \mathbf{\Sigma}_{\mathrm{res}} &= \mathbf{\Gamma}_0\mathbf{\Gamma}_0^T\sigma_0^2 + \mathbf{\Gamma}\mathbf{\Gamma}^T\sigma^2. \end{aligned}$$

When σ_0 is small the first sample PC direction evidently provides the best estimate of $\mathcal{S}_{\mathbf{\Gamma}}$. When σ_0 is large, we can gain information on $\mathcal{S}_{\mathbf{\Gamma}}$ from the smallest PC or the largest PFC. Evidently, the error variation that comes with σ_0 causes the smallest PC to be less reliable than the largest PFC. When $\sigma_0^2 = \sigma^2 + \sigma_Y^2$, the PC directions provide no information on $\mathcal{S}_{\mathbf{\Gamma}}$, but the PFC and residual directions can both provide information. These results suggest that restricting attention to the eigenstructure of any single matrix, as SIR and standard PC methodology do, may result in nontrivial loss of information in some regressions, and that worthwhile gains are possible by pursuing the maximum likelihood estimator of $\mathcal{S}_{\mathbf{\Gamma}}$.

To gain insights about the potential advantages of pursuing the maximum likelihood estimator, we used a gradient algorithm for Grassman manifolds (Edelman, et al. 1998) to find a local maximum of the likelihood, starting with the best direction from the full candidate set. The local likelihood solution resulted in improvements all along the "$\mathrm{PFC}_{\mathrm{all}}$" curve, with the greatest improvement in a neighborhood around $\sigma_0 = 1.5$. For instance, at $\sigma_0 = 1.5$, the average angle for the local likelihood solution was about 11 degrees, which is quite close to the average angles for SIR and OLS, while the average angle shown in Figure 2d for the full candidate set is about 15 degrees. On balance, the local likelihood gave solutions that were never worse and were sometimes much better than those of the three competing methods.

SIR and RMAVE have been seen as robust nonparametric methods of dimension reduction for regression. Normally some loss would be expected relative to likelihood-based methods when the model holds. But the magnitude of the loss in this simulation, particularly in Figures 1a and 1b, is surprising. Some authors (see, for example, L. Li and H. Li 2004, and Chiaromonte and Martinelli 2002) have used principal components to reduce the dimension of the predictor vector prior to a second round of reduction using SIR. The results of this simulation raise questions regarding such methodology generally. If we are in a situation like Figure 1a or 1b where principal components do well, the motivation for switching to SIR seems unclear. On the other hand, if we are in a situation like Figure 1c with $\sigma_Y \approx \sqrt{2}$ then there seems little justification for using principal components in the first place.

7 PC and PFC Models with Unstructured Errors

Suppose that $\mathbf{X}_y$ follows the general PC model

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y + \sigma\boldsymbol{\Delta}^{1/2}\boldsymbol{\varepsilon}, \tag{15}$$

where the parameters and the error $\boldsymbol{\varepsilon}$ have the same structure as in model (2) and the conditional covariance matrix $\mathrm{Var}(\mathbf{X}_y) = \sigma^2\boldsymbol{\Delta} > 0$. Then we have the following:

Proposition 6 *Under model (15), the distribution of $Y|\mathbf{X}$ is the same as the distribution of $Y|\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$ for all values of $\mathbf{X}$.*

We first consider implications of this proposition when $\boldsymbol{\Delta}$ is known, and then turn to case in which $\mathrm{Var}(\mathbf{X}_y)$ is unknown.

7.1 $\boldsymbol{\Delta}$ known

The essential variance condition in the PC model (2) and the PFC model (5) is that $\mathrm{Var}(\mathbf{X}_y) = \sigma^2\boldsymbol{\Delta}$, where $\boldsymbol{\Delta}$ is known but not necessarily the identity. According to Proposition 6 a sufficient reduction for (15) is $\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$. Letting $\mathbf{Z}_y = \boldsymbol{\Delta}^{-1/2}\mathbf{X}_y$, we have

$$\begin{aligned}\mathbf{Z}_y &= \boldsymbol{\Delta}^{-1/2}\boldsymbol{\mu} + \boldsymbol{\Delta}^{-1/2}\boldsymbol{\Gamma}\boldsymbol{\nu}_y + \sigma\boldsymbol{\varepsilon} \\ &= \boldsymbol{\mu}^* + \boldsymbol{\Gamma}^*\boldsymbol{\nu}_y^* + \sigma\boldsymbol{\varepsilon},\end{aligned}$$

where the columns of $\boldsymbol{\Gamma}^*$ are an orthonormal basis for $\mathrm{Span}(\boldsymbol{\Delta}^{-1/2}\boldsymbol{\Gamma})$ and $\boldsymbol{\nu}_y^*$ is the corresponding coordinate function. It follows that $\mathcal{S}_{\boldsymbol{\Gamma}} = \boldsymbol{\Delta}^{1/2}\mathcal{S}_{\boldsymbol{\Gamma}^*}$ and thus that the coordinates of all sufficient reduction are in $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}} = \boldsymbol{\Delta}^{-1/2}\mathcal{S}_{\boldsymbol{\Gamma}^*}$. In short, the required reductive subspace $\mathcal{S}_{\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}}$ is estimated by the span of $\boldsymbol{\Delta}^{-1/2}$ times the first d

eigenvectors of $\boldsymbol{\Delta}^{-1/2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}^{-1/2}$. An implication of this result is that principal components computed in the standardized $\mathbf{Z}$ scale are appropriate reductions for both $\mathbf{X}$ and $\mathbf{Z}$, because

$$\boldsymbol{\Gamma}^{*T}\mathbf{Z} = (\boldsymbol{\Gamma}^{*T}\boldsymbol{\Delta}^{-1/2})(\boldsymbol{\Delta}^{1/2}\mathbf{Z}) = (\boldsymbol{\Delta}^{-1/2}\boldsymbol{\Gamma}^{*})^{T}\mathbf{X}.$$

Turning to the general PFC model

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y + \sigma\boldsymbol{\Delta}^{1/2}\boldsymbol{\varepsilon}, \tag{16}$$

and following the discussion of model (15), $\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$ is again a sufficient reduction. The maximum likelihood estimate of the reductive subspace is the span of $\boldsymbol{\Delta}^{-1/2}$ times the first d eigenvalues of $\boldsymbol{\Delta}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\text{fit}}\boldsymbol{\Delta}^{-1/2}$.

