Sufficient Reductions in Regressions With Exponential Family Inverse Predictors

Efstathia Bura[*], Sabrina Duarte[†] and Liliana Forzani[†]

[*]Department of Statistics, The George Washington University, Washington, DC 20052

[†]Departamento de Matemática, Facultad de Ingeniería Química (UNL) and Instituto Matemática Aplicada Litoral (CONICET-UNL), Santa Fe, Argentina

[1]The author order is alphabetical

Abstract

We develop methodology for identifying and estimating sufficient reductions in regressions with predictors that, given the response, follow a multivariate exponential family distribution. This set-up includes regressions where predictors are all continuous, all categorical or mixtures of categorical and continuous. We derive the minimal sufficient reduction of the predictors and its maximum likelihood estimator by modeling the conditional distribution of the predictors given the response. Whereas nearly all extant estimators of sufficient reductions are linear and only partly capture the sufficient reduction, our method is not limited to linear reductions. It also provides the exact form of the sufficient reduction, which is exhaustive, its ML estimates via an IRLS estimation algorithm, and asymptotic tests for the dimension of the regression.

1 Introduction

The goal of dimension reduction methodology in regression is the reduction of the dimension of the predictor vector $\mathbf{X} \in \mathbb{R}^p$ without loss of information about the response Y. Formally this amounts to finding a function $\mathbf{R}(\mathbf{X}) \in \mathbb{R}^d$, $d \leq p$, such that $F(Y|\mathbf{X}) = F(Y|\mathbf{R}(\mathbf{X}))$, where $F(\cdot|\cdot)$ signifies the conditional cumulative distribution function (cdf) of the first given the second argument. Most of the methodology in Sufficient Dimension Reduction (SDR) is based on the inverse regression of $\mathbf{X}$ on Y. Li (1991) introduced the concept of inverse regression as a dimension reduction tool in Sliced Inverse

Regression (SIR). Cook and his collaborators formalized the field in several papers (e.g. Cook and Weisberg 1991; Cook 1994, 1998b, 2007; Cook and Lee 1999; Bura and Cook 2001a,b; Cook and Yin 2001; Chiaromonte et al. 2002; Cook and Ni 2005; Cook and Forzani 2008, 2009) and a book (Cook 1998a), where much of its terminology was introduced.

The estimation in SDR was originally based on moments or functions of moments of the conditional distribution of $\mathbf{X}|Y$ (SIR, Li 1991; SAVE, Cook and Weisberg 1991; pHd, Li 1992; PIR, Bura and Cook 2001; MAVE, Xia et al. 2002; Li et al. 2005; Cook and Ni 2005; Zhu and Zheng 2006; Cook and Li 2002; DR, Li and Wang 2007). It required continuous predictors, conditions on their moments, and typically captured *only part* of the reduction.

Chiaromonte et al. (2002) made the first attempt to extend the SDR moment-based approach to regressions with both quantitative and categorical predictors. The two types of predictors were treated independently, with the categorical ones, of finite many values, used to define subpopulations for the continuous predictors given the response. Reductions of the continuous predictors were computed for each of the subpopulations separately. The sum of these reductions was shown to be a reduction of the form $(\boldsymbol{\alpha}^T\mathbf{X}, W)$ for the regression of $Y|(\mathbf{X}, W)$, where $\mathbf{X}$ are the continuous and W the categorical predictors, respectively. This results in reductions where the categorical and the continuous predictors do not mix, i.e. they are not allowed to be correlated, and are not necessarily minimal. Their estimation algorithm, partial SIR, computes a reduction in each category so that when the number of categories is large or there are several categorical predictors, substantially large samples are required to yield reliable estimates. Partial SIR also requires homogeneous predictor covariances across the subpopulations, but this was later relaxed in Wen and Cook (2007). We refer to all methods that require moment conditions on the marginal distribution of the predictors, including Chiaromonte et al.'s partial SIR, as *moment-based* SDR.

Cook (2007) connected dimension reduction methodology in regression with the classical statistical concept of sufficiency. He introduced *model-based* inverse regression in order to avoid the restrictive conditions on the marginal distribution of the predictors required in moment-based SDR. The main attraction of model-based approaches is the identification of *sufficient reductions* of the predictors in the sense that they are *exhaustive* for the regression of Y on $\mathbf{X}$; that is, they contain all the information in $\mathbf{X}$ relevant to Y. Inability to ascertain exhaustive estimation of the sufficient reduction was one of the main limitations of moment-based SDR methods. Moreover, maximum likelihood estimators of the sufficient reductions can be obtained that are optimal in terms of efficiency under the exact model.

Cook (2007) and Cook and Forzani (2008) considered the case of normal $\mathbf{X}|Y$ in developing their Principal Fitted Components (PFC) and Likelihood Acquired Directions (LAD) methodology (Cook and Forzani 2009). Assuming normal $\mathbf{X}|Y$ with constant variance, they obtained *minimal* sufficient reductions that were *linear* and *exhaustive* for the regression of Y on $\mathbf{X}$. When the constant variance assumption is not satisfied, though, the linear reduction is no longer necessarily a minimal sufficient reduction (Cook and Forzani 2009).

The first attempt to develop SDR methodology where both categorical and continuous predictors are jointly considered under *model-based* inverse regression was by Cook and Li (2009). Their approach posits that $X_j|Y$ follows a one-parameter exponential family distribution for each coordinate X_j of $\mathbf{X}$ and requires the components be independent given the response Y. Under this model, they obtained a minimal sufficient reduction that is linear in the predictors for the regression of Y on $\mathbf{X}$. This is a specific and restrictive form of conditional independence that limits the applicability of this method in real data analyses. For example, it excludes conditionally independent normal predictors whose both first two moments depend on Y, as we show in Section 3.1. Noorbaloochi and Nelson (2008), whose framework follows Cook's (2007), also considered the problem of finding sufficient reductions in exponential families. They found that the space spanned by the logarithm of the density ratios, $f(\mathbf{x}|Y = y)/f(\mathbf{x}|Y = y_0)$, with y, y_0 varying in the support of Y, provides a linear representation for a minimal dimension reduction subspace.

A defining feature of sufficient reductions in the majority of SDR methods, and in particular in all SDR methods discussed above, is that they are projections of the predictor vector on lower dimensional subspaces. Therefore, sufficient reductions have been by default *linear* in the predictors and SDR methodology has been mostly *linear*.

In this paper, we consider the problem of identifying sufficient reductions in regressions with predictors that can be all continuous, all categorical, or mixtures of continuous and categorical variables. We follow a model-based approach and assume that the distribution of $\mathbf{X}|Y$ is an exponential family, but we require no conditional independence of the inverse predictors or any other condition. We identify the *minimal sufficient reduction* for the regression of Y on $\mathbf{X}$, and show that it is *linear not in the predictors* but in $\mathbf{T}(\mathbf{X})$, the minimal sufficient statistic for the $\mathbf{X}|Y$ exponential family, where Y is considered as a parameter. Depending on the form of $\mathbf{T}(\mathbf{X})$, the minimal sufficient reduction may very well be non-linear in the predictors, as in the case of Bernoulli inverse predictors in Section 7.1.1. Our approach fills in a void in SDR as there were no methods for regressions with only categorical predictors without

imposing unrealistic and untestable conditions.

The linearity of the sufficient reduction in the sufficient statistic $\mathbf{T}(\mathbf{X})$ in this very broad family of regression problems is important because of its simplicity and interpretability. The estimation of the sufficient reduction is done via fitting exponential family regression. We propose an iterated reweighted least squares algorithm (IRLS) to compute the maximum likelihood estimate (MLE) of the sufficient reduction. Asymptotic tests for the dimension of the reduction are derived. Our approach subsumes both Cook and Li (2009) and Noorbaloochi and Nelson (2008) set-ups and related methodologies, and also the normal-based approaches of Principal Fitted Components (Cook and Forzani 2008), Likelihood Acquired Directions (Cook and Forzani 2009) and the normal inverse predictor model in Bura and Forzani (2015), since the normal distribution is the only common member of the exponential and the elliptically contoured families of distributions. Furthermore, our IRLS estimation algorithm is fast and easy to implement and can be used in place of Grassmann optimization for the estimation of sufficient reductions in all model-based SDR approaches.

The rest of the paper is organized as follows. In Section 2 we provide a short review of the multivariate exponential family as it relates to our methodology. Sufficient reductions for general $\mathbf{X}|Y$ exponential family distributions are presented in Section 3 and the minimal sufficient reduction is derived in Section 3.1. Maximum likelihood estimates of the sufficient reductions are derived in Section 4 and asymptotic tests for dimension of the sufficient reduction in Section 5. Section 6 relates and compares our methodology to existing approaches including kernel-based nonlinear dimensionality reduction methods. In Section 7 we turn to the important regression problem with multivariate Bernoulli predictors, which has many applications and wide appeal to the machine learning community. In the same section, a simulation study and a real data analysis are carried out. We conclude in Section 8.

2 Exponential Family

Let $\mathbf{X} = (X_1, \ldots, X_p)^T$ be a p-dimensional random vector with a distribution $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k$, where $\boldsymbol{\Theta}$ is an open and connected subset of $\mathbb{R}^k$. The family of distributions $\{P_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ for $\mathbf{X} = (X_1, \ldots, X_p)^T$ is said to belong to the k-parameter exponential family if the family $\mathcal{F} = \{f(\cdot; \boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ of probability density functions (pdf), or probability mass functions (pmf), of $\mathbf{X}$ are of the form

$$f(\mathbf{x}|\boldsymbol{\theta}) = e^{\boldsymbol{\eta}(\boldsymbol{\theta})^T \mathbf{T}(\mathbf{x}) - \psi(\boldsymbol{\theta})} h(\mathbf{x}) \tag{1}$$

where the natural parameters of the family, $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_k(\boldsymbol{\theta}))^T$, are twice continuously differentiable functions of $\boldsymbol{\theta}$ with Jacobian of full rank, $\mathbf{T}(\mathbf{x}) = (T_1(\mathbf{x}), \ldots, T_k(\mathbf{x}))^T$ is a vector of known real-valued functions, $h(\mathbf{x}) \geq 0$ is a known function, and $\boldsymbol{\psi}(\boldsymbol{\theta})$ is such that $f(\cdot)$ is a proper pdf or pmf. We assume that $\mathbf{T}$ is the minimal complete and sufficient statistic for the family and denote that a random vector follows such distribution with $\mathbf{X} \sim \mathcal{F}_{\eta,\mathbf{T},\psi}$. The natural parameter space

$$\mathbf{H} = \{\boldsymbol{\eta} \in \mathbb{R}^k : e^{\psi(\boldsymbol{\eta})} = \int e^{\boldsymbol{\eta}^T \mathbf{T}(\mathbf{x})} h(\mathbf{x}) dx < \infty\},$$

where the integral is replaced by a sum when $\mathbf{X}$ is discrete, describes the largest set of values of $\boldsymbol{\eta}$ for which the density (or pmf) can be defined. The moments of $\mathbf{T}$ are related to the parameter function ψ in (1) with the following equations (see Lindsay 1997; Nelder and Wedderburn 1972; McCulloch and Nelder 1989),

$$\mathrm{E}_{\boldsymbol{\eta}}(\mathbf{T}(\mathbf{X})) = \left(\frac{\partial \psi(\boldsymbol{\eta})}{\partial \eta_j}\right), \quad \mathrm{var}_{\boldsymbol{\eta}}(\mathbf{T}(\mathbf{X})) = \left(\frac{\partial^2 \psi(\boldsymbol{\eta})}{\partial \eta_j \partial \eta_k}\right)$$

for any $\boldsymbol{\eta} \in \mathbf{H}$. In particular, $\mathrm{E}_{\boldsymbol{\eta}}(T_i(\mathbf{X})) = \partial \psi(\boldsymbol{\eta})/\partial \eta_i$, $\mathrm{cov}_{\boldsymbol{\eta}}(T_i(\mathbf{X}), T_j(\mathbf{X})) = \partial^2 \psi(\boldsymbol{\eta})/\partial \eta_i \partial \eta_j$. A more recent thorough treatment of the exponential family can be found in Bickel and Doksum (2006).

A *generalized linear model* (GLM) (Lindsay 1997; Nelder and Wedderburn 1972; McCulloch and Nelder 1989) hypothesizes that the k-vector of natural parameters $\boldsymbol{\eta}$ is a known linear function $\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\nu}$ of a vector $\boldsymbol{\nu} \in \mathbb{R}^d$, $d \leq k$, where $\boldsymbol{\Gamma}$ is a full rank $k \times d$ matrix.

3 Sufficient Reductions for $\mathbf{X}|Y$ in the Exponential Family

The success of Sufficient Dimension Reduction (SDR) approaches based on inverse regression derives from the following fact. Assume $(Y, \mathbf{X})$ has a joint distribution and let $\mathbf{R}(\mathbf{X})$ be a measurable function of the predictor vector. Then,

$$F(Y|\mathbf{X}) = F(Y|\mathbf{R}(\mathbf{X})) \quad \text{iff} \quad \mathbf{X}|(\mathbf{R}(\mathbf{X}), Y) \stackrel{d}{=} \mathbf{X}|\mathbf{R}(\mathbf{X}) \tag{2}$$

Cook (2007) stated (2) and recognized that the equivalence in (2) means that if one finds a *sufficient statistic* for Y using the distribution of $\mathbf{X}|Y$, where Y is considered as a parameter, then this sufficient statistic is also a *sufficient reduction* for $\mathbf{X}$ in the forward regression of Y on $\mathbf{X}$. Cook's idea yields a

genial tool for obtaining sufficient reductions in regression. It also reveals that the intrinsic dimension of the regression of Y on $\mathbf{X}$ is the dimension of the sufficient statistic for Y in the inverse model $\mathbf{X}|Y$. Estimation proposals for $\mathbf{R}(\mathbf{X})$ constitute different SDR methods.

Because of (2), model-based approaches such as ours provide *sufficient reductions* of the predictors $\mathbf{R}(\mathbf{X})$ in the sense that they are *exhaustive* for the regression of Y on $\mathbf{X}$. That is, the reduction $\mathbf{R}(\mathbf{X})$ satisfies $F(Y|\mathbf{X}) = F(Y|\mathbf{R}(\mathbf{X}))$, where $F(\cdot|\cdot)$ signifies the conditional cdf of the first given the second argument. This statement means that $\mathbf{R}(\mathbf{X})$ is *all that is needed* for the regression of Y on $\mathbf{X}$, and thus $\mathbf{R}(\mathbf{X})$ is exhaustive for the regression of Y on $\mathbf{X}$. Herein, we focus on identifying sufficient reductions for the regression of a response Y on continuous, categorical, or a mix of both types of predictors whose conditional distribution $\mathbf{X}|Y$ belongs to the exponential family.

We assume that $\mathbf{X}_y = \mathbf{X}|(Y = y) = (X_{yj})_{j=1,\ldots,p} \sim \mathcal{F}_{\eta_y,\mathbf{T},\psi}$; that is, $\mathbf{X}_y$ has a pdf (pmf) given by

$$f(\mathbf{x}|\boldsymbol{\eta}_y, Y = y) = e^{\boldsymbol{\eta}_y^T \mathbf{T}(\mathbf{x}) - \psi(\boldsymbol{\eta}_y)} h(\mathbf{x}) \tag{3}$$

The natural parameters $\eta_{yj}, j = 1, \ldots, k$, are functions of y as indicated by the first subscript, and $\boldsymbol{\eta}_y = (\eta_{y1}, \ldots, \eta_{yk})^T$. The dimension of the natural parameter vector typically satisfies $k \geq p$. For example, if $\mathbf{X}_y$ were p-variate normal with mean $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Sigma}_y$, then $k = p+p(p+1)/2$.

