New Estimation and Inference Procedures for a Single-Index Conditional Distribution Model

We consider the conditional distribution $F_Y(y|x)$ of a real-valued response Y given continuous covariates $X = x$, where $X = (X_1, \cdots, X_d)^T$ and $x = (x_1, \cdots, x_d)^T$. In regression analysis, a wide cross-section of research interests is pursued in the study of the conditional mean $E[Y|x]$. A more complete methodology and theoretical framework related to fully nonparametric and semiparametric distribution models still remains and a further investigation is necessary. As one can see, with a large number of covariatess, a fully nonparametric distribution usually suffers from the curse of dimensionality (Bellman (1961)). Although parametric models have played prominent roles in applications, they are frequently detected to be inadequate in many studies. Consequently, a more flexible semiparametric model becomes a great interest to characterize the dependence of Y on X and avoids the impact of misspecification of parametric models and the difficulty in the estimation of nonparametric distributions.

One of the most popular extension of parametric models is the single-

index (SI) conditional distribution model:

$$F_Y(y|x) = G(y, x_{\theta_0}), \tag{1}$$

where $G(\cdot,\cdot)$ is an unknown bivariate function, $x_\theta = x_1 + (x_2, \cdots, x_d)^T\theta$, and θ_0 is a vector of true index coefficients. The most significant covariate is assumed, without loss of generality, to be X_1 and the setting of its coefficient is mainly to deal with the problem of identifiability. When the conditional mean exists, it can be easily obtained from the above model that $E[Y|x] = m(x_{\theta_0})$ with $m(\cdot)$ being some unspecified function. Based on the conditional mean model, Powell, Stock, and Stoker (1989) utilized the estimation of the density-weighted average derivative to estimate θ_0. Although the estimator was shown to be $\sqrt{n}$-consistent, asymptotically normal, and computationally simple, the numerical instability is usually seen as a consequence of high-dimensional kernel smoothing. To overcome such a weakness with practice, Ichimura (1993) developed a semiparametric least squares approach and derived its asymptotic properties. Meanwhile, Härdle, Hall, and Ichimura (1993) recommended a cross-validation criterion to simultaneously estimate bandwidths and index coefficients. Under the validity of model (1) with a continuous response, Delecroix, Härdle, and Hristache (2003) introduced the pseudo likelihood (PL) estimation for θ_0. Without moment and continuous conditions on Y, Hall and Yao (2005) suggested an estimation criterion on the basis of the average squared difference between the empirical estimator and the model-based estimator of the joint probability of $(Y, X^T)^T$. As one can see, the good performance of their estimation procedure is con-

nected to an appropriate number of spheres and the corresponding radii used in the integral approximation. Currently, there is still no standard rule to determine the values of these two quantities. Furthermore, the established algorithm is often computationally slow and intensive, especially in high-dimensional covariate spaces. Confronted with these problems, we proposed a new type of estimation criterion, which is simple and easily implemented, for θ_0. The basic rationale behind this approach is to define the response process $N(y) = I(Y \leq y)$ and to directly use the difference between $N(y)$ and its conditional mean $G(y, x_{\theta_0})$ over the support of Y. Under some suitable conditions, the asymptotic distribution of the PLISE is derived to be multivariate normal. To make inferences related to θ_0, the frequency distributions of its bootstrap analogues are used to estimate the asymptotic variance of the PLISE because a sandwich-type estimator tends to provide a very poor approximation. With the proposed residual process, the method of Xia (2009) is extended to establish a test rule to check the adequacy of model (1). There are two features of the PLISE: Firstly, our estimation approach can be applied to different types of response variable and outperforms the existing ones; secondly, the foregoing inferences can be easily adopted and generalized to the considered problems in this article.