7.2 $\mathrm{Var}(\mathbf{X}_y)$ unknown

We now turn to the PFC model (16) with $\mathrm{Var}(\mathbf{X}_y)$ unknown. For notational convenience, redefine $\boldsymbol{\Delta} = \mathrm{Var}(\mathbf{X}_y)$, absorbing the scale parameter σ^2 into the definition of $\boldsymbol{\Delta}$. Maximizing the log likelihood over $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$, the resulting partially maximized log likelihood

$$(-n/2)\log|\mathbf{D}| - (n/2)\,\mathrm{trace}[\mathbf{D}^{-1/2}\widehat{\boldsymbol{\Sigma}}\mathbf{D}^{-1/2} - P_{\mathbf{D}^{-1/2}\mathbf{G}}\mathbf{D}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\text{fit}}\mathbf{D}^{-1/2}] \tag{17}$$

is a function of possible values $\mathbf{D}$ and $\mathbf{G}$ for $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$. For fixed $\mathbf{D}$ this function is maximized by choosing $\mathbf{D}^{-1/2}\mathbf{G}$ to be a basis for the span the first d eigenvectors of $\mathbf{D}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\text{fit}}\mathbf{D}^{-1/2}$, yielding another partially maximized log likelihood $K(\mathbf{D})$:

$$\begin{aligned}
K(\mathbf{D})/n =& (-1/2)\log|\mathbf{D}| - (1/2)\{\,\mathrm{trace}[\mathbf{D}^{-1/2}\widehat{\boldsymbol{\Sigma}}\mathbf{D}^{-1/2}] - \sum_{i=1}^{d}\lambda_i(\mathbf{D}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\text{fit}}\mathbf{D}^{-1/2})\}\\
=& (-1/2)\log|\mathbf{D}| - (1/2)\{\,\mathrm{trace}[\mathbf{D}^{-1/2}(\widehat{\boldsymbol{\Sigma}}_{\text{fit}} + \widehat{\boldsymbol{\Sigma}}_{\text{res}})\mathbf{D}^{-1/2}] - \sum_{i=1}^{d}\lambda_i(\mathbf{D}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\text{fit}}\mathbf{D}^{-1/2})\}\\
=& (-1/2)\log|\mathbf{D}| - (1/2)\{\,\mathrm{trace}[\mathbf{D}^{-1}\widehat{\boldsymbol{\Sigma}}_{\text{res}}] + \sum_{i=d+1}^{p}\lambda_i(\mathbf{D}^{-1}\widehat{\boldsymbol{\Sigma}}_{\text{fit}})\},
\end{aligned}$$

where $\lambda_i(\mathbf{A})$ indicates the i-th eigenvalue of $\mathbf{A}$. The first two terms alone

$$(-1/2)\log|\mathbf{D}| - (1/2)\,\mathrm{trace}[\mathbf{D}^{-1}\widehat{\boldsymbol{\Sigma}}_{\text{res}}]$$

are maximized by $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Sigma}}_{\text{res}}$ and this is a consistent estimator of $\mathbf{D}$. The final term

$$-(1/2)\sum_{i=d+1}^{p}\lambda_i(\mathbf{D}^{-1}\widehat{\boldsymbol{\Sigma}}_{\text{fit}})$$

reflects that fact that we may have $r > d$ and thus that there is an error component in $\widehat{\mathbf{\Sigma}}_{\rm fit}$ due to overfitting. Since it is assumed that $\mathrm{rank}(\widehat{\mathbf{\Sigma}}_{\rm fit}) \geq d$, this term will not be present if $r = d$, and use of $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Sigma}}_{\rm res}$ should be reasonable if r isn't much larger than d. However, the final term may be important if r is substantially larger than d.

Once $\widehat{\mathbf{\Delta}}$ is determined, the estimate of the reductive subspace is the span of $\widehat{\mathbf{\Delta}}^{-1/2}$ times the first d eigenvectors of $\widehat{\mathbf{\Delta}}^{-1/2}\widehat{\mathbf{\Sigma}}_{\rm fit}\widehat{\mathbf{\Delta}}^{-1/2}$. This is the same as the estimator in the case where $\mathbf{\Delta}$ is known, except $\widehat{\mathbf{\Delta}}$ is substituted for $\mathbf{\Delta}$.

7.3 Standardization

Reduction by principal components is often based on the marginal correlation matrix of the predictors rather than on $\widehat{\mathbf{\Sigma}}$ (see, for example, Jolliffe, 2002, p. 169; L. Li and H. Li, 2004). In this section we provide an intuitive explanation at the population level of why this practice is not supported by any of the results in this article.

The contours of the conditional covariance matrix are spherical in the first PC model (2), $\mathrm{Var}(\mathbf{X}_y) = \sigma^2 I_p$, and the contours of $\mathbf{\Sigma} = \sigma^2 I_p + \mathbf{\Gamma}\mathrm{Var}(\boldsymbol{\nu}_Y)\mathbf{\Gamma}^T$ are elliptical. Generally, the method of maximum likelihood treats $\mathrm{Var}(\mathbf{X}_y)$ as a reference point, with the impact of the response being embodied in the eigenvectors of $\mathbf{\Sigma}$ relative to $\mathrm{Var}(\mathbf{X}_y)$. In the context of model (2), these eigenvectors are the same as the eigenvectors of $\mathbf{\Sigma}$. When passing from $\mathrm{Var}(\mathbf{X}_y) = \sigma^2 I_p$ to $\mathbf{\Sigma}$, the response distorts the conditional variance by "pulling" its spherical contours parallel to the reductive subspace, which is then spanned by the first d principal components. The same ideas work for the first PFC model (5), except more information is supplied to the mean function $\mathrm{E}(\mathbf{X}_y)$ with the consequence that the marginal covariance matrix of the predictors $\mathbf{\Sigma}$ is replaced by the covariance matrix of the fitted values $\mathbf{\Sigma}_{\rm fit}$.

Likelihood estimation in all of the other normal models considered in this article can be interpreted similarly, but the calculations become more involved because the reference point $\mathrm{Var}(\mathbf{X}_y)$ is no longer spherical. Consider first the general PC model (15) with $\mathbf{\Delta}$ known. The reference point $\mathrm{Var}(\mathbf{X}_y) = \sigma^2\mathbf{\Delta}$ no longer has spherical contours, so the eigenstructure of $\mathbf{\Sigma}$ is not sufficient to find the reductive subspace. To find the eigenvectors of $\mathbf{\Sigma}$ relative to $\mathrm{Var}(\mathbf{X}_y)$, first construct the standardized predictors $\mathbf{Z} = \mathbf{\Delta}^{-1/2}\mathbf{X}$. The reductive subspace in the $\mathbf{Z}$-scale is then the span of the first d eigenvectors of $\mathrm{Var}(\mathbf{Z}) = \mathbf{\Delta}^{-1/2}\mathbf{\Sigma}\mathbf{\Delta}^{-1/2}$, because the contours of $\mathrm{Var}(\mathbf{Z}_y) = \sigma^2 I_p$ are spherical. The final step is to return to the $\mathbf{X}$-scale by multiplying these eigenvectors by $\mathbf{\Delta}^{-1/2}$. The general PFC model (16) follows the same pattern, as may be seen from the discussion at the end of Section 7.2.

Reductive subspaces and their estimates considered here are scale dependent. If we change the units of a predictor or more generally perform a full rank linear transformation of $\mathbf{X}$, the reductive subspace will change. In this regard, reductive subspaces

behave like coefficient vectors in linear models. On the other hand, sufficient reductions – $\boldsymbol{\Gamma}^T\mathbf{X}$ or $\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\mathbf{X}$ – behave like fitted values in linear models and are invariant under full-rank linear transformations of $\mathbf{X}$.

It may now be clear why the approach in this article provides no support for the common practice of scaling the predictors so that the covariance matrix of the new predictors $\mathbf{W} = \mathrm{diag}(\boldsymbol{\Sigma})^{-1/2}\mathbf{X}$ is a correlation matrix. Basing reduction on the eigenvalues of $\mathrm{Var}(\mathbf{W})$ requires that $\mathrm{Var}(\mathbf{W}_y)$ be spherical, but this requirement will not generally be met. As a consequence, the eigenvalues of $\mathrm{Var}(\mathbf{W})$ may have little useful relation to the reductive subspace.