Following the GLM formulation, we assume that the natural parameter is a linear function of $\boldsymbol{\nu}_Y$,

$$\boldsymbol{\eta}_Y = \bar{\boldsymbol{\eta}} + \mathbf{A}\boldsymbol{\nu}_Y$$

where $\bar{\boldsymbol{\eta}} = \mathrm{E}(\boldsymbol{\eta}_Y)$, $\mathbf{A} \in \mathbb{R}^{k\times d}$ is full rank semi-orthogonal so that its columns form a basis for $\mathcal{S}_{\mathbf{A}} = \mathrm{span}\{\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}} \colon Y \in \mathcal{S}_Y\}$, with $\mathcal{S}_Y$ denoting the sample space of Y and $\mathrm{E}(\boldsymbol{\nu}_Y) = 0$. This implies that $\boldsymbol{\nu}_Y = \mathbf{A}^T(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}})$, a vector of dimension d.

In order to accommodate the dependence of $\boldsymbol{\nu}_Y$ on Y, we assume that the coordinate vectors are modeled as $\boldsymbol{\nu}_Y = \mathbf{C}^T(\mathbf{f}_Y - \mathrm{E}(\mathbf{f}_Y))$, where $\mathbf{f}_Y \in \mathbb{R}^r$ is a known vector-valued function of Y, and $\mathbf{C} \in \mathbb{R}^{r\times d}$, is an unrestricted matrix of rank $d \leq \min(k, r)$. This formulation is similar to the principal fitted components of Cook (2007) and Cook and Forzani (2008), and parametric inverse regression (PIR) of Bura and Cook (2001) for continuous covariates. Under this model each coordinate $\boldsymbol{\eta}_{Yj}$, $j = 1, \ldots, k$, follows a linear model with predictor vector $\mathbf{f}_Y$. Consequently, we can use inverse response plots (see Cook 1998a, chap.10) of X_j versus Y to gain information about suitable choices for $\mathbf{f}_Y$. This visual tool is not generally available in the forward regression of Y on a multi-component $\mathbf{X}$.

Choosing appropriate $\mathbf{f}_Y$'s is discussed in Cook and Forzani (2008) and Adragni and Cook (2009).

Let $\mathbf{D} = \mathbf{A}\mathbf{C}^T$ with $\mathbf{A} : k \times d$, $\mathbf{C} : r \times d$ and $\mathbf{f}_Y \in \mathbb{R}^r$ known functions of Y, so that

$$\boldsymbol{\eta}_Y = \bar{\boldsymbol{\eta}} + \mathbf{A}\mathbf{C}^T(\mathbf{f}_Y - \bar{\mathbf{f}}) \tag{4}$$

$$= \bar{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_Y - \bar{\mathbf{f}}) \tag{5}$$

We refer to (5) as the unrestricted generalized linear model, and to (4) as the reduced-rank generalized linear model.

3.1 Minimal sufficient reduction

Theorem 1 *If $\mathbf{X}|Y$ has density in the exponential family given by (3), then the minimal sufficient reduction for the regression of $Y|\mathbf{X}$ is given by*

$$\mathbf{R}(\mathbf{X}) = \boldsymbol{\alpha}^T(\mathbf{T}(\mathbf{X}) - \mathrm{E}(\mathbf{T}(\mathbf{X}))$$

where $\boldsymbol{\alpha} = \mathrm{span}\{(\boldsymbol{\eta}_Y - \mathrm{E}_Y(\boldsymbol{\eta}_Y) = (\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}), Y \in \mathcal{S}_Y\}$, and $\mathbf{T}(\mathbf{X})$ is a sufficient statistic for the natural parameters of the exponential family defined by (3).

PROOF. By the Lehmann-Scheffé Theorem (see also Cook 2007), that $\boldsymbol{\alpha}^T(\mathbf{T}(\mathbf{X}) - \mathrm{E}(\mathbf{T}(\mathbf{X}))$ is a minimal sufficient reduction for Y derives from showing the equivalence of (i) $\log(f_{\mathbf{X}|Y}(\mathbf{x})/f_{\mathbf{X}|Y}(\mathbf{z}))$ is independent of Y, and (ii) $\boldsymbol{\alpha}^T(\mathbf{T}(\mathbf{z}) - \mathrm{E}(\mathbf{T}(\mathbf{z}))) = \boldsymbol{\alpha}^T(\mathbf{T}(\mathbf{x}) - \mathrm{E}(\mathbf{T}(\mathbf{x})))$, or equivalently, $\boldsymbol{\alpha}^T\mathbf{T}(\mathbf{z}) = \boldsymbol{\alpha}^T\mathbf{T}(\mathbf{x})$.

Under (3), the independence of $\log(f_{\mathbf{X}|Y}(\mathbf{x})/f_{\mathbf{X}|Y}(\mathbf{z}))$ from Y is equivalent to

$$\log\frac{h(\mathbf{x})}{h(\mathbf{z})} + \big(\mathbf{T}(\mathbf{x}) - \mathbf{T}(\mathbf{z})\big)^T\boldsymbol{\eta}_Y = c \tag{6}$$

where c does not depend on Y. Taking the expectation with respect to Y yields that (6) is equivalent to

$$\big(\mathbf{T}(\mathbf{x}) - \mathbf{T}(\mathbf{z})\big)^T(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}) = 0 \ \text{ for all } y \in \mathcal{S}_Y. \tag{7}$$

If $\mathbf{T}(\mathbf{x}) \in \mathbb{R}^k$, let $\boldsymbol{\alpha} \in \mathbb{R}^{k\times d}$ be a basis for $\mathrm{span}(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}, Y \in \mathcal{S}_Y)$. Then, $\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}} = \boldsymbol{\alpha}\boldsymbol{\nu}_Y$ for some $\boldsymbol{\nu}_Y \in \mathbb{R}^d$, and (7) is equivalent to $(\mathbf{T}(\mathbf{x}) - \mathbf{T}(\mathbf{z}))^T\boldsymbol{\alpha}\boldsymbol{\nu}_Y = 0$, or $\boldsymbol{\alpha}^T\mathbf{T}(\mathbf{x}) = \boldsymbol{\alpha}^T\mathbf{T}(\mathbf{z})$. Therefore, $\boldsymbol{\alpha}^T\left(\mathbf{T}(\mathbf{X}) - \mathrm{E}(\mathbf{T}(\mathbf{X})\right)$ is the minimal sufficient reduction. □

The quadratic exponential family density depends only on its first two moments and is a subclass of the exponential family. In particular, the distribution of $\mathbf{X}|Y$ belongs to the quadratic exponential family if its pdf (pmf) is given by

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\eta}_y, Y=y) &= e^{\boldsymbol{\eta}_y^T \mathbf{T}(\mathbf{X}) - \boldsymbol{\psi}(\boldsymbol{\eta}_y)} h(\mathbf{x}) \\ &= e^{\boldsymbol{\eta}_{y,1}^T \mathbf{X} + \boldsymbol{\eta}_{y,2}^T \mathrm{vec}(\mathbf{X}\mathbf{X}^T) - \boldsymbol{\psi}(\boldsymbol{\eta}_y)} h(\mathbf{x}) \end{aligned} \tag{8}$$

Corollary 1 *If $\mathbf{X}|Y$ belongs to the quadratic exponential family with pdf (pmf) (8), the minimal sufficient reduction for the regression of $Y|\mathbf{X}$ is given by*

$$\mathbf{R}(\mathbf{X}) = \left(\boldsymbol{\alpha}_1^T(\mathbf{X} - \mathrm{E}(\mathbf{X})), \boldsymbol{\alpha}_{20}^T(\mathrm{vec}(\mathbf{X}\mathbf{X}^T) - \mathrm{E}(\mathrm{vec}(\mathbf{X}\mathbf{X}^T)))\right)$$

where $\mathrm{span}(\boldsymbol{\alpha}_1) = \mathrm{span}(\boldsymbol{\eta}_{Y,1} - \mathrm{E}(\boldsymbol{\eta}_{Y,1}), Y \in \mathcal{S}_Y)$ *and* $\boldsymbol{\alpha}_{20} = \mathrm{span}(\boldsymbol{\alpha}_2) \ominus \mathrm{span}(\boldsymbol{\alpha}_1 \otimes \boldsymbol{\alpha}_1)$ *with* $\mathrm{span}(\boldsymbol{\alpha}_2) = \mathrm{span}\left(\boldsymbol{\eta}_{Y,2} - \mathrm{E}(\boldsymbol{\eta}_{Y,2}), Y \in \mathcal{S}_Y\right)$.

The proof of the Corollary follows directly from Theorem 1, the fact that the sufficient statistic for the quadratic exponential family is $\mathbf{T}(\mathbf{X}) = \left(\mathbf{X}^T, \mathrm{vec}^T(\mathbf{X}\mathbf{X}^T)^T\right)$, and that if $\mathrm{span}(\boldsymbol{\alpha}_2) \subset \mathrm{span}(\boldsymbol{\alpha}_1) \otimes \mathrm{span}(\boldsymbol{\alpha}_1)$, then the quadratic part is absorbed in the linear part.

3.2 Some known examples

Here we revisit previous results for quadratic exponential family distributions, such as the normal, and describe how they relate to our results.

a. **Normal with constant variance: $\mathbf{X}|Y \sim N(\boldsymbol{\mu}_Y, \boldsymbol{\Delta})$**

Since $\mathbf{X}|Y$ follows a quadratic exponential family, we apply Corollary 1 to obtain that $\boldsymbol{\alpha}_1 = \boldsymbol{\Delta}^{-1}\mathrm{span}\{\boldsymbol{\mu}_Y - \boldsymbol{\mu}, Y \in \mathcal{S}_Y\}$ and $\alpha_2 = 0$, which yields that $\boldsymbol{\alpha}_1^T(\mathbf{X} - \mathrm{E}(\mathbf{X}))$ is the minimal sufficient reduction as in Cook (2007) and Cook and Forzani (2008).

b. **Normal with variance depending on Y by a multiplicative effect: $\mathbf{X}|Y \sim N(\boldsymbol{\mu}_Y, c_Y\boldsymbol{\Delta})$**

The pdf of $\mathbf{X}|Y$ is

$$f_{\mathbf{X}|Y=y}(\mathbf{x}) = \frac{e^{-\frac{1}{2}(\boldsymbol{\mu}_y - \boldsymbol{\mu})^T c_y \boldsymbol{\Delta}(\boldsymbol{\mu}_y - \mu)}}{2\pi^{p/2}|c_y\boldsymbol{\Delta}|^{1/2}} e^{-\frac{1}{2}\mathrm{vec}((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T)^T \mathrm{vec}(c_y^{-1}\boldsymbol{\Delta}^{-1}) + (\mathbf{x}-\boldsymbol{\mu})^T c_y^{-1}\boldsymbol{\Delta}^{-1}(\boldsymbol{\mu}_y - \boldsymbol{\mu})}$$

Then, according to Corollary 1, $\boldsymbol{\alpha}_1 = \boldsymbol{\Delta}^{-1}\text{span}((\boldsymbol{\mu}_Y - \boldsymbol{\mu}), Y \in \mathcal{S}_Y)$ and $\boldsymbol{\alpha}_2 = \text{vec}(\boldsymbol{\Delta}^{-1})$, so that the minimal sufficient reduction is $\{\boldsymbol{\alpha}_1^T(\mathbf{X} - \boldsymbol{\mu}), (\mathbf{X} - \boldsymbol{\mu})^T\boldsymbol{\Delta}^{-1}(\mathbf{X} - \boldsymbol{\mu})\}$, as in Bura and Forzani (2015).

c. **Normal with non-constant variance:** $\mathbf{X}|Y \sim N(\boldsymbol{\mu}_Y, \boldsymbol{\Delta}_Y)$

Similarly to b., we obtain that the sufficient statistic is $\mathbf{T}(\mathbf{X}) = (\mathbf{X}-\boldsymbol{\mu}, \text{vec}(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T)$, and $\boldsymbol{\eta}_Y = (\boldsymbol{\Delta}_Y^{-1}(\boldsymbol{\mu}_Y - \boldsymbol{\mu}), -\text{vec}(\boldsymbol{\Delta}_Y^{-1} - \text{E}(\boldsymbol{\Delta}_Y^{-1}))/2)$. Let $\boldsymbol{\alpha}_1 = \text{span}(\boldsymbol{\Delta}_Y^{-1}(\boldsymbol{\mu}_Y - \boldsymbol{\mu}))$ and consider the $p^2 \times q$ matrix $\mathbf{S} = (\mathbf{S}_j)_{j=1}^q$, such that $\text{span}(\mathbf{S}) = \text{span}(\text{vec}(\boldsymbol{\Delta}_Y^{-1} - \text{E}(\boldsymbol{\Delta}_Y^{-1})))$. By rearranging in $p \times p$ matrices the columns of $\mathbf{S}$, the symmetric matrices $\mathbf{S}_j$ are formed for $j = 1, \ldots, q$. The minimal sufficient reduction is $\{\boldsymbol{\alpha}_1^T(\mathbf{X} - \boldsymbol{\mu}), (\mathbf{X} - \boldsymbol{\mu})^T\mathbf{S}_1(\mathbf{X} - \boldsymbol{\mu}), \ldots, (\mathbf{X} - \boldsymbol{\mu})^T\mathbf{S}_q(\mathbf{X} - \boldsymbol{\mu})\}$, as in Forzani (2007).

Cook and Forzani (2009) also considered this case, i.e. $\mathbf{X}|Y \sim N(\boldsymbol{\mu}_Y, \boldsymbol{\Delta}_Y)$. They computed a linear sufficient reduction $\boldsymbol{\alpha}^T\mathbf{X}$ with

$$\text{span}(\boldsymbol{\alpha}) \subset \{\boldsymbol{\Delta}^{-1}(\boldsymbol{\mu}_Y - \boldsymbol{\mu}), \boldsymbol{\Delta}_Y^{-1} - \boldsymbol{\Delta}^{-1}\} \tag{9}$$

Corollary 1 shows that the linear reduction (9) is not necessarily minimal. For example, in b) the minimal *linear* sufficient reduction obtained in (9) is $\mathbb{R}^p$, while the minimal reduction is $\big(\text{span}(\boldsymbol{\Delta}^{-1}(\boldsymbol{\mu}_Y - \boldsymbol{\mu}))^T\mathbf{X}, (\mathbf{X} - \boldsymbol{\mu})^T\boldsymbol{\Delta}^{-1}(\mathbf{X} - \boldsymbol{\mu})\big)$ by Corollary 1.

Cook and Li (2009) considered the special case of independent normal $X_j|Y$ with mean μ_{jY} and constant variance σ_j^2, $j = 1, \ldots, p$. Thus, in Corollary 1, $\boldsymbol{\eta}_{Y,2} = \boldsymbol{\eta}_2$ and $\mathbf{R}(\mathbf{X})$ becomes $\boldsymbol{\alpha}_1^T(\mathbf{X} - \boldsymbol{\mu})$, and, therefore, there is no quadratic contribution to the sufficient reduction.

The case of conditional independence of the predictors with variances depending on Y is not covered by Cook and Li (2009). To see this, observe that when $X_j|Y \sim N(\mu_Y, \sigma_Y^2)$ and X_j is independent of X_k given Y,

$$\begin{aligned} f(\mathbf{x}|y) = \prod f(x_j|y) &= \frac{1}{(2\pi)^{p/2}} \frac{1}{\prod_{j=1}^p \sigma_{jy}} \exp\left(-\frac{1}{2}\Big[\sum_{j=1}^p \frac{x_j^2}{\sigma_{jy}^2} - 2\sum_j \frac{x_j\mu_{jy}}{\sigma_{jy}^2} + \sum_j \frac{\mu_{jj}^2}{\sigma_{jy}^2}\Big]\right) \\ &= \frac{1}{(2\pi)^{p/2}} \frac{1}{\prod_{j=1}^p \sigma_{jy}} \exp\left(-\frac{1}{2}\sum_j \frac{\mu_{jj}^2}{\sigma_{jy}^2}\right) \exp\left(-\frac{1}{2}\sum_{j=1}^p \frac{x_j^2}{\sigma_{jy}^2} + \sum_j \frac{x_j\mu_{jy}}{\sigma_{jy}^2}\right). \end{aligned}$$

Hence, the sufficient reduction comprises of $\boldsymbol{\eta}_{Y,1} = \boldsymbol{\Delta}_Y^{-1}\boldsymbol{\mu}_Y$, $\boldsymbol{\eta}_{Y,2} = -\boldsymbol{\Delta}_Y^{-1}/2$, and $\boldsymbol{\Delta}_Y =$

$\mathrm{diag}(\sigma^2_{1Y}, \ldots, \sigma^2_{pY})$ by Corollary 1.

