Three Essays on Application of Semiparametric Regression: Partially Linear Mixed Effects Model and Index Model

Dissertation

zur Erlangung des wirtschaftswissenschaftlichen Doktorgrades der Wirtschaftswissenschaftlichen Fakultät der Universität Göttingen

vorgelegt von

OHINATA Ren

aus Kawasaki, Japan

Göttingen, 2012

Erstgutachter: Prof. Dr. Stefan Sperlich

Zweitgutachter: Prof. Stephan Klasen, Ph.D.

Drittgutachter: Prof. Dr. Thomas Kneib

Tag der mündlichen Prüfung: 3 Mai, 2012

Acknowledgments

I owe it to countless people that I finally stand at the end of this journey.

First and foremost, I am sincerely and heartily grateful to my first supervisor Prof. Dr. Stefan Sperlich for his guidance and encouragement. I am grateful for the dissertation topics he suggested to me which keep fascinating me. I wish to thank him for his utmost generosity and patience I surely stretched to the limit. I am truly indebted and thankful to my second supervisor Prof. Stephan Klasen, Ph.D., who was the supervisor of my master's thesis as well, for his advice and his way of working I learned a lot from. Without his generosity and patience throughout, this dissertation would not have seen daylight. I would also like to extend my gratitude to Prof. Dr. Thomas Kneib, who generously accepted my request at short notice that he act as my third examiner.

I would like to offer my special thanks to Prof. Dr. Walter Zucchini for initiating me into statistics. I am proud of having taken virtually all the available courses of his at University of Göttingen, which built the base of this dissertation. Even years later, I still feel the excitement and the heat of the battle in the lecture room. I wish to thank Prof. Dr. María José Lombardía for valuable advice and discussions. I am most grateful for her warm welcome to me on my research trip to Universidade da Coruña. She kindly introduced me to Prof. Dr. Mario Francisco Fernández, the author of one of the most relevant papers to my work. I also wish to thank him for invaluable discussions. I would like to express my appreciation to Prof. Dr. Carmen María Cadarso for her assistance. She gave me a precious opportunity to present my work to her staff members at Universidade de Santiago de Compostela. I would like to extend my deep gratitude to Prof. Dr. SENGA Shigeyoshi, who was the supervisor of my bachelor's thesis at Yokohama City University, for encouragement I always felt during my doctoral study.

I wish to thank Dr. Boris Branisa for many hours of valuable conversations and suggestions. I would like to acknowledge the help provided by Dr. Nils-Hendrik Klann, who took over indispensable and yet the most tedious work of data preparation. I also thank Tatiyana Apanasovich, Ph.D. and Dr. Antonello Maruotti for helpful discussions.

I am much obliged to my colleagues at the Institute for Statistics and Econometrics and the Center for Statistics at University of Göttingen. I have been very lucky to share with my foreign colleagues good times and hard times of studying at university as foreigners. My special thanks go to Dr. Jing Dai and Dr. Duygu Savaşci for many productive conversations, Dr. Yesilda Balavarca for her assistance in preparing for the dissertation defense, Tinoush Jamali for providing me with knowledge of computer tools that have greatly facilitated my work, and Dr. Ta-Chao Kao for being the best office mate during a long doctoral study. I also thank Daniel Adler for his friendliness and expert programming knowledge he shared with me.

I am aware that there are many more people I should express my gratitude to, and that I am even unaware of invaluable support I received from people to whom I owe this dissertation. Lastly I express my gratefulness to my family. Without their support I couldn't have arrived at the end of this long journey. Just before beginning and ending this journey, I lost two of my family members. My grandfather passed away a few months before I came to Göttingen and my grandmother passed on one and a half months after I defended this dissertation in the grandfather's suit. This dissertation is dedicated to them.