When the true underlying model has a sparse representation, identifying significant covariates becomes an important issue to enhance the accuracy in prediction. The traditional best-subset selection algorithms are usually computationally infeasible in the presence of a potentially high-dimensional covariate space. The ridge regression estimation is another variance-stabilizing technique, which shrinks the least square estimator toward zero but not iden-

tifies significant covariates cleverly. To simultaneously select significant variables and to estimate the parameters in regression models, Tibshirani (1996) introduced a least absolute shrinkage and selection operator (Lasso). Since Lasso variable selection might be inconsistent, Fan and Li (2001) and Zou (2006) proposed a smoothly clipped absolute deviation (SCAD) penalty and an adaptive Lasso instead. In their model specifications, the adaptive Lasso further avoids the problem of nonconcavity in the SCAD penalty although both of the procedures enjoy the oracle properties. By extending the adaptive Lasso in generalized linear models to our framework, we propose the penalized pseudo least integrated squares estimator (PPLISE) and derive the corresponding oracle properties. Moreover, in a small sample size scenario, a multi-stage adaptive Lasso estimation procedure is proposed to improve the possible selection inconsistency and predictive inaccuracy in the PPLISE.

For each fixed (y, x_θ), the approach of Hall, Wolff, and Yao (1999) can be applied for the estimation of $G(y, x_\theta)$. Let $K(u)$ denote a kernel density, h be a positive-valued bandwidth, $N_l(y, X_{i\theta}) = \sum_{j\neq i} N_j^l(y) K_h(X_{j\theta} - X_{i\theta})/(n-1)$, $i = 1, \cdots, n$, $l = 0, 1$, and $K_h(u) = h^{-1}K(u/h)$. The Nadaraya-Watson estimator for $G(y, X_{i\theta})$ is given by $\widehat{G}(y, X_{i\theta}) = N_1(y, X_{i\theta})/N_0(y, X_{i\theta})$. By using the response process $N(y)$ and a consistent estimator of $G(y, x_\theta)$, the pseudo least integrated squares estimator (PLISE) $\widehat{\theta}$ is proposed to be a minimizer of the pseudo sum of integrated squares (PSIS):

$$\mathrm{SS}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{Y}} e_i^2(y;\theta) dW_{ni}(y), \tag{2}$$

where $\mathcal{Y}$ is the support of Y or the interval of interest, $e_i(y;\theta) = N_i(y) -$

$\widehat{G}(y, X_{i\theta})$, and $W_{ni}(y)$ is a non-negative weight function. In practical implementation, $\widehat{G}(y, x_\theta)$ is set to be zero if the denominator $N_0(y, x_\theta)$ is zero. Although a local linear estimator of $G(y, x_\theta)$ can be used in the PSIS, it does not share the properties of a cumulative distribution function and might cause some complications in the above estimation procedure.

It follows from a direct algebraic calculation that

$$E[(N(y) - G(y, X_\theta))^2] = E[(N(y) - F_Y(y|X))^2] + E[(F_Y(y|X) - G(y, X_\theta))^2]. \quad (3)$$

Since the first term at the right-hand side of (3) does not depend on θ, both minimizers of $E[\int_{\mathcal{Y}}(N(y) - G(y, X_\theta))^2 dW(y)]$ and $E[\int_{\mathcal{Y}}(F_Y(y|X) - G(y, X_\theta))^2 dW(y)]$ can be shown to be θ_0 under the validity of model (1), where $W(y)$ is a convergent function of $W_n(y)$. Thus, minimizing $\mathrm{SS}(\theta)$ is on average approximated by minimizing $E[\int_{\mathcal{Y}}(F_Y(y|X) - G(y, X_\theta))^2 dW(y)]$ with respect to θ. In our theoretical development and numerical implementation, the quartic kernel $K(u) = (15/16)(1 - u^2)^2 I(|u| \leq 1)$ is specified. The advantage of such a density function is that $\widehat{\theta}$ can achieve the $\sqrt{n}$-consistency. As a spacial case, the uniform distribution or the empirical distribution of Y can be specified for $W_{ni}(y)$'s in (2). In the case where $G(y, x_\theta)$ is known, the optimal weight for $w_{ni}(y) = dW_{ni}(y)/dy$ is proportional to $\{G(y, X_{i\theta})(1 - G(y, X_{i\theta}))\}^{-1}$, the reciprocal of the conditional variance of $N_i(y)$, at each fixed y. We further replace $G(y, x_\theta)$ by a consistent estimator $\widehat{G}(y, x_{\widehat{\theta}})$ and iteratively update the weight estimation. The resulting estimator coincides with the maximizer of the following log-pseudo likelihood

function for a random sample $\{N_i(y) : 1 \leq i \leq n\}$:

$$\mathrm{l}_p(\theta) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Y}} (N_i(y)\ln(\widehat{G}(y, X_{i\theta})) + (1 - N_i(y))\ln(1 - \widehat{G}(y, X_{i\theta}))dy. \quad (4)$$

