

# Dimension Reduction in Regressions with Exponential Family Predictors \*

R. Dennis Cook      Lexin Li

February 19, 2009

## Abstract

We present first methodology for dimension reduction in regressions with predictors that, given the response, follow one-parameter exponential families. Our approach is based on modeling the conditional distribution of the predictors given the response, which allows us to derive and estimate a sufficient reduction of the predictors. We also propose a method of estimating the forward regression mean function without requiring an explicit forward regression model. Whereas nearly all existing estimators of the central subspace are limited to regressions with continuous predictors only, our proposed methodology extends estimation to regressions with all categorical or a mixture of categorical and continuous predictors. Supplementary materials including the proofs and the computer code are available from the JCGS website.

---

\*R. Dennis Cook is Professor, School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street SE, Minneapolis, MN 55455. email: dennis@stat.umn.edu. Lexin Li is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695. email: li@stat.ncsu.edu. Research for this article was supported in part by National Science Foundation Grant DMS-0405360 awarded to RDC, and Grant DMS-0706919 awarded to LL.

*Key words and phrases:* Central subspace, Grassmann manifolds, inverse regression, sufficient dimension reduction.

## 1 Introduction

Many different statistical contexts have been developed to aid in studying the conditional distribution of a response  $Y$  given a predictor vector  $\mathbf{X} \in \mathbb{R}^p$ , including generalized additive models, projection pursuit, inverse regression methods, as well as long-standing methods like principal component regression and partial least squares. These and other regression methods employ some type of dimension reduction for  $\mathbf{X}$ , either estimated or imposed as an intrinsic part of the method. Nearly all of the methods can be viewed as instances of the following definition (Cook, 2007):

**Definition 1** *A reduction  $R : \mathbb{R}^p \rightarrow \mathbb{R}^q$ ,  $q \leq p$ , is sufficient if it satisfies one of the following three statements: (i) inverse reduction,  $\mathbf{X}|(Y, R(\mathbf{X})) \sim \mathbf{X}|R(\mathbf{X})$ ; (ii) forward reduction,  $Y|\mathbf{X} \sim Y|R(\mathbf{X})$ ; (iii) joint reduction,  $\mathbf{X} \perp\!\!\!\perp Y|R(\mathbf{X})$ , where  $\perp\!\!\!\perp$  indicates independence and  $\sim$  means identically distributed.*

If we consider a generic statistical problem and reinterpret  $\mathbf{X}$  as the total data  $D$  and  $Y$  as the parameter  $\theta$ , then condition (i) for inverse reduction becomes  $D|(\theta, R) \sim D|R$  so that  $R$  is a sufficient statistic. In this way, the definition of a sufficient reduction encompasses Fisher's (1922) classical definition of sufficiency. One difference is that sufficient statistics are observable, while a sufficient reduction may contain unknown parameters and thus need to be estimated. There are several reasons why sufficient reductions may be useful in practice, including the possibilities of mitigating the effects of collinearity, facilitating model development by allowing visualization of the regression in low dimensions (Cook, 1998) and providing a relatively small set of composite variables  $R(\mathbf{X})$  on which to base prediction or interpretation.

The choice of a reductive paradigm depends on the stochastic nature of  $\mathbf{X}$  and  $Y$ . If the values of  $\mathbf{X}$  are fixed by design, then forward regression (*ii*) seems the natural choice. In discriminant analysis  $\mathbf{X}|Y$  is a random vector of features observed in one of a number of subpopulations indicated by  $Y$ . If the values of  $Y$  are fixed by design then inverse regression (*i*) is perhaps the only reasonable reductive route. The three statements in Definition 1 are equivalent when  $(Y, \mathbf{X})$  has a joint distribution. In that case we are free to estimate a reduction inversely or jointly and then pass the estimated reduction to the forward regression without additional structure. We assume that  $Y$  and  $\mathbf{X}$  are jointly distributed in this article.

Inspired by the pioneering methods of sliced inverse regression (SIR; Li, 1991) and sliced average variance estimation (SAVE; Cook and Weisberg, 1991), there has been considerable interest in model-free dimension reduction methods, including minimum discrepancy estimation (Cook and Ni, 2005), contour regression (Li, Zha, and Chiaromonte, 2005), directional regression (Li and Wang, 2007), minimum average variance estimation (MAVE, Xia, Tong, Li and Zhu, 2002) and single-index direction estimation (Yin and Cook, 2005), among others. Nearly all of these methods require the predictors to be continuous and impose additional constraints on their marginal distribution. Normality is typically required to achieve the best performance. In most cases the reductive parameter is taken to be the central subspace  $\mathcal{S}_{Y|\mathbf{X}}$ , defined as the intersection of all subspaces  $\mathcal{S}_{\mathbf{B}} \subseteq \mathbb{R}^p$  such that  $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^T\mathbf{X}$ , where the columns of  $\mathbf{B}$  are a basis for the subspace; see Cook and Ni (2005) for a review of this literature. Chiaromonte, Cook and Li (2002), Li, Cook and Chiaromonte (2003) and Lique and Saracco (2007) allow for the presence of categorical predictors, but the dimension reduction exercise is still restricted to the continuous predictors; no reduction is attempted for the categorical predictors themselves. To our knowledge, no sufficient dimension reduction methods have been developed for application to categorical predictors, which is a limitation that has impeded the appli-

cation of dimension reduction methods in some areas. In this article we present a first solution to address this issue, leading to methods for dimension reduction in regressions with all categorical or mixed types of predictors.

More specifically, we develop reductive methodology for regressions in which the individual conditional predictors in  $\mathbf{X}|Y$  are distributed according to exponential families. Reductions  $R(\mathbf{X})$  are pursued inversely following route (i) and may be used subsequently for graphics, for the development of forward models for  $Y|R(\mathbf{X})$  and, as discussed in Section 4.3, for the estimation of  $E(Y|\mathbf{X})$  without explicitly developing a forward model. In Section 2 we develop a conditional independence model on which our primary methodology is based. Estimation is considered in Section 3, and inference is discussed in Section 4. Some numerical studies are reported in Section 5. We assume in Sections 2–5 that the predictors are conditionally independent given the response. This working independence assumption has been commonly used since the pioneering work of Liang and Zeger (1986). In Section 6 we propose a moment-based estimation method that allows for conditionally dependent predictors. Our general conclusion is that methodology based on conditional independence should be useful even with conditionally dependent predictors, as long as the dependence is moderate or the sample size is not large enough to allow good estimation of the dependence structure. This conclusion is also in agreement with Prentice and Zhao (1991, p.830) who recommended in a related context that independence models should be adequate for a broad range of applications in which the dependencies are not strong. Technical proofs are available from the JCGS website.

## 2 Generalized Principal Fitted Components

Let  $\mathbf{X}_y = (X_{yj})$ ,  $j = 1, \dots, p$ , denote a random vector distributed as  $\mathbf{X}|(Y = y)$ . We assume that the conditional predictors  $X_{yj}$  are independent and that  $X_{yj}$  is distributed

according to a one-parameter exponential family with density or mass function of the form

$$f_j(x|\eta_{yj}, Y = y) = a_j(\eta_{yj})b_j(x) \exp(x\eta_{yj}). \quad (1)$$

The natural parameters  $\eta_{yj}$ ,  $j = 1, \dots, p$ , are a function of  $y$  as indicated by the first subscript. Let  $\boldsymbol{\eta}_y \in \mathbb{R}^p$  have elements  $\eta_{yj}$ ,  $j = 1, \dots, p$ , let  $\bar{\boldsymbol{\eta}} = \mathbf{E}(\boldsymbol{\eta}_Y)$ , and let  $\mathcal{S}_{\boldsymbol{\Gamma}} = \text{span}\{\boldsymbol{\eta}_y - \bar{\boldsymbol{\eta}} | y \in S_Y\}$ , where  $S_Y$  denotes the sample space of  $Y$  and the columns of the semi-orthogonal matrix  $\boldsymbol{\Gamma} \in \mathbb{R}^{p \times d}$  form a basis for the  $d$ -dimensional subspace  $\mathcal{S}_{\boldsymbol{\Gamma}}$ . Then

$$\boldsymbol{\eta}_y = \bar{\boldsymbol{\eta}} + \boldsymbol{\Gamma}\boldsymbol{\nu}_y \quad (2)$$

$$\eta_{yj} = \bar{\eta}_j + \boldsymbol{\gamma}_j^T \boldsymbol{\nu}_y, \quad j = 1, \dots, p, \quad (3)$$

where  $\boldsymbol{\nu}_y = \boldsymbol{\Gamma}^T(\boldsymbol{\eta}_y - \bar{\boldsymbol{\eta}})$ , and  $\boldsymbol{\gamma}_j^T$  is the  $j$ -th row of  $\boldsymbol{\Gamma}$ . Equation (2) represents  $\boldsymbol{\eta}_y$  in terms of coordinates  $\boldsymbol{\nu}_y$  relative to the basis  $\boldsymbol{\Gamma}$ , while (3) is in terms of the individual natural parameters. The matrix  $\boldsymbol{\Gamma}$  is not identified in this model, but  $\mathcal{S}_{\boldsymbol{\Gamma}}$  is identified and estimable. The parameter space for  $\mathcal{S}_{\boldsymbol{\Gamma}}$  is thus a Grassmann manifold  $\mathcal{G}_{(d,p)}$  of dimension  $d$  in  $\mathbb{R}^p$ .  $\mathcal{G}_{(d,p)}$  is an analytic manifold of dimension  $d(p-d)$  (Chikuse, 2002, p. 9), which is the number of reals needed to specify a single subspace. For a sample of size  $n$ , the total number of real parameters in model (2) is thus  $d(p-d) + nd$ . Because this count increases with  $n$ , estimation and inference under model (2) will encounter special problems. In many regressions it will be possible and desirable to model  $\boldsymbol{\nu}_y$  and thereby increase efficiency by reducing the number of parameters that need to be estimated.