4 Estimation

Suppose that a random sample of size n is drawn from the response Y and the p-dimensional covariate vector $\mathbf{X} = (X_1, \ldots, X_p)^T$. Assuming the data follow (3) with $\boldsymbol{\eta}_{Y_i}$ satisfying (4) or (5), $\mathrm{span}(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}, Y \in \mathcal{S}_Y)$ needs to be estimated. Since, for $i = 1, \ldots, n$, $\boldsymbol{\eta}_{Y_i} = \bar{\boldsymbol{\eta}} + \mathbf{A}\mathbf{C}^T(\mathbf{f}_{Y_i} - \bar{f}) = \bar{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_{Y_i} - \bar{f})$, the estimation of $\bar{\boldsymbol{\eta}}$ and $\mathbf{A}$ and $\mathbf{C}$ is required when the rank of $\mathbf{D}$ is known. If the rank of $\mathbf{D}$ is unknown, we need to estimate $\bar{\boldsymbol{\eta}}$ and $\mathbf{D}$.

In both cases, we use maximum likelihood estimation. Since the dimension will typically be unknown, if we use the first approach we need to estimate $\bar{\boldsymbol{\eta}}$, and $\mathbf{A}$, $\mathbf{C}$ for $d = 0, \ldots, \min(r, k)$, and then compare the models, for example by AIC or BIC, in order to estimate $\mathrm{span}(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}, Y \in \mathcal{S}_Y)$ and its dimension. If the rank of $\mathbf{D}$ is unknown, once the MLEs of $\bar{\boldsymbol{\eta}}$ and $\mathbf{D}$ are obtained, the rank d of the latter is estimated using the asymptotic tests in Section 5, and the first $\hat{d}$ eigenvectors of $\widehat{\mathbf{D}}$ form the MLE of $\mathrm{span}(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}, Y \in \mathcal{S}_Y)$.

The solution of the maximum likelihood equations has been shown to be equivalent to the iterative reweighted least-squares (see, e.g., McCullough and Searle 2001, p.143; Nelder and Wedderburn 1972; Yee and Hastie 2003). We can estimate the parameters of interest using the alternating estimation algorithm proposed by Yee and Hastie (2003) that extends the standard IRLS algorithm, which is applicable when the rank of $\mathbf{D}$ is unrestricted. We describe the IRLS algorithm proposed by Yee and Hastie (2003), noting that the algorithm reduces to usual IRLS when $\mathbf{C}^T$ is a constant, in which case the algorithm estimates $\mathbf{D}$.

In the reduced-rank formulation, where $\mathbf{D} = \mathbf{A}\mathbf{C}^T$, and $\mathbf{C}$ has fixed rank d, the IRLS algorithm has the following steps:

1. Given current estimates $\mathbf{A}^{(t)}, \bar{\boldsymbol{\eta}}^{(t)}, \mathbf{C}^{(t)}$, $\mathbf{W}_i^{(t)} = D^2\psi(\boldsymbol{\eta}_{y_i}(\mathbf{A}^{(t)}, \bar{\boldsymbol{\eta}}^{(t)}, \mathbf{C}^{(t)}))$, set

$$\mathrm{vec}(\mathbf{C}^{(t+1)}) = \left(\sum_{i=1}^{n} \mathbf{A}^{(t)T} \mathbf{W}_i^{(t)} \mathbf{A}^{(t)} \otimes (\mathbf{f}_{y_i} - \bar{\mathbf{f}})(\mathbf{f}_{y_i} - \bar{\mathbf{f}})^T \right)^{-1} \cdot \sum_{i=1}^{n} \mathbf{A}^{(t)T} \mathbf{W}_i^{(t)} (\mathbf{z}_i^{(t)} - \bar{\boldsymbol{\eta}}^{(t)}) \otimes (\mathbf{f}_{y_i} - \bar{\mathbf{f}})$$

where

$$\mathbf{z}_i^{(t)} - \bar{\boldsymbol{\eta}}^{(t)} = \mathbf{A}^{(t)}\mathbf{C}^{(t)T}(\mathbf{f}_{y_i} - \bar{\mathbf{f}}) + (\mathbf{W}_i^{(t)})^{-1}(\mathbf{T}(\mathbf{X}_i) - \nabla\psi(\boldsymbol{\eta}_{y_i}(\mathbf{A}^{(t)}, \bar{\boldsymbol{\eta}}^{(t)}, \mathbf{C}^{(t)}))),$$

using $\partial l/\partial \boldsymbol{\eta}_y = \mathbf{T}(\mathbf{X}) - \nabla\psi(\boldsymbol{\eta}_y)$, where $l_i = \log(\exp(\boldsymbol{\eta}_{y_i}\mathbf{T}(\mathbf{x}_i) - \psi(\boldsymbol{\eta}_{y_i}))h(\mathbf{x}_i))$.

2. Let $\mathbf{H}_i^{(t)} = (1, (\mathbf{f}_{y_i} - \bar{\mathbf{f}})^T\mathbf{C}^{(t+1)})^T \otimes \mathbf{I}_k$. Then,

$$\mathrm{vec}(\bar{\boldsymbol{\eta}}^{(t+1)}, \mathbf{A}^{(t+1)}) = \left(\sum_{i=1}^n \mathbf{H}_i^{(t)}\mathbf{W}_i^{(t)}\mathbf{H}_i^{(t)T}\right)^{-1} \sum_{i=1}^n \mathbf{H}_i^{(t)}\mathbf{W}_i^{(t)}\mathbf{s}_i^{(t)}$$

where

$$\mathbf{W}_i^{(t)} = D^2\psi(\boldsymbol{\eta}_{y_i}(\mathbf{A}^{(t)}, \bar{\boldsymbol{\eta}}^{(t)}, \mathbf{C}^{(t+1)}))$$

and

$$\mathbf{s}_i^{(t)} = \bar{\boldsymbol{\eta}}^{(t)} + \mathbf{A}^{(t)}\mathbf{C}^{(t+1)T}(\mathbf{f}_{y_i} - \bar{\mathbf{f}}) + (\mathbf{W}_i^{(t)})^{-1}(\mathbf{T}(\mathbf{X}_i) - \nabla\psi(\boldsymbol{\eta}_{y_i}(\mathbf{A}^{(t)}, \bar{\boldsymbol{\eta}}^{(t)}, \mathbf{C}^{(t+1)}))),$$

since $\mathbf{H}_i^{(t)}\mathrm{vec}(\bar{\boldsymbol{\eta}}^{(t)}, \mathbf{A}^{(t)}) = \bar{\boldsymbol{\eta}}^{(t)} + \mathbf{A}^{(t)}\mathbf{C}^{(t+1)T}(\mathbf{f}_{y_i} - \bar{\mathbf{f}})$.

The choice of starting values can vary from example to example. We propose to obtain the initial coordinates of $\bar{\boldsymbol{\eta}}$ and $\mathbf{AC}^T$ of $\boldsymbol{\eta}_y$ that correspond to X_j, $j = 1, \ldots, p$, by fitting one dimensional exponential family regressions. In the Appendix we derive the initial values we use in the Australian athletes data analysis example in Section 6.1, and adapt the standard multivariate version of the IRLS algorithm when the rank of $\mathbf{D}$ is not fixed. The nuisance parameters are estimated at the outset and then remain constant.

We refer to our approach of obtaining the minimal sufficient reduction of the regression of Y on $\mathbf{X}$, where $\mathbf{X}|Y$ is in the exponential family, as *Exponential Family Dimension Reduction* (EF-DR).

5 Asymptotic Tests for Dimension

The dimension of the regression of Y on $\mathbf{X}$ is $d = \dim(\mathrm{span}(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}))$, where the inverse predictor $\mathbf{X}|Y$ has an exponential family distribution with pdf (or pmf) given by (3), and $\boldsymbol{\eta}_Y = \overline{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_Y - \bar{\mathbf{f}})$, with $\overline{\boldsymbol{\eta}} = \mathrm{E}(\boldsymbol{\eta}_Y) \in \mathbb{R}^k$ and $\mathbf{D} \in \mathbb{R}^{k\times r}$.

The rank d of $\mathbf{D}$, or equivalently, the dimension of the regression of Y on $\mathbf{X}$ can be estimated via

model selection criteria such as AIC and BIC, where the values of $d = 0, \ldots, \min(r, k)$ are compared by fitting the reduced-rank regression models in Section 4. In this section, we also derive asymptotic tests to estimate the dimension. The standard IRLS algorithm is used to fit the generalized linear model $\boldsymbol{\eta}_y = \overline{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_Y - \bar{\mathbf{f}})$, with $\mathbf{D} \in \mathbb{R}^{k \times r}$. Using the fact that the IRLS estimate of $\mathbf{D}$ is also the MLE and asymptotically normal, we can estimate the dimension of $\mathbf{D}$ using either of the two tests in Bura and Yang (2011). These tests require knowledge of the asymptotic variance of $\widehat{\mathbf{D}}$.

We express the natural parameters as $\boldsymbol{\eta}_Y = \overline{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_Y - \bar{\mathbf{f}}) = (\Gamma_1, \mathbf{D})(1, (\mathbf{f}_Y - \bar{\mathbf{f}})^T)^T = \boldsymbol{\Gamma}\mathbf{F}_Y$, with $\mathbf{D} = \boldsymbol{\Gamma}\mathbf{M}^T$, where $\mathbf{M} = (\mathbf{0}, \mathbf{I}_r)$ and $\mathbf{F}_Y = (1, (\mathbf{f}_Y - \bar{\mathbf{f}})^T)^T$. The solution of the ML equations is equivalent to the iterative re-weighted least-squares procedure given in Section 4 (see, e.g. McCullough and Searle 2001, p. 143; Nelder and Wedderburn 1972; Yee and Hastie 2003). Thus, $\widehat{\mathbf{D}}_{MLE} = \widehat{\mathbf{D}}_{IRLS}$ with the same asymptotic normal distribution. The computation of the asymptotic variance for $\mathbf{D}$ requires the information matrix for $\widehat{\boldsymbol{\Gamma}}$. From (3), the log-likelihood of $\mathbf{X}|(Y = y)$ with natural parameters modeled as $\boldsymbol{\eta}_y = \boldsymbol{\Gamma}\mathbf{F}_y$ is given by

$$l = (\boldsymbol{\Gamma}\mathbf{F}_y)^T\mathbf{T}(\mathbf{x}) - \psi(\boldsymbol{\Gamma}\mathbf{F}_y)$$

Since $(\boldsymbol{\Gamma}\mathbf{F}_y)^T\mathbf{T}(\mathbf{x}) = \text{vec}^T(\mathbf{T}(\mathbf{x})\mathbf{F}_y^T)\text{vec}(\boldsymbol{\Gamma})$ and $\boldsymbol{\Gamma}\mathbf{F}_y = (\mathbf{F}_y^T \otimes \mathbf{I}_k)\text{vec}(\boldsymbol{\Gamma})$, we have

$$\begin{aligned}
\frac{\partial l}{\partial \text{vec}^T(\boldsymbol{\Gamma})} &= \text{vec}(\mathbf{T}(\mathbf{x})\mathbf{F}_y^T) - (\mathbf{F}_y \otimes \mathbf{I}_k)\nabla\psi(\boldsymbol{\Gamma}\mathbf{F}_y) \\
\frac{\partial^2 l}{\partial \text{vec}(\boldsymbol{\Gamma})\partial \text{vec}^T(\boldsymbol{\Gamma})} &= -(\mathbf{F}_y \otimes \mathbf{I}_k)\nabla^2\psi(\boldsymbol{\Gamma}\mathbf{F}_y)(\mathbf{F}_y^T \otimes \mathbf{I}_k)
\end{aligned}$$

Therefore, $\sqrt{n}\text{vec}(\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}) \Rightarrow N_{r \times k}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\Gamma}})$, with $\mathbf{V}_{\boldsymbol{\Gamma}} = \left[(\mathbf{F}_y \otimes \mathbf{I}_k)\nabla^2\psi(\boldsymbol{\Gamma}\mathbf{F}_y)(\mathbf{F}_y^T \otimes \mathbf{I}_k)\right]^{-1}$. As a consequence, $\sqrt{n}\ \text{vec}(\widehat{\mathbf{D}} - \mathbf{D}) \Rightarrow N_{r \times k}(\mathbf{0}, \mathbf{V}_{\mathbf{D}})$, with $\mathbf{V}_{\mathbf{D}} = (\mathbf{M} \otimes \mathbf{I}_k)\mathbf{V}_{\boldsymbol{\Gamma}}(\mathbf{M}^T \otimes \mathbf{I}_k)$. In the sample, $\widehat{\mathbf{V}}_{\mathbf{D}}$ is obtained by replacing $\mathbf{V}_{\boldsymbol{\Gamma}}^{-1}$ by its sample mean, i.e.

$$\widehat{\mathbf{V}}_{\boldsymbol{\Gamma}}^{-1} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{F}_{y_i} \otimes \mathbf{I}_k)\nabla^2\psi(\widehat{\boldsymbol{\Gamma}}\mathbf{F}_{y_i})(\mathbf{F}_{y_i} \otimes \mathbf{I}_k)$$

Since $d = \text{rank}(\mathbf{D})$ and $\widehat{\mathbf{D}}$ is asymptotically normal, we can use either an asymptotic weighted chi-squared test or a Wald-type asymptotic chi-squared test based on the smallest singular values of $\widehat{\mathbf{D}}$ developed by Bura and Yang (2011).

Let $\widehat{\mathbf{D}} = \widehat{\mathbf{U}}^T \widehat{\mathbf{\Delta}} \widehat{\mathbf{R}}$ be the SVD ecomposition of $\widehat{\mathbf{D}}$, with $\widehat{\mathbf{\Delta}} = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_{\min(r,k)})$, where $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_{\min(r,k)}$ are the singular values of $\widehat{\mathbf{D}}$, $\widehat{\mathbf{U}}$ is the $k \times k$ matrix of its left singular vectors and $\widehat{\mathbf{R}}$ is the $r \times r$ matrix of its right singular vectors. For $m = 0, \ldots, \min(r, k)$, let $\widehat{\mathbf{U}}_0$ be the $k \times (k - m)$ sub-matrix of the first $k - m$ columns of $\widehat{\mathbf{U}}$, and $\widehat{\mathbf{R}}_0$ be the $r \times (r - m)$ sub-matrix of the first $r - m$ columns of $\widehat{\mathbf{R}}$. Also, let $\widehat{\mathbf{Q}} = (\widehat{\mathbf{R}}_0^T \otimes \widehat{\mathbf{U}}_0^T)\widehat{\mathbf{V}}_{\mathbf{D}}(\widehat{\mathbf{R}}_0 \otimes \widehat{\mathbf{U}}_0)$ and $\widehat{\mathbf{D}}_0 = \widehat{\mathbf{U}}_0^T \widehat{\mathbf{D}} \widehat{\mathbf{R}}_0$.

