Computational Statistics and Data Analysis 111 (2017) 131–144

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Sufficient dimension reduction constrained through sub-populations

Elias Al-Najjar, Kofi P. Adragni *

Department of Mathematics & Statistics, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, United States

ARTICLE INFO

Article history:
Received 20 July 2014
Received in revised form 1 August 2016
Accepted 4 February 2017
Available online 14 February 2017

Keywords:
Dimension reduction
Principal fitted components
Partial reduction
Principal components

ABSTRACT

Most methodologies for sufficient dimension reduction (SDR) in regression are limited to continuous predictors, although many data sets do contain both continuous and categorical variables. Application of these methods to regressions that include qualitative predictors such as gender or species may be inappropriate. Regressions that include a set of qualitative predictors W in addition to a vector X of many-valued predictors and a response Y are considered. Using principal fitted components (PFC) models, a likelihood-based SDR method, a sufficient dimension reduction of X that is constrained through the sub-populations established by W is sought. An estimator of the sufficient reduction subspace is provided and its use is demonstrated through applications.

1. Introduction

We consider a regression problem of a univariate response Y on a p-vector X of continuous predictors. When p is large, it is always worthwhile to find a reduction $R(X)$ of dimension less than p that captures all regression information of Y on X. Replacing X by a lower dimensional function $R(X)$ is called dimension reduction. When $R(X)$ retains all of the relevant information about Y, it is referred to as a sufficient reduction. For our purposes, we can replace X with a sufficient reduction $R(X)$, improving the ability to visualize data and predict future observations, and mitigating dimensionality issues when carrying out further analysis. Cook (2007) formally defined a reduction $R : \mathbb{R}^p \to \mathbb{R}^d$, $d \leq p$, to be sufficient if it satisfies one of the following three conditions: (i) $Y|X \sim Y|R(X)$, (ii) $X|(Y, R(X)) \sim X|R(X)$, and (iii) $X \perp\!\!\!\perp Y|R(X)$. The symbol $\perp\!\!\!\perp$ stands for statistical independence, and $U \sim V$ stands for U and V having identical distribution. Condition (i) holds in a forward regression while condition (ii) holds in an inverse regression setup. Under a joint distribution of (Y, X) the three conditions are equivalent. Thus, we can use an inverse regression to obtain a sufficient reduction of X and use this reduction in lieu of X when modeling $Y|X$.

Although it is difficult to deal with dimension reduction in general, much progress has been made by restricting attention to linear subspaces of $\mathbb{R}^p$. If $R(X) = \eta^T X$ is a sufficient linear reduction, then so is $R(X) = (\eta A)^T X$ for any $d \times d$ full-rank matrix A. Consequently, only the subspace spanned by the columns of η can be identified, and this subspace is called a dimension-reduction subspace. If $\text{span}(\eta_1)$ and $\text{span}(\eta_2)$ are both dimension-reduction subspaces, then under mild conditions so is their intersection. Thus, the inferential target in sufficient dimension reduction is often taken to be the central subspace $\mathcal{S}_{Y|X}$, defined as the intersection of all dimension-reduction subspaces (Cook, 1998). A minimal sufficient linear reduction is then of the form $R(X) = \eta^T X$, where the columns of η form a basis for $\mathcal{S}_{Y|X}$.

* Corresponding author.
E-mail address: kofi@umbc.edu (K.P. Adragni).

http://dx.doi.org/10.1016/j.csda.2017.02.008
0167-9473/

Several sufficient dimension methods have been proposed in the literature since the seminal sliced inverse regression (SIR; Li, 1991). Most of these existing methods, including inverse regression estimation (Cook and Ni, 2005) and directional regression (Wang and Li, 2007) are distribution-free. A recent class of methodologies called principal fitted components (PFC) proposed by Cook (2007) and further elaborated by Cook and Forzani (2008) is likelihood-based.

Nearly all methodologies for sufficient dimension reduction are limited to continuous predictors, although many data sets do contain both continuous and categorical variables. Application of these methods to regressions that include categorical predictors may be inappropriate, because of the questionable relevance of linear combinations involving qualitative variables. Chiaromonte et al. (2002) described how the theory of dimension reduction can be extended to regression analyses involving both quantitative predictors X, and a categorical predictor, say W. They developed their methodology by extending the formulation of SIR using the concept of a partial dimension reduction subspace, defined as any subspace $\mathcal{S}$ that satisfies the conditional independence statement $Y \perp\!\!\!\perp X|(P_{\mathcal{S}}X, W)$. The methodology of Chiaromonte et al. (2002) assumed that the sub-populations are homogeneous in terms of their variances. Ni and Cook (2006) expanded the methodology of Chiaromonte et al. (2002) to heterogeneous sub-populations. Recently, Kim (2011) developed partial PFC models to reduce the dimension of one set of predictors, say X, while maintaining another set of predictors W. In her work, the set W was assumed random and continuous.

We herein assume that W is a variable which indicates relatively few categories. Our goal is to develop partial PFC, in order to obtain a reduction of X which is constrained through the sub-populations indexed by the values of W. Chiaromonte et al. (2002) and Li et al. (2003) have established the general framework with the extended moment-based method SIR. As a model-based method, PFC provides a wider range of possibilities than SIR. It has been shown that when the response Y is categorical, SIR and PFC estimate the same minimal sufficient reduction subspace. When Y is continuous, SIR discretizes the response through a slicing procedure and can leave intra slice information behind. On the other hand, the use of flexible basis functions in PFC can potentially help avoid such loss of information (Cook, 2007; Cook and Forzani, 2008).