We next model  $\boldsymbol{\nu}_y$  as  $\boldsymbol{\nu}_y = \boldsymbol{\beta}\{\mathbf{f}_y - \mathbf{E}(\mathbf{f}_Y)\}$ , where  $\boldsymbol{\beta} \in \mathbb{R}^{d \times r}$  has rank  $d \leq \min(p, r)$  and  $\mathbf{f}_y \in \mathbb{R}^r$  is a known function of  $y$ . Each coordinate  $X_{yj}$ ,  $j = 1, \dots, p$ , of  $\mathbf{X}_y$  follows a generalized linear model as a function of  $y$ . Consequently, we are able to use known graphical methods (Cook, 1998) to aid in the selection of  $\mathbf{f}_y$ , an ability that is not generally

available in the forward regression of  $Y$  on  $\mathbf{X}$ . There may be a natural choice for  $\mathbf{f}_y$  in some regressions. For instance, suppose that  $Y$  is categorical, taking values in one of  $h$  categories  $C_k$ ,  $k = 1, \dots, h$ . In that case we can set  $r = h - 1$  and specify the  $k$ -th element of  $\mathbf{f}_y$  to be  $J(y \in C_k)$ , where  $J$  is the indicator function. We can also consider selecting  $\mathbf{f}_y$  to contain a reasonably flexible set of basis functions when  $Y$  is continuous, e.g., a polynomial in  $Y$ , which may be important when it is impractical to apply graphical methods to all of the predictors. Another option consists of “slicing” the observed values of  $Y$  into  $h$  bins (categories)  $C_k$ ,  $k = 1, \dots, h$ , and then specifying the  $k$ -th coordinate of  $\mathbf{f}_y$  as for the case of a categorical  $Y$ . This has the effect of approximating each natural parameter  $\eta_{yj}$  as a step function of  $y$  with  $h$  steps. Piecewise polynomials could also be used. It is also interesting to note that the choice of  $\mathbf{f}_y$  connects the method proposed here with parametric inverse regression development in Bura and Cook (2001) and Yin and Cook (2002).

The resulting model is then obtained by substituting  $\beta\mathbf{f}_y$  for  $\nu_y$  in (2):

$$\boldsymbol{\eta}_y = \boldsymbol{\mu} + \Gamma\beta\mathbf{f}_y, \quad (4)$$

where  $\boldsymbol{\mu} = \bar{\boldsymbol{\eta}} - \Gamma\beta\mathbf{E}(\mathbf{f}_Y)$ . If  $d = \min(p, r)$  then  $\Gamma\beta$  becomes an unconstrained matrix parameter  $\boldsymbol{\Phi} \in \mathbb{R}^{p \times r}$ , and we refer to the resulting model  $\boldsymbol{\eta}_y = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{f}_y$  as the *full model*. The full model provides a benchmark for assessing models with  $d < \min(p, r)$ , as discussed later.

The following proposition states that  $\Gamma^T\mathbf{X}$  is a minimal sufficient reduction, which implies that  $\mathcal{S}_\Gamma$  is the central subspace.

**Proposition 1** *Let  $R(\mathbf{X}) = \Gamma^T\mathbf{X}$  and let  $T(\mathbf{X})$  be any sufficient reduction. Then, under models (2) and (4),  $R$  is a sufficient reduction and  $R$  is a function of  $T$ .*

Cook (2007) studied the special case of model (4) in which  $\mathbf{X}_y$  is normally distributed

with mean  $\boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y$  and variance  $\sigma^2\mathbf{I}_p$ . He showed that the maximum likelihood estimator of  $R$  is of the form  $\widehat{R} = (\widehat{\mathbf{v}}_1^T \mathbf{X}, \dots, \widehat{\mathbf{v}}_d^T \mathbf{X})^T$ , where  $\widehat{\mathbf{v}}_j$  is the  $j$ -th eigenvector of the sample covariance matrix of the fitted vectors from the multivariate linear regression of  $\mathbf{X}_y$  on  $\mathbf{f}_y$ . The components  $\widehat{\mathbf{v}}_j^T \mathbf{X}$  were called *principal fitted components (PFC)* and the normal model giving rise to them was called a PFC model. We follow this terminology, referring to model (4) as a generalized PFC model.

Model (4) assumes that the predictors, given the response, are conditionally independent. This assumption seems restrictive, but turns out to be very useful in many regressions. As we will demonstrate in Section 6, model (4) with conditionally independent predictors is useful for conditionally dependent predictors as well, provided the dependency among data is moderate or the sample size is not large. In addition, there has recently been a surge of applications of principal component methodology in genome-wide association studies and microarray gene expression analyses to correct for population stratification and heterogeneity; see for example, Price, et al. (2006) and Leek and Storey (2007). Conditionally independent predictors are commonly assumed in these applications. Judging from their wide-spread use, we expect that model (4) with conditionally independent predictors will be a reasonable choice in many applications.

### 3 Estimation

Maximum likelihood estimation of the sufficient reduction  $R(\mathbf{X}) = \boldsymbol{\Gamma}^T \mathbf{X}$  requires estimation of  $\boldsymbol{\Gamma}$ . However, as mentioned previously,  $\boldsymbol{\Gamma}$  is not identified but  $\mathcal{S}_{\boldsymbol{\Gamma}}$  is identified and estimable. Any estimated basis for  $\mathcal{S}_{\boldsymbol{\Gamma}}$  will then serve to construct an estimated reduction.

Allowing for replication in the observed  $y$ 's, we assume that they take one of  $h$  distinct values  $v_i$  with frequencies  $n_i$ ,  $i = 1, \dots, h$ , and we let  $X_{v_i j k}$  denote the  $k$ -th observation

on the  $j$ -th predictor at  $v_i$ . The log likelihood  $L$  can then be constructed from (1):

$$L(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\beta}) = \sum_{i=1}^h \left\{ n_i \sum_{j=1}^p [\log\{a_j(\mu_j + \boldsymbol{\gamma}_j^T \boldsymbol{\beta} \mathbf{f}_{v_i})\}] + \mathbf{W}_{v_i}^T (\boldsymbol{\mu} + \boldsymbol{\Gamma} \boldsymbol{\beta} \mathbf{f}_{v_i}) \right\} + C \quad (5)$$

where  $\mathbf{W}_{v_i} \in \mathbb{R}^p$  has elements  $X_{v_i j \bullet} = \sum_k X_{v_i j k}$ ,  $j = 1, \dots, p$ , and  $C$  does not depend on the parameters. When  $d < \min(p, r)$ , (5) is to be maximized over  $\boldsymbol{\mu} \in \mathbb{R}^p$ ,  $\boldsymbol{\beta} \in \mathbb{R}^{d \times r}$  and  $\mathcal{S}_{\boldsymbol{\Gamma}} \in \mathcal{G}_{(d,p)}$ .

Under the full model,  $d = \min(p, r)$ . If  $d = p \leq r$ , then we can take  $\boldsymbol{\Gamma} = \mathbf{I}_p$  and  $\boldsymbol{\beta} \in \mathbb{R}^{p \times r}$ . If  $d = r \leq p$  then  $\boldsymbol{\beta} \in \mathbb{R}^{r \times r}$  and  $\boldsymbol{\Gamma} \boldsymbol{\beta} \in \mathbb{R}^{p \times r}$  is unconstrained. Consequently, when  $d = \min(p, r)$  we replace  $\boldsymbol{\Gamma} \boldsymbol{\beta}$  with the unconstrained matrix  $\boldsymbol{\Phi} \in \mathbb{R}^{p \times r}$  and  $\boldsymbol{\gamma}_j^T \boldsymbol{\beta}$  with the  $j$ -th row  $\boldsymbol{\phi}_j^T$  of  $\boldsymbol{\Phi}$ . In this case maximizing the log likelihood  $L(\boldsymbol{\mu}, \boldsymbol{\Phi})$  corresponds to fitting  $p$  separate regressions of the individual predictors on  $\mathbf{f}_y$ , the coefficients of  $\mathbf{f}_y$  for the  $j$ -th predictor  $X_j$  being  $\boldsymbol{\phi}_j^T$ .