Contents

List of Figures

List of Tables

1 Introduction

Whether parametric or nonparametric, statistical models have three fundamental aspects: flexibility, dimensionality and interpretability. In general, parametric regression models have an advantage in estimability and interpretability of parameters of research interests. However, this advantage is conditional on appropriate assumptions about the model, particularly the functional form of the relationship between the response variable and the regressors. Nevertheless knowledge of the true functional relationship is seldom available. A misspecified model may incur severe bias and thus invalid inference. This problem motivates nonparametric regression, which has the advantage of flexibility in model specification. However, it has an inherent problem with dimensionality, known in literature as the "curse of dimensionality".

Semiparametric regression embodies the strength of nonparametric and parametric regression models. In this dissertation I study two types of semiparametric regression: partially linear model and index model. The former circumvents the curse of dimensionality by additively combining a linear parametric component and a nonparametric component. This is a natural way to maintain interpretability of parametrics and flexibility of nonparametrics to avoid specification bias. The latter model evades the curse by reducing the dimension of the design space. This approach is especially useful when data are effectively concentrated in a space of a reasonably small number of dimensions. Index model estimation centers on a subspace (index space) spanned by a set of orthogonal index vectors and a nonparametric function linking the subspace and the response.

The dissertation consists of three essays presented in the subsequent three chapters. Chapter 2 presents a survey essay "Some Recent Advances in Modeling with Mixed Effects for Small Areas, Multi-level and Panel Models", which is a joint work with Prof. Dr. Stefan Sperlich. Mixed effects models, often known by different names in different scientific disciplines, are popular in various fields and rich in model extensions. A wide variety of data types and research interests have promoted development of model extensions and relaxation of rigid model assumptions. The mixed effects model deals with correlations in data by explicitly modeling random effects, not by incorporating the correlations into the covariance structure of the regression error.

Thus the model renders itself convenient not only for practical model extension but even for the purpose of smoothing a function estimate in the spline regression framework. Among others, an important direction of recent advances is semiparametric modeling, which is the topic of Chapter 3. In view of extensive developments in the last few decades, we believe that a survey conducted in an interdisciplinary manner will benefit researchers in diverse scientific communities.

Chapter 3 is devoted to an essay on "Partially Linear Mixed Effects Model without Distributional Assumptions". This model integrates two classes of regression: parametric linear mixed effects model and nonparametric regression model. There has been a well-studied estimation method based on the penalized spline with normality assumptions for the random terms. I propose an alternative estimation which does not rely on distributional assumptions. The new approach faces a series of challenges including bias in parameter estimators due to estimation of a nonparametric component, bandwidth selection in the presence of correlations in data, the test of significance of regression coefficients, and computational difficulties in practice. The essay addresses these challenges and provides a simulation study and practical applications, which demonstrate improvement over the standard linear mixed effects model and another semiparametric estimation recently proposed. A program package `plmm` is provided in the statistical software R to implement the procedures discussed in the essay.

Chapter 4 presents the third essay "Comparison of Principal Component and Directed Principal Component Index Models: An Empirical Study" in cooperation with Dr. Nils-Hendrik Klann, who processed and provided household survey data for this study. In this essay, index models are applied to construct an indicator of the household's welfare status. Theoretical interests lie in comparison of index models differentiated by the types of index vectors, specifically, principal components and "directed principal components". Principal components can be used to reduce the dimension of the design space. However, they span a subspace which certainly contains a maximal amount of information in terms of regressors, but not necessarily information relevant to the functional relationship between the response and the regressors. In contrast, directed principal components are estimated simultaneously with the link function that models the relationship. Consequently the index space relates the most to the response, and thereby the quality of the indicator is expected to improve. Prior to analysis, an important data-type problem is discussed: how to apply to categorical data the index model with directed principal components, which requires continuous data. As a practical solution I suggest using a data transformation method proposed for principal component analysis. The empirical comparison illustrates the potential of an index model with directed principal components as a

tool of exploratory, preliminary analysis.

The empirical results presented in Chapter 3 and 4 are obtained using R and Stata. Program codes are available on request. The R package `plmm` is available at `http://cran.r-project.org/` .