Following Bura and Yang (2011), when $m = \text{rank}(\mathbf{D})$, the test statistic $\Lambda_1(m) = n \sum_{i=m+1}^{\min(r,k)} \hat{\lambda}_i^2$, $m = 0, \ldots, \min(r, k)$ has an asymptotic weighted chi-squared distribution with weights the ordered eigenvalues of $\mathbf{Q}$, the population counterpart to $\widehat{\mathbf{Q}}$.

Also, when $m = \text{rank}(\mathbf{D})$, the test statistic $\Lambda_2(m) = n\text{vec}(\widehat{\mathbf{D}}_0^T)\widehat{\mathbf{Q}}^+\text{vec}(\widehat{\mathbf{D}}_0)$, with $\widehat{\mathbf{Q}}^+$ the Moore-Penrose inverse of $\widehat{\mathbf{Q}}$, is asymptotically chi-squared with $\min(\text{rank}(\text{cov}(\widehat{\mathbf{D}})), (r - m)(k - m))$ degrees of freedom. The dimension is estimated to be the first value of m for which $d = m$ cannot be rejected at a prespecified level α when carrying out sequential tests of $d = m$ versus $d > m$ for $m = 0, \ldots, \min(r, k)$.

The hypotheses in the sequential testing are hierarchically ordered in the sense that in order to test $d = m + 1$ one has to first reject $d = m$. Such sequential application of unadjusted tests that requires $\min(r, k)$ steps to test $\min(r, k)$ null hypotheses is called serial gatekeeping and is analogous to union-intersection tests (Berger 1982). An important feature of this process is that each null is tested sequentially at the overall level α, which is controlled for the entire sequential testing (see Dmitrienko et al. 2010, sec. 5.3; Westfall and Krishen 2001).

6 Connection with Other SDR Methods for Exponential Family Inverse Predictors

The first attempt to extend sufficient dimension reduction methodology to regressions with both continuous and categorical predictors, was partial dimension reduction by Chiaromonte et al. (2002). Their set-up, except for the continuous predictors $\mathbf{X} \in \mathbb{R}^p$, also requires a qualitative predictor W with values $w = 1, \ldots, C$, which identify C subpopulations. Chiaromonte et al. (2002) proved that if $Y \perp\!\!\!\perp (\mathbf{X}, W) | (\boldsymbol{\zeta}^T \mathbf{X}, W)$, then $(\widetilde{\boldsymbol{\zeta}}^T \mathbf{X}, W)$ partially recovers the reduction for the regression of Y on $(\mathbf{X}, W)$, where $\widetilde{\boldsymbol{\zeta}} = \text{span} \bigoplus_{w=1}^{C} \boldsymbol{\zeta}_w$, with $\boldsymbol{\zeta}_w$ such that $\boldsymbol{\zeta}_w^T \mathbf{X}$ is a component, not necessarily exhaustive, of the reduction for the regression $Y | \mathbf{X}_w$, and $\mathbf{X}_w$ denotes the restriction of the continuous covariates

in subpopulation $w = 1, \ldots, C$.

Partial SIR, their estimation algorithm, was an effective way to accommodate both categorical and continuous predictors utilizing contemporaneous moment-based SDR methodology, which could only handle continuous predictors. With this approach, a component of the sufficient reduction for the regression of Y on $(\mathbf{X}, W)$ was identified. As it does not allow mixing among continuous and categorical variables, a main disadvantage is redundancy, especially if C is large, as each subpopulation is considered separately. Moreover if the number of categories is large, the sample size in each category defined by W can be very small, leading to pronounced inaccuracy in estimation.

We showed in Section 3.2 that Cook and Li's (2009) formulation is a special case of our general approach. Their estimation of the reduction is done via optimization over Grassmann manifolds, a specialized optimization algorithm which requires knowledge of the full likelihood and all its partial derivatives and fairly accurate starting values.

Our method, EF-DR, allows for inter-dependence among the inverse predictors, which are modeled as multiparameter exponential family distributed random vectors. EF-DR identifies the minimal sufficient reduction and computes its MLE. Our IRLS estimation algorithm is fast and easy to implement since it only requires the derivatives of the likelihood, which, in the case of the exponential family, can be computed without computing the likelihood, and is a quadratic optimization problem in each iteration. It can be used in place of Grassmann optimization in the Cook and Li (2009) framework. For example, we can quickly reproduce Cook and Li's (2009) Table 1 simulation results under their set-up of conditionally independent inverse Bernoulli predictors.

We illustrate the connections, differences and advantages of our EF-DR approach and both previous methods, as well as kernel-based methods, using the Australian athletes data set in Section 6.1.

6.1 Australian Athletes

To illustrate partial SIR, Chiaromonte et al. (2002) analyzed a data set discussed by Cook and Weisberg (1994, 1999). Lean body mass (LBM), Y, is regressed on the logarithms of height, weight, red cell count, white cell count and hemoglobin, plus an indicator for gender, for a sample of 202 individuals training at the Australian Institute of Sport. Chiaromonte et al. (2002) considered five many-valued predictors that comprise $\mathbf{X}$, and one qualitative predictor, gender, denoted by W, with $W = 0$ for males and $W = 1$ for females. They concluded that a single linear combination of the continuous predictors together with gender were sufficient to describe the regression of Y on $(\mathbf{X}, W)$. The reduction in $\mathbf{X}$ was

estimated with the first partial SIR predictor $\widehat{\boldsymbol{\zeta}}^T\mathbf{X}$. The plot of Y versus the first partial SIR predictor is given in Figure 1 with the two genders identified by different plotting symbols. Using this summary plot to inform their forward regression model, they inferred that the relationship between Y and $\mathbf{X}$ is linear in the first partial SIR predictor but also that the two straight lines differ for the two sexes both in terms of rate (slope) and starting point (intercept), and they proceeded to fit a model for $Y|(\mathbf{X}, G)$ as linear the first partial SIR direction, with two different straight lines for sex (G).

To compare with Cook and Li (2009), we assume that $(\mathbf{X}, W)|Y$ has an exponential family distribution with all components of $\mathbf{X}$ and W independent given Y. That is,

$$f_{(\mathbf{X},W)|Y} = f_{\mathbf{X}|Y} f_{W|Y} = f_{W|Y} \prod_{i=1}^{p} f_{X_i|Y}$$

By surveying the marginal plots, the normal distribution appears to be a reasonable model for all $X_j|Y$. The variances are assumed to be unequal but independent of Y. We also assume that the distribution of $W|Y$ is Bernoulli with $f_{W=w|Y=y} = p_y^w(1-p_y)^{1-w}$, $w = 0, 1$. In Figure 2, we plot Y versus the first two sufficient reductions under this model. The plot of the first direction looks very similar to the partial SIR plot but we note that in this case the reduction is of the form $R(\mathbf{X}, W)$ and not $(R(\mathbf{X}), W)$ as in partial SIR. The plots suggest we need both directions of the reduction to model Y. This result is also supported by both asymptotic tests for dimension, with the weighted chi-squared test estimating the dimension to be 2, i.e. separate linear models for the two sexes, at 0.05 level with a p-value of 0.02, and the chi-squared test estimating the dimension as greater than 2, with a very small p-value. The latter, though, is known to be a very conservative test so we typically use the weighted chi-squared test to estimate dimension. As this is a two-dimensional problem and it is difficult to visualize the relationship between the response and the two reductions, we simply fit a linear model in the two reductions and report results from this model in the next section where we compare predictive accuracy.

Next, we assume that the *joint* distribution of $(\mathbf{X}, W)|Y$ has an exponential family distribution. Since $f_{(\mathbf{X},W)|Y} = f_{\mathbf{X}|(Y,W)} f_{W|Y}$, we model $\mathbf{X}|(Y, W)$ as multivariate normal with constant variance, based on marginal plots of $\mathbf{X}$ versus Y for the two genders, and $W|Y$ as Bernoulli(p_Y). The plots of the $\mathbf{X}$ components versus Y indicate that $\mathrm{E}(\mathbf{X}|(W, Y))$ depends on Y in a linear fashion within the two sexes, i.e. $\mathrm{E}(\mathbf{X}|(W, Y)) = \boldsymbol{\mu}_X + \mathbf{b}_1(f_Y - \bar{f}_Y) + \mathbf{b}_2(W - \mu_W)$, with $f_Y = Y$, $\bar{f}_Y = \mathrm{E}(f_Y)$,

$\mu_W = \mathrm{E}(W)$. Therefore, we model the conditional joint distribution of $(\mathbf{X}, W)$ given Y as

$$\begin{aligned} f(\mathbf{x}, w|y) &= \frac{1}{(2\pi)^{p/2}|\mathbf{\Delta}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_x-\mathbf{b}_1(f_y-\bar{f}_y)-\mathbf{b}_2(w-\mu_w))^T\mathbf{\Delta}^{-1}(\mathbf{x}-\mu_x-\mathbf{b}_1(f_y-\bar{f}_y)-\mathbf{b}_2(w-\mu_w))} \\ &\quad \times e^{(w-\mu_w)\log\frac{p_y}{1-p_y}+\log(1-p_y)+\mu_w\log\frac{p_y}{1-p_y}} \\ &= e^{\boldsymbol{\eta}_y^T\mathbf{T}(\mathbf{x},w)-\psi(\boldsymbol{\eta}_y)}h(\mathbf{x}, w) \end{aligned} \tag{10}$$

The pdf form in (10) belongs to the exponential family with

$$\begin{aligned} \mathbf{T}(\mathbf{x}, w) &= (\mathbf{x}-\mu_x, w-\mu_w) \\ \boldsymbol{\eta}_y = (\boldsymbol{\eta}_1, \eta_2) &= \left(\mathbf{\Delta}^{-1}\mathbf{b}_1(f-\bar{f}_y), \log\frac{p_y}{1-p_y} - \mathbf{b}_2^T\mathbf{\Delta}^{-1}\mathbf{b}_1(f-\bar{f}_y)\right) \\ \psi(\boldsymbol{\eta}_y) &= \frac{1}{2}\left(\mathbf{b}_1\left(f_y-\bar{f}_y\right)\right)^T\mathbf{\Delta}^{-1}\left(\mathbf{b}_1\left(f_y-\bar{f}_y\right)\right) - \log(1-p_y) - \mu_g\log\frac{p_y}{1-p_y} \\ h(\mathbf{x}, w) &= \frac{1}{(2\pi)^{p/2}|\mathbf{\Delta}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_x)^T\mathbf{\Delta}^{-1}(\mathbf{x}-\boldsymbol{\mu}_x)-\frac{1}{2}(w-\mu_w)^2\mathbf{b}_2^T\mathbf{\Delta}^{-1}\mathbf{b}_2+(w-\mu_w)\mathbf{b}_2^T\mathbf{\Delta}^{-1}(\mathbf{x}-\boldsymbol{\mu}_x)} \end{aligned}$$

where $\boldsymbol{\eta}_1 = \mathbf{\Delta}^{-1}\mathbf{b}_1(f-\bar{f}_y)$ is $p \times 1$, and $\eta_2 = \log(p_y/(1-p_y)) - \mathbf{b}_2^T\mathbf{\Delta}^{-1}\mathbf{b}_1(f_y-\bar{f}_y)$ is a scalar. Letting $\boldsymbol{\alpha} = \mathrm{span}\{\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}} : Y \in \mathcal{S}_Y\}^T$, $\boldsymbol{\alpha}^T\mathbf{T}(\mathbf{X}, W)$ is the minimal sufficient reduction for the regression of Y on $\mathbf{X}$ and W.

The estimation algorithm for the sufficient reduction is given in the Appendix. Using the tests for dimension in Section 5, the dimension was estimated to be one using both the weighted chi-squared and the chi-squared tests at 5% level. The weighted chi-squared test had a p-value of 0 for testing dimension 0 versus 1 and 0.83 for 1 versus 2, and the chi-squaredd test had a p-value of 0 for 0 versus 1 and 0.053 for 1 versus 2. In Figure 3, the response, LBM, is plotted versus the sufficient reduction estimated by EF-DR. The plot indicates that there is no difference between the two sexes and the same quadratic function in $R(\mathbf{X}, W)$ can be fitted to both sexes to predict LBM when the possible dependence among the predictors, both quantitative and qualitative, is accounted for. That is, the regression model for predicting LBM is

$$E(Y|\mathbf{X}, W) = E(Y|R(\mathbf{X}, W)) = \gamma_1 + \gamma_2 R(\mathbf{X}, W) + \gamma_3 R^2(\mathbf{X}, W)$$

This result is in contrast to both the partial SIR analysis in Chiaromonte et al. (2002) and Cook and Li (2009) who estimate the dimension of the reduction to be 2. By accounting for the dependence structure,

EF-DR not only provides a more accurate characterization of the relationship between the response and the predictors, but also reduces the complexity of the regression.

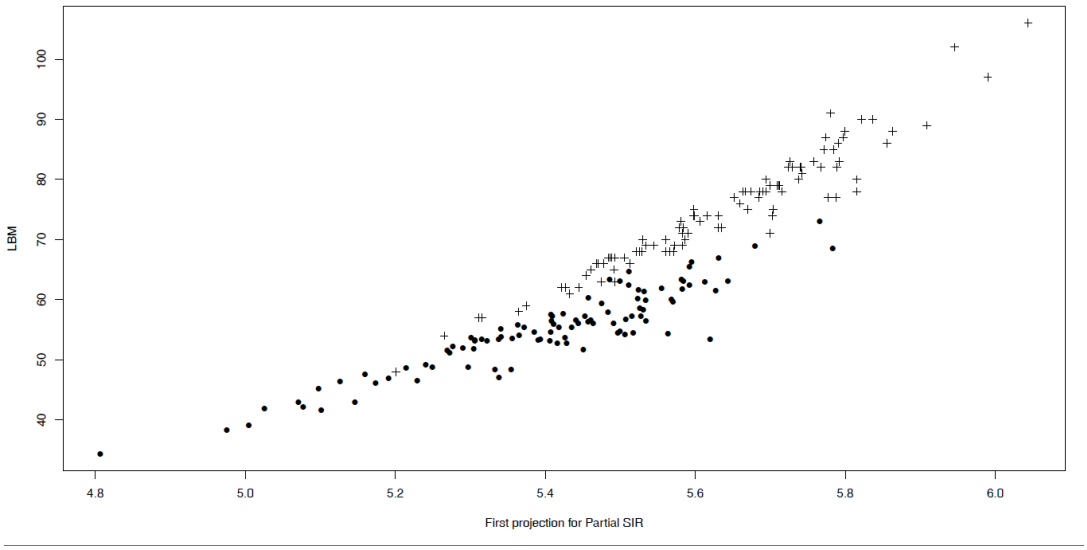


Figure 1: LBM versus the Partial SIR (Chiaromonte et al. 2002) estimated sufficient reduction (Dots for females, crosses for males)

To examine how the dependence structure affects the performance of EF-DR, we carried out additional simulations using the set-up of the Australian athletes data varying the correlation of the inverse predictors. That is, we generated a normal response and one binary and several continuous predictors satisfying (10). The results, which are not reported, largely agree with our observations in the Australian athletes example. EF-DR exhibited overall superior performance across dependence structures with respect to both estimation and prediction.

We also studied how EF-DR behaves as the sample size and the number of predictors increases, again using the Australian athletes set-up. As expected, increased sample sizes result in better performance. Also, the higher the correlation among the predictors, the better EF-DR performed across sample size and number of predictors.