7.4 Connection with OLS

Suppose that we adopt model (16) with $\mathbf{f}_y = (y - \bar{y})$, so $d = r = 1$, $\boldsymbol{\beta}$ is a scalar, and $\boldsymbol{\Gamma} \in \mathbb{R}^p$. Letting $\widehat{\mathbf{C}} = \widehat{\mathrm{Cov}}(\mathbf{X}, Y)$,

$$\widehat{\boldsymbol{\Sigma}}_{\text{fit}} = \frac{\widehat{\mathbf{C}}\widehat{\mathbf{C}}^T}{\hat{\sigma}_Y^2},$$

which is of rank 1. Consequently, it follows from the previous section that

$$\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Sigma}}_{\text{res}} = \widehat{\boldsymbol{\Sigma}} - \frac{\widehat{\mathbf{C}}\widehat{\mathbf{C}}^T}{\hat{\sigma}_Y^2}$$

and that

$$\widehat{\boldsymbol{\Delta}}^{-1} = \widehat{\boldsymbol{\Sigma}}^{-1} + \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{C}}[1 - \widehat{\mathbf{C}}^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{C}}/\hat{\sigma}_Y^2]^{-1}\widehat{\mathbf{C}}^T\widehat{\boldsymbol{\Sigma}}^{-1}/\hat{\sigma}_Y^2.$$

The estimate of $\widehat{\mathcal{S}}_{\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}}$ is $\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}$ times the eigenvector corresponding to the largest eigenvalue of $\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\text{fit}}\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}$. But $\widehat{\boldsymbol{\Sigma}}_{\text{fit}}$ has rank 1, and consequently the first (non-normalized) eigenvector is $\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}\widehat{\mathbf{C}}$. Multiplying this by $\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1/2}$ we have that $\widehat{\mathcal{S}}_{\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}} = \mathrm{Span}(\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1}\widehat{\mathbf{C}})$. Next,

$$\widehat{\boldsymbol{\Sigma}}_{\text{res}}^{-1}\widehat{\mathbf{C}} = \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{C}} \times \text{constant}.$$

And it follows that $\widehat{\mathcal{S}}_{\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}} = \mathrm{Span}(\hat{\boldsymbol{\beta}})$, so the inverse method yields OLS under model (16) with $\mathbf{f}_y = (y - \bar{y})$. The constant above is $\widehat{\mathbf{C}}^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{C}}/(\hat{\sigma}_Y^2 - \widehat{\mathbf{C}}^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{C}})$.

This connection with OLS require that $d = r = 1$ and that $\mathbf{f}_y = (y - \bar{y})$. Estimators based on $r \geq d > 1$ may thus be regarded as a generalization of OLS.

7.5 Connection with SIR

As in Section 6.3, we must use the slice basis function (4) to relate the SIR estimator of the reductive subspace $\mathcal{S}_{\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}}$ to the estimator obtained from model (16) using $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Sigma}}_{\rm res}$ as the estimator of $\boldsymbol{\Delta}$. Letting $\hat{\ell}_i$ denote an eigenvector of the normalized SIR matrix (cf. section 6.3)

$$\widehat{\boldsymbol{\Sigma}}_{\rm sir} = \widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{\rm fit}\widehat{\boldsymbol{\Sigma}}^{-1/2},$$

$\widehat{\boldsymbol{\Sigma}}^{-1/2}\hat{\ell}_i$ is an non-normalized eigenvector of $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\rm fit}$ with the same eigenvalues as $\widehat{\boldsymbol{\Sigma}}_{\rm sir}$. The subspace spanned by the first d eigenvectors of $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\rm fit}$ give the SIR estimate of $\mathcal{S}_{\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}}$. Similarly, the subspace spanned by the first d eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{\rm res}^{-1}\widehat{\boldsymbol{\Sigma}}_{\rm fit}$ is the estimate of the reductive subspace subspace from model (16), still using $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Sigma}}_{\rm res}$. These estimators are identical provided $\widehat{\boldsymbol{\Sigma}}_{\rm res} > 0$, because then the eigenvectors of $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\rm fit}$ and $\widehat{\boldsymbol{\Sigma}}_{\rm res}^{-1}\widehat{\boldsymbol{\Sigma}}_{\rm fit}$ are identical, with corresponding eigenvalues λ_i and $\lambda_i/(1-\lambda_i)$ (See Appendix G).

7.6 Simulations with unstructured errors

To help fix ideas and provided results to direct further discussion, consider data simulated from the model

$$\mathbf{X}_y = \boldsymbol{\Gamma} y + \boldsymbol{\Delta}^{1/2}\boldsymbol{\varepsilon} \tag{18}$$

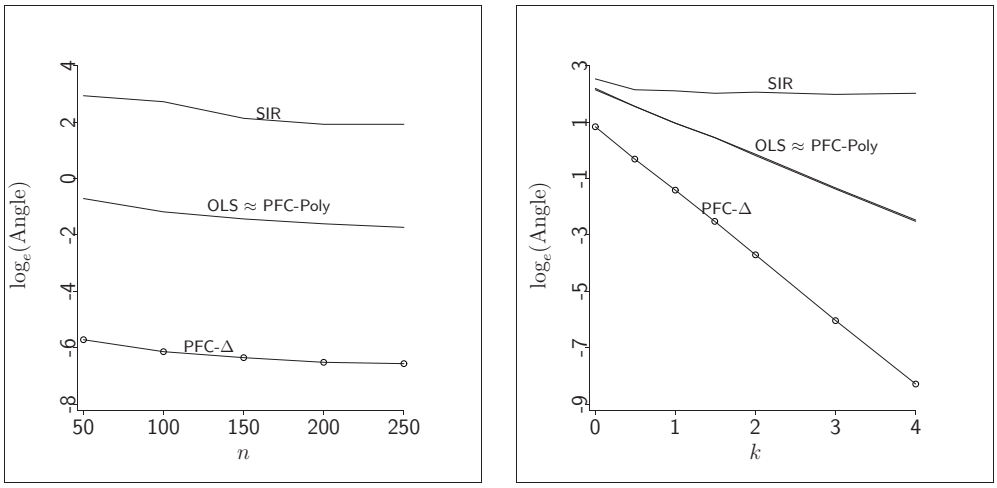
where $p = 10$, Y is a normal random variable with mean 0 and standard deviation $\sigma_Y = 15$, $\boldsymbol{\Gamma} = (1, \ldots, 1)^T/\sqrt{10}$, and $\boldsymbol{\Delta}$ was generated once as $\boldsymbol{\Delta} = \mathbf{A}^T\mathbf{A}$, where $\mathbf{A}$ is a $p \times p$ matrix of independent standard normal random variables, yielding predictor variances of about 10 and correlations ranging between 0.75 and $-.67$. Four estimators of the sufficient reduction subspace $\mathcal{S}_{\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}}$ were computed for each data set generated in this way:

1. The OLS estimator. This is the same as the PFC estimator with $\mathbf{f}_y = (y - \bar{y})$ (cf. Section 7.4).

2. The PFC estimator with a third degree polynomial in y for $\mathbf{f}_y$, designated PFC-Poly in later plots.

3. The SIR estimator with 8 slices. This is the same as the PFC estimator with the slicing option for $\mathbf{f}_y$ and $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Sigma}}_{\rm res}$ (cf. Section 7.5).

4. The PFC estimator using the slicing construction for $\mathbf{f}_y$ with 8 slices and the true $\boldsymbol{\Delta}$, designated PFC-Δ in later plots (cf. Section 7.2).



a. $\log_e(\text{Angle})$ versus n b. $\log_e(\text{Angle})$ versus k

Figure 3: Natural logarithm of the average angle versus (a) sample size and (b) k for four estimators based on simulation model (18) with different covariance matrices $\mathbf{\Delta}$.