2 Survey of Mixed Effects Model

Some Recent Advances in Modeling with Mixed Effects for Small Areas, Multi-level and Panel Models

Ren Ohinata[a] and **Stefan Sperlich**[b]

a) Institut für Statistik und Ökonometrie, Georg-August Universität Göttingen

b) Département des sciences économiques, Université de Genève

Abstract

While mixed effects models are widely available effective tools in small area estimation, the complexity of real data structure and the necessity of models that are specifically tailored for the objectives of data analysis require a variety of mixed effects model extensions. We review extensions of the classical linear mixed model and bring together knowledge of different research fields where those extensions are frequently used. Our focus is mainly set on giving an overview of a variety of major model extensions rather than of ongoing researches of their asymptotic properties. The survey concentrates on typical relaxation of distributional assumptions and of the classical covariance structure for the error terms, and making the functional form more flexible. This survey includes parametric and nonparametric approaches.

Key words: Mixed effects models; Small area statistics; Longitudinal data; Repeated Measurement; Semiparametric regression.

2.1 Introduction

The linear mixed effects model (LMM) has been established as a linear regression model which takes correlation between observations into account. In a wide range of research fields LMMs have been developed and extended to accommodate data of different types found, for example, in biomedical, forestry, agricultural, economic and social sciences. LMMs are especially popular for panel data analysis; see Laird and Ware (1982), Ghosh et al. (1996) and Diggle et al. (2002). More recently, they have attracted a considerable attention in small area statistics; see for example Ghosh and Rao (1994), Rao(1999, 2003), Pfeffermann (2002), Jiang and Lahiri (2006), Jiang and Ge (2006) and Datta (2009). Recent development is centered around the mixed effects models which are categorized into the model-based approach. Fay and Herriot (1979) is a seminal paper based on a model-based mixed model. See Longford (2010) for some new development of design-based approaches. In the Bayesian framework Ghosh et al. (2006) studied cases where covariates were measured with errors in the small area estimation context. Their predictor of the small area means was further developed to realize more efficient use of data by Torabi et al. (2009).

While different research areas favor different terminologies, for example, small area statistics, multi-level (regression) models or repeated measurement problems (mostly in biology and medicine), it seems to us that little effort has been spent until now on bringing together these different areas although many of the statistical problems of modeling, estimation and testing are basically the same. We believe that the potential synergy is enormous since, to our understanding, most of the differences arise mainly in the subsequent inferences. Different research areas have in common that they try to account for certain clustering, may it be due to space, time, climate, administrative areas or districts, villages or even large families, genetic groups or species.

The above mentioned research fields are even less connected to the more recent phenomena of using mixed effects models in nonparametric statistics as a kind of smart (mostly spline) smoothing; see Ruppert et al. (2003) and Wand (2003). Semiparametric Bayesian methods using mixed models should be considered as a special case since they treat functions and parameters as random which would otherwise be considered as fixed, see Adebayo and Fahrmeir (2005), Kneib and Fahrmeir (2006) and Fahrmeir and Lang (2001) among others. In probably most of the literature on semiparametric models the idea has always been to separate the nonparametric function into a deterministic (fixed effects) and a random part (random effects) so that the smoothing parameter of a spline estimator can be written in terms of the variances of the random effects and the error term. Considerations of additional random effects

can be found especially in the Bayesian literature. The recent non-Bayesian literature often concentrates on longitudinal studies with functions (like varying coefficients) of time, see for example Wu and Zhang (2006). Only recently has literature come out on smooth function estimation of covariate impacts in longitudinal studies (Gu and Ma, 2005), repeated measurement data (Lin and Carroll, 2006), with a focus on small areas (Lombardía and Sperlich, 2008 or Opsomer et al., 2008) and testing (Sperlich and Lombardía, 2010).