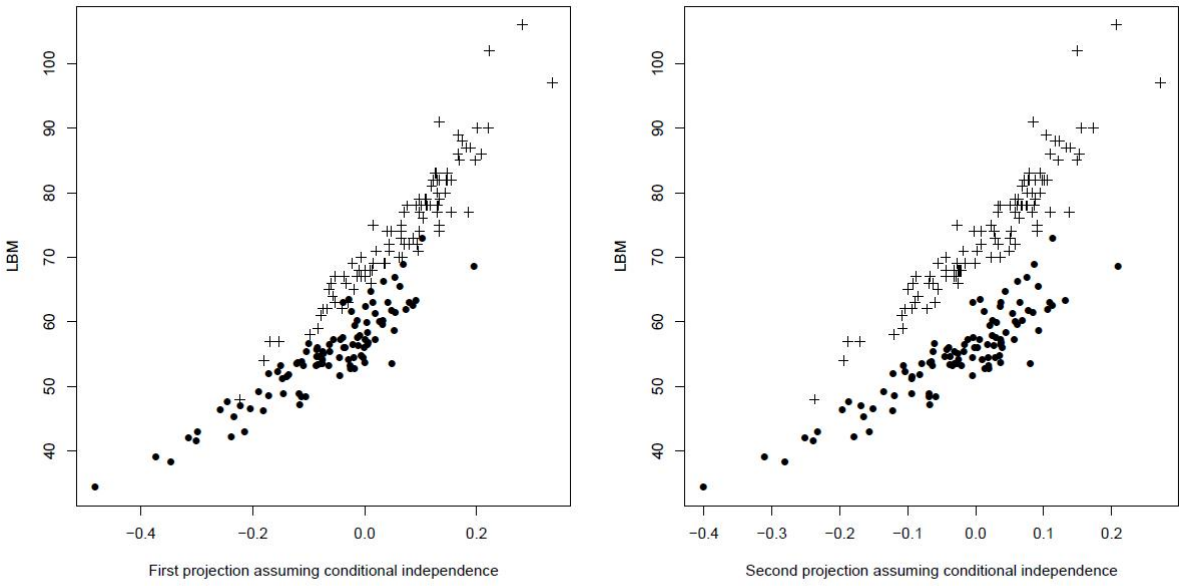


Figure 2: LBM versus the two Cook and Li (2009) estimated sufficient reductions (Dots for females, crosses for males)

6.2 Comparison with Kernel-based SDR Methods

In attempts to overcome the linear nature of most SDR methods, several recent papers have combined sufficient dimension reduction and reproducing kernel Hilbert spaces (Akaho 2001; Bach and Jordan 2002; Fukumizu et al. 2004, 2009; Fukumizu et al. 2007; Wu 2008; Wu et al. 2008; Hsing and Ren 2009; Yeh et al. 2009; Zhu and Li 2011). Along the same lines, Li et al. (2011) used support vector machines to estimate functions of the sufficient reductions and called their method Principal Support Vector Machine (PSVM). Kernel PSVM, its nonlinear version, compares well with linear SDR methods in specific set-ups but it is computationally demanding and requires some judicious choices by the user at the outset. Kim and Pavlovic (2013) developed Covariance Operator Inverse Regression (COIR), a generalized version of linear discriminant analysis.

Kernel-based methods belong to what is called “manifold learning" or “nonlinear dimension reduction," where the data are considered to be a sample from a manifold that is embedded in a higher-dimensional space (see Parviainen 2011). Manifolds generalize linear subspaces by introducing curvilinear coordinates. The manifold coordinates can be approximated by studying the data in a local scale, and then combining local linear views into a global nonlinear shape. Local views are obtained by re-

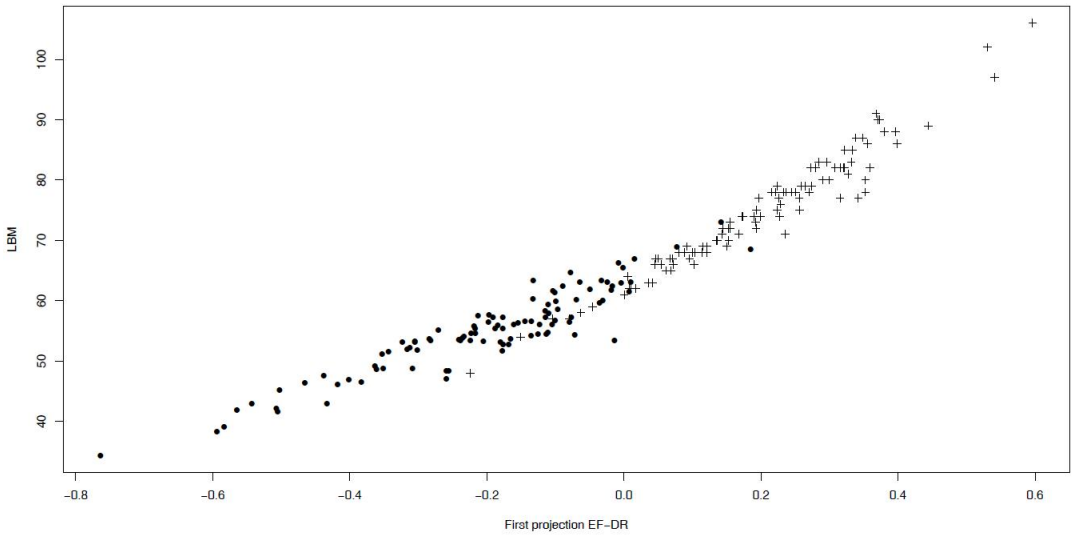


Figure 3: LBM versus the EF-DR estimated sufficient reduction (Dots for females, crosses for males)

stricting the analysis to the nearest neighbors of a data point. Kernel methods assume existence of a low-dimensional structure and try to find it, instead of transforming high-dimensional data into a low-dimensional form. Many manifold methods are reviewed in van der Maaten, Postma and van den Herik (2009).

We apply KSIR (Kernel SIR, Wu 2008), KDR (Kernel Dimensionality Reduction, Fukumizu et al. 2004) and COIR (Kim and Pavlovic 2013) to the Australian athletes data. COIR is a modification of KSIR for univariate response, where, instead of slicing, a non-parametric model is used in the inverse regression of $\mathbf{X}|Y$. The code for kernel PSVM (Li et al. 2011) is not publicly available. Except for KDR, the other kernel-based SDR methods produce nonlinear reductions of the predictors that do not have an explicit form. KSIR, KDR and COIR provide a representation of the capabilities of kernel-based SDR methods for data such as the Australian athletes.

There are no tests for dimension for these kernel-based methods. Yeh et al. (2009) rank the importance of the KSIR predictors by the associated eigenvalues and suggest using either the first one or the first two after inspection of the eigenvalues. Fukumizu et al. (2009) do not consider the problem of inferring the dimension of the KDR reduction important because KDR is often used in the context of graphical exploration of data, where a data analyst may wish to explore views of varying dimensionality. Both KSIR and KDR reductions are consistent (Wu et al. 2008; Fukumizu et. al. 2009). There are no published consistency results for COIR (Kim and Pavlovic 2013).

In Figures 4, 5 and 6 we plot the response versus the first KSIR, KDR and COIR predictors applied to the continuous predictors, respectively. The Gaussian kernel was used for all three, and the choice of optimal bandwidth was highly computationally intensive. Based on these summary plots and on the plot of Y versus those predictors, we model Y as in the case of partial-SIR with

$$E(Y|\mathbf{X}, W) = E(Y|R(\mathbf{X}), W) = (\alpha_1 + \alpha_2 I(W = 1)) + (\beta_1 + \beta_2 I(W = 1))R(\mathbf{X})$$

where W is gender (1 for women, 0 for men), I is the indicator function, and $R(\mathbf{X})$ is the reduction for each method.

In Table 1 we report the coefficient vector $\widehat{\boldsymbol{\alpha}}$ of the reduction for the methods we are able to obtain explicit solutions: For partial SIR and KDR, the reduction is $\widehat{\boldsymbol{\alpha}}^T\mathbf{X}$, and for Cook and Li (2009) and our EF-DR, $\widehat{\boldsymbol{\alpha}}^T(\mathbf{X}, W)$. In Table 2 we report the leave-one-out squared error loss, $\sum_{i=1}^{n}(y_i - \hat{y}_{(-i)})^2/n$, where $\hat{y}_{(-i)}$ indicates prediction of the ith response using all data except for the ith observation. EF-DR has the best performance with respect to prediction error, followed by partial SIR and then KDR, KSIR and COIR.

We also included kernel methods in the unreported simulations, using the set-up of the Australian athletes data varying the correlation of the inverse predictors. We found EF-DR to exhibit overall better estimation and prediction performance. This is not surprising as the data were simulated to comply with the exponential family inverse predictor model, which is the premise of our methodology, and kernel-based approaches are not tailored to specific distributional models. We note that the computational cost for applying kernel methods in this set-up was significantly higher as compared with EF-DR: the time required for KSIR or KDR ranged from 150 to about 1000 times that of EF-DR. We did not include COIR in the computations as it generally gave the worst results at high computational cost.

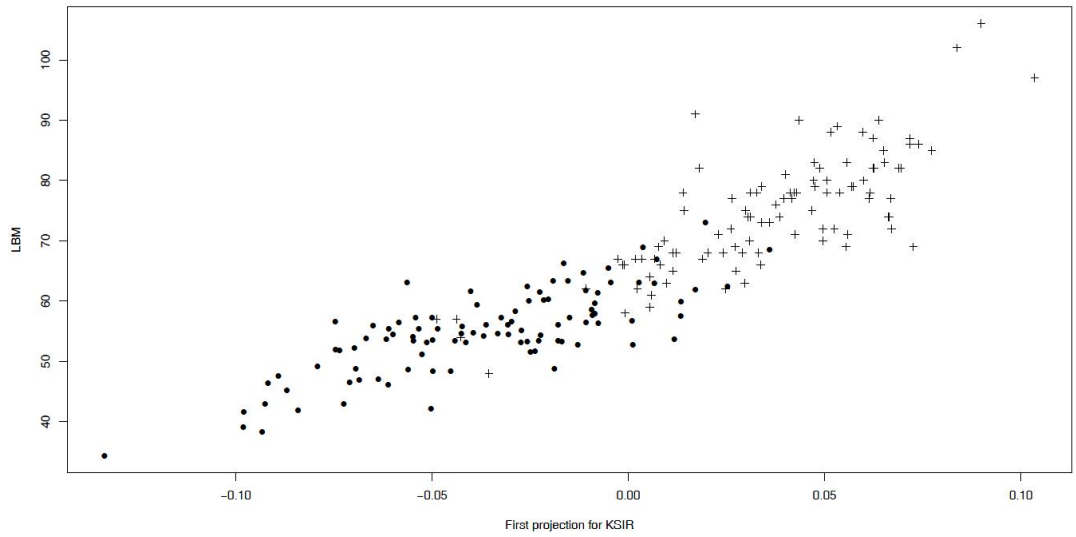


Figure 4: LBM versus the first KSIR predictor (Dots for females, crosses for males)

7 Multivariate Bernoulli Inverse Predictors

In many fields, predictors of a target variable are binary or categorical. Examples include gene association studies (Peng et al. 2009; Wang et al. 2011), image processing (Hassner and Sklansky 1980; Woods 1978), natural language processing (Manning and Schutze 1999), social networks (Wasserman and Pattison 1996; Handcock 2003), spatial statistics (Besag 1974). In Economics it is often necessary to work around the proliferation of dichotomic predictors such as in deterministic weekly seasonal adjustment (Pierce et al. 1984), or when the empirical analysis of a large number of interactions is needed in interaction based models (Brock and Durlauf 2001).

The multivariate Bernoulli distribution (Whittaker 1990; Dai et al. 2013) models potentially dependent binary variables and is a member of the exponential family. Because unordered categorical variables can be represented by binary variables, the multivariate Bernoulli can be used to model both binary and categorical variables. In this section we turn our attention to regressions with jointly Bernoulli inverse predictors and apply our methodology in order to identify and estimate the sufficient reductions.

7.1 The Ising Model

The probability mass function of the multivariate Bernoulli involves terms representing third and higher order moments. Graphical models have been used to represent the joint distribution of categorical

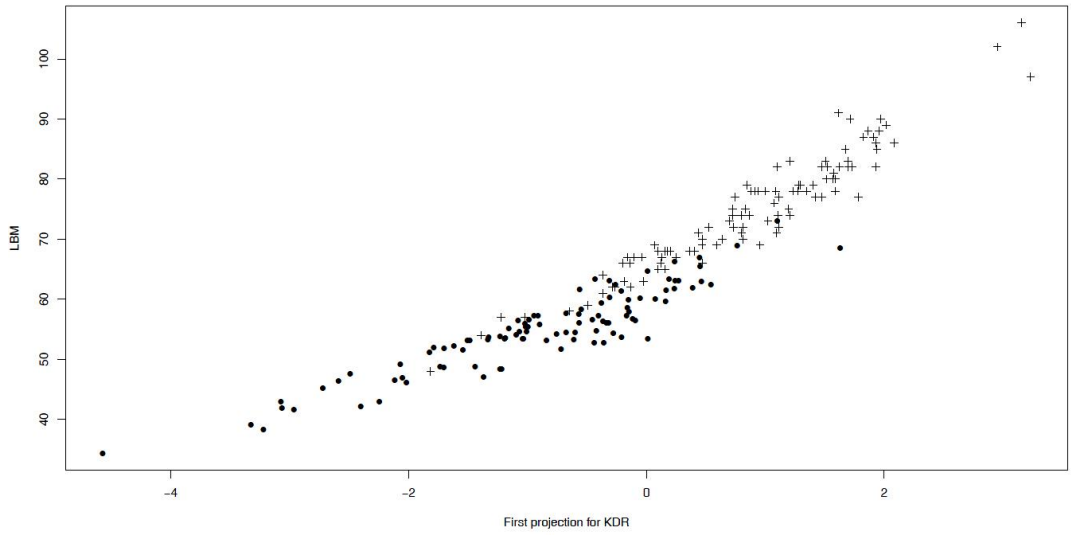


Figure 5: LBM versus the first KDR predictor (Dots for females, crosses for males)

variables. While graphical models can easily capture pairwise correlations, higher order interactions are extremely complex to accommodate. Moreover, the estimation is computationally infeasible for realistic size networks as the evaluation of the likelihood is required.

The Ising model (Ising 1925) is an undirected graphical model that allows up to pairwise interaction effects and it has been used to model multivariate binary data at great extent. Even though the Ising model is a special case of the multivariate Bernoulli distribution, Wainwright and Jordan (2008) showed that it can accommodate more general dependence structures as higher order interactions can be converted to pairwise ones through the introduction of additional variables (also see Ravikumar et al. 2010).

We model binary inverse predictors using the Ising model. Suppose we have p binary inverse predictors $X_1|Y, \ldots, X_p|Y$, with $X_j|Y \in \{1, 0\}, 1 \leq j \leq p$, whose joint distribution has density function,

$$p(x_1, \ldots, x_p|Y = y) = \frac{1}{Z(\mathbf{\Theta}(y))} \exp\left(\sum_{j=1}^{p} \theta_{jj}(y)x_j + \sum_{1\leq j\leq j'\leq p} \theta_{jj'}(y)x_j x_{j'}\right) \tag{11}$$

where $\mathbf{\Theta}(y) = (\theta_{jj'}(y))_{p\times p}$ is a symmetric matrix specifying the network structure. The partition function $Z(\boldsymbol{\theta}(y)) = \sum_{X_j\in\{0,1\},1\leq j\leq p} \exp(\sum_{j=1}^{p} \theta_{jj}(y)X_j + \sum_{1\leq j<j'\leq p} \theta_{jj'}(y)X_j X_{j'})$ ensures that the density function in (11) is proper and integrates to one.