To help fix ideas, we consider an example where the coordinates of  $\mathbf{X}_y$  are independent Bernoulli random variables. We can then express the model in terms of a multivariate logit defined coordinate-wise as  $\text{multlogit}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma} \boldsymbol{\beta} \mathbf{f}_y$ . Since the predictors are conditionally independent,

$$\Pr(\mathbf{X} = \mathbf{x} | Y = y) = \prod_{j=1}^p p_j(y)^{x_j} q_j(y)^{1-x_j},$$

where  $\mathbf{x} = (x_j)$ ,  $p_j(y) = \Pr(X_j = x_j | Y = y)$ ,  $q_j(y) = 1 - p_j(y)$ , and  $\log(p_j/q_j) = \mu_j + \boldsymbol{\gamma}_j^T \boldsymbol{\beta} \mathbf{f}_y$ . The log likelihood is therefore

$$L(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\beta}) = \sum_{i=1}^n \left\{ \sum_{j=1}^p q_j(y_i) + \mathbf{X}_{y_i}^T (\boldsymbol{\mu} + \boldsymbol{\Gamma} \boldsymbol{\beta} \mathbf{f}_{y_i}) \right\}, \quad (6)$$

which is to be maximized over  $\boldsymbol{\mu} \in \mathbb{R}^p$ ,  $\boldsymbol{\beta} \in \mathbb{R}^{d \times r}$  and  $\mathcal{S}_{\boldsymbol{\Gamma}} \in \mathcal{G}_{(d,p)}$ .

In the remainder of this section we give two algorithms for maximizing (5), one for continuous responses without replication and one for categorical responses with replication. The algorithms are similar but make special use of the nature of the response.

### 3.1 Continuous responses

With a continuous response, the log likelihood can be obtained from (5) by setting  $h = n$ ,  $v_i = y_i$  and  $\mathbf{W}_{v_i} = \mathbf{X}_{y_i}$ . We have found the following algorithm for maximizing  $L$  with a continuous response to be quite stable and fast. Let  $\mathbf{B} = (\boldsymbol{\mu}^T, \text{vec}(\boldsymbol{\beta})^T)^T$  and let  $\mathbf{Z}_{y_i j} = (\mathbf{e}_j^T, \mathbf{f}_{y_i}^T \otimes \boldsymbol{\gamma}_j^T)^T$ , where  $\mathbf{e}_j \in \mathbb{R}^p$  is a vector of zeros except for a 1 in the  $j$ -th position. Then log likelihood (5) can be written as

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\beta}) &= \sum_{i=1}^n \sum_{j=1}^p \{ \log\{a_j(\mu_j + \boldsymbol{\gamma}_j^T \boldsymbol{\beta} \mathbf{f}_{y_i})\} + X_{y_i j}(\mu_j + \boldsymbol{\gamma}_j^T \boldsymbol{\beta} \mathbf{f}_{y_i}) \} + C \\ &= \sum_{i=1}^n \sum_{j=1}^p \{ \log\{a_j(\mathbf{Z}_{y_i j}^T \mathbf{B})\} + X_{y_i j} \mathbf{Z}_{y_i j}^T \mathbf{B} \} + C. \end{aligned} \quad (7)$$

With a starting value for  $\boldsymbol{\Gamma}$ , an algorithm for maximizing  $L(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\beta})$  is as follows:

1. Perform the possibly mixed-family regression of  $X_{y_i j}$  on  $\mathbf{Z}_{y_i j}$  through the origin. The first  $p$  elements of the resulting coefficient vector is the current estimate  $\hat{\boldsymbol{\mu}}$  of  $\boldsymbol{\mu}$ , the last  $dr$  elements give the current estimate  $\text{vec}(\hat{\boldsymbol{\beta}})$  of  $\text{vec}(\boldsymbol{\beta})$ .
2. Maximize  $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\Gamma}, \hat{\boldsymbol{\beta}})$  over the Grassmann manifold  $\mathcal{G}_{(d,p)}$ , and let  $\hat{\boldsymbol{\Gamma}}$  denote a resulting semi-orthogonal basis (Liu, Srivastava and Gallivan, 2004).
3. Return to step 1 using the current estimates as starting values until convergence.

### 3.2 Categorical responses

When  $Y$  is categorical with finite sample space  $S_Y = \{v_1, \dots, v_h\}$  we set  $\boldsymbol{\beta} = (\boldsymbol{\nu}_{v_1}, \dots, \boldsymbol{\nu}_{v_h})$  and choose  $\mathbf{f}_{v_i}$  to be an indicator vector with a 1 in the  $i$ -th position and 0's elsewhere. In effect  $Y$  is used only to identify subpopulations. The log likelihood for this setting can be written as

$$L(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\beta}) = \sum_{i=1}^h L_i(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\nu}_{v_i}) + C, \quad (8)$$

where  $C$  does not depend on the parameters and

$$L_i(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\nu}_{v_i}) = n_i \sum_{j=1}^p [\log\{a_j(\mu_j + \boldsymbol{\gamma}_j^T \boldsymbol{\nu}_{v_i})\} + \bar{X}_{v_i j \bullet}(\mu_j + \boldsymbol{\gamma}_j^T \boldsymbol{\nu}_{v_i})].$$

This log likelihood is to be maximized over  $\boldsymbol{\mu} \in \mathbb{R}^p$ ,  $\boldsymbol{\nu}_{v_i} \in \mathbb{R}^d$ ,  $i = 1, \dots, h$ , and  $\mathcal{S}_{\boldsymbol{\Gamma}} \in \mathcal{G}_{(d,p)}$ .

When  $h \ll n$  the maximization can be carried out by using an alternating algorithm:

Beginning with starting values for  $\boldsymbol{\mu}$  and  $\boldsymbol{\Gamma}$ ,

1. For  $i = 1, \dots, h$ , maximize  $L_i(\boldsymbol{\mu}, \boldsymbol{\Gamma}, \boldsymbol{\nu}_{v_i})$  over  $\boldsymbol{\nu}_{v_i}$ . For each  $v_i$  this corresponds to fitting a possibly mixed-family generalized linear model with known offsets  $\boldsymbol{\mu}_j$ , “predictors”  $\boldsymbol{\gamma}_j$  and regression coefficient  $\boldsymbol{\nu}_{v_i}$ . Let  $\hat{\boldsymbol{\nu}}_{v_i}$  denote the value of  $\boldsymbol{\nu}_{v_i}$  that maximizes  $L_i$ .
2. For  $j = 1, \dots, p$ , find

$$\hat{\mu}_j = \arg \max_{\mu_j} \sum_{i=1}^h n_i [\log\{a_j(\mu_j + \boldsymbol{\gamma}_j^T \hat{\boldsymbol{\nu}}_{v_i})\} + \bar{X}_{v_i j \bullet}(\mu_j + \boldsymbol{\gamma}_j^T \hat{\boldsymbol{\nu}}_{v_i})].$$

This corresponds to fitting  $p$  single-family generalized linear models with known offsets  $\boldsymbol{\gamma}_j^T \hat{\boldsymbol{\nu}}_{v_i}$  and intercepts  $\mu_j$ .

3. Maximize  $L(\hat{\boldsymbol{\mu}}, \boldsymbol{\Gamma}, \hat{\boldsymbol{\beta}})$  over  $\mathcal{G}_{(d,p)}$ . Let  $\hat{\boldsymbol{\Gamma}}$  denote a resulting semi-orthogonal basis.

4. If  $L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\beta}})$  has not converged, return to step 1 using the current estimates as starting values.

This algorithm works well because the effect of replication is to smooth the variation in the observed  $\mathbf{X}_y$ 's with their averages  $\bar{\mathbf{X}}_{v_i}$  over the replicates. With sufficiently large  $n$ ,  $\bar{\mathbf{X}}_{v_i}$  will be approximately normal, yielding a stable algorithm.

## 4 Inference

The inference issues that would normally be associated with analyses based on a generalized PFC model include inference on  $d$ , predictor tests and predictor selection. We next address these issues. Since the likelihood function effectively replaces the kernel matrix in the model-free dimension reduction, our proposed inference methods are likelihood based. Additionally, we also propose an estimator of  $E(Y|\mathbf{X})$  without explicitly developing a forward regression model.