The main aim of this review article is to bring together these different research fields. This implies (a) that we will concentrate on aspects which we believe are well known - if at all - only to one of these statistical sub-communities but should be of interest for all, and (b) that we put less emphasis on topics of interest for only one of the research areas. More specifically, we concentrate on the following three aspects: extensions in distributional assumptions, extensions in covariance modeling, and extensions of the functional form. As the estimation of the mean squared prediction error (MSPE) is of strong interest in small area statistics, we will briefly address this aspect for each of the reviewed papers and methods.

Since random effects modeling with longitudinal data in biometrics seems somewhat better known to the statistical community than small area statistics, we would like to add some comments only to the latter. Small area estimation (SAE), especially the model-based approach has received considerable attention during the last two decades. The term "small area" may refer to a small geographical area, (county, district or neighborhood) but it may equally well describe a small domain like a specific group of people or a climatic cluster. SAE makes use of LMMs by specifying a general model as fixed, adding afterward random effects related to the specific area. Note that the interest is not directed toward the estimation of model parameters but area parameters such as the area mean or certain quantiles (for example for poverty mapping). For the prediction of these area parameters the random effects and their predictors are explicitly needed. The squared prediction error estimation is an additional challenge which we do not include in this review due to space restrictions and also because it is of less interest for the other research areas. In the exclusively model-based framework, interesting researches have been done using both Bayesian and frequentist methodologies. In official administrative statistics, small area statistics is standard practice. Indeed, since 2003, the member states of the European Union have been required to supply Eurostat with small area statistics, on provinces, districts, departments, etc. In the USA and Canada the statistical bureaus use this technique for more than a decade. For example, the US Department of Agriculture publishes annual estimates of farm real estate values for 48 states, based on the Agricultural and Land Values Survey, which is characterized by a low

response rate. Thus the topic has become a focus of statistical research, see Battese et al. (1988) and Pfeffermann and Barnard (1991) for early studies in this filed.

In this essay we review extensions of the basic LMM, whose assumptions are often too restrictive in practice. The focus of this review is set on the specification and estimation of extended models. We decided to leave out detailed discussion of the properties of the estimators and predictors of the models, partly for briefness and partly because research is still going on in this direction. Even for a widely known method such as the generalized linear mixed effects model (GLMM), the asymptotic properties of the maximum likelihood (ML) estimators in general still seem to need to be established (Jiang and Ge, 2006). As we will see, although the MSPE is one of the most important topics, it is also open to research in the context of model extension discussed in this review. Statistical inference including hypothesis tests, confidence or prediction interval building, model diagnostics and model selection are also left out for brevity. Instead we refer to the recent articles of Claeskens and Hart (2009) for a review on testing distributional assumptions and Sperlich and Lombardía (2010) for functional form specification tests. There is a large amount of literature on computational methods needed to implement model extensions. They are mentioned only briefly in relation to individual models in question. While some simulation techniques used typically in the Bayesian framework are often mentioned, Bayesian mixed models are left out of scope, except for in the GLMM where Bayesian approaches have certain technical advantages over frequentist approaches.

The rest of the essay is organized as follows. In Section 2.2 we set out the Gaussian linear mixed model which serves as the basic model and is extended in the following sections. It is linear in the fixed and random effects. The normal distribution is assumed for the random effects and the regression error. The covariance structure of those random terms is based on homogeneity and absence of between-cluster correlation. Extensions of the basic model toward more flexible distributional assumptions will be introduced in Section 2.3. Section 2.4 deals with the error term structure especially in the light of longitudinal data. Some approaches are presented which model correlations between within-cluster errors and allow for heterogeneity of the error variance. In Section 2.5 we turn to the functional form of the model. We discuss the GLMM, which allows the response variable of non-continuous types such as count or binomial. The semiparametric LMM and the semiparametric GLMM are also reviewed. Throughout our review, we are interested only in models where the impact of random effects is explicitly modeled. This means that, for example, the generalized estimating equations (GEE) remains out of scope. For the details of GEE, we refer, for example, to Diggle et al. (1994).

