ARTICLE IN PRESS

Journal of Statistical Planning and Inference xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Quantile-slicing estimation for dimension reduction in regression

Hyungwoo Kim [a], Yichao Wu [b], Seung Jun Shin [a,*]

[a] Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 02841, South Korea
[b] Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, 851 S. Morgan Street Chicago, IL 60607, USA

ARTICLE INFO

Article history:
Received 22 March 2017
Received in revised form 2 February 2018
Accepted 5 March 2018
Available online xxxx

MSC:
00-01
99-00

Keywords:
Heteroscedasticity
Kernel quantile regression
Quantile-slicing estimation
Sufficient dimension reduction

ABSTRACT

Sufficient dimension reduction (SDR) has recently received much attention due to its promising performance under less stringent model assumptions. We propose a new class of SDR approaches based on slicing conditional quantiles: quantile-slicing mean estimation (QUME) and quantile-slicing variance estimation (QUVE). Quantile-slicing is particularly useful when the quantile function is more efficient to capture underlying model structure than the response itself, for example, when heteroscedasticity exists in a regression context. Both simulated and real data analysis results demonstrate promising performance of the proposed quantile-slicing SDR estimation methods.

1. Introduction

In high-dimensional data analysis, it is often a primary goal to reduce the dimensionality of data without losing much information of interest. The well-known principal component analysis (PCA) is a canonical example. In the regression context, PCA fails to exploit information about association between the response and predictors. Penalization-based variable selection methods such as LASSO (Tibshirani, 1996) or SCAD (Fan and Li, 2001) can be regarded as another type of dimension reduction. However, many variable selection methods rely on stringent parametric assumptions which may often be unrealistic in practice.

Sufficient dimension reduction (SDR) has received much attention in statistical community. In a regression framework, SDR reduces the predictor dimension by seeking a matrix $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_d) \in \mathbb{R}^{p\times d}$ that satisfies

$$Y \perp \mathbf{X} | \mathbf{B}^T\mathbf{X}, \tag{1}$$

where Y and $\mathbf{X} = (X_1, \ldots, X_p)^T$ are the univariate response and p-dimensional predictor, respectively. The SDR model (1) is quite flexible since it does not impose any type of link function on the relationship between Y and $\mathbf{X}$. Yet SDR preserves information about association between Y and $\mathbf{X}$, which differs from PCA. The space spanned by the columns of $\mathbf{B}$, denoted as span($\mathbf{B}$), is called dimension reduction subspace (DRS). DRS is not unique and thus not identifiable. So is $\mathbf{B}$. To impose identifiability, we define the central subspace, denoted by $\mathcal{S}_{Y|\mathbf{X}}$, as the intersection of all DRSs that satisfy (1). It is shown that $\mathcal{S}_{Y|\mathbf{X}}$ exists uniquely under mild conditions (Cook, 1996). We finally assume that span($\mathbf{B}$) $= \mathcal{S}_{Y|\mathbf{X}}$ to make $\mathbf{B}$ (or more

* Corresponding author.
E-mail address: sjshin@korea.ac.kr (S.J. Shin).

https://doi.org/10.1016/j.jspi.2018.03.001
0378-3758/

precisely span(**B**)) an identifiable target in SDR. The dimension of $\mathcal{S}_{Y|\mathbf{X}}$, d is called the structural dimension and is another important quantity of interest to be estimated from the data.

Since the seminal work of sliced inverse regression (SIR, Li, 1991) and sliced average variance estimation (SAVE, Cook and Weisberg, 1991) both of which are based on inverse moments, a variety of SDR methods have been developed. Li and Wang (2007) proposed the directional regression based on empirical directions of Y that generalizes the idea of inverse-moment. The inverse-moment-based methods often require to slice the support of Y, and the selection of slices may affect the finite sample performance. To tackle this issue, Cook and Zhang (2014) proposed a fusing method and Zhu et al. (2010) develop a cumulative slicing estimation. Li et al. (2005) proposed an alternative method for SDR called contour regression, and this motivates the principal support vector machine (Li et al., 2011), a unified framework to handle both linear and nonlinear SDR.

The SDR model (1) can be viewed as a semi-parametric model for the conditional distribution function of Y given $\mathbf{X}$, denoted by $F_{y|\mathbf{x}}$. Xia et al. (2002) proposed the minimum average variance estimation (MAVE) which recovers $\mathcal{S}_{Y|\mathbf{X}}$ by estimating the conditional expectation, $E(Y|\mathbf{X}) = \int y dF_{y|\mathbf{x}}$ which is the same as $\mathrm{E}(Y|\mathbf{B}^T\mathbf{X})$ under (1). Motivated by MAVE, related variants have been developed. See for example, Xia (2007), Wang and Xia (2008), and Yin et al. (2011). Zhu and Zeng (2006) proposed an estimator of $\mathcal{S}_{Y|\mathbf{X}}$ by estimating the gradient of $F_{y|\mathbf{x}}$ using the Fourier transformation. Kong and Xia (2014) exploited the gradient of quantile function, instead of $F_{y|\mathbf{x}}$, and proposed the adaptive composite quantile outer product of gradients method. Ma and Zhu (2012) derived the space of influence functions of the estimator of $\mathcal{S}_{Y|\mathbf{X}}$, and Ma and Zhu (2013) further proposed an efficient estimator of $\mathcal{S}_{Y|\mathbf{X}}$ and established its asymptotic properties. Recently, Huang and Chiang (2017) proposed an alternative semi-parametric estimator for SDR that can estimate **B** and d simultaneously.