Shown in Figure 3a are the natural logarithms of the average angles from 500 replications of each sampling configuration. The logarithms were necessary since the angles varied over several orders of magnitude. The OLS and PFC-Poly estimators were essentially indistinguishable and are represented by a single curve in Figure 3a. These estimators seemed to do quite well, with the average angle varying between 0.51 for $n = 50$ and 0.18 for $n = 250$. The performance of the PFC estimator using the true $\mathbf{\Delta}$ was exceptional, the average angle varying between 0.0034 and 0.0014. SIR performed the worst, its average angle varying between 19.2 and 8 degrees.

PFC did exceptionally well when using the true $\mathbf{\Delta}$, while SIR's performance was considerably worse. The reason for this difference seems to be that SIR's slicing estimate of $\mathbf{\Delta}$ is biased. In the context of this example, $\mathbf{\Delta} = \mathbf{\Sigma} - \mathbf{\Gamma}\mathbf{\Gamma}^T\sigma_Y^2$. But when using the slicing construction for $\mathbf{f}_y$, $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Sigma}}_{\text{res}}$ is a consistent estimator of

$$\mathbf{\Delta}_{\text{sir}} = \mathbf{\Sigma} - \text{Var}(\text{E}(\mathbf{X}|Y \in H_k)) = \mathbf{\Sigma} - \mathbf{\Gamma}\mathbf{\Gamma}^T\text{Var}(\text{E}(Y|Y \in H_k))$$

where H_k indicates slice k, as defined in Section 4. The difference between these two population covariance matrices is

$$\begin{aligned} \mathbf{\Delta} - \mathbf{\Delta}_{\text{sir}} &= \mathbf{\Gamma}\mathbf{\Gamma}^T(\sigma_Y^2 - \text{Var}(\text{E}(Y|Y \in H_k))) \\ &= \mathbf{\Gamma}\mathbf{\Gamma}^T\text{E}(\text{Var}(Y|Y \in H_k)) \end{aligned}$$

which is nonzero for continuous responses. One consequence of this bias is that SIR may not be able to find an exact or near exact fit. An exact fit occurs in the context of model (16) if $\boldsymbol{\Gamma}^T\boldsymbol{\Delta} = 0$, so $\boldsymbol{\Gamma}^T\mathbf{X}$ is a deterministic function of y. To illustrate this phenomenon we generated data from simulation model (18) with

$$\boldsymbol{\Delta} = (cI - \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)\sigma_Y^2$$

and $c > 1$. Here $\boldsymbol{\Sigma} = c\sigma_Y^2 I$ and an exact fit occurs if $c = 1$. Values of $c < 1$ are not allowed since then $\boldsymbol{\Delta}$ will not be positive definite. With $c = 1 + .1/10^k$, Figure 3b shows the natural logarithm of the average simulation angle versus k for the four estimators used in Figure 3a with $n = 50$. Clearly, SIR's response to increasing k is negligible. At $k = 4$ the average angles for SIR, OLS and PFC-Δ were about 7.7, 0.085 and 0.0002.

A general conclusion here is that substantial gains can be realized from careful selection of $\mathbf{f}_y$, and that the slicing construction for $\mathbf{f}_y$ may impose inherent limitations on the analysis under model (16). The same limitation do no occur for any of the other inverse models discussed in this article, because for them $\mathcal{S}_{\boldsymbol{\Gamma}}$ is the reductive subspace, which does not require a direct estimate of $\boldsymbol{\Delta}$.

8 Discussion

Fisher believed that if a statistic is ancillary then inferences should be made from the conditional distribution of the data given that statistic. As consequence of this logic, many of us have been taught and still practice what has become effectively a first principle of regression: Inference should be conditioned on the observed values of the predictors, even if Y and $\mathbf{X}$ are both random. It may be, however, that this working principle has forced a myopic view of regression methodology. The inverse models studied in the previous sections describe the conditional distribution of $\mathbf{X}$ given Y and thereby make explicit use of the randomness in $\mathbf{X}$. With these inverse models, we were able to achieve results that are superior to those from standard methods, and to those from a recent dimension reduction methods. The success of these inverse models depends in part on imposing an appropriately restrictive structure on the conditional variances $\mathrm{Var}(\mathbf{X}_y)$. They were presented, not as models for shotgun application, but as illustrations of potential, recalling Fisher's view that model specification is a matter for the practical statistician. In contrast, if we condition on the observed values of the predictors, they become known constants and the possibility of utilizing knowledge of $\mathrm{Var}(\mathbf{X}_y)$ is lost to us.

The reductive subspace provides a connection between forward and inverse regressions. Starting with the PC model (2) and passing through a series of extensions, the last two PFC models considered were the extended PFC model (12), $\mathbf{X}_y =$

$\boldsymbol{\mu}+\boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y+\boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\varepsilon}_0+\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\varepsilon}$, and the general PFC model (16), $\mathbf{X}_y = \boldsymbol{\mu}+\boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y+\boldsymbol{\Delta}^{1/2}\boldsymbol{\varepsilon}$, in which $\boldsymbol{\Delta}$ is unstructured and unknown. This final model connects inverse and forward regression methodology, since it is here that certain forward and inverse estimates of $\mathcal{S}_{\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma}}$ are the same. While the error structure in model (12) is restrictive it may be useful in some applications. Perhaps more importantly, we should be aware of the possibility to develop models "between" (12) and (16) that allow us to infer simultaneously about the reductive subspace and about the relevant structure of $\mathrm{Var}(\mathbf{X}_y)$.

The rationale for employing principal components in regression has been rather uneven. Tied closely to the presence of collinearity, reduction by principal components has been seen as a way to compensate for variance inflation in the estimates of the regression coefficients. However, collinearity played no notable role in the developments of this article, suggesting that the utility of principal component reduction is broader than previously seen.

Dimension reduction seems particularly important in regressions where "$n < p$". Many available methods encounter problems at an operational level because of the need to compute $\widehat{\boldsymbol{\Sigma}}^{-1}$. However, with the exception of Section 7, the methods described in this article do not require the computation of an inverse, and may therefore have value in regressions where n is not large relative to p. First simulation results sustain this conjecture, particularly when the methods are used in conjunction with prior model-based predictor selection.

In this article the dimension d of the reductive subspace was assumed known. While this limited the scope of the discussion, there does not seems to be any particular road block to the development of methods for inference on d. Likelihood methods are a natural first choice. Another possibility is to adapt the known methods that have been developed in the context of model-free dimension reduction discussed in Section 8.2.

8.1 Model-based sufficient dimension reduction

At the outset of his book *Statistical Methods for Research Workers*, Fisher (1941, p. 1) offered the following definition:

> Statistics may be regarded as (i.) as the study of **populations**, (ii.) as the study of **variation**, (iii.) as the study of methods of the **reduction of data**.

In this statement and in his commentary that follows, Fisher seems to suggest that "reduction of data" may encompass more than just sufficient statistics; for instance, efficient statistics may be adequate for such purposes. For this reason I imagine that Fisher would not have objected to the notion of a sufficient reduction as defined in this article.