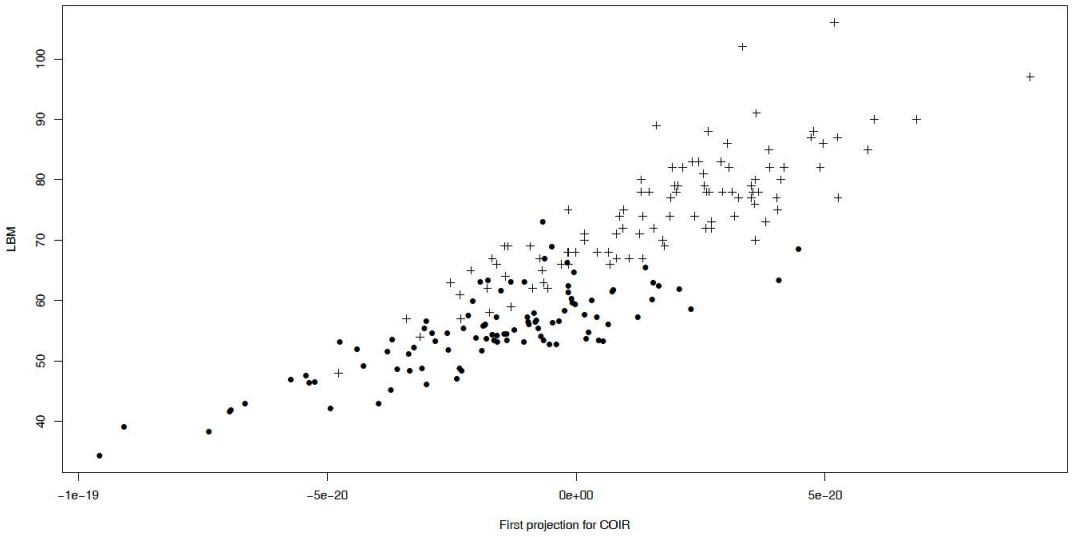


Figure 6: LBM versus the first COIR predictor (Dots for females, crosses for males)

The term $\theta_{jj}(y)$, $1 \leq j \leq p$, corresponds to the main effect for variable $X_j|Y$, and $\theta_{jj'}(y)$, $1 \leq j < j' \leq p$, corresponds to the interaction effect between $X_j|Y$ and $X_{j'}|Y$. The $\theta_{jj'}(y)$s reflect the structure of the underlying network and Cheng et al. (2012) showed that they can be connected to conditional log-odds via

$$\log \frac{P(X_j = 1|\mathbf{X}_{-j}, y)}{1 - P(X_j = 1|\mathbf{X}_{-j}, y)} = \theta_{jj}(y) + \sum_{j' \neq j} \theta_{jj'}(y) X_{j'}$$

where $\mathbf{X}_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$. Also, by conditioning on $\mathbf{X}_{-j,-j'} = \mathbf{0}$, Cheng et al. (2012) obtained

$$\theta_{jj'}(y) = \log \frac{P(X_j = 1, X_{j'} = 1|\mathbf{X}_{-j,-j'}, Y) P(X_j = 0, X_{j'} = 0|\mathbf{X}_{-j,-j'}, Y)}{P(X_j = 1, X_{j'} = 0|\mathbf{X}_{-j,-j'}, Y) P(X_j = 0, X_{j'} = 1|\mathbf{X}_{-j,-j'}, Y)}$$

which implies that X_j and $X_{j'}$ are conditionally independent given Y and all other $\mathbf{X}$-variables if and only if $\theta_{jj'}(y) = 0$, and hence their corresponding nodes are not connected.

	$\widehat{\boldsymbol{\alpha}}$	W
Partial SIR		
	$(0.18, 0.98, 0.01, 0.02, 0.12)$	
Cook and Li		
	$(0.88, 0.45, 0.10, 0.02, 0.11)$	0.04
	$(0.92, 0.37, 0.07, 0.02, 0.08)$	0.10
EF-DR		
	$(0.27, 0.92, -0.03, -0.04, 0.21)$	-0.17
KDR		
	$(-0.29, -0.91, -0.03, 0.08, -0.29)$	

Table 1: Reduction Coefficients

partial SIR	Cook-Li	EF-DR	KDR	KSIR	COIR
2.660	3.563	2.573	3.160	5.580	8.804

Table 2: Prediction Errors

7.1.1 Estimation of the Sufficient Reduction

The Ising probabibility function (11) belongs to the exponential family with natural parameter vector $\boldsymbol{\eta}_y = (\theta_{11}(y), \theta_{22}(y), \ldots, \theta_{pp}(y), \theta_{12}(y), \ldots, \theta_{p-1,p}(y))^T$, and sufficient statistic

$$\mathbf{T}(\mathbf{X}) = (X_1, \ldots, X_p, X_1X_2, \ldots, X_1X_p, \ldots, X_{p-1}X_p)^T, \tag{12}$$

both with $p + p(p-1)/2$ elements. Applying Theorem 1, the sufficient reduction for the regression of Y on $\mathbf{X}$ is $\boldsymbol{\alpha}^T(\mathbf{T}(\mathbf{X}) - \mathrm{E}(\mathbf{T}(\mathbf{X}))$, where $\boldsymbol{\alpha} = \mathrm{span}(\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}})$.

We use multivariate logistic regression to estimate the natural parameters $\boldsymbol{\eta}_Y$. Suppose n samples are drawn from Y and $\mathbf{X}$, denoted by $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. The log-likelihood for the generalized linear model of the Ising distribution is

$$l(\mathbf{x}, y) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{p} x_{ij}\theta_{jj} + \sum_{j<j'} \theta_{jj'}x_{ij}x_{ij'} - \log\left\{ 1 + \sum_{j=1}^{p} \exp(\theta_{jj} + \sum_{j'<j} \theta_{j'j}) \right\} \right\} \tag{13}$$

The maximization of the log-likelihood in (13) with respect to the coefficients of the $\mathbf{f}_y$ components in the generalized linear model for the natural parameters $\boldsymbol{\eta}_y$ can be done as in Section 4, and the

IRLS algorithm can be used to estimate $\boldsymbol{\alpha}$ via fitting a multivariate logistic regression model. The dimension of $\boldsymbol{\alpha}$ can be estimated either with the asymptotic tests in Section 5, or with an information based criterion such as AIC or BIC. The sufficient reduction for the regression of Y on the binary $\mathbf{X}$'s is then $(\widehat{\boldsymbol{\alpha}}_1, \ldots, \widehat{\boldsymbol{\alpha}}_{\hat{d}})^T \left(\mathbf{T}(\mathbf{X}) - \bar{\mathbf{T}}(\mathbf{X})\right)$, where $\mathbf{T}(\mathbf{X})$ is given in (12), $\bar{\mathbf{T}}$ is the sample mean of $\mathbf{T}(\mathbf{X})$, and $\hat{d}$ is the estimated dimension. In the next two sections, 7.2 and 7.3, we illustrate how to apply our methodology in regressions with Bernoulli inverse predictors.

7.2 Simulation Experiments

Suppose Y is normal with zero mean and standard deviation 0.5. Given Y, we let $\mathbf{X} = (X_1, \ldots, X_p)^T$ be a Bernoulli vector with p components with pairwise correlation structure among contiguous pairs as follows, (X_1, X_2), $(X_3, X_4), \ldots, (X_{p-1}, X_p)$, and all other interactions of all orders are zero. Therefore, $\boldsymbol{\eta}_Y = (\theta_{11}, \theta_{22}, \ldots, \theta_{pp}, \theta_{12}, \theta_{34}, \ldots, \theta_{(p-1)p})$.

The sufficient statistic is $\mathbf{T}(\mathbf{X}) = (X_1, X_2, \ldots, X_p, X_1X_2, X_3X_4, \ldots, X_{p-1}X_p)$. The natural parameter $\boldsymbol{\eta}_Y$ is generated as

$$\boldsymbol{\eta}_Y = \mathbf{A}\mathbf{C}^T(\mathbf{f}_Y - \mathrm{E}\mathbf{f}_Y)$$

where $\mathbf{C} = 1$, $\mathbf{f}_Y = Y$. For $p = 4$, $\mathbf{A} = (1, 1, 1, 1, 10, 10)^T/\sqrt{204}$, and for $p > 4$, we set $A_5 = \ldots = A_p = 0$, so that, for all p, the minimal sufficient reduction is $[(X_1 - \mathrm{E}(X_1)) + (X_2 - \mathrm{E}(X_2)) + (X_3 - \mathrm{E}(X_3)) + (X_4 - \mathrm{E}(X_4)) + 10(X_1X_2 - \mathrm{E}(X_1X_2)) + 10(X_3X_4 - \mathrm{E}(X_3X_4))]/\sqrt{204}$.

We compute the accuracy of the estimation of the sufficient reduction in three cases: (a) assuming the true correlation structure; (b) assuming the $\mathbf{X}$ components are independent given Y, which is Cook and Li's (2009) approach; and (c) assuming that all pairwise interactions are present as in the full Ising model. Accuracy is measured as the maximum angle between the true subspace spanned by $\boldsymbol{\eta}_Y - \bar{\boldsymbol{\eta}}$, $Y \in \mathcal{S}_Y$, and the estimated one. We report results for $n = 100, 200, 300, 500$, and $p = 4, 6, 10$.

The estimation under (b) and (c) is carried out using the R package MVB developed by Dai (2013). Dai et al. (2013) studied the multivariate Bernoulli distribution as well as estimation in a generalized linear model that includes higher order interactions among the covariates. We applied our algorithm for the estimation of the natural parameter vector under the true model (a). We note that the current version of the MVB package does not allow setting specific individual second order, or any other higher order interactions, to zero. To the best of our knowledge, our IRLS algorithm is the only computational tool

for fitting the true model.

In Table 4 we report the time it takes to compute the reductions under (a), (b) and (c) for increasing values of sample size and number of binary predictors in one replication. As expected, the full Ising model is the most computionally demanding, followed by our method under the true correlation structure. The conditional independence model (b) is by far the fastest, since much fewer quantities are computed.

	$n = 100$	$n = 200$	$n = 300$	$n = 500$	N
$p = 4$					
(a)	35.12 (16.65)	26.29 (12.14)	21.94 (8.50)	17.49 (8.09)	100
(b)	43.69 (16.30)	39.29 (12.52)	37.64 (10.49)	37.19 (7.33)	100
(c)	45.81 (15.47)	41.12 (13.78)	30.95 (13.62)	29.56 (12.24)	100
$p = 6$					
(a)	40.42 (12.10)	34.82 (13.15)	28.76 (16.96)	23.31 (13.24)	50
(b)	64.24 (15.84)	63.16 (23.91)	58.08 (25.59)	55.42 (30.12)	50
(c)	51.06 (14.56)	42.51 (13.25)	38.21 (13.78)	36.80 (12.89)	50
$p = 10$					
(a)	50.38 (10.78)	40.43 (10.27)	33.97 (10.47)	28.84 (9.16)	50
(b)	75.23 (10.86)	72.16 (11.93)	73.14 (12.11)	74.45 (9.90)	50
(c)	57.12 (13.41)	55.63 (12.71)	50.22 (12.03)	47.01 (10.82)	50

Table 3: Mean angles and their standard deviations in parentheses between the true and estimated subspaces

The average angles and their standard deviations, in parentheses, between the true and estimated reductions under (a), (b) and (c) are reported in Table 3. The column headed by N reports the number of replications for each variable number-sample size combination. From Table 3 we see that when we account for the true correlation structure of the inverse predictors, the accuracy of the estimation of the sufficient reduction rapidly increases as the sample size increases and is uniformly much better than under both (b) and (c) schemes across sample sizes. It is also noteworthy that the over-parametrized full Ising model (c) that contains all second order interactions yields better estimates compared to the simple yet naive model (b) of conditional independence.

When the correlation structure of the inverse predictors is incorporated in the estimation process, the accuracy improves significantly and is much better than both competing approaches. The question

	$n = 100$	$n = 200$	$n = 300$	$n = 500$
$p = 4$				
(a)	20.99	42.09	63.94	105.9
(b)	1.98	4.15	5.8	11.14
(c)	30.12	52.23	89.21	200.98
$p = 6$				
(a)	87.36	183.76	292.87	492.14
(b)	2.35	5.32	8.38	12.85
(c)	102.23	207.69	387.21	605.41
$p = 10$				
(a)	2264.08	3888.49	5327.56	11962.96
(b)	2.73	3.34	8.25	13.1
(c)	3675.45	4671.90	6713.13	13945.12
$p = 12$				
(a)	6158.65	13925.21	18243.74	26580.66
(b)	3.26	5.74	8.89	14.50
(c)	8712.21	15092.01	20193.12	28143.81

Table 4: Computational time (in seconds) for one replication of the reduction estimation under (a), (b) and (c) for increasing n and p.

then arises how one can deduce the correlation structure in real data analysis problems with binary predictors. An approach, also supported by the simulation results in Table 3, is to fit a full Ising model, which has $p + p(p-1)/2$ parameters to estimate, and test for effect significance. For large p, this would be practically impossible. For example, even for a relatively small number of predictors such as $p = 10$, there are 55 parameters to estimate and 210 for $p = 20$.

Alternatively, one can screen the inverse predictor network. If the interaction terms in a multivariate Bernoulli distribution are zero, the corresponding terms are independent (Dai et al. 2013). Using this fact, one can control the dependence structure among the components of a Bernoulli vector by setting the corresponding elements of the natural parameter vector to zero. This approach can be very useful when the Bernoulli vector has large dimension and the number of parameters to be estimated makes the problem infeasible for realistic sample sizes. Cheng et al. (2012) proposed a l_1-based penalization for interaction and variable selection in Ising models. We apply their method in order to identify which main and second order effects to include in the generalized linear model and then use our IRLS estimation

algorithm in the Zoo data analysis that follows.

7.3 Zoo Data

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 subjects 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 fifteen predictors are dichotomous, and the legs variable is polychotomous.

Cook and Li (2009) analyzed this data set. The fifteen dichotomous predictors, $\mathbf{X}$, classify animals into the seven animal categories, Y. The polychotomous legs variable is omitted from the analysis in order to focus on comparing the classification accuracy under the assumption of independent and dependent inverse Bernoulli predictors.

Under conditional independence of the predictors $\mathbf{X}$ given the response Y, the multivariate Bernoulli logistic model can be easily fitted via iterative re-weighted least squares. Since the response is categorical, we define a function $\mathbf{f}_y$ with kth coordinate

$$f_{yk} = I(y = k) - \frac{n_k}{n} \quad \text{for } k = 1, \ldots, r,$$

$r = 7 - 1 = 6$, where 7 is the number of distinct values of the response. The function $I(\cdot)$ is the indicator function and n_k is the number of observations in category k (see, also, Cook 2007). The natural parameters $\theta_{jj}(y)$ in (11) can then be written as a linear function of $\mathbf{f}_y$ using

$$\theta_{jj}(y) = \gamma_{0j} + \gamma_{1j} f_{y1} + \ldots + \gamma_{kj} f_{yr} \tag{14}$$

where $\boldsymbol{\gamma}_j = (\gamma_{0j}, \ldots, \gamma_{rj})^T$ is the coefficient vector to be estimated for $j = 1, \ldots, p$. There are $7 \times 15 = 105$ coefficients to be estimated in the simple case of independence.