### 4.1 Inference on $d$

The dimension  $d = \dim(\mathcal{S}_{\mathbf{r}})$  is essentially a model selection parameter and we have obtained good results using an information criterion as the basis of its estimation. These estimates are of the form

$$\hat{d} = \arg \min_m \{-2L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Gamma}}, \hat{\boldsymbol{\beta}}|d = m) + \phi(n)g(m)\}, \quad (9)$$

where the minimum is computed over  $m \in \{0, 1, \dots, p\}$ , and the conditioning in  $L(\cdot|d = m)$  indicates that the estimates are computed with  $d = m$ . In the penalty  $\phi(n)g(m)$ ,  $\phi(n) = 2$  for AIC,  $\phi(n) = \log(n)$  for BIC and  $g(m)$  denotes the number of real parameters used in fitting with  $d = m$ :  $g(m) = p + mr + m(p - m)$ . This count arises because  $\boldsymbol{\mu} \in \mathbb{R}^p$

and  $\boldsymbol{\beta} \in \mathbb{R}^{m \times r}$  are unconstrained, and identifying a single subspace in  $\mathcal{G}_{(m,p)}$  requires specification of  $m(p - m)$  real numbers.

Our ability to use the likelihood as a basis for hypothesis tests on  $d$  is limited because models with different values for  $d$  are not necessarily nested. However, all models with  $d < \min(p, r)$  are properly nested within the full model with  $d = \min(p, r)$ , and may be tested against it by using a likelihood ratio statistic. Thus, with  $m < \min(p, r)$ , the hypotheses  $H_0 : d = m$  versus  $H_1 : d > m$  can be tested by using the likelihood ratio statistic

$$\Lambda_m \equiv -2\{L(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\beta}}|d = m) - L(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Phi}}|d = \min(p, r))\}, \quad (10)$$

which has an asymptotic chi-squared distribution with  $(p - m)(r - m)$  degrees of freedom under  $H_0$ . A series of such tests can be used to construct another estimate of  $d$ : Starting with  $m = 0$ , test the hypothesis  $H_0 : d = m$ . If the hypothesis is rejected, increment  $m$  by one and test again. The estimate  $\widehat{d}$  of  $d$  is then the value of  $m$  that gives the first nonsignificant result. If  $d < m$  then asymptotically the likelihood ratio test of  $H_0 : d = m$  will have power 1, implying that all of the estimation error comes from the common level  $\alpha$  of the test. Asymptotically this procedure should then give  $\widehat{d} = d$  with probability  $1 - \alpha$  and  $\widehat{d} > d$  with probability  $\alpha$ . Bonferroni's inequality could be used to control the global error rate.

## 4.2 Inference about predictors

The next proposition characterizes conditional independence hypotheses for predictor tests in terms of model (4):

**Proposition 2** *Partition  $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$ , where  $\mathbf{X}_k \in \mathbb{R}^{p_k}$ ,  $k = 1, 2$ , and conformably partition  $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1^T, \boldsymbol{\Gamma}_2^T)^T$ . Then under model (4),  $Y \perp\!\!\!\perp \mathbf{X}_2 | \mathbf{X}_1$  if and only if  $\boldsymbol{\Gamma}_2 = 0$ .*

Accordingly, given  $d$ , we can test a conditional independence hypothesis  $Y \perp\!\!\!\perp \mathbf{X}_2 | \mathbf{X}_1$  by testing if the corresponding sub-matrix  $\mathbf{\Gamma}_2 \in \mathbb{R}^{p_2 \times d}$  of  $\mathbf{\Gamma}$  equals 0. This hypothesis is clearly false if  $p_1 < d$ , so for meaningful hypotheses we must have  $p_1 \geq d$ ; equivalently,  $p_2 \leq p - d$ . The likelihood ratio then provides a straightforward test statistic:

$$\Lambda_{\mathbf{\Gamma}_2} \equiv -2\{L(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Gamma}}, \hat{\boldsymbol{\beta}} | \mathbf{\Gamma}_2 = 0, d) - L(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Gamma}}, \hat{\boldsymbol{\beta}} | d)\}, \quad (11)$$

which is distributed asymptotically as a chi-squared random variable with  $dp_2$  degrees of freedom when  $\mathbf{\Gamma}_2 = 0$ . (We used the same notation for the estimators in (10) and (11), although they are different estimators since they are computed under different hypotheses.) The algorithm sketched in Section 3 can be modified straightforwardly to fit under the null hypothesis that  $\mathbf{\Gamma}_2 = 0$ . This involves replacing  $\mathbf{\Gamma}$  with  $\mathbf{\Gamma}_s = (\mathbf{\Gamma}_1^T, 0)^T$  in step 1, leading to maximization of  $L(\hat{\boldsymbol{\mu}}, \mathbf{\Gamma}_s, \hat{\boldsymbol{\beta}} | \mathbf{\Gamma}_2 = 0)$  over  $\mathcal{S}_{\mathbf{\Gamma}_1} \in \mathcal{G}_{(d, p_1)}$  in step 2. Here the subscript  $s$  indicates that a subset  $\mathbf{\Gamma}_2$  of  $p_2$  rows of  $\mathbf{\Gamma}$  has been set to 0.

Predictor selection can be based on minimizing the following information criterion over subsets of size  $p_2$  not greater than  $p - d$ :

$$-2L(\boldsymbol{\mu}, \hat{\mathbf{\Gamma}}_s, \hat{\boldsymbol{\beta}} | \mathbf{\Gamma}_2 = 0) + \phi(n)g(p_2), \quad (12)$$

where  $g(p_2) = p + dr + d(p - d - p_2)$  and as before  $\phi(n) = 2$  for AIC and  $\phi(n) = \log n$  for BIC. This criterion assumes that  $d$  is known, which may be estimated by using the inference tools described in Section 4.1.

In cases where  $d$  is unknown, as will generally happen when initial predictor screening is desirable, the following full-model criterion may be useful, even when  $d < \min(p, r)$ . Partition  $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1^T, \boldsymbol{\Phi}_2^T)^T$  to correspond to the partition  $\mathbf{\Gamma} = (\mathbf{\Gamma}_1^T, \mathbf{\Gamma}_2^T)^T$ . Then, since  $\text{rank}(\boldsymbol{\beta}) = d$ ,  $\boldsymbol{\Phi}_2 = 0$  if and only if  $\mathbf{\Gamma}_2 = 0$  regardless of the value of  $0 < d \leq \min(p, r)$ .

Consequently, we can test  $\Gamma_2 = 0$  by testing  $\Phi_2 = 0$  with the likelihood ratio statistic  $\Lambda_{\Phi_2} \equiv -2\{L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Phi}}|\Phi_2 = 0) - L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Phi}})\}$ . This statistic, which is based on a fit of the full model, is distributed asymptotically as a chi-squared random variable with  $rp_2$  degrees of freedom. A disadvantage of this approach is that it may be relatively inefficient when  $d$  is small. Predictor selection without knowledge of  $d$  can also be based on minimizing the information criterion  $-2L(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Phi}}_s) + \phi(n)g(p_2)$ , where now  $g(p_2) = p + (p - p_2)r$ , and  $\Phi_s$  is defined as  $\Gamma_s$  is defined.

### 4.3 Estimating $E(Y|\mathbf{X})$

Estimates of the forward mean function  $E(Y|\mathbf{X})$  are often of interest when  $Y$  is quantitative. Since  $E(Y|\mathbf{X}) = E\{Y|R(\mathbf{X})\}$ , an estimate of the mean function could be constructed by building a model for the regression of  $Y$  on the estimated reduction  $\hat{R}$ . This has the advantage of providing an explicit function for the mean, but it may require substantial additional effort. Alternatively, an estimate of  $E\{Y|R(\mathbf{X})\}$  can be constructed from the inverse regression without explicitly building a forward model by noting that:

$$E(Y|\mathbf{X} = \mathbf{x}) = E\left\{\prod_{j=1}^p Y f_j(x_j|Y)\right\} / E\left\{\prod_{j=1}^p f_j(x_j|Y)\right\},$$

where  $f_j$  is as defined in (1) and  $\mathbf{x} = (x_j)$ . This can be estimated consistently by using its sample version which is a weighted average of the observed responses  $\{y_i\}$ :

$$\begin{aligned} \hat{E}(Y|\mathbf{X} = \mathbf{x}) &= \sum_{i=1}^n y_i \hat{w}_i, \\ \hat{w}_i &= \frac{\prod_{j=1}^p a_j(\hat{\eta}_{y_{ij}}) \exp(x_j \hat{\eta}_{y_{ij}})}{\sum_{i=1}^n \prod_{j=1}^p a_j(\hat{\eta}_{y_{ij}}) \exp(x_j \hat{\eta}_{y_{ij}})}, \end{aligned} \tag{13}$$

and  $\hat{\eta}_{y_{ij}}$  is the estimator of  $\eta_{y_{ij}}$  coming from the inverse model. For example, in the context of the generalized PFC model we have from (4) that  $\hat{\eta}_{y_{ij}}$  is the  $j$ -th element

of  $\widehat{\boldsymbol{\mu}} + \widehat{\boldsymbol{\Gamma}}\widehat{\boldsymbol{\beta}}\mathbf{f}_{y_i}$ . Moreover, estimating  $E(Y|\mathbf{X} = \mathbf{x})$  for each value  $\mathbf{x}_i$  of  $\mathbf{X}$  in the data gives residuals  $y_i - \widehat{E}(Y|\mathbf{X} = \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , which can in turn be used in graphical diagnostic checks of the forward mean function that is implied by the inverse model.