Let $y_{(1)} < \cdots < y_{(m)}$ denote the distinct order statistics of $\{Y_1, \cdots, Y_n\}$ and $W_{(j)} = \int_{y_{(j)}}^{y_{(j+1)}} dW_{ni}(y)$. Since $N_i(y)$'s are zero-one processes and $\widehat{G}(y, X_{i\theta})$'s are step functions with jumps occurring at $\{y_{(1)}, \cdots, y_{(m)}\}$, the PSIS in (2) has a computationally more attractive alternative as follows:

$$\mathrm{SS}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m-1} e_i^2(y_{(j)}; \theta)W_{(j)}. \quad (5)$$

In contrast, the estimation procedure of Hall and Yao (2005) is often computationally intensive. When the response variable Y is discrete and has a finite support, the above estimation criterion can also be applied. As for the binary response with values in $\{0, 1\}$, the PSIS will automatically reduce to the sum of squares in Ichimura (1993). In kernel estimation, a criterion for bandwidth selection is provided via generalizing the most commonly used "leave one subject out" cross-validation procedure of Rice and Silverman (1991). The optimal bandwidth h_{cv} is naturally defined to be the unique minimizer of

$$\mathrm{CV}_1(h) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j}^{m-1} e_i^2(y_{(j)}; \widehat{\theta}_i)W_{(j)} \quad (6)$$

with $\widehat{\theta}_i = \arg\min\{(n-1)^{-1}\sum_{l\neq i}\sum_{j=1}^{m-1} e_l^2(y_{(j)}; \theta)W_{(j)}\}$. Another criterion developed by Härdle, Hall, and Ichimura (1993) is further adopted and ex-

tended to our framework. The estimators of h and θ_0 can be simultaneously obtained via minimizing $\mathrm{CV}_2(h,\theta) = \mathrm{SS}(\theta)$

The steps and contributions of this research are stated as follows:

- We will present an appealing estimation procedure for index coefficients and show that it outperforms the existing ones.

- Compared with the PMLE, an important advantage of the PLISE is that it only requires a lower-order kernel in a one-dimensional bandwidth space.

- The modified cross-validation scores and residual process are provided for bandwidth selection and model checking.

- We employ random weighted bootstrap analogues of the asymptotic variance of the PLISE.

- The L^1-penalty with random weights is further adopted into the PLISE criterion to improve estimation and variable selection simultaneously in sparse high-dimensional models. Under the partial orthogonality condition of Huang, Ma, and Zhang (2008), our PPLISE still enjoys the oracle property when the number of covariates increases exponentially with the sample size.

- In some applications, the predictive abilities of covariates might depend on the values of a response variable. It is more realistic to consider the following varying-index model:

$$F_Y(y|x) = G(y, x_{\theta_0(y)}), \tag{7}$$

where $\theta_0(y)$ is a vector of index coefficient functions of y. This modelling approach is especially useful to handle an ordinal response variable and for quantile forecasting. The PSIS in (5) and the PPSIS can be modified as

$$\mathrm{SS}(\theta(y)) = \frac{1}{n}\sum_{i=1}^{n}(N_i(y) - \widehat{G}(y, X_{i\theta(y)}))^2 \tag{8}$$

and

$$\mathrm{PSS}(\theta(y)) = SS(\theta(y)) + \lambda_y \sum_{k=2}^{d} \frac{|\theta_k(y)|}{|\widehat{\theta}_k(y)|}. \tag{9}$$

- In survival analysis, the response measurement represents the time to a particular event. It is worthy to note that the considered model includes more acceptable proportional hazards and accelerated failure time models. A major challenge in dealing with this issue is that the failure times of some individuals might not be available due to censoring. Our results should be valuable in the development of related inferences.

- The simulation experiments are conducted and the proposed approaches are applied to two empirical examples.