Nearly all of the sufficient reductions described in the previous sections are special cases of a general reductive paradigm that emerges from the following definition: A

function $R : \mathbb{R}^p \to \mathbb{R}^q$, $q \leq p$, is a *sufficient reduction* if at least one of the following three statements holds:

Reductive forms

(i) $\mathbf{X}|(Y, R(\mathbf{X})) \sim \mathbf{X}|R(\mathbf{X})$

(ii) $Y|\mathbf{X} \sim Y|R(\mathbf{X})$

(iii) $Y \perp\!\!\!\perp \mathbf{X}|R(\mathbf{X})$

Statement (i) corresponds to inverse regression and requires only the conditional distribution of $\mathbf{X}|Y$. For instance, if Y is a categorical response indicating one of two populations and $\mathbf{X}|Y$ is normal with mean $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Delta}$, then Fisher's linear discriminant function is a sufficient reduction (Rao 1962). Statement (ii) corresponds to forward regression and requires only the conditional distribution of $Y|\mathbf{X}$, while statement (iii) requires the joint distribution of $(Y, \mathbf{X})$. The key point for present purposes is that these three forms are equivalent if Y and $\mathbf{X}$ are both random. For example, we may determine a sufficient reduction from $\mathbf{X}|Y$ (i) and then pass that reduction to the forward regression (ii) or the joint distribution (iii) without specifying the marginal distribution of Y or the conditional distribution of $Y|\mathbf{X}$. This is how all sufficient reductions were determined in the previous sections, the methods of derivation being essentially the same as the methods available for determining sufficient statistics.

The connection with sufficient statistics goes further: If we set $\mathbf{X}$ equal to the the data, $\mathbf{X} = D$, and set Y equal to the parameters, $Y = \theta$, then statement (i) becomes $D|(\theta, R) \sim D|R$ and we are lead back to Fisher's concept of sufficiency (1). In this way, the notion of a sufficient reduction encompasses sufficient statistics as well.

While the reductions discussed in the previous sections are linear functions of $\mathbf{X}$, sufficient reductions do not have to be linear. If $\mathbf{X}_y$ is normally distributed with mean 0 and $\mathrm{Var}(\mathbf{X}_y) = \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T + (1+\nu_y)\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$, $\boldsymbol{\Gamma} \in \mathbb{R}^p$, then $(\boldsymbol{\Gamma}^T\mathbf{X})^2$ is a sufficient reduction.

8.2 Model-free sufficient dimension reduction

In this article I have adopted a largely Fisherian perspective: (1) Find an adequate solution to the problem of specification for the conditional distribution of $\mathbf{X}|Y$, (2) use the inverse model for $\mathbf{X}|Y$ to estimate a sufficient reduction $R(\mathbf{X})$, and then, as described in Section 8.1, (3) pass the estimated reduction to the forward regression. Lacking an inverse model, these ideas are not applicable because then there is no probability structure with which to determine a sufficient reduction. However, progress is still possible by restricting the search for a sufficient reduction to a specific functional form. In

view of their prevalence in the previous sections, linear reductions form a natural and potentially useful class.

Consider then the reductive forms of Section 8.1 with $R(\mathbf{X}) = \mathbf{G}^T\mathbf{X}$, $\mathbf{G} \in \mathbb{R}^{p\times k}$, $k \leq p$. One linear reduction always exists because statements (i)-(iii) are trivially true with $\mathbf{G} = I_p$. If $\mathbf{G}^T\mathbf{X}$ is a linear reduction then so is $\mathbf{A}^T\mathbf{G}^T\mathbf{X}$ for any full rank $\mathbf{A} \in \mathbb{R}^{k\times k}$, suggesting again that interest center on the *dimension reduction (DR) subspace* $\mathcal{S}_{\mathbf{G}}$. If $\mathcal{S}_{\mathbf{G}}$ is a DR subspace and $\mathcal{S}_{\mathbf{G}} \subseteq \mathcal{S}$, then $\mathcal{S}$ is also a DR subspace. These results suggest that in any particular problem there may be infinitely many DR subspaces, and it therefore becomes necessary to consider the "smallest" subspace.

There are at least two ways to define a smallest DR subspace. One way is to require a subspace $\mathcal{S}_{\min} = \min_k \mathcal{S}_{\mathbf{G}}$ with the smallest dimension. However, such subspaces are not necessarily unique and, even if they were, they do not impose sufficient structure on the regression for progress in theory or uncomplicated application in practice. Another way is to restrict attention to the class of regressions in which the intersection $\mathcal{S}_{Y|\mathbf{X}} = \cap\mathcal{S}_{\mathbf{G}}$ of all DR subspaces is itself a DR subspace. The *central subspace* $\mathcal{S}_{Y|\mathbf{X}}$ (Cook 1994, 1998) then becomes the parameter of interest. The reductive subspaces $\mathcal{S}_{\mathbf{\Gamma}}$ and $\mathcal{S}_{\mathbf{\Delta}^{-1}\mathbf{\Gamma}}$ encountered in previous sections are instances of model-based central subspaces. Since generally there is no forward or inverse model to tie up loose ends like high-order conditional moments, it has proven quite hard to estimate the entire central subspace without some restrictions on the regression. Nevertheless, there are successful methods that can provide useful estimates of $\mathcal{S}_{Y|\mathbf{X}}$ under conditions that are weak relative to a fully specified forward or inverse model.

An introduction to model-free sufficient dimension reduction via central subspaces is available from Cook (1998) and the references contained therein. See Cook and Ni (2005) for recent methodology and references to recent literature.

APPENDIX

A Propositions 1, 3 and 6

We first demonstrate Proposition 6. Propositions 1 and 3 will then follow as special cases.

To demonstrate Proposition 6 we first show that the distribution of $\mathbf{X}|(\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{X}, Y = y)$ is the same as the distribution of $\mathbf{X}|\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{X}$ for all y. According to model (15), $\mathbf{X}_y$ is normally distributed with mean $\boldsymbol{\mu}_y = \boldsymbol{\mu} + \mathbf{\Gamma}\boldsymbol{\nu}_y$ and constant variance. Thus

$\mathbf{X}|(\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{X}, Y = y)$ is normally distributed with constant variance and mean

$$\begin{aligned}
\mathrm{E}(\mathbf{X}|\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{X}, Y = y) &= \boldsymbol{\mu}_y + P^T_{\mathbf{\Delta}^{-1}\mathbf{\Gamma}(\mathbf{\Delta})}(\mathbf{X} - \boldsymbol{\mu}_y) \\
&= (I - P^T_{\mathbf{\Delta}^{-1}\mathbf{\Gamma}(\mathbf{\Delta})})\boldsymbol{\mu} + P^T_{\mathbf{\Delta}^{-1}\mathbf{\Gamma}(\mathbf{\Delta})}\mathbf{X} + (I - P^T_{\mathbf{\Delta}^{-1}\mathbf{\Gamma}(\mathbf{\Delta})})\mathbf{\Gamma}\boldsymbol{\nu}_y \\
&= (I - P^T_{\mathbf{\Delta}^{-1}\mathbf{\Gamma}(\mathbf{\Delta})})\boldsymbol{\mu} + P^T_{\mathbf{\Delta}^{-1}\mathbf{\Gamma}(\mathbf{\Delta})}\mathbf{X}
\end{aligned}$$

where$P_{\mathbf{\Delta}^{-1}\mathbf{\Gamma}(\mathbf{\Delta})}$ is the operator that projects onto $\mathrm{Span}(\mathbf{\Delta}^{-1}\mathbf{\Gamma})$ in the $\mathbf{\Delta}$ inner product. From this the last term in the second equation is 0. Since $\mathbf{X}|(\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{X}, Y = y)$ is normally distributed with mean and variance that do not depend on y, it follows that the distribution of $\mathbf{X}|(\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{X}, Y = y)$ is the same as the distribution of $\mathbf{X}|\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{X}$ for all y. Consequently, $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{X}$, which implies that $Y|\mathbf{X}$ and $Y|\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{X}$ have identical distributions.