## 5 Numerical Studies

In this section we report simulation results to provide numerical support for the theoretical and computational conclusions stated previously, and some qualitative information on the finite sample performance of the proposed generalized PFC approach.

### 5.1 Estimation of $\mathcal{S}_{\boldsymbol{\Gamma}}$

We first examine PFC in terms of estimation accuracy of  $\mathcal{S}_{\boldsymbol{\Gamma}}$ , assuming that  $d$  is known. We start with a regression with all binary predictors. Specifically, the response  $Y$  was generated as a normal random variable with mean 0 and variance  $\sigma_Y^2$ , and  $\mathbf{X}|(Y = y)$  was generated as a Bernoulli random vector with the natural parameter  $\boldsymbol{\eta}_y \in \mathbb{R}^p$  and the canonical link function  $g$ ,

$$\boldsymbol{\eta}_y = g\{E(\mathbf{X}|Y = y)\} = \boldsymbol{\Gamma}\boldsymbol{\beta}\mathbf{f}_y. \quad (14)$$

We considered two cases. First we took  $d = 1$ ,  $p = 20$ ,  $\boldsymbol{\Gamma} = (1, \dots, 1, 0, \dots, 0)^T / \sqrt{10}$  with 10 ones and 10 zeros,  $\boldsymbol{\beta} = 1$ , and  $\mathbf{f}_y = y$ . In the second case,  $d = 2$ ,  $p = 20$ ,  $\boldsymbol{\Gamma} = ((1, \dots, 1, 0, \dots, 0)^T, (0, \dots, 0, 1, \dots, 1)^T) / \sqrt{10}$  with 10 ones in each direction,  $\boldsymbol{\beta} = \text{diag}(1, 0.1)$ , and  $\mathbf{f}_y = (y, y^2)^T$ . We fixed the sample size at  $n = 200$ . The data were fitted by using the method outlined in Section 3, with an intercept  $\boldsymbol{\mu}$  and  $\mathbf{f}_y = (y, y^2, y^3)^T$ . Thus the  $\mathbf{f}_y$  used for fitting was “larger” than the  $\mathbf{f}_y$  used to generate the data. The first two rows of Table 1 give the average angles in degree between  $\mathcal{S}_{\boldsymbol{\Gamma}}$  and  $\mathcal{S}_{\widehat{\boldsymbol{\Gamma}}}$  out of

Table 1: Average angles between  $\mathcal{S}_\Gamma$  and  $\mathcal{S}_{\hat{\Gamma}}$  as  $\sigma_Y$  varies in simulation model (14).

$\sigma_Y$	1	2	3	4	5	10
all binary predictors ( $d = 1$ )	31.92	18.39	13.48	10.71	9.61	8.39
all binary predictors ( $d = 2$ )	74.66	56.00	33.39	24.94	20.75	13.40
mixed predictors ( $d = 1$ )	22.09	11.79	8.59	6.99	6.21	9.10
mixed predictors ( $d = 2$ )	75.35	37.82	20.99	14.12	11.47	6.79

50 data replications, as  $\sigma_Y$ , the term that controls the “signal” strength, increases. For comparison, the mean angle between a fixed direction and a randomly chosen direction in  $\mathbb{R}^{20}$  is about 80 degrees. It is clearly seen that the estimation accuracy of PFC improved quickly as the signal strengthened.

We also considered a regression with a mixture of categorical and continuous predictors. We still adopted simulation model (14) with  $p = 20$ . For the  $d = 1$  case, we took  $\Gamma = (1, \dots, 1, 0.5, \dots, 0.5, -0.5, \dots, -0.5, -1, \dots, -1)^T / \sqrt{12.5}$ , where the first five predictors followed a Bernoulli distribution, the next five followed a Poisson, and the last ten predictors were normal with standard deviation one. For  $d = 2$ , we chose  $\Gamma = ((1, \dots, 1, 0, \dots, 0)^T, (0, \dots, 0, 1, \dots, 1)^T) / \sqrt{10}$ ,  $\beta = \text{diag}(0.5, 0.1)$ , where the first five predictors were binary and the remaining 15 were normal with standard deviation one. The rest setup is the same as before. The average angles over 50 replications for both cases are reported in the last two rows of Table 1. Again it is seen that PFC performs well in the presence of both categorical and continuous predictors. The increase in the angle at  $\sigma_Y = 10$  in the third row of Table 1 was caused by the tendency of the Poisson to occasionally give observations that are orders of magnitude larger than the rest of the data for very large signals. Robust versions of our estimates might mitigate such behavior, but are outside the scope of this report.

To further evaluate the effects of the modeled  $\mathbf{f}_y$  on the performance of PFC, we adopted model (14), but generated the response  $Y$  as a uniform random variable ranging

between 0 and 4, and set  $\mathbf{f}_y = \exp(y)$ . We then fitted PFC with the working  $\mathbf{f}_y = (y, y^2, \dots, y^k)^T$ , and  $k$  varied from 1 to 5. For this working  $\mathbf{f}_y$ ,  $\beta$  is a  $1 \times k$  vector, while  $\beta$  is a scalar for the true  $\mathbf{f}_y$ . We stopped at  $k = 5$  since a fifth degree polynomial provides a close approximation to  $\exp(y)$  for  $y \in (0, 4)$ . We took the all binary predictors case with  $d = 1$  and  $n = 200$ . Figure 1 shows boxplot of the angles between the true and the estimated  $\Gamma$  for each value of  $k$ . The last boxplot shows the results for fitting the true  $\mathbf{f}_y = \exp(y)$ . The estimation accuracy for  $k = 2$  is noticeably better than that for  $k = 1$ , but there seems little to distinguish the other settings. We note also that the median performance for  $k = 1$  was reasonable, although there are more extreme deviations. In general, as long as the canonical correlations between the working  $\beta\mathbf{f}_y$  and its true counterpart are sufficiently high, we expect PFC to work well. Meanwhile we emphasize that graphical methods of Cook (1998) can be employed to aid the selection of a reasonable  $\mathbf{f}_y$ .

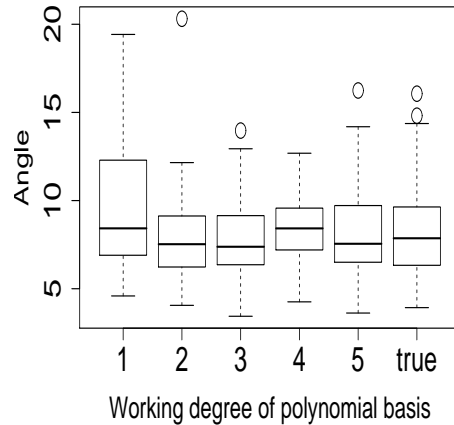


Figure 1: Boxplot of angles between  $\mathcal{S}_\Gamma$  and  $\mathcal{S}_{\hat{\Gamma}}$  as the degree of polynomial basis in the working  $\mathbf{f}_y$  increases.

## 5.2 Prediction and estimation of $E(Y|\mathbf{X})$

We next evaluate prediction and mean function estimation as proposed in Section 4.3. We focused our attention on the all binary predictors case as an illustration. We generated  $n = 200$  observations according to (14) and then obtained the PFC-based estimate of  $\widehat{E}(Y|\mathbf{X} = \mathbf{x})$  following (13). We next generated 100 additional observations  $(\mathbf{X}_l, Y_l)$  and estimated the scaled prediction error as

$$\widehat{\text{PE}}^2(\sigma_Y) \equiv \sum_{l=1}^{100} \frac{\{Y_l - \widehat{E}(Y|\mathbf{X}_l)\}^2}{100\sigma_Y^2},$$

This procedure was repeated 50 times for each selected value of  $\sigma_Y$ , and the average of the resulting 50 scaled prediction errors were plotted (solid line). As a comparison, the results based on the true parameter values and thus the true weights  $w_i$  are also plotted (dashed line). Additionally, when using the sample mean for prediction,  $\widehat{E}(Y|\mathbf{X}_l) = \bar{Y}$ ,  $E\{\widehat{\text{PE}}^2(\sigma_Y)\} = 1.005$ . Figure 2 shows the results for  $d = 1$  and  $d = 2$  respectively. It is seen from the plot that the PFC-based prediction is quite accurate and is close to using the true weights in prediction.