With n samples on Y and $\mathbf{X}$, (14) can be written as

$$\boldsymbol{\eta}_n = \mathbf{F}_n \boldsymbol{\Gamma} \tag{15}$$

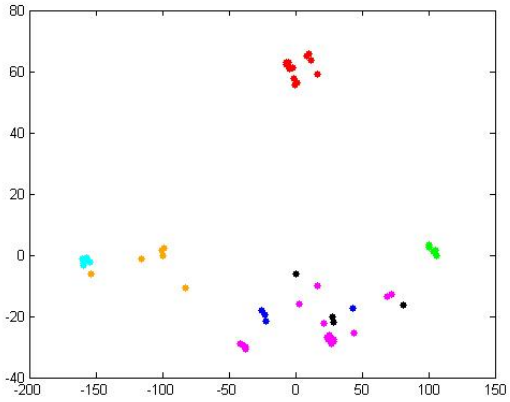
where $\boldsymbol{\eta}_n = (\theta_{i(jj)})$ is an $n \times p$ random matrix with $\theta_{i(jj)} = \theta_{jj}(y_i)$, $\mathbf{F}_n = (f_{il}) = (f_{y_i,l})$, an $n \times (r+1)$

fixed matrix, and $\mathbf{\Gamma} = (c_{lj})$, the $(r+1) \times p$ matrix of coefficients. Observe that $\text{rank}(\mathbf{F}_n\mathbf{\Gamma}) = \text{rank}(\mathbf{\Gamma}^T\mathbf{F}_n\mathbf{F}_n^T\mathbf{\Gamma}) = \text{rank}(\mathbf{\Gamma})$, since $\mathbf{F}_n^T\mathbf{F}_n$ is a positive definite matrix (see section A4.4 of Seber (1977)). From (15), the natural parameter space is spanned by the rows of $\mathbf{F}_n\mathbf{\Gamma}$, i.e. $\text{span}(\boldsymbol{\eta}_y) = \text{span}(\mathbf{\Gamma}^T\mathbf{F}_n^T) = \text{span}(\mathbf{\Gamma}^T)$. In consequence, inference on the dimension d of $\text{span}(\boldsymbol{\eta}_y)$ can be based solely on $\mathbf{\Gamma}$ in the sense that an estimate of the rank of $\mathbf{\Gamma}$ constitutes an estimate of the dimension of $\text{span}(\boldsymbol{\eta}_y)$. The d eigenvectors of the IRLS estimate of $\mathbf{\Gamma}^T$ that correspond to its d largest eigenvalues, yield estimates of the basis vectors of $\text{span}(\boldsymbol{\eta}_y)$, and the resulting d linear combinations of the centered $\mathbf{T}(\mathbf{X})$ are the sufficient predictors.

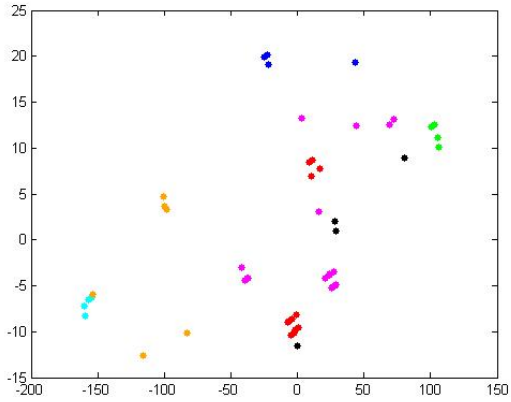
The MVB package yields parameter estimates for $\mathbf{\Gamma}$, which we expect to be somewhat inaccurate since $n = 101$. Under conditional independence, the MVB package yields the same estimates as Cook and Li (2009) and the two methods can be used interchangeably. In Figure 7, the sufficient reductions 1, 3, 5, and 6 are plotted. The colors serve to identify the different categories with blue for amphibian ($Y = 1$), red for bird ($Y = 2$), green for fish ($Y = 3$), cyan for insect ($Y = 4$), magenta for mollusc.et.al ($Y = 5$), gold for mammal ($Y = 6$), and black for reptile ($Y = 7$). We see that under conditional independence of the Bernoulli predictors given the response, the first versus the third in panel (a) can separate red (bird), green (fish), but even though cyan (insect) and gold (mammal) appear as separate, one gold is covered by cyan. The first versus the sixth in panel (b) can separate green (fish) and blue (amphibian) easily. Also, there is separation between red (bird) and black (reptile), as one can draw a closed curve around all black points without intersecting any other group. This can be done using machine learning algorithms for pattern classification that are based on image segmentation and optimization techniques used in level sets (see, for example, Varshney and Willsky 2010; Cai and Sowmya 2007). The third versus the fifth direction in panel (c) can separate red (bird), magenta (molusc et al.), cyan (insect) and green (fish) easily, and one can draw closed curves around the gold (mammal) and either blue (amphibian) or black (reptile) points, but with blue and black being too close for differentiation.

Even though we report results for some reductions, the singular values of $\widehat{\mathbf{\Gamma}}^T$ are 120.87, 100.81, 67.03, 48.25, 35.52, 22.75, all so substantially far from zero to indicate that the dimension of the problem is six. The sufficient reductions that were plotted in Figure 7 were selected as they provide the clearest visual separation. We note that the reductions that can separate the classes are not ordered according to their separation potential and several are needed.

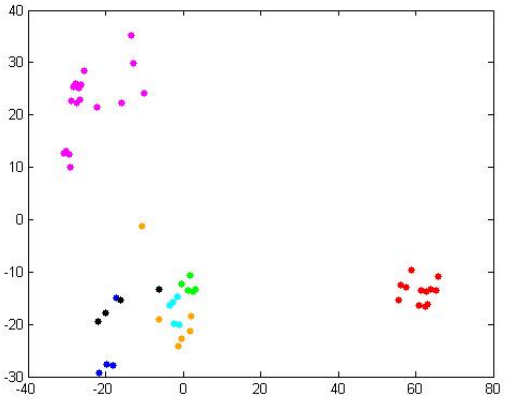
With $n = 101$ it is impossible to fit the full pairwise dependence Ising model, as the number



(a) SR_1 versus SR_3 under Independence



(b) SR_1 versus SR_6 under Independence



(c) SR_3 versus SR_5 under Independence

Figure 7: Scatterplots of Sufficient Reductions 1, 3, 5, 6 under conditional independence of $\mathbf{X}$ given Y.

of natural parameters in this model is $p + \binom{p}{2} = 120$ and fitting the corresponding model to (14) would require the estimation of $7 \times 120 = 840$ parameters with 101 observations. For this reason, we assume that the Ising model is sparse; that is, that some natural parameters $\theta_{jk}(y)$ are zero, for some $j \leq k = 1, \ldots, p$.

The l_1 penalties of Cheng et al. (2012) induce sparsity both in the natural parameter vector $\boldsymbol{\eta}_y = (\theta_{11}(y), \ldots, \theta_{pp}(y), \theta_{12}(y), \ldots, \theta_{p-1,p}(y))^T$, and in the (adjusted) coefficient matrix $\boldsymbol{\Gamma}$, which is now of dimension $(r+1) \times \left(p + \binom{p}{2}\right) = 7 \times 120$ in the accordingly enlarged model (14). We applied the *joint* estimation algorithm of Cheng et al. (2012) to the Zoo data. The Matlab code was provided by Dr. Cheng. The singular values of the estimated adjusted $\boldsymbol{\Gamma}$ matrix are 12.07, 10.09, 9.91, 9.62, 7.91,

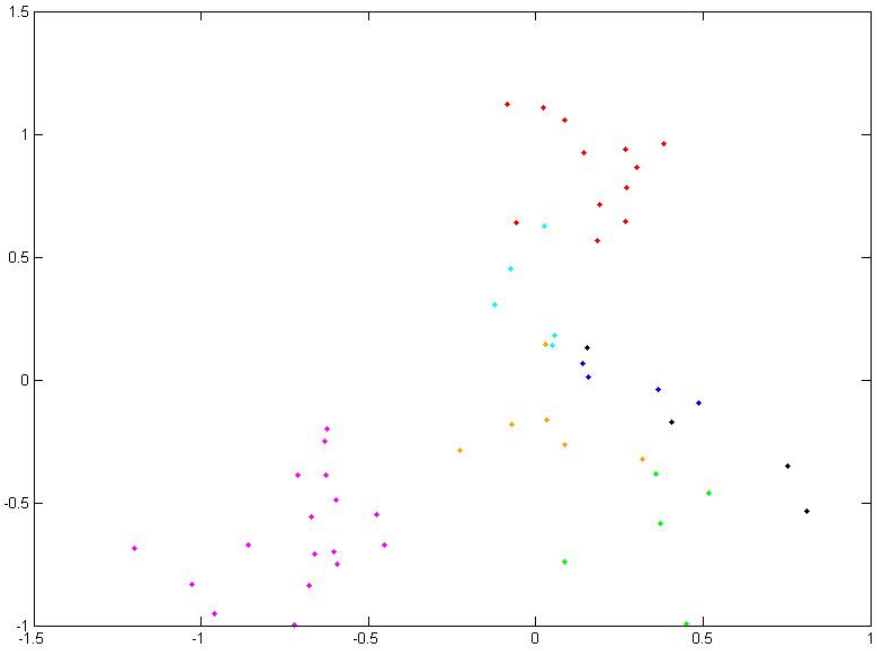


Figure 8: EF-DR: SR_1 versus SR_2 under Dependence

7.46. The third and fourth and the fifth and sixth singular values are very similar, indicating that the corresponding left singular vectors define the same eigenspace.

The reductions are of the form $\widehat{\boldsymbol{\alpha}}^T(\mathbf{T}(\mathbf{X}) - \bar{\mathbf{T}}(\mathbf{X}))$. At the start, i.e. prior to applying Cheng et al.'s (2012) variable selection, $\mathbf{T}(\mathbf{X}) = (X_1, \ldots, X_{15}, X_1X_2, \ldots, X_{14}X_{15})$ with 120 terms. After screening for sparsity, 66 of its terms were retained in the computation of the sufficient reduction. As so many are left, we do not explicitly report either $\widehat{\boldsymbol{\alpha}}$ or the sufficient reduction, but note that the X_1, X_5, X_7, X_{11} and X_{14} main effects are active in the sufficient reduction. They are indicators of whether the animals have hair, are airborne, predatory, venomous and domestic, respectively. All other 61 effects are interactions, suggesting that the predictors are dependent.

In Figure 8 we plot the first two sufficient reductions. We can see that all colors are separated by simple closed curves, the black and blue points included. That is, two sufficient reductions can separate all seven classes of points, and, therefore, we estimate the dimension of the regression to be two, in contrast to six under the independence assumption. For the Zoo data, EF-DR offers perfect in-sample classification and significantly reduces complexity.

We also applied KDR with Gaussian kernel to the Zoo data. The seven classes cannot be separated

with closed curves in the plot of the first two KDR directions, as in EF-DR. In order to compare numerically the classification accuracy of EF-DR and KDR, we used linear discriminant analysis (LDA) and the naive Bayes version of quadratic discriminant analysis (dQDA), which uses diagonal covariance matrix estimates, since some class covariance matrices were singular. This form of QDA classifier operates under the assumption of class-conditional independence between variables, which, even though is not true in general, has been found to work well in practice on many data sets. KDR was applied directly to the data without imposing any sparsity constraints such as in the implementation of EF-DR. It can be seen from Table 5, where the LDA and diagonal QDA misclassification rates of the two methods and two dimension values for the reduction ($d = 2, 3$) are reported, that EF-DR has better classification accuracy with both classifiers.

	EF-DR	KDR
$d = 2$		
LDA	0.139	0.297
dQDA	0.158	0.257
$d = 3$		
LDA	0.119	0.158
dQDA	0.109	0.139

Table 5: LDA and dQDA misclassification errors

8 Discussion

Until very recently, sufficient dimension reduction meant finding linear combinations of the predictors that are sufficient for the forward regression. Reproducing kernel Hilbert spaces (RKHS)-based methods, in which the predictors are immersed in a higher dimensional space so that linear sufficient dimension methodology can be applied, went a step further to obtain functions of the predictors that are sufficient but not necessarily linear in the original predictors, though for most it is not possible to extract an explicit form of the reduction. The generality of kernel-based methods, as they are applicable to any types of data, is their most appealing feature. For the same reason, they can be less efficient than methods that take advantage of the distribution of the data when the latter is known or easy to deduce.

Model based SDR emerged from the connection that Cook (2007) drew between sufficient statistics and sufficient reductions. We make use of this connection to compute sufficient reductions in regressions

with exponential family inverse predictors. The most attractive features of the sufficient reductions we derive with EF-DR are that (1) they are exhaustive, (2) they are linear functions of the sufficient statistics, which can be both linear and nonlinear functions of the predictors, and (3) they have explicit functional forms. Moreover, the estimates of the sufficient reductions are MLEs and hence efficient.

Our results hold true when the response and the predictors have a joint exponential family distribution, since the conditional distribution of any subvector of $(Y, \mathbf{X})$ given the rest is also in the exponential family. Our methodology easily extends to the case of a vector response.

A potentially challenging aspect of our methodology stems from the fact that we regress a multivariate vector of exponential family distributed predictors, which can be any mixture of continuous and/or categorical variables, on functions of the response that requires the adjustment of the IRLS algorithm to the specific joint distribution of $\mathbf{X}|Y$. We touch upon the potential computational difficulties in the analysis of the Australian athletes data example in Section 6.1.

Acknowledgment:

We would like to thank Prof. Levina and Dr. Cheng for providing us with the Matlab code that implements their sparsity inducing algorithm in the Ising model. We also thank two anonymous referees and the associate editor whose comments and suggestions helped us improve the paper.

Appendix:

Initial values and Estimation Algorithm for the EF-DR sufficient reduction in the Australian Athletes Data Analysis

We assume that for each $Y = y$, $\boldsymbol{\eta}_y = \bar{\boldsymbol{\eta}} + \mathbf{D}(\mathbf{f}_y - \bar{\mathbf{f}}_y) \in \mathbb{R}^{p+1}$. Under this setup, we let $\boldsymbol{\Gamma}^T = (\bar{\boldsymbol{\eta}}, \mathbf{D})$ be the $(r+1) \times (p+1)$ matrix of unknown coefficients in the GLM model (5), and write $\boldsymbol{\gamma} = \text{vec}(\boldsymbol{\Gamma}) = \text{vec}\begin{pmatrix} \bar{\boldsymbol{\eta}}^T \\ \mathbf{D}^T \end{pmatrix} \in \mathbb{R}^{(p+1)(r+1)\times 1}$, so that

$$\boldsymbol{\eta}_y = \begin{pmatrix} \boldsymbol{\eta}_1 \\ \eta_2 \end{pmatrix} = \left(\mathbf{I}_{p+1} \otimes (1, (\mathbf{f}_y - \bar{\mathbf{f}}_y)^T)\right) \boldsymbol{\gamma} = \mathbb{F}_y \boldsymbol{\gamma}$$

with $\boldsymbol{\eta}_1 \in \mathbb{R}^{p\times 1}$ and $\eta_2 \in \mathbb{R}$, and $\mathbb{F}_y = \mathbf{I}_{p+1} \otimes (1, (\mathbf{f}_y - \bar{\mathbf{f}}_y)^T) : (p+1) \times (p+1)(r+1)$. Note that $\bar{\boldsymbol{\eta}}$

is $(p+1) \times 1$ and $\mathbf{D}$ is of order $(p+1) \times r$, where p is the number of $\mathbf{X}$ predictors and the additional dimension is due to the binary variable W.

The parameters of interest are $\mathbf{D}$ and $\bar{\boldsymbol{\eta}}$, whereas $\boldsymbol{\Delta}$ and $\mathbf{b}_2$ are nuisance parameters. The latter are fixed at their initial values whereas $\mathbf{D}$ and $\bar{\boldsymbol{\eta}}$ are updated in each iteration of the estimation algorithm.

Initial values and algorithm:

1. (a) Regress $\mathbf{X}_{ctd} = \mathbf{X} - \bar{\mathbf{X}} : p \times 1$ on the centered $\mathbf{f}_y : r \times 1$ and centered W,

$$\mathbf{X}_{ctd} = \mathbf{b}_1(\mathbf{f}_y - \bar{\mathbf{f}}_y) + \mathbf{b}_2(W - \bar{W}) + \boldsymbol{\epsilon} \tag{16}$$

   to obtain the least squares estimates $\widehat{\mathbf{b}}_2 : p \times 1$ and $\widehat{\mathbf{b}}_1 : p \times r$. Let $\widehat{\boldsymbol{\Delta}}$ denote the covariance matrix of the residuals from fitting (16).