2.2 The "Base Model" and Literature in Brief

2.2.1 Basic linear mixed effects model

Although the basic LMM is well known and studied, we start with a brief introduction to clarify notation and summarize the assumptions typically made. It is exactly this set of assumptions that motivate extensions we will review and discuss in subsequent sections.

Let y_{ij} be the jth observed response of the ith cluster. The notation "cluster" is exchangeable with terms such as "subject", "group", "block" or "area", which are used in different statistical contexts. It simply indicates a set of observations that are correlated even after being conditioned on the covariates. Suppose that there are m clusters with n_i observations in the ith cluster, and $n = \sum_{i=1}^{m} n_i$. The basic LMM for the observations in the ith cluster is specified as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i;\ i = 1, \ldots, m\ , \tag{2.1}$$

where $\mathbf{y}_i$ is an n_i-dimensional vector of the response variable and $\mathbf{X}_i$, $\mathbf{Z}_i \subset \mathbf{X}_i$ are design matrices with conforming dimensions, and $\mathbf{e}_i$ is the vector of regression errors with covariance $\mathbf{R}_i = \sigma_e^2\mathbf{I}_{n_i}$. Further, $\boldsymbol{\beta}$ is a p-dimensional vector of parameters and $\mathbf{u}_i$ is a q-dimensional vector of unobservable random effects with zero mean and covariance $\mathbf{D}_i$. In matrix notation, by setting $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_m^T)^T$, $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_m^T)^T$, $\mathbf{Z} = \mathrm{diag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_m)$, $\mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_m^T)^T$ and $\mathbf{e} = (\mathbf{e}_1^T, \ldots, \mathbf{e}_m^T)^T$, (2.1) can equivalently be given as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$.

For the basic model it is assumed that these random effects are independent and identically distributed (i.i.d.) for each cluster with covariance $\mathbf{D}_i$, and are independent of $\mathbf{X}_i$ and $\mathbf{e}_i$. Often, $\mathbf{e}_i$ and $\mathbf{u}_i$ are assumed to be normally distributed as follows:

$$\begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix}\right)\ , \tag{2.2}$$

where $\mathbf{D} = \mathrm{diag}(\mathbf{D}_1, \ldots, \mathbf{D}_m)$ and $\mathbf{R} = \sigma_e^2\mathrm{diag}(\mathbf{I}_{n_1}, \ldots, \mathbf{I}_{n_m})$. Let $\mathbf{V}$ denote the variance of $\mathbf{y}$ given $\mathbf{X}$, $\mathrm{Var}[\mathbf{y}|\mathbf{X}]$. $\mathbf{V}$ can be written as $\mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}^T + \mathbf{R}$, where $\mathbf{V}$ is assumed as a function of a parameter vector $\boldsymbol{\varphi}$, i.e. $\mathbf{V} = \mathbf{V}(\boldsymbol{\varphi})$. Implications of the above assumptions are:

A1 Normal distributions of $\mathbf{u}$ and $\mathbf{e}$ with mean zero.

A2 $\mathrm{Cov}[\mathbf{u}_i, \mathbf{e}_k] = \mathbf{0}$ for k $(= 1, \ldots, m)$, and $\mathrm{Cov}[\mathbf{u}, \mathbf{X}] = \mathbf{0}$, and no correlation in random effects between clusters.

A3 $\mathrm{Cov}[e_{ij}, e_{il}] = 0 \ \ (j \neq l)$, $\mathrm{Cov}[\mathbf{e}_i, \mathbf{e}_k] = \mathbf{0} \ \ (i \neq k)$ and $\mathrm{Cov}[\mathbf{e}, \mathbf{X}] = \mathbf{0}$.

A4 Homogeneous within-cluster variance for $\mathbf{e}$, i.e. $\sigma_e^2 \mathbf{I}_n$.

A5 Homogeneous random effects variances with respect to clusters.

A6 The functional form of the conditional expectation of $\mathbf{y}$ is linear additive separable.