## 5.3 Estimation of $d$

We next consider AIC and BIC as methods for estimating  $d = \dim(\mathcal{S}_{\mathbf{F}})$ . Our goal here is to demonstrate that an information criterion can be useful for estimation of  $d$ ; we do not address the issue of selecting a “best” criterion for the models proposed. The simulation setup of (14) was employed, using  $d = 1$  and  $d = 2$  and a series of  $\sigma_Y$  values. Table 2 reports the percentage of the times out of 50 replications that the estimate  $\hat{d}$  takes value among  $\{0, 1, 2, 3\}$ . The PFC estimate of  $d$  as given in (9) became increasingly accurate as the the signal strength, as controlled by  $\sigma_Y$ , increased.

Tests of dimension hypotheses based on PFC, as sketched in Section 4.1 (10), may

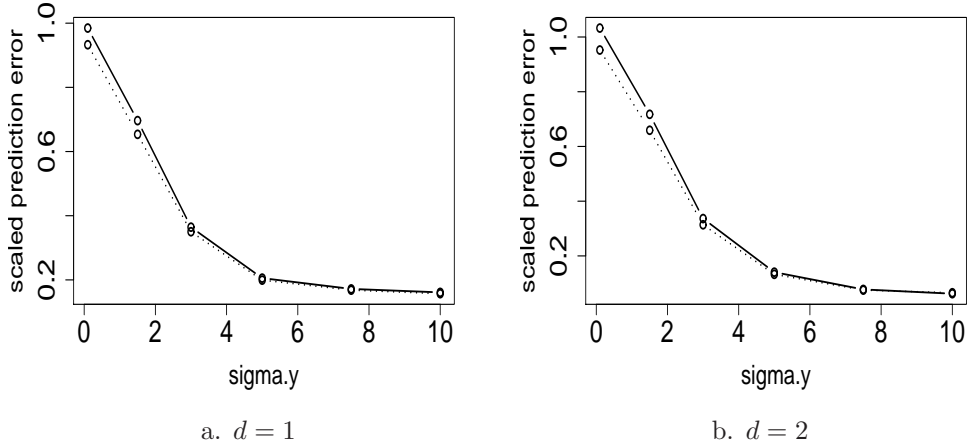


Figure 2: Prediction error for simulation model (14). The solid line is for generalized PFC and the dashed line is for the true weights.

also be employed to estimate  $d$  in a sequential manner. Although such tests may be useful from time to time for different practical reasons, we prefer information criteria (9) because they have well-known properties and they avoid the choice of a nominal level and multiple testing issues. Nevertheless, we examined the actual level and the power of the proposed chi-squared test by comparing the test statistic  $\Lambda_m$  to a chi-squared distribution with  $(p - m)(r - m)$  degrees of freedom and thereby obtaining a p-value. We repeated this process 1000 times. For model (14) with  $d = 1$  and  $\sigma_Y = 4$ , we found the test for hypotheses  $d = 0$  versus  $d = 1$  yielded an observed power of 1. For model (14) with  $d = 2$  and  $\sigma_Y = 4$ , the observed power of the test of  $d = 1$  versus  $d = 2$  was again 1. Figure 3a shows the empirical cumulative distribution function of p-values for  $d = 2$  versus  $d > 2$ . The results agree well with our conclusion in Section 4.1 that the proposed test statistic has a chi-squared distribution.

Table 2: Estimation of  $d = \dim(\mathcal{S}_\Gamma)$  based on (9) for simulation model (14).

	$\hat{d}$	$d = 1$				$d = 2$			
		0	1	2	3	0	1	2	3
$\sigma_Y = 1$	AIC	0.02	<b>0.92</b>	0.06	0.00	0.02	0.98	<b>0.00</b>	0.00
	BIC	1.00	<b>0.00</b>	0.00	0.00	1.00	0.00	<b>0.00</b>	0.00
$\sigma_Y = 2$	AIC	0.00	<b>0.96</b>	0.04	0.00	0.00	0.66	<b>0.34</b>	0.00
	BIC	0.00	<b>1.00</b>	0.00	0.00	0.00	1.00	<b>0.00</b>	0.00
$\sigma_Y = 3$	AIC	0.00	<b>0.96</b>	0.04	0.00	0.00	0.00	<b>0.98</b>	0.02
	BIC	0.00	<b>1.00</b>	0.00	0.00	0.00	0.78	<b>0.22</b>	0.00
$\sigma_Y = 4$	AIC	0.00	<b>0.92</b>	0.08	0.00	0.00	0.00	<b>1.00</b>	0.00
	BIC	0.00	<b>1.00</b>	0.00	0.00	0.00	0.02	<b>0.98</b>	0.00
$\sigma_Y = 5$	AIC	0.00	<b>0.96</b>	0.04	0.00	0.00	0.00	<b>1.00</b>	0.00
	BIC	0.00	<b>1.00</b>	0.00	0.00	0.00	0.00	<b>1.00</b>	0.00
$\sigma_Y = 10$	AIC	0.00	<b>1.00</b>	0.00	0.00	0.00	0.00	<b>1.00</b>	0.00
	BIC	0.00	<b>1.00</b>	0.00	0.00	0.00	0.00	<b>1.00</b>	0.00

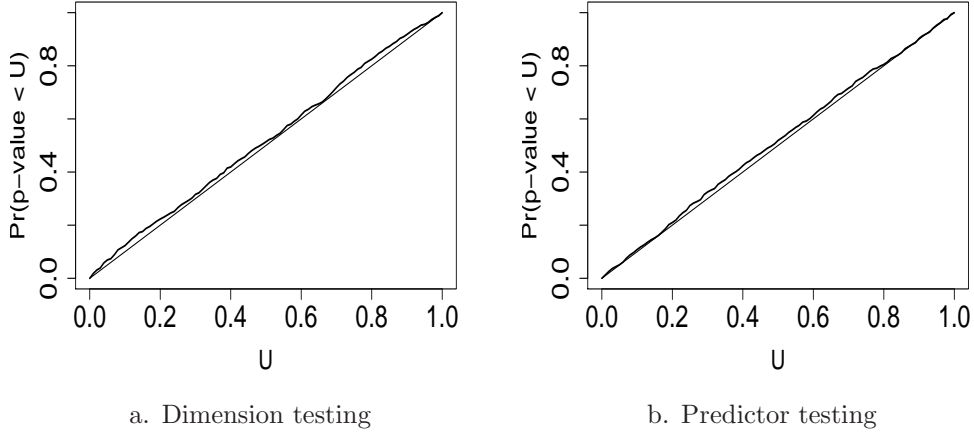


Figure 3: Empirical distributions of p-values for (a) testing  $d = 2$  vs.  $d > 2$  with statistic (10) in simulations from model (14) with  $d = 2$ , and (b) testing the last 10 predictors with statistic (11) in simulations from (14) with  $\Gamma_{(1)}$ .

## 5.4 Predictor selection

We illustrate our simulation results on the predictor selection methods of Section 4.2 by using model (14) with known  $d = 1$  and two  $\Gamma$ 's:  $\Gamma_{(1)} = (1, \dots, 1, 0, \dots, 0)^T / \sqrt{10}$

Table 3: Predictor selection based on (12) for simulation model (14) with two forms of  $\mathbf{\Gamma}$ . Evaluation criteria include the angle (in degree) between the estimated direction  $\hat{\mathbf{\Gamma}}$  and the true  $\mathbf{\Gamma}$  (ANG), the number of non-zero components in  $\hat{\mathbf{\Gamma}}$  (NUM), the true positive rate (TPR), and the false positive rate (FPR).