In practice, data often display heteroscedastic variance which can be of scientific importance. Note that the SDR model (1) requires the conditional independence only. In principal it has no difficulty to encompass underlying heteroscedasticity in the data. However, most of SDR methods are designed to focus primarily on conditional mean relation and can be inefficient to identify such heteroscedasticity, as illustrated by a toy example coming up next. See Kong and Xia (2014) and Wang et al. (2018) for difficulties in SDR with heteroscedasticity.

The quantile regression is a popular alternative to the conventional mean regression when homoscedastic error assumption is violated. The quantile regression seeks the τth conditional quantile of $Y|\mathbf{X}$ denoted by $f_\tau(\mathbf{X})$ that satisfies

$$P(Y \le f_\tau(\mathbf{X})|\mathbf{X}) = \tau$$

for a given quantile level $\tau \in (0, 1)$. Notice that the conditional distribution of $Y|\mathbf{X}$ possesses all the information about Y. By stacking together all the conditional quantile functions of $Y|\mathbf{X}$ at different quantile levels, we define $Q_{\mathbf{X}} \equiv Q_{\mathbf{X}}(\tau) = f_\tau(\mathbf{X})$ as a function of the quantile level τ. The stacked conditional quantile function $Q_{\mathbf{X}}$ contains complete information about the conditional distribution of $Y|\mathbf{X}$.

It can be shown that the stacked conditional quantile function $Q_{\mathbf{X}}$ contains same amount of information on $\mathcal{S}_{Y|\mathbf{X}}$ as Y does. Namely, $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{S}_{Q_{\mathbf{X}}|\mathbf{X}}$ where $\mathcal{S}_{Q_{\mathbf{X}}|\mathbf{X}}$ denotes the central subspace for the "regression" of $Q_{\mathbf{X}}$ on $\mathbf{X}$ and is defined accordingly. This motivates us to develop a new SDR approach that slices conditional quantiles of $Y|\mathbf{X}$ instead of the response Y itself to estimate $\mathcal{S}_{Y|\mathbf{X}}$. Finally two versions of estimators based on quantile-slicing are developed: QUantile-slicing Mean Estimation (QUME) and QUantile-slicing Variance Estimation (QUVE).

In practice, $Q_{\mathbf{X}}$ is an unknown quantity and should be inferred from the data. Toward this, we propose to use the kernel quantile regression (KQR, Takeuchi et al., 2006; Li et al., 2007). The KQR is a flexible nonparametric method showing promising performance in high-dimensional data due to the use of the kernel trick (Zhang, 2002). The KQR solution as a function of τ is piecewise linear in $\tau \in (0, 1)$ (Takeuchi et al., 2009). This enables us to estimate the stacked quantile function $Q_{\mathbf{X}}$ completely from a finite sample.

As a simple illustration, we consider a toy example of simple regression with heteroscedastic error: $Y = X_1 + \exp(X_2)\epsilon$, where $X_1, X_2, X_3 \overset{i.i.d.}{\sim} N(0, 1)$ and $\epsilon \sim N(0, 0.2^2)$ are independent of each other. Notice that $\mathbf{B} = (\mathbf{e}_1, \mathbf{e}_2)$ where $\mathbf{e}_1 = (1, 0, 0)^T$ and $\mathbf{e}_2 = (0, 1, 0)^T$ and hence X_1 and X_2 are the two sufficient predictors, X_1 for mean and X_2 for variance of the response. We apply SIR and QUME to a random sample of size 100 generated from this simple model. Fig. 1 depicts the estimated $\mathcal{S}_{Y|\mathbf{X}}$ by SIR (red plane with a finer mesh) and QUME (back plane with a coarser mesh) on the three dimensional predictor space. The table at the bottom-right corner reports the distance between true and estimated $\mathcal{S}_{Y|\mathbf{X}}$ in terms of the criterion defined in (10). One can see that QUME outperforms SIR for identifying $\mathcal{S}_{Y|\mathbf{X}}$. In particular, SIR shows insufficient accuracy to identify the direction associated with X_2 that controls variance of the response.

The rest of article is organized as follows. In Section 2, we first show the equivalence between $\mathcal{S}_{Y|\mathbf{X}}$ and $\mathcal{S}_{Q_{\mathbf{X}}|\mathbf{X}}$, and then propose two versions of quantile-slicing scheme, QUME and QUVE. In Section 3, the finite-sample implementation of QUME and QUVE via solution paths of the KQR is described in details. Additional issues including estimation of the structural dimension and tuning parameter selection in KQR are discussed in Section 4. Illustration to both simulated and real data are presented in Sections 5 and 6, respectively. Concluding remarks are in Section 7. All the proofs are relegated to the Appendix.

2. Quantile-slicing estimation

Slicing scheme has been regarded as a standard approach in SDR. We propose a new SDR method based on slicing conditional quantiles instead of the observed response. We first establish that $Q_{\mathbf{X}}$ contains the same amount of information as Y for $\mathcal{S}_{Y|\mathbf{X}}$, and then propose two versions of the quantile-slicing schemes: QUME and QUVE.

Please cite this article in press as: Kim H., et al., Quantile-slicing estimation for dimension reduction in regression. J. Statist. Plann. Inference (2018), https://doi.org/10.1016/j.jspi.2018.03.001.