Proposition 1 follows immediately because $\mathbf{\Delta} = I$. Under Proposition 3,

$$[\mathrm{Var}(\mathbf{X}_y)]^{-1} = \mathbf{\Gamma}_0\mathbf{\Omega}_0^{-2}\mathbf{\Gamma}_0^T + \mathbf{\Gamma}\mathbf{\Omega}^{-2}\mathbf{\Gamma}^T.$$

Consequently, $[\mathrm{Var}(\mathbf{X}_y)]^{-1}\mathbf{\Gamma} = \mathbf{\Gamma}\mathbf{\Omega}^{-2}$ and Proposition 3 follows.

B Proposition 2

We use a different approach to demonstrate this proposition. Suppose that under model (3) the joint density or mass function $f(\mathbf{x}|y)$ of $\mathbf{X}|(Y = y)$ can be written as

$$f(\mathbf{x}|y) = h(\mathbf{x})g(\mathbf{\Gamma}^T\mathbf{x}, y)$$

where h is a function that does not depend on y and g is a function that depends on $\mathbf{x}$ only through $\mathbf{\Gamma}^T\mathbf{x}$. It would then follow from the usual factorization theorem for sufficiency that the distribution of $\mathbf{X}|(\mathbf{\Gamma}^T\mathbf{X}, Y = y)$ is the same as the distribution of $\mathbf{X}|\mathbf{\Gamma}^T\mathbf{X}$ for all y, and thus $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{\Gamma}^T\mathbf{X}$.

To demonstrate the required factorization, let x_j be the jth element of $\mathbf{x}$ and, using the conditional independence of the predictors, write

$$\begin{aligned}
f(\mathbf{x}|y) &= \prod_{j=1}^{p} a_j(\eta_{yj})b_j(x_j)\exp\{x_j\eta_{yj}\} \\
&= \prod_{j=1}^{p} a_j(\eta_{yj})b_j(x_j)\exp\{x_j(\mu_j + \boldsymbol{\gamma}_j^T\boldsymbol{\nu}_y)\} \\
&= \left[\prod_{j=1}^{p} b_j(x_j)\exp(x_j\mu_j)\right]\left[\exp\{\boldsymbol{\nu}_y^T\mathbf{\Gamma}^T\mathbf{x}\}\prod_{j=1}^{p} a_j(\eta_{yj})\right] \\
&= h(\mathbf{x})g(\mathbf{\Gamma}^T\mathbf{x}, y).
\end{aligned}$$

C Equation 6

Substituting $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$ and $\hat{\boldsymbol{\beta}}^T = (\mathbf{F}^T\mathbf{F})^-\mathbf{F}^T\mathbb{X}\mathbf{G}$ into the log likelihood, we need to maximize

$$\begin{aligned}
M_{\text{PFC}} =& (-np/2)\log(s^2) - (1/2s^2)\sum_y \|\mathbf{X}_y - \bar{\mathbf{X}} - P_{\mathbf{G}}\mathbb{X}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^-\mathbf{f}_y\|^2 \\
=& (-np/2)\log(s^2) - (1/2s^2)\{\,\text{trace}[\sum_y (\mathbf{X}_y - \bar{\mathbf{X}})^T(\mathbf{X}_y - \bar{\mathbf{X}})] \\
&- \text{trace}[P_{\mathbf{G}}\mathbb{X}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^-\sum_y \mathbf{f}_y(\mathbf{X}_y - \bar{\mathbf{X}})^T] \\
&- \text{trace}[\sum_y (\mathbf{X}_y - \bar{\mathbf{X}})\mathbf{f}_y^T(\mathbf{F}^T\mathbf{F})^-\mathbf{F}^T\mathbb{X}P_{\mathbf{G}}] \\
&+ \text{trace}[(\mathbf{F}^T\mathbf{F})^-\mathbf{F}^T\mathbb{X}P_{\mathbf{G}}\mathbb{X}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^-\sum_y \mathbf{f}_y\mathbf{f}_y^T]\}
\end{aligned}$$

But $\sum_y \mathbf{f}_y(\mathbf{X}_y - \bar{\mathbf{X}})^T = \mathbf{F}^T\mathbb{X}$ and $\sum_y \mathbf{f}_y\mathbf{f}_y^T = \mathbf{F}^T\mathbf{F}$. Thus

$$\begin{aligned}
M_{\text{PFC}} =& (-np/2)\log(s^2) - (1/2s^2)\{\,\text{trace}[\mathbb{X}^T\mathbb{X}] - \text{trace}[P_{\mathbf{G}}\mathbb{X}^T P_{\mathbf{F}}\mathbb{X}] \\
&- \text{trace}[\mathbb{X}^T P_{\mathbf{F}}\mathbb{X}P_{\mathbf{G}}] + \text{trace}[\mathbb{X}^T P_{\mathbf{F}}\mathbb{X}P_{\mathbf{G}}]\} \\
=& (-np/2)\log(s^2) - (1/2s^2)\{\,\text{trace}[\mathbb{X}^T\mathbb{X}] - \text{trace}[P_{\mathbf{G}}\mathbb{X}^T P_{\mathbf{F}}\mathbb{X}]\} \\
=& (-np/2)\log(s^2) - (n/2s^2)\{\,\text{trace}[\widehat{\boldsymbol{\Sigma}}] - \text{trace}[P_{\mathbf{G}}\widehat{\boldsymbol{\Sigma}}_{\text{fit}}P_{\mathbf{G}}]\}
\end{aligned}$$

D Proposition 4

Requiring the errors to be uncorrelated but not necessarily normal, it is know that $\widehat{\boldsymbol{\Sigma}} \xrightarrow{p} \boldsymbol{\Sigma}$, and

$$\begin{aligned}
\boldsymbol{\Sigma} &= \text{Var}(\mathbf{X}) \\
&= \text{E}(\text{Var}(\mathbf{X}|Y)) + \text{Var}(\text{E}(\mathbf{X}|Y)) \\
&= \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^2\boldsymbol{\Gamma}_0^2 + \boldsymbol{\Gamma}\boldsymbol{\Omega}^2\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}\boldsymbol{\beta}\text{Var}(\mathbf{f}_Y)\boldsymbol{\beta}^T\boldsymbol{\Gamma}^T
\end{aligned}$$