		$\mathbf{\Gamma}_{(1)}$				$\mathbf{\Gamma}_{(2)}$			
		ANG	NUM	TPR	FPR	ANG	NUM	TPR	FPR
$\sigma_Y = 4$	-	11.29	20.00	1.00	1.00	11.11	20.00	1.00	-
	AIC	9.69	10.82	1.00	0.08	11.11	20.00	1.00	-
	BIC	9.21	10.00	1.00	0.00	11.11	20.00	1.00	-
$\sigma_Y = 10$	-	8.74	20.00	1.00	1.00	8.12	20.00	1.00	-
	AIC	8.58	11.20	1.00	0.12	8.12	20.00	1.00	-
	BIC	8.49	10.02	1.00	0.00	8.12	20.00	1.00	-

and  $\mathbf{\Gamma}_{(2)} = (1, \dots, 1, 0.5, \dots, 0.5, -0.5, \dots, -0.5, -1, \dots, -1)^T / \sqrt{12.5}$  with each element repeated 5 times. Only the first 10 predictors are active in  $\mathbf{\Gamma}_{(1)}$ , but all predictors are active in  $\mathbf{\Gamma}_{(2)}$ . For each case, two  $\sigma_Y$  values were examined. To evaluate the selection accuracy, the following criteria were computed: the angle (in degree) between the estimated direction  $\hat{\mathbf{\Gamma}}$  and the true  $\mathbf{\Gamma}$ , the number of non-zero components in  $\hat{\mathbf{\Gamma}}$ , the true positive rate, and the false positive rate. Table 3 reports the average of the above criteria based on 50 replications. For each setup, the first row reports results with no predictor selection, and the next two rows report results based on AIC and BIC as specified in (12). Again we found that PFC-based selection criteria worked well. The basis estimate was accurate, yielding small average angle between  $\hat{\mathbf{\Gamma}}$  and  $\mathbf{\Gamma}$ . The active predictors were always selected, resulting in an average true positive rate equal to one. In addition, BIC achieved a very low false positive rate, while AIC selected more false positives.

PFC-based conditional independence hypotheses test, as sketched in Section 4.2, may also be used to infer predictor contributions. For model (14) with  $\mathbf{\Gamma}_{(1)}$  and  $\mathbf{\Gamma}_{(2)}$ , we focused on the null hypothesis that the last ten rows of  $\mathbf{\Gamma}$  are all zeros. Figure 3b shows the empirical distributions of the p-values out of 1000 replications for  $\mathbf{\Gamma}_{(1)}$ . Again the

chi-squared distribution provided a good approximation. In addition, the observed power of the test for  $\Gamma_{(2)}$  was 1.

## 5.5 Zoo data

To compliment our use of a continuous response in the simulations and facilitate the presentation of graphics, we present a brief analysis of a regression with a categorical response.

The zoo data consist of 101 animals classified into 7 categories: amphibian, bird, fish, insect, invertebrate, mammal, and reptile. The animals are distributed unevenly among the classes, the number of instances in the respective classes being 4, 20, 13, 8, 10, 41, and 5. Sixteen categorical predictors were measured on each animal, including hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic, cat-size, and legs. The first 15 predictors are dichotomous, and the leg variable is polychotomous. Aiming to differentiate the 7 groups of animals, we applied PFC and treated the first 15 predictors as Bernoulli and the last predictor as Poisson. The information criteria both estimated  $d = 3$ , and predictor inference suggested that all 16 attributes are important to distinguish the species. Figure 4 shows two views in two-dimensional space of the three-dimensional summary plot of  $\hat{\Gamma}^T \mathbf{X}$  by animal classes. Except for amphibians (class 1) and reptiles (class 7), all species seem well-separated in Figure 4(a), while amphibians and reptiles are separated in Figure 4(b). One invertebrate (class 5) is mixed with insects (class 4). This invertebrate is the scorpion which, like all insects, is an arthropod. In Figure 4(b) there is a reptile (class 7) that is mixed with the fish (class 3). It turns out that the reptile is a sea snake, and biologically it is not far from fish. Since the generalized PFC methodology is designed to estimate only the subspace  $\mathcal{S}_{\Gamma}$  we generally do not attempt to interpret or reify the PFCs  $\hat{\Gamma}^T \mathbf{X}$  without a

firm substantive context or a forward model. Of course, reification is not necessary for prediction.

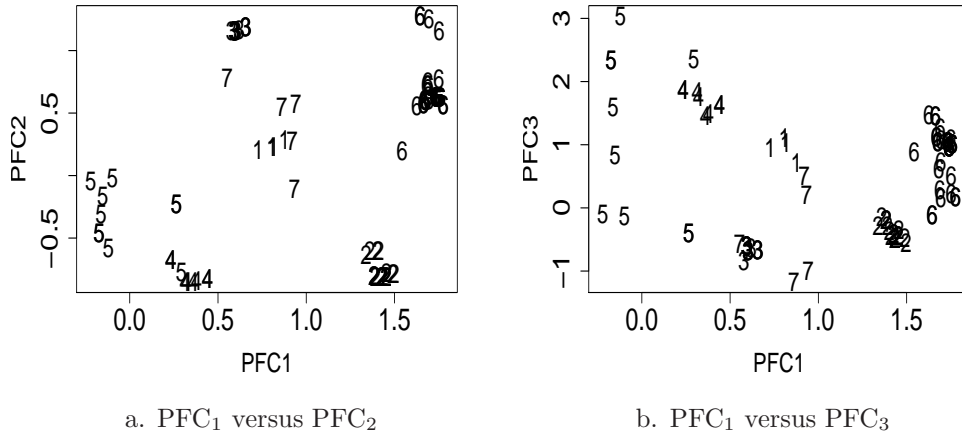


Figure 4: Zoo data: two 2D views of the 3D summary plot of  $\hat{\Gamma}^T \mathbf{X}$  by animal classes. Numbers 1 to 7 correspond to the 7 species, amphibian, bird, fish, insect, invertebrate, mammal, and reptile, respectively.

## 6 Conditionally Dependent Predictors

Our main goal in this section is to demonstrate that the methodology based on conditional independence can still be useful when the predictors are conditionally dependent. To aid in this demonstration, we first give a minimal sufficient reduction under a model that allows dependence and a corresponding moment-based method of estimation.

Assume that  $\mathbf{X}|Y$  follows a quadratic exponential model as studied by Gouriéroux, Montfort and Trognon (1984) and by Prentice and Zhao (1991):

$$f(\mathbf{x}|\boldsymbol{\theta}_y, Y = y) = a(\boldsymbol{\theta}_y, \boldsymbol{\lambda})b(\mathbf{x}) \exp(\mathbf{x}^T \boldsymbol{\theta}_y + \mathbf{w}^T \boldsymbol{\lambda}), \quad (15)$$

where  $\mathbf{w} = \text{vech}(\mathbf{xx}^T)$ , “vech” is the vector-half operator that maps the unique elements of a symmetric  $p \times p$  matrix to  $\mathbb{R}^{p(p+1)/2}$ ,  $\boldsymbol{\lambda} \in \mathbb{R}^{p(p+1)/2}$  is a vector of parameters that may be modified depending on the family. For instance, if  $\mathbf{x}$  consists of binary variables  $x_j$  taking values 0 and 1 then the elements of  $\boldsymbol{\lambda}$  corresponding to the quadratic terms  $x_j^2$  should be set to 0 since  $x_j = x_j^2$ . Let the columns of the  $p \times d$  matrix  $\Theta$  be a basis for  $\mathcal{S}_\Theta = \text{span}\{\boldsymbol{\theta}_y - \bar{\boldsymbol{\theta}}|y \in S_Y\}$ , where  $\bar{\boldsymbol{\theta}} = \text{E}(\boldsymbol{\theta}_Y)$ . Under this model we have

**Proposition 3** *Let  $R(\mathbf{X}) = \Theta^T \mathbf{X}$  and let  $T(\mathbf{X})$  be any sufficient reduction. Then, under model (15),  $R$  is a sufficient reduction and  $R$  is a function of  $T$ .*

The proof of this proposition, which shows that  $\Theta^T \mathbf{X}$  is a minimal sufficient reduction under model (15), follows that of Proposition 1 and is omitted. Next we propose a relatively straightforward method of estimating  $\mathcal{S}_\Theta$  that uses fits from a working independence model.

Consider fitting the full model with the working assumption that the predictors are conditionally independent and with  $\Phi = \Gamma\beta \in \mathbb{R}^{p \times r}$  unconstrained, even if  $d < \min(p, r)$ . We assume that the fitted mean function is correct, but the predictors may still be conditionally dependent. It follows from Prentice and Zhao (1991) that  $\hat{\Phi}$  is a consistent estimator of  $\Phi$  and that  $\sqrt{n}(\hat{\Phi} - \Phi)$  is asymptotically normal. For this to be useful, we must have a way of turning  $\hat{\Phi}$  into an estimator of  $\mathcal{S}_\Theta$ , which necessitates estimation of the conditional dependence structure. When the response is continuous we estimate this structure nonparametrically using slicing.