Principal fitted component models are a likelihood-based approach to dimension reduction via inverse regression. Let X_y denote the p-dimensional random variable distributed as $X|(Y = y)$ and let $\bar{\mu} = \mathrm{E}(X)$, $\mu_y = \mathrm{E}(X_y)$. The models are based on the assumption that X_y has a multivariate normal distribution, and are therefore only appropriate for many-valued, quantitative, continuous or nearly-continuous predictors. It is assumed that $\mu_y - \bar{\mu}$ falls in a subspace $\mathcal{S}$ of dimension d in $\mathbb{R}^p$ as y varies in its sample space. Let $\Gamma \in \mathbb{R}^{p\times d}$ denote a basis of $\mathcal{S}$. We can then write $X_y \sim N(\bar{\mu} + \Gamma\nu_y, \Delta)$ where $\nu_y = \Gamma^T(\mu_y - \bar{\mu})$ is a function of y. The conditional variance Δ is assumed to be independent of Y. Once the response values are observed, the unknown function ν_y can be modeled as $\nu_y = \beta(\mathrm{f}_y - \mathrm{E}[\mathrm{f}_Y])$, where β is an unknown, unconstrained parameter and f_y is a flexible basis function. The subsequent model, written as

$$X_y = \mu + \Gamma\beta\mathrm{f}_y + \Delta^{1/2}\varepsilon, \tag{1}$$

with $\mu = \bar{\mu} + \Gamma\beta\mathrm{E}(\mathrm{f}_Y)$, where the error term $\varepsilon \sim N(0, \mathrm{I})$, is called a PFC model. Several basis functions have been suggested, including the polynomial, piecewise constant and piecewise polynomial, among others (Cook, 2007; Cook and Forzani, 2008; Adragni and Cook, 2009). A number of variance structures have been proposed to model the conditional dependence among the predictors. These structures include the isotropic ($\Delta = \delta^2\mathrm{I}$), the anisotropic [$\Delta = \mathrm{diag}(\delta_1^2, \ldots, \delta_p^2)$], and the structured, which can accommodate conditional dependency among groups of predictors. Under PFC model with conditional variance Δ, a minimal sufficient reduction is $R(X) = \Gamma^T\Delta^{-1}X$ and the central subspace is obtained as $\mathcal{S}_{Y|X} = \Delta^{-1}\mathcal{S}_\Gamma$, where $\mathcal{S}_\Gamma$ is the subspace spanned by the columns of Γ.

In the following, we seek the reduction of the predictor vector $X \in \mathbb{R}^p$, while W is an additional predictor that is not to be included. We assume that W represents one or more categorical variables that identify $w = 1, \ldots, c$ sub-populations. In Section 2, we provide the models and describe the methodology. Section 3 gives the maximum likelihood estimation of the parameters involved in the models. Section 4 shows how to predict future observations by inverting the inverse regression mean function $\mathrm{E}(X|Y)$ to obtain $\mathrm{E}(Y|X)$. We show two applications in Section 5, then provide some discussions.

2. Class-based principal fitted components

For a given level w of the categorical predictor W, we assume that as y varies in its sample space, the curve $\mu_{yw} - \bar{\mu}_w = \mathrm{E}(X|Y = y, W = w) - \mathrm{E}(X|W = w)$ falls in a subspace $\mathcal{S}_{\Gamma_w}$ of $\mathbb{R}^p$. We first assume that $\mathrm{Var}(X|Y = y, W = w) = \Delta_w$ is constant for any y within each sub-population indexed by $w = 1, \ldots, c$. The term Γ_w is a $p\times d_w$ semi-orthogonal matrix, that is $\Gamma_w^T\Gamma_w = I_{d_w}$ so that its columns span $\mathcal{S}_{\Gamma_w}$. We thus have within each class, $X_{yw} \sim N(\bar{\mu}_w + \Gamma_w\nu_{yw}, \Delta_w)$, $w = 1, \ldots, C$, where ν_{yw} is an unknown function of y for the particular class w. This model is essentially a principal component model as developed by Cook (2007), specifically for class w. The model assumes different means μ_w, different reduction kernel matrices Γ_w, and different covariance Δ_w for the c sub-populations. Given the response y, we model the unknown function ν_{yw} by $\nu_{yw} = \beta_w(\mathrm{f}_{yw} - \mathrm{E}[\mathrm{f}_{Yw}])$ where f_{yw} is a set of flexible known basis functions and $\beta_w \in \mathbb{R}^{d_w\times r}$ is an unrestricted rank d_w matrix. For simplicity we assume that the same basis functions are used for each class, but this assumption is not required; from this point on we will write f_y instead of f_{yw}. With $\mu_w = \bar{\mu}_w + \Gamma_w\beta_w\mathrm{E}(\mathrm{f}_Y|W = w)$, the subsequent model for each sub-population is a principal fitted component model (Cook, 2007; Cook and Forzani, 2008).

$$X_{yw} = \mu_w + \Gamma_w\beta_w\mathrm{f}_y + \Delta_w^{1/2}\varepsilon, \quad w = 1, \ldots, c. \tag{2}$$