Computational Statistics and Data Analysis 56 (2012) 1837–1853

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

# Estimation for the single-index models with random effects

Zhen Pang [a,*], Liugen Xue [b]

[a] *Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, 637371, Singapore*
[b] *College of Applied Sciences, Beijing University of Technology, Beijing, 100124, China*

ARTICLE INFO

*Article history:*
Received 1 July 2011
Received in revised form 10 October 2011
Accepted 5 November 2011
Available online 15 November 2011

*Keywords:*
Single-index models
Mixed-effects models
Local linear smoother
Pooled estimator
Variance components

ABSTRACT

In this paper, we generalize the single-index models to the scenarios with random effects. The introduction of the random effects raises interesting inferential challenges. Instead of treating the variance matrix as the tuning parameters in the nonparametric model of Gu and Ma (2005), we propose root-$n$ consistent estimators for the variance components. Furthermore, the single-index part in our model avoids the curse of dimensionality and makes our model simpler. The variance components also cannot be treated as nuisance parameters and are canceled in the estimation procedure like Wang et al. (2010). A new set of estimating equations modified for the boundary effects is proposed to estimate the index coefficients. The link function is estimated by using the local linear smoother. Asymptotic normality is established for the proposed estimators. Also, the estimator of the link function achieves optimal convergence rate. These results facilitate the construction of confidence regions and hypothesis testing for the parameters of interest. Simulations show that our methods work well for high-dimensional $p$.



## 1. Introduction

Mixed effects models are widely used for the analysis of correlated data such as longitudinal data and repeated measures. The linear and nonlinear mixed effects models can be found in, for example, Harville (1977), Lindstrom and Bates (1990) and Ke and Wang (2001). For cross-sectional data, Gu and Ma (2005) proposed the nonparametric mixed effects model

$$Y = \eta(X) + Z^T b + \varepsilon,$$

where the regression function $\eta(x)$ is assumed to be a smooth function on a generic domain $\mathcal{X}$. The model terms $\eta(x)$ was estimated using the penalized (unweighted) least squares method and the variance matrix, together with the tuning parameter $\lambda$ in spline smoothing, are treated as tuning parameters that are not estimated. Furthermore, when the dimension of $X$ is high, the "curse of dimensionality" will occur. To improve on these two aspects, we consider the single-index model with random effects

$$Y_{ij} = g(X_{ij}^T \beta_0) + Z_{ij}^T b_i + \varepsilon_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, m, \tag{1}$$

where $\beta_0$ is a $p \times 1$ index coefficients vector associated with the covariates $X_{ij}$, the $b_i$ are independent $q \times 1$ vectors of random effects with mean 0 and covariance matrix $D$, $g(\cdot)$ is an unknown link function, $\varepsilon_{ij}$'s are independent mean zero random variables with variance $\sigma_\varepsilon^2 > 0$. Here $D$ is a positive definite matrix depending on a parameter vector $\phi$; $X_{ij}$ and $Y_{ij}$ are the observable random variables, and $Z_{ij}$ is a fixed matrix. We assume $b_i$ and $\varepsilon_{ij}$ to be mutually independent. For the sake of identifiability, it is often assumed that $\|\beta_0\| = 1$ and the first nonzero component of $\beta_0$ is positive, where $\|\cdot\|$ denotes

* Corresponding author.
*E-mail address:* zpang@ntu.edu.sg (Z. Pang).

0167-9473/$ – see front matter 
doi:10.1016/j.csda.2011.11.007

the Euclidean metric. We address the general problem of estimating the parameter $\beta_0$ and the function $g(\cdot)$ as well as the variance components in $D$ and $\sigma_\varepsilon^2$ simultaneously.

The single-index model is an important tool in multivariate nonparametric regression, which can avoid the so-called "curse of dimensionality" by searching for a univariate index of the multivariate covariate $X$ to capture important features of high-dimensional data. The single-index models have been applied in a variety of fields, such as discrete choice analysis in econometrics and dose-response models in biometrics (Härdle et al., 1993). In cross-sectional data, many authors have studied the statistical inference problems for single-index models, and reported many results, for example, Li (1991), Ichimura (1993), Xia and Li (1999), Naik and Tsai (2000), Hristache et al. (2001), Xia et al. (2002), Xia (2006) and Xue and Zhu (2006). These reported methods have been proven to be useful and effective for independent data. On the other hand, to our knowledge, a method to treat data with random effects, which are commonly seen in econometrics and biometrics, is lacking in literature. In this paper, such models will be developed and reported.