To find the limiting value of $\widehat{\boldsymbol{\Sigma}}_{\text{fit}} = \mathbb{X}^T P_{\mathbf{F}}\mathbb{X}/n$, use model (12) to write

$$\mathbf{X}_y^T - \bar{\mathbf{X}}^T = (\boldsymbol{\mu}^T - \bar{\mathbf{X}}^T) + \mathbf{f}_y^T\boldsymbol{\beta}^T\boldsymbol{\Gamma}^T + \boldsymbol{\varepsilon}_0^T\boldsymbol{\Omega}_0^T\boldsymbol{\Gamma}_0^T + \boldsymbol{\varepsilon}^T\boldsymbol{\Omega}^T\boldsymbol{\Gamma}^T$$

and thus

$$\begin{aligned}
\mathbb{X} &= 1_n(\boldsymbol{\mu}^T - \bar{\mathbf{X}}^T) + \mathbf{F}\boldsymbol{\beta}^T\boldsymbol{\Gamma}^T + \mathbf{R}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T + \mathbf{R}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T \\
P_{\mathbf{F}}\mathbb{X} &= \mathbf{F}\boldsymbol{\beta}^T\boldsymbol{\Gamma}^T + P_{\mathbf{F}}\mathbf{R}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T + P_{\mathbf{F}}\mathbf{R}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T
\end{aligned}$$

where $\mathbf{R}_0$ is the $n \times p - d$ with rows $\boldsymbol{\varepsilon}_0^T$ and $\mathbf{R}$ is the $n \times d$ matrix with rows $\boldsymbol{\varepsilon}^T$. From this, $\mathbb{X}^T P_{\mathbf{F}} \mathbb{X}/n$ contains 9 terms. The first of these is $\boldsymbol{\Gamma}\boldsymbol{\beta}(\mathbf{F}^T\mathbf{F}/n)\boldsymbol{\Gamma}^T\boldsymbol{\beta}^T \xrightarrow{p} \boldsymbol{\Gamma}\boldsymbol{\beta}\mathrm{Var}(\mathbf{f}_Y)\boldsymbol{\Gamma}^T\boldsymbol{\beta}^T$. The remaining terms involve products containing either or both of the factors $\mathbf{F}^T\mathbf{R}/n$ and $\mathbf{F}^T\mathbf{R}_0/n$. Recalling that the rows (corresponding to samples) of $\mathbf{R}$ and $\mathbf{R}_0$ are independent, these factors both converge to 0 and this property forces each of the 8 remaining terms to converge to 0. For instance, the last quadratic term can be written $\boldsymbol{\Gamma}\boldsymbol{\Omega}(\mathbf{R}^T P_{\mathbf{F}} \mathbf{R}/n)\boldsymbol{\Omega}\boldsymbol{\Gamma}^T$, and

$$\mathbf{R}^T P_{\mathbf{F}} \mathbf{R}/n = (\mathbf{R}^T\mathbf{F}/n)(\mathbf{F}^T\mathbf{F}/n)^-(\mathbf{F}^T\mathbf{R}/n)$$

which converges to 0 in probability by Slutsky's theorem.

Since $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}} + \widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}$, take the difference of the limiting values for $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{fit}}$ to confirm the limiting value for $\widehat{\boldsymbol{\Sigma}}_{\mathrm{res}}$.

E Equation 14

Let $\mathbf{O} = (\mathbf{O}_1, \mathbf{O}_2)$ be a partitioned $p \times p$ orthogonal matrix. Then

$$\begin{aligned}
|\widehat{\boldsymbol{\Sigma}}| =& |\mathbf{O}^T\widehat{\boldsymbol{\Sigma}}\mathbf{O}| = \begin{vmatrix} \mathbf{O}_1^T\widehat{\boldsymbol{\Sigma}}\mathbf{O}_1 & \mathbf{O}_1^T\widehat{\boldsymbol{\Sigma}}\mathbf{O}_2 \\ \mathbf{O}_2^T\widehat{\boldsymbol{\Sigma}}\mathbf{O}_1 & \mathbf{O}_2^T\widehat{\boldsymbol{\Sigma}}\mathbf{O}_2 \end{vmatrix} \\
=& |\mathbf{O}_1^T\widehat{\boldsymbol{\Sigma}}\mathbf{O}_1||\mathbf{O}_2^T\widehat{\boldsymbol{\Sigma}}\mathbf{O}_2 - \mathbf{O}_2^T\widehat{\boldsymbol{\Sigma}}\mathbf{O}_1(\mathbf{O}_1^T\widehat{\boldsymbol{\Sigma}}\mathbf{O}_1)^{-1}\mathbf{O}_1^T\widehat{\boldsymbol{\Sigma}}\mathbf{O}_2| \\
\leq& |\mathbf{O}_1^T\widehat{\boldsymbol{\Sigma}}\mathbf{O}_1||\mathbf{O}_2^T\widehat{\boldsymbol{\Sigma}}\mathbf{O}_2|
\end{aligned}$$

F Proposition 5

Let $\mathbf{C} = \boldsymbol{\beta}\mathrm{Var}(\mathbf{f}_Y)\boldsymbol{\beta}^T$ and $\mathbf{H} = \mathbf{G}^T\boldsymbol{\Gamma}\mathbf{C}^{1/2}$. Then it follows from Proposition 4 that

$$\begin{aligned}
\widetilde{L}_{\mathrm{PFC}}(\mathbf{G}) =& \log|\mathbf{G}_0^T\boldsymbol{\Sigma}\mathbf{G}_0| + \log|\mathbf{G}^T\boldsymbol{\Sigma}_{\mathrm{res}}\mathbf{G}| \\
=& \log|\mathbf{G}_0^T\boldsymbol{\Sigma}_{\mathrm{res}}\mathbf{G}_0 + \mathbf{G}_0^T\boldsymbol{\Gamma}\mathbf{C}\boldsymbol{\Gamma}^T\mathbf{G}_0| + \log|\mathbf{G}^T\boldsymbol{\Sigma}_{\mathrm{res}}\mathbf{G}| \\
=& \log|\mathbf{G}_0^T\boldsymbol{\Sigma}_{\mathrm{res}}\mathbf{G}_0| + \log|I_d + \mathbf{H}^T(\mathbf{G}_0^T\boldsymbol{\Sigma}_{\mathrm{res}}\mathbf{G}_0)^{-1}\mathbf{H}| + \log|\mathbf{G}^T\boldsymbol{\Sigma}_{\mathrm{res}}\mathbf{G}| \\
\geq& \log|\mathbf{G}_0^T\boldsymbol{\Sigma}_{\mathrm{res}}\mathbf{G}_0| + \log|\mathbf{G}^T\boldsymbol{\Sigma}_{\mathrm{res}}\mathbf{G}| \\
\geq& \log|\boldsymbol{\Sigma}_{\mathrm{res}}| \\
=& \log|\boldsymbol{\Gamma}\boldsymbol{\Omega}^2\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^2\boldsymbol{\Gamma}_0^T| \\
=& \log|\boldsymbol{\Omega}_0^2||\boldsymbol{\Omega}^2|
\end{aligned}$$

In addition,

$$\begin{aligned}
\widetilde{L}_{\mathrm{PFC}}(\boldsymbol{\Gamma}) =& \log|\boldsymbol{\Gamma}_0^T\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0| + \log|\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}_{\mathrm{res}}\boldsymbol{\Gamma}| \\
=& \log|\boldsymbol{\Omega}_0^2||\boldsymbol{\Omega}^2|
\end{aligned}$$

Therefore $\widetilde{L}_{\mathrm{PFC}}(\mathbf{G}) \geq L(\mathbf{\Gamma})$ for all $\mathbf{G}$. The minimizing argument yields a unique subspace if $\widetilde{L}_{\mathrm{PFC}}(\mathbf{G}) > \widetilde{L}_{\mathrm{PFC}}(\mathbf{\Gamma})$ for all $\mathbf{G}$ such that $\mathbf{G}^T\mathbf{G} = I$, $\dim(\mathcal{S}_{\mathbf{G}}) = d$ and $\mathcal{S}_{\mathbf{G}} \neq \mathcal{S}_{\mathbf{\Gamma}}$.