   (b) Fit a logistic regression for W on $\mathbf{f}_y - \bar{\mathbf{f}}_y$,

$$\log \frac{p_y}{1-p_y} = h_1 + \mathbf{h}_2^T(\mathbf{f}_y - \bar{\mathbf{f}}_y)$$

   and obtain the IRLS estimates $\widehat{h}_1$ and $\widehat{\mathbf{h}}_2$, or, equivalently, the MLEs of h_1 and $\mathbf{h}_2$, respectively.

2. The initial value for $\boldsymbol{\gamma}$ is

$$\boldsymbol{\gamma}_0 = \text{vec}\begin{pmatrix} \bar{\boldsymbol{\eta}}_0^T \\ \mathbf{D}_0^T \end{pmatrix}$$

   where

   (a) $\bar{\boldsymbol{\eta}}_0 = (\mathbf{A}_{10}^T, A_{20})^T$, $\mathbf{A}_{10} : p \times 1$, $A_{20} \in \mathbb{R}$, and $\widehat{\mathbf{D}}_0 = (\mathbf{D}_{10}, \mathbf{D}_{20})^T$, with $\mathbf{D}_{10} : r \times p$, $\mathbf{D}_{20} : r \times 1$

   (b) $\mathbf{A}_{10} = \mathbf{0}$ and $\widehat{\mathbf{D}}_{10}^T = \widehat{\boldsymbol{\Delta}}^{-1}\widehat{\mathbf{b}}_1$

   (c) $A_{20} = \widehat{h}_1$ and $\mathbf{D}_{20} = \widehat{\mathbf{h}}_2 - \widehat{\mathbf{b}}_2^T\widehat{\mathbf{D}}_{10}^T$

3. We update $\boldsymbol{\gamma}$ by

$$\boldsymbol{\gamma}^{(t+1)} = \left(\sum_{i=1}^n \mathbb{F}_{y_i}^T \mathbf{W}_{y_i}^{(t)} \mathbb{F}_{y_i}\right)^{-1} \sum_{i=1}^n \mathbb{F}_{y_i}^T \mathbf{W}_{y_i}^{(t)} \left(\mathbb{F}_{y_i}\boldsymbol{\gamma}^{(t)} + (\mathbf{W}_{y_i}^{(t)})^{-1}(\mathbf{T}(\mathbf{x}, w) - d_{y_i}^{(t)})\right)$$

where for each $i = 1, \ldots, n$,

(a)

$$d_{y_i}^{(t)} = \begin{pmatrix} \widehat{\boldsymbol{\Delta}} \boldsymbol{\eta}_{1i}^{(t)} + \widehat{\mathbf{b}}_2 (p_{y_i}^{(t)} - \bar{w}) \\ p_{y_i}^{(t)} - \bar{w} \end{pmatrix}$$

using that $\mathrm{E}(\mathbf{T}(\mathbf{X}, W)|Y) = \nabla\psi(\boldsymbol{\eta}_y)$, and $\partial l / \partial \boldsymbol{\eta}_y = \mathbf{T}(\mathbf{X}, W) - \nabla\psi(\boldsymbol{\eta}_y)$,

(b)

$$\mathbf{W}_{y_i}^{(t)} = \begin{pmatrix} \widehat{\boldsymbol{\Delta}} + \widehat{\mathbf{b}}_2 \widehat{\mathbf{b}}_2^T p_{y_i}^{(t)} (1 - p_{y_i}^{(t)}) & p_{y_i}^{(t)} (1 - p_{y_i}^{(t)}) \widehat{\mathbf{b}}_2 \\ \widehat{\mathbf{b}}_2^T p_{y_i}^{(t)} (1 - p_{y_i}^{(t)}) & p_{y_i}^{(t)} (1 - p_{y_i}^{(t)}) \end{pmatrix}$$

using $\mathrm{var}(\mathbf{T}(\mathbf{X}, W)|Y) = \partial^2 \psi(\boldsymbol{\eta}_y) / \partial \boldsymbol{\eta}_y^T \partial \boldsymbol{\eta}_y$, where

- $\boldsymbol{\eta}_{y_i}^{(t)} = \begin{pmatrix} \boldsymbol{\eta}_{1i}^{(t)} \\ \eta_{2i}^{(t)} \end{pmatrix} = \mathbb{F}_{y_i} \boldsymbol{\gamma}^{(t)}$, $\boldsymbol{\eta}_{1i}^{(t)} \in \mathbb{R}^{p+1}$ and $\eta_{2i}^{(t)} \in \mathbb{R}$
- $p_{y_i}^{(t)} = \dfrac{e^{\eta_{2i}^{(t)} + \widehat{\mathbf{b}}_2^T \boldsymbol{\eta}_{1i}^{(t)}}}{1 + e^{\eta_{2i}^{(t)} + \widehat{\mathbf{b}}_2^T \boldsymbol{\eta}_{1i}^{(t)}}}$

4. Repeat step 3 until convergence.

References

[1] Adragni, K. and Cook, R. D. (2009), "Sufficient Dimension Reduction and Prediction in Regressions," *Phil. Trans. Royal Soc A*, 367(1906): 4385 - 4405.

[2] Akaho, S. (2001), "A kernel method for canonical correlation analysis," In *Proceedings of the International Meeting of the Psychometric Society (IMPS2001)*. Springer, Tokyo.

[3] Bach, F. R. and Jordan, M. I. (2002), "Kernel independent component analysis," *J. Mach. Learn. Res.*, 3, 1-48.

[4] Berger, R.L. (1982), "Multiparameter hypothesis testing and acceptance sampling," *Technometrics*, 24, 295-300.

[5] Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society, Series B*, 36, 192-236.

[6] Bickel, P. J. and Doksum, K. (2006), *Mathematical Statistics, Basic Ideas and Selected Topics,* Vol I, Prentice Hall, Saddle River, NJ.

[7] Brock, W. and Durlauf, S. (2000), “Interactions Based Models” in *Handbook of Econometrics*, ed. by J. Heckman and E. Learner. Amsterdam: North-Holland.

[8] Bura, E. and Cook, R.D. (2001a), "Estimating the structural dimension of regressions via parametric inverse regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, 393–410.

[9] Bura, E. and Cook, R.D. (2001b), “Extending SIR: The Weighted Chi-Square Test," *Journal of the American Statistical Association*, 96, 996-1003.

[10] Bura, E. and Forzani, L. (2015), "Sufficient reductions in regressions with elliptically contoured inverse predictors," *Journal of the American Statistical Association*, 110, 420-434.

[11] Bura, E. and Yang, J. (2011), "Dimension Estimation in Sufficient Dimension Reduction: A Unifying Approach," *Journal of Multivariate Analysis*, 102, 130-142.

[12] Cai, X. and Sowmya, A. (2007), “Level Learning Set: A novel classifier based on active contour models," J.N. Kok et al. (Eds.): ECML 2007, LNAI 4701, pp. 79–90, Berlin: Springer-Verlag.

[13] Cheng, J., Levina, E., Wang, P., and Zhu, J. (2012), “A sparse Ising model with covariates," *Biometrics*, 70, 943-953.

[14] Chiaromonte, F., Cook, R.D. and Li, B. (2002), "Sufficient dimension reduction in regression with categorical predictors," *The Annals of Statistics*, 30, 475-497.

[15] Cook, R. D. (1994), "Using dimension-reduction subspaces to identify important inputs in models of physical systems," In *Proc. Sect. Phys. Eng. Sc.* , p. 18-25. Alexandria, VA: American Statistical Association

[16] Cook, R. D. (1998a), *Regression Graphics: Ideas for studying regressions through graphics*, New York: Wiley.

[17] Cook, R. D. (1998b), "Principal Hessian directions revisited (with discussion)," *Journal of the American Statistical Association,* 93, 84-94.

[18] Cook, R. D. (2007), "Fisher lecture: Dimension reduction in regression," *Statistical Science*, 22, 1-26.

[19] Cook, R.D., and Forzani, L. (2008), "Principal Fitted Components for Dimension Reduction in Regression," *Statistical Science*, 23, 485-501.

[20] Cook, R. D. and Forzani, L. (2009), "Likelihood-Based Sufficient Dimension Reduction," *Journal of the American Statistical Association*, 104, 197-208.

[21] Cook, R. D. and Lee, H. (1999), "Dimension reduction in binary response regression," *Journal of the American Statistical Association*, 94, 1187-1200.

[22] Cook, R. D. and Li, B. (2002), "Dimension reduction for the conditional mean in regression," *The Annals of Statistics*, 30, 455 - 474.

[23] Cook, R.D., and Li, L. (2009). "Dimension reduction in regressions with exponential family predictors," *Journal of Computational and Graphical Statistics*, 18, 774-791.

[24] Cook, R.D. and Ni, L. (2005), "Sufficient dimesion reduction via inverse regression: A minimum discrepancy approach," *Journal of the American Statistical Association*, 100, 410-428.

[25] Cook, R.D., and Weisberg, S. (1991), "Discussion of *Sliced inverse regression for dimension reduction,*" *Journal of the American Statistical Association*, 86, 328-332.

[26] Cook, R. D. and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* Wiley, New York.

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

[28] Cook, R. D. and Yin, X. (2001), "Dimension-reduction and visualization in discriminant analysis (Invited with discussion)," *Australia & New Zealand Journal of Statistics*, 43, 147-200.

[29] Dai, B. (2013), *MVB: Mutivariate Bernoulli log-linear model,* R package version 1.1.

[30] Dai, B., Din, S. and Wahba, G. (2013), "Multivariate Bernoulli distribution," *Bernoulli*, 19(4), 1465-1483.

[31] Dmitrienko, A., Tamhane, A. C. and Bretz, F. (editors) (2010), "*Multiple Testing Problems in Pharmaceutical Statistics*," Chapman& Hall/CRC: Boca Raton, Florida.

[32] Forzani, L. (2007), "Sufficient dimension reduction based on normal and Wishart inverse models," Ph.D. thesis, School of Statistics, University of Minnesota.

[33] Fukumizu, K., Bach, F. R. and Gretton, A. (2007), "Statistical consistency of kernel canonical correlation analysis," *J. Mach. Learn. Res.*, 8, 361-383.

[34] Fukumizu, K., Bach, F. R. and Jordan, M. I. (2004), "Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces," *Journal of Machine Learning Research*, 5, 73-99.

[35] Fukumizu, K., Bach, F. R. and Jordan, M. I. (2009), "Kernel Dimension Reduction in Regression," *The Annals of Statistics*, 37, 1871-1905.

[36] Handcock, M. (2003), "Statistical models for social networks: inference and degeneracy," in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, National Academies Press, pp. 191-203.

[37] Hassner, M. and Sklansky, J. (1980), "The use of markov random fields as models of texture," *Computer Graphics Image Processing*, 12, 357-370.

[38] Hsing, T. and Ren, H. (2009), "An RKHS formulation of the inverse regression dimension- reduction problem," *The Annals of Statistics*, 37, 726-755.

[39] Ising, E. (1925), "Beitrag zur Theorie des Ferromagnetismus [Contribution to the theory of ferromagnetism]," *Zeitschrift für Physik*, 31, 253-258.

[40] Kim, M. and Pavlovic, V. (2013), "Central Subspace Dimensionality Reduction Using Covariance Operators," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4), 657-670.

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

[42] Li, K. C. (1992), "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma," *Journal of the American Statistical Association*, 87, 1025-1039.

[43] Li, B., Artemiou, A. and Li, L. (2011), "Principal support vector machines for linear and nonlinear sufficient dimension reduction," *The Annals of Statistics*, 39(6), 3182-3210.

[44] Li, B. and Wang, S. (2007), "On directional regression for dimension reduction," *Journal the American Statistical Association*, 102, 997-1008.

[45] Li, B., Zha, H. and Chiaromonte, F. (2005), "Contour Regression: A general approach to dimension reduction," *The Annals of Statistics*, 33(4), 1580-1616.

[46] Lindsey, J. K. (1997), *Applying Generalized Linear Models,* New York: Springer.

[47] van der Maaten, L., Postma, E. and van den Herik, H. (2009), "Dimensionality reduction: A comparative review," Technical Report TiCC-TR 2009-005, Tilburg University.

[48] Manning, C. and Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT press.

[49] McCulloch, P. and Nelder, J. A. (1989), *Generalized linear models. Statistics in the 21st Century,* CRC Press LLC and American Statistical Association, Boca Raton, Alexandria, 387-396.

[50] McCulloch, C. E., and Searle, S. R. (2001), *Generalized, Linear, and Mixed Models,* New York: Wiley.

[51] Nelder, J. A and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society. Series A (General)*, 135(3), 370-384.

[52] Noorbaloochi, S. and Nelson, D. (2008), "Conditionally specified models and dimension reduction in the exponential families," *Journal of Multivariate Analysis*, 99(8), 1574 - 1589.

[53] Parviainen, E. (2011), "Studies on dimension reduction and feature spaces," Ph.D. Thesis, Aalto University.

[54] Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009), "Partial correlation estimation by joint sparse regression model," *Journal of the American Statistics Association*, 104, 735-746.

[55] Pierce, D. A., Grupe, M. R., and Cleveland, W. P. (1984), "Seasonal Adjustment of Weekly Monetary Aggregates: A Model-Based Approach," *Journal of Business and Economics Statistics*, 2, 260-270.

[56] Ravikumar, P., Wainwright, M., and Lafferty, J. (2010), "High-dimensional Ising model selection using l_1-regularized logistic regression," *The Annals of Statistics*, 38, 1287-1319.

[57] Seber, G.A.F. (1977), *Linear Regression Analysis,* New York: Wiley.

[58] Varshney, K. R. and Willsky, A. S. (2010), “Classification using geometric level sets," *Journal of Machine Learning Research*, 11, 491–516.

[59] Wainwright, M. and Jordan, M. (2008), “Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, 1, 1-305.

[60] Wang, P., Chao, D., and Hsu, L. (2011), "Learning networks from high dimensional binary data: An application to genomic instability data," *Biometrics*, 67, 164-173.

[61] Wasserman, S. and Pattison, P. (1996), “Logit models and logistic regressions for social networks. I: An introduction to Markov graphs and $p^{\star}$, *Psychometrika*, 60, 401-425.

[62] Wen, X. M. and Cook, R. D. (2007), “Optimal sufficient dimension reduction for regressions with categorical predictors," *Journal of Statistical Planning and Inference*, 137, 1961-1978.

[63] Westfall, P.H. and Krishen, A. (2001), “Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures," *Journal of Statistical Planning and Inference*, 99 , 25-40.

[64] Whittaker, J. (1990), *Graphical Models in Applied Mathematical Multivariate Statistics*, New York: Wiley.

[65] Woods, J. (1978), “Markov image modeling," *IEEE Transactions on Automatic Control*, 23, 846-850.

[66] Wu, H.-M. (2008), "Kernel sliced inverse regression with applications to classification," *Journal of Computational and Graphical Statististics*, 17, 590-610.

[67] Wu, Q., Liang, F. and Mukherjee, S. (2008), "Regularized sliced inverse regression for kernel models," Technical report, Duke Univ., Durham, NC.

[68] 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.

[69] Yee, T.W. and Hastie, T.J. (2003), "Reduced-rank vector generalized linear models," *Statistical Modelling*, 3, 15-41.

[70] Yeh, Y. R., Huang, S. Y. and Lee, Y. Y. (2009), "Nonlinear dimension reduction with kernel sliced inverse regression," *IEEE Transactions on Knowledge and Data Engineering*, 21, 1590-1603.

[71] Zhu, H. and Li, L. (2011), "Biological pathway selection through nonlinear dimension reduction," *Biostatistics*, 12, 429-444.

[72] Zhu, Y. and Zeng, P. (2006), "Fourier methods for estimating the central subspace and the central mean subspace in regression," *Journal of the American Statistical Association*, 101, 1638-1651.