There is already extensive literature on the generalized linear, nonparametric and semiparametric methods for the mixed effects models, see, for example, Zeger and Diggle (1994), Ruckstuhl et al. (2000), Ke and Wang (2001), Wu and Zhang (2002), Hall and Maiti (2006), Jiang (2007) and Field et al. (2008), among others. However, literature on the applications of single-index models with random effects is limited. Honorá and Kyriazidou (2000) and Carro (2007) proposed some estimating methods for dynamic panel data discrete choice models. Bai et al. (2009) studied the single-index models for longitudinal data, where they proposed a procedure to estimate the single-index component and the link function based on the combination of the penalized splines and quadratic inference functions. Liang and Zeger (1986) proposed an extension of the generalized linear models and introduced the generalized estimating equations (GEEs) that gave consistent estimates of the index coefficients and their variances under mild assumptions on the time dependence. The GEEs were derived without specifying the joint distribution of a subject's observations yet they reduced to the score equations for multivariate Gaussian outcomes.

In this paper, we apply the idea of GEEs to the single-index models with random effects. To estimate the index coefficients $\beta_0$, we propose a new set of estimating equations which take the boundary effects and the constraint $\|\beta_0\| = 1$ into account. The estimators based on these estimating equations outperform previous ones, as summarized below. First, our estimation procedure does not specify a form for both the distribution of the random effects and the joint distribution of the repeated measurements. Second, we introduce estimating equations that give a root-$n$ consistent estimate of $\beta_0$ under weak assumptions on the joint distribution. In particular, the estimator of $\beta_0$ has smaller asymptotic variance when compared to the least squares method proposed by Härdle et al. (1993), see, for example Wang et al. (2010) and Cui et al. (2011). Third, we construct the root-$n$ consistent estimates of the variance components in $D$ and $\sigma_\varepsilon^2$. It allows us to deal with further statistical inferences such as the construction of confidence regions and hypothesis testing for $\beta_0$. Lastly, we also obtain the asymptotic normality and the uniform convergence rate of the estimator of $g(\cdot)$. Our algorithm is numerically fast and stable.

The rest of the paper is organized as follows. In Section 2, we elaborate on the methodology. Section 3 presents the asymptotic properties for all proposed estimators. Section 4 reports the results of simulation studies and one real example. The proofs of the main theorems are relegated to the Appendix.

## 2. Estimation method

### 2.1. Estimations of the parametric and nonparametric components

In this section, we first focus on the estimation of the index coefficients $\beta_0$. Let $Y_i = (Y_{i1}, \ldots, Y_{im})^T, X_i = (X_{i1}, \ldots, X_{im})^T, Z_i = (Z_{i1}, \ldots, Z_{im})^T, \varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})^T$ and $G(X_i\beta_0) = (g(X_{i1}^T\beta_0), \ldots, g(X_{im}^T\beta_0))^T$. Model (1) can be represented as

$$Y_i = G(X_i\beta_0) + Z_ib_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

The model implies that the $Y_i$ are independent with $E(Y_i|X_i) = G(X_i\beta_0)$ and $\mathrm{cov}(Y_i|X_i) = Z_iDZ_i^T + \sigma_\varepsilon^2 I_m$, where $I_m$ is the $m \times m$ identity matrix. A naive idea to estimate $\beta_0$ is to minimize

$$R_n(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n}\{Y_i - G(X_i\beta)\}^T W(X_i\beta)V_i^{-1}\{Y_i - G(X_i\beta)\} \tag{2}$$

for $\beta$ with $\|\beta\| = 1$, where $V_i = Z_iDZ_i^T + \sigma_\varepsilon^2 I_m$ is the variance–covariace matrix, $W(X_i\beta) = \mathrm{diag}\{w(X_{i1}\beta), \ldots, w(X_{im}\beta)\}$ and $w(\cdot)$ is a bounded weight function with a bounded support $\mathcal{U}_w$, which is introduced to control the boundary effects in the estimations of $g(\cdot)$ and $g'(\cdot)$, For simplicity and convenience, we assume that $w(\cdot)$ is the indicator function on $\mathcal{U}_w$.

However, in (2), the link function $g$, variance components $\phi$ in $D$ and $\sigma_\varepsilon^2$ are unknown and we cannot get $\beta$ directly. We will show in the next section that the variance components $\phi$ in $D$ and $\sigma_\varepsilon^2$ can be estimated at root-$n$ rates and hence can effectively be treated as known for the purpose of developing and analyzing estimators of $\beta$ and $g(\cdot)$. We then propose an alternating procedure to estimate $\beta$ and $g(\cdot)$ iteratively.

If $g$ is known, (2) is a restricted least squares problem because there is the constraint $\|\beta_0\| = 1$ and the function $g(X_{ij}^T\beta)$ does not have derivative at point $\beta_0$. Wang et al. (2010) used the delete-one-component to transfer the restricted least squares to the unrestricted least squares in the Euclidean space $R^{p-1}$. We also use this method, but note that they deal