G Eigenvectors of $\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}$ and $\widehat{\mathbf{\Sigma}}_{\mathrm{res}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}$

$$\begin{aligned}
\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}\ell = \lambda\ell \quad &\Leftrightarrow \quad \widehat{\mathbf{\Sigma}}_{\mathrm{fit}}\ell = \lambda\widehat{\mathbf{\Sigma}}\ell \\
&\Leftrightarrow \quad \widehat{\mathbf{\Sigma}}_{\mathrm{fit}}\ell = \lambda\widehat{\mathbf{\Sigma}}_{\mathrm{res}}\ell + \lambda\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}\ell \\
&\Leftrightarrow \quad (1-\lambda)\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}\ell = \lambda\widehat{\mathbf{\Sigma}}_{\mathrm{res}}\ell \\
&\Leftrightarrow \quad \widehat{\mathbf{\Sigma}}_{\mathrm{res}}^{-1}\widehat{\mathbf{\Sigma}}_{\mathrm{fit}}\ell = (\lambda/(1-\lambda))\ell
\end{aligned}$$

The conclusion follows because $\widehat{\mathbf{\Sigma}}_{\mathrm{res}} = \widehat{\mathbf{\Sigma}} - \widehat{\mathbf{\Sigma}}_{\mathrm{fit}} > 0$ and $\lambda/1-\lambda$ is a strictly monotonic function of λ,

References

Adcock, R.J. (1878). A problem in least squares. *The Analyst* **5**, 53-54.

Alter, O., Brown, P. and Botstein, D. (2000). Singular value decomposition for genewide expression data processing and modeling. *Proceedings of the National Academy of Sciences* **97**, 10101–10106.

Anderson, T.W. (1984). Estimating linear statistical relationships. *The Annals of Statistics*, **12**, 1-45.

Anscombe, F.J. (1961). Examination of residuals. *Proceedings of the Fourth Berkeley Symposium*, **1**, 1–36.

Anscombe, F.J. and Tukey, J.W. (1963). The examination and analysis of residuals. *Technometrics*, **5**, 141-160.

Box G.E.P. (1979). Robustness in the strategy of scientific model building. In R. Launer and G. Wilkinson (eds.), *Robustness in Statistics*, pp. 201–235. New York: Academic Press.

Box, G.E.P. (1980). Sampling and Bayes inference in scientific modeling and robustness (with discussion). *Journal of the Royal Statistical Society, A.*, **134**, 383–430.

Bura, E. and Cook, R.D. (2001). Extending sliced inverse regression: the weighted chi-squared test. *Journal of the American Statistical Association* **96**, 996–1003.

Bura, E and Pfeiffer, R.M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics* **19**, 1252–1258.

Breiman, L. (2001). Statistical modeling: The two cultures (with discussion). *Statistical Science* **16**, 199-226.

De Leeuw, L. (2006). Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data Analysis* **50**, 21-39.

Cook, R.D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, B*, **48**, 133-155.

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*, pp. 18-25. Alexandria VA: American Statistical Association.

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

Cook, R.D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association* **100**, 410–428.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.

Cook, R.D. and Weisberg, S. (1994). *An Introduction to Regression Graphics.* New York: Wiley.

Chiaromonte, F. and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences* **176**, 123–144.

Christensen, R. (2001). *Advanced Linear Modeling, Second Ed.*, New York: Springer.

Cox, D.R. (1968). Notes on some aspects of regression analysis. *Journal of the Royal Statistical Society, ser. A* **131**, 265–279.

Cox, D.R. (1990). Role of models in statistical analysis. *Statistical Science*, **5**, 169–174.

Edelman, A., Tomás, A.A., Smith, S.T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal of Matrix Analysis and Applications* **20**, 303–353.

Edgeworth, F. Y. (1884). On the reduction of observations. *Philosophical Magazine*, 135-141.

Eubank, R.L. (1999). *Spline Smoothing and Nonparametric Regression.* New York: Marcel Dekker.

Fisher, R.J. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Statistical Society, A*. **222**, 309–368.

Fisher, R.J. (1936). Uncertain inference. *Proceedings of the American Academy of Arts and Sciences*, **71**, 245–258.

Fisher, R.J. (1924). The influence of rainfall on the yield of wheat at Rothamsted. *Philosophical Transactions of the Royal Statistical Society, B*, **213**, 89–142.

Fisher, R.J. (1941). *Statistical Methods for Research Workers*, eighth ed., London: Oliver and Boyd.

George, E.I. and Oman, S.D. (1996). Multiple-shrinkage principal component regression. *The Statistician* **45**, 111-124.

Gould, S.J. (1981). *The Mismeasure of Man*. New York: W.W. Norton.

Hadi, A.S. and Ling, R.F. (1998). Some cautionary notes on the use of principal components regression. *The American Statistician* **52**, 15–19.

Hawkins, D.M. and Fatti, L.P. (1984). Exploring multivariate data using the minor principal components. *The Statistician* **33**, 325–338.

Helland, I.S. (1992). Maximum likelihood regression on relevant components. *Journal of the Royal Statistical Society, Series B* **54**, 637-647.

Helland, I.S., and Almoy, T. (1994). Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association* **89**, 583-591.

Hocking, R.R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1-49.

Hotelling, H. (1933). Analysis of a comples statistical variable into its principal components. *Journal of Educational Psychology* **24**, 417-441.

Hotelling, H. (1957). The relationship of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology* **10**, 69-79.

Hwang, J.T.G., and Nettleton, D. (2003). Principal components regression with data-chosen components and related methods. *Technometrics* **45**, 70–79.

Jolliffe, I.T. (1982). A note on the use of principal components in regression. *Applied Statistics*, **31** 300–303.

Jolliffe, I.T. (2003). *Principal Component Analysis, Second Ed.*, New York: Springer.

Jong, J. and Kotz, S. (1999). On a relation between principal components and regression analysis. *The American Statistician* **53**, 349–352.

Kendall, M.G.(1957). *A Course in Multivariate Analysis*. London: Griffin.

Lehmann, E.L. (1990). Model specification: The views of Fisher, Neyman, and later developments. *Statistical Science*, **5**, 160–168.

Li, K-C. (1991), Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**,316-342.

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

Maronna, R. (2005). Principal components and orthogonal regression based on robust scales. *Technometrics*, **47**, 264-273.

McCullagh, P. (2002). What is a statistical model? (with discussion), *The Annals of Statistics*, **30**, 1225–1300.

Oman, S.D. (1991). Random calibration with many measurements: An application of Stein estimation. *Technometrics* **33**, 187-195.

Pearson, K. (1901). On lines and planes of closest fit to a system of points in space. *Philosophical Magazine (6)* **2**, 559-572.

Rao, C.R. (1962). Use of discriminant and allied functions in multivariate analysis. *Sankhyā*, **23**, 149-154.

Savage, L.J. (1976). On reading R.A. Fisher. *The Annals of Statistics*, **4**, 441–500.

Scott, D. (1992). *Multivariate Density Estimation*. New York: Wiley.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology*, **15**, 201–293.

Stigler, S.M. (1973). Studies in the history of probability and statistics. XXXII: Laplace, Fisher and the discovery of the concept of sufficiency. *Biometrika*, **60**, 439–445.

Stigler, S.M. (1976). Discussion of "On reading R.A. Fisher" by L.J. Savage. *Annals of Statistics*, **4**, 441–500.

Stigler, S.M. (2005). Fisher in 1921. *Statistical Science*, **20**, 32–49.

Tipping, M.E. and Bishop, C.M. (1999). Probabilistic principal components. *Journal of the Royal Statistical Society, ser B* **61**, 611-622.

Xia, Y., Tong, H., Li, W. K., and Zhu, L-X. (2002), An adaptive estimation of dimension reduction space (with discussion), *Journal of the Royal Statistical Society*, Ser. B, **64**, 363-410.

Zhao, P. L., and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642-648.