Specifically, for a continuous response, partition the range of  $Y$  into  $h$  slices  $H_s$ ,  $s = 1, \dots, h$ . Let  $\Sigma_s = \text{Var}(\mathbf{X}|Y \in H_s)$ , let  $\mathbf{W}_s = \text{Var}(\mathbf{f}_Y|Y \in H_s)$  and let  $\mathbf{G}_s \in \mathbb{R}^{p \times p}$  be a diagonal matrix with diagonal elements  $\dot{g}_j = \partial g_j(\mu_j)/\partial \mu_j$  evaluated at  $\text{E}(\mathbf{X}|Y \in H_s)$ . We assume that  $\boldsymbol{\theta}_y$  is a smooth function of  $y \in H_s$  and that  $\text{Cov}(\theta_{Y_j}, Y|Y \in H_s) \neq 0$ ,  $s = 1, \dots, h$ ,  $j = 1, \dots, p$ . Then for sufficiently narrow slices,  $\mathcal{S}_\Theta$  can be approximated

accurately in population as the span of the eigenvectors corresponding to the nonzero eigenvalues of the kernel matrix  $\mathbf{M} = \text{E}[\boldsymbol{\Sigma}_s^{-1} \mathbf{G}_s^{-1} \boldsymbol{\Phi} \mathbf{W}_s \boldsymbol{\Phi}^T \mathbf{G}_s^{-1} \boldsymbol{\Sigma}_s^{-1}]$ , where the expectation is over slices (see Appendix). A consistent sample version  $\widehat{\mathbf{M}}$  of  $\mathbf{M}$  can be constructed by substituting sample versions of  $\boldsymbol{\Sigma}_s$ ,  $\mathbf{G}_s$  and  $\mathbf{W}_s$ , and replacing  $\boldsymbol{\Phi}$  with its estimate from the fit of the full model under the working assumption of independence. The estimate  $\widehat{\mathcal{S}}_{\Theta}$  of  $\mathcal{S}_{\Theta}$  is then the span of the first  $d$  eigenvectors of  $\widehat{\mathbf{M}}$ . Since inference methods for  $d$  are not yet available, a scree plot may be useful for choosing a specific value of  $d$ .

We conducted a small simulation study to gain qualitative information about the performance of the proposed dependence estimator and the independence estimator under dependence. Data were simulated from the quadratic exponential family (15) with mixtures of 10 normal and 5 Bernoulli variables. We chose  $\boldsymbol{\theta}_y = \boldsymbol{\Gamma} \boldsymbol{\beta} \mathbf{f}_y$ , with  $\boldsymbol{\Gamma} = (1, 1, 1, 1, 1, 0, \dots, 0, 1, 1, 1, 1, 1)^T / \sqrt{10}$ ,  $\boldsymbol{\beta} = 1$ ,  $\mathbf{f}_y = y$ , and  $Y$  was normal with mean 0 and standard deviation 4. The Bernoulli predictors were conditionally dependent, while the normal predictors remain conditionally independent. We used the average odds ratio of all pairs of Bernoulli variables to summarize the strength of association among those variables. We examined four scenarios where the average odds ratio were about 1.07, 0.72, 0.42, and 0.25, so the simulations range from the conditional independence case to relatively strong dependence. Two estimation methods, one based on the conditional independence model in Section 3, and the other based on the conditional dependence model as described in this section, were compared at five sample sizes. The average angles based on 50 data replications are reported in Table 4. As a comparison, the mean angle between a fixed direction and a randomly chosen direction in  $\mathbb{R}^{15}$  is about 78 degrees. We first note that the estimate based on the conditional dependence model performed similarly across all strengths of association, and its accuracy improved as the sample size increased. This agrees with our expectation, since like all other methods that deal with dependence, the method requires a relatively large sample size to estimate

Table 4: Average angles between  $\mathcal{S}_{\Theta}$  and  $\mathcal{S}_{\hat{\Theta}}$  based on the conditional independence model (1) and the conditional dependence model (15).

odds ratio	working model	Sample size $n$				
		100	200	400	800	1000
1.07	(1)	10.06	6.84	4.65	3.38	3.01
	(15)	44.20	30.22	22.50	17.98	16.54
0.72	(1)	14.66	12.25	11.36	10.93	11.03
	(15)	42.66	29.10	22.09	15.74	13.99
0.42	(1)	23.49	21.72	21.22	20.45	20.54
	(15)	45.33	31.45	22.81	16.37	13.91
0.25	(1)	26.94	25.84	25.54	25.02	25.04
	(15)	46.46	34.79	24.37	16.04	15.84

the dependence structure. Secondly, the estimate based on the conditional independence model degraded as the odds ratio decreased, as one would expect. However, this estimator performed reasonably well when the dependency among predictors was moderate, and it performed as well as or better than the estimate based on the dependence model when  $n \leq 400$ . This likely reflects a bias-variance tradeoff. Although the estimate based on the conditional independence model may be biased, it has relatively small variance. As such, the estimation approach assuming the conditional independence model seems to be a reasonable choice as long as the dependence is modest or the sample size is not large.

## 7 Discussion

In this article we have proposed methodology for dimension reduction in regressions with all categorical or mixed continuous and categorical predictors. Our results extend current sufficient dimension reduction development in at least two ways. First, Propositions 1 and 3 show that  $\mathcal{S}_{\Gamma}$  and  $\mathcal{S}_{\Theta}$  are central subspaces. This indicates that the model-free definition of a central subspace extends immediately to some regressions with mixed

predictors. Secondly, the likelihood effectively replaces a kernel matrix and eliminates the need for the linearity condition.

Inverse regression has gained considerable interest and success in recent regression research; see Cook and Ni (2005) and references therein. Notably, inverse regression can exploit regression information from the marginal distribution of  $\mathbf{X}$  that is excluded in forward regression when conditioning on  $\mathbf{X}$ . As such, inverse regression may gain efficiency over forward regression methods in some applications even when the forward model is parametric and known. Given a joint distribution for  $Y$  and  $\mathbf{X}$ , the forward and inverse approaches are connected with sufficient reductions, as defined in this article.

Other dimension reduction methods like SIR (Li, 1991) or MAVE (Xia, Tong, Li and Zhu, 2002) could be applied as they stand in regressions with categorical or mixture predictors, although we know of no theory to support such application for SIR, and we expect that the kernel estimation used in MAVE would not perform satisfactorily with categorical predictors. In some cases, direct application of existing methods may provide about the same numerical estimates as we obtain with PFC, while in other cases there would be relevant differences. But finding predictive conditions for this is nontrivial and outside the scope of this article. Even with such knowledge, special inference procedures would need to be derived for dimension tests and predictor selection, like those in Section 4.

We have focused on the conditional independence model, since this case is an important first step in the development of dimension reduction with exponential family predictors. As we have demonstrated in Section 6, it can yield useful estimates of the sufficient reduction even when the predictors are conditionally dependent. Nevertheless, complete methodology that allows for arbitrary conditional dependence is important and is currently under investigation.

## 8 Supplementary Materials

**Appendix:** This pdf file includes the proofs of Propositions 1 and 2, and the derivation of  $M$ . It also contains the sample R code to call functions written for the methods proposed in this article. The link to the zoo data used in Section 5.5 is given.

**R code:** All the computer code in R can be found in `sdrepx.R`.

## References

- Bura, E. and Cook, R.D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of Royal Statistical Society, Series B.* **63**, 393-410.
- Chiaromonte, F., Cook, R.D. and Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *Annals of Statistics*, **30**, 475–497.
- Cook, R.D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. New York: Wiley.
- Cook, R.D. (2007). Fisher Lecture: Dimension reduction in regression (with discussion). *Statistical Science*, **22**, 1–26.
- 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. (1991). Comment on ‘Sliced inverse regression for dimension reduction’ by K.C. Li. *Journal of the American Statistical Association*, **86**, 328-332.
- Chikuse, Y. (2002). *Statistics on Special Manifolds*. New York: Springer.

- Cox, D.R. and Hinkley, D. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- Fisher, R.J. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Statistical Society, A*, **222**, 309–368.
- Li, B., Cook, R.D. and Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *Annals of Statistics*, **30**, 1636–1668.
- Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, **102**, 997-1008.
- Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Annals of Statistics*, **33**, 1580–1616.
- Li, K-C. (1991), Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316-342.
- Liang, K-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.
- Liquet, B. and Saracco, J. (2007). Pooled marginal slicing approach via  $SIR_\alpha$  with discrete covariables. *Computational Statistics*, **22**, 599–617.
- Liu, X., Srivastava, A. and Gallivan, K. (2004). Optimal linear representations of images for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**, 662–666.
- Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis, *PLoS Genetics*, **3**, 1724-1735.
- Prentice, R.L. and Zhao, L.P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics*, **47**, 825–838.
- Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904-909.

- 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, Series B*, **64**, 363-410.
- Yin, X., and Cook, R.D. (2002). Dimension reduction for the conditional  $k$ th moment in regression. *Journal of Royal Statistical Society, Series B*, **64**, 159-175.
- Yin, X., and Cook, R.D. (2005). Direction estimation in single-index regressions. *Biometrika*, **92**, 371-384.