A7 The random terms $\mathbf{u}$ and $\mathbf{e}$ are independent from the covariates of the model.

Certainly, the normality is often replaced by different distributional assumptions. Extensions to relax assumption (A1), affecting also (A4) and (A5), will be discussed in Section 2.4. Section 2.3 studies more in detail extensions of assumptions (A2), (A3), (A4) and partly (A5). Assumption (A6) will be relaxed in Section 2.5 with some discussion of perspectives for (A2). Finally, assumption (A7) is hardly discussed in the literature although fundamental for the use of mixed effects models, see Lombardía and Sperlich (2012) for a first rigorous attempt.

There are several ways of fixed parameter estimation and random effects prediction that lead to basically the same outcomes. We present here Henderson's method based on the joint distribution of $\mathbf{y}|\mathbf{u}$ and $\mathbf{u}$. The best linear unbiased estimator (BLUE) $\tilde{\boldsymbol{\beta}}$ and the best linear unbiased predictor (BLUP) $\tilde{\mathbf{u}}$ are given by

$$\begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{pmatrix} = \underset{\boldsymbol{\beta},\mathbf{u}}{\operatorname{argmin}} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\right)^T \mathbf{R}^{-1} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\right) + \mathbf{u}^T \mathbf{D}^{-1} \mathbf{u} \ . \tag{2.3}$$

It follows that $\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ and $\tilde{\mathbf{u}} = \mathbf{D}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$. "best" means that they minimize the mean squared error (MSE) of the estimator of $\boldsymbol{\beta}$ and that of prediction of $\mathrm{E}[\mathbf{u}|\mathbf{y}]$. The best predictor coincides with the best linear predictor in the case of normality but otherwise not necessarily.

Note that, even though normality for the random and error terms is assumed for convenience, the BLUE and BLUP can be derived without assuming the distributional family for either the random effects or the error term, see for example Searle et al. (1992).

The normality assumptions (or alternatively, moment methods) allow us further to estimate the elements of $\mathbf{V}$, which are usually unknown. Under the distributional assumptions the (profiled) log-likelihood function for $\mathbf{V}$ can be maximized with $\boldsymbol{\beta}$ replaced by the BLUE estimator $\tilde{\boldsymbol{\beta}}$, i.e.

$$l_p(\mathbf{V}) = -\frac{1}{2}\left\{n \, \log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\right\} \tag{2.4}$$

with respect to $\boldsymbol{\varphi}$. When $\mathbf{V}$ is estimated and plugged in the BLUE, the empirical

BLUE (EBLUE) $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{y}$ and the empirical BLUP (EBLUP) $\hat{\mathbf{u}} = \mathbf{D}\mathbf{Z}^T\widehat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ are obtained.

Another likelihood-based estimation is the restricted maximum likelihood (REML) estimation. It is based on linear transformation of data with a matrix of error contrasts $\mathbf{K}$ of dimension n by $(n - p)$ such that $\mathbf{K}'\mathbf{X} = \mathbf{0}$ and thus $\mathrm{E}[\mathbf{K}'\mathbf{y}] = \mathbf{0}$. Then, under the normality assumption, the variance components can be estimated without knowing (or any prior knowledge of) $\boldsymbol{\beta}$. The REML automatically accounts for the loss of degrees of freedom due to the estimation of these unknown parameters.

2.2.2 Existing reviews

We discuss briefly some of the existing reviews, may they be articles or books, about mixed effects models with a special focus on small area statistics and longitudinal data. In the past fifteen years the mixed effects model has been a hot topic in applied statistics so that a new book came out almost every year. However, little effort has been spent so far to bring together different research areas in statistics working with basically the same model.

Ghosh and Rao (1994) presented a comprehensive overview of the SAE. They reviewed the EBLUE, the empirical Bayes and the hierarchical Bayes methods, including their extensions. Model diagnostics of model-based methods was also within the scope of their survey.

Verbeke and Molenberghs (2000) provided a thorough exposition of the LMM for longitudinal data. Their discussion included potential problems arising from the classical normality assumptions as well as the use of mixture distributions as an approach to the problems. Topics such as data exploration, model building, and missing data handling were also discussed.

McCulloch and Searle (2001) studied in detail linear models, LMMs and GLMMs with a main focus on likelihood-based methods, including a brief discussion about nonlinear mixed models.

From a perspective of SAE, Rao (2003) gave an extensive account of the LMM including some basic parametric extensions. Specifically, the MSE (including its estimation) of small area predictions was discussed in detail throughout. The empirical Bayes and hierarchical Bayes methods were given a thorough examination.

Longford (2005) provided a comprehensive description of the SAE based on likelihood-based techniques with detailed case studies. The book also dealt with practical issues such as small-sample properties of small area estimators as well as missing data and model selection.

In the context of SAE, Jiang and Lahiri (2006) gave an extensive review of mixed effects models of which the area level and unit level models were special cases. They surveyed a broad range of variance components estimators of LMMs and their asymptotic variances as well as prediction of the random effects and/or the mixed effects. They also discussed approaches to estimate the MSPE of the EBLUP, which is of paramount importance for SAE. Difficulties involved in estimation, inference, prediction and MSPE of the GLMM were also given detailed description. The scope of the review further included topics such as interval prediction (especially its asymptotic behavior), model building, model selection and model diagnostics.

Similarly to Jiang and Lahiri (2006), but from a more general perspective, Jiang and Ge (2006) reviewed mixed effects models including the nonlinear mixed effects models of which the GLMM is a subset. In addition, they briefly covered semiparametric relaxation of the assumptions on the random effects distribution.

Jiang (2007) gave a comprehensive exposition of the LMM and GLMM including inference methods, model diagnostics, model selection and the MSPE. In addition to likelihood-oriented methods, Bayesian methods were covered. Jiang paid a special attention to non-Gaussian LMMs which assume knowledge of the mean and covariance structure of the random effects and the error but not their distributional families.

2.3 Relaxation of the Distributional Assumptions

This section is dedicated to some of the contributions to relax the distributional assumptions typically imposed on the random effects and error terms. Certainly, in many cases, no distributional assumption is necessary for a feasible generalized least squares estimation of the fixed effects or for moment estimators of the variance components. One could rely, for example, on the so-called Henderson's 3rd (or fitting of constants) method, for which an early reference is Fuller and Battese (1973) for the homoscedastic case and Stukel and Rao (1997) for the heteroscedastic case. Nevertheless, not only for estimation but even more for further inference, we should also care about a sensible likelihood formulation of the problem. In some contexts such as SAE, prediction is the main goal of analysis and thus MSPE is of great interest. MSPE estimation without distributional assumption has been proposed by Hall and Maiti (2006a) who used bootstrap methods to estimate the MSPE.

2.3.1 Parametric extensions of the distributional assumption

The "convenience" of assumption (A1) consists in the fact that, even if it may be too restrictive in practice, it has a mathematical advantage that given (A1) the marginal distribution of the response as well as the conditional distribution of the random effects are normal, which simplifies the maximization of the likelihood and the prediction of the random effects enormously. Certainly, similar things can be said about several of the other assumptions.

Obviously, the normality assumption may lead to misspecification of the model. The prediction of the random effects often strongly depend on distributional assumptions. As far as the estimation of the fixed effects is concerned, the wrongly assumed normal distribution for the random effects has much weaker impact on their estimation, see Butler and Louis (1992), Neuhaus et al. (1992) and Verbeke and Lesaffre (1996). In spite of robustness, the correct specification of the random effects is important not only for efficient estimation but moreover for correct inference for all parameter estimates, see Verbeke and Molenberghs (2000), Butler and Louis (1992), or Ghidey et al. (2004).

Another reason for the need of relaxation of (A1) is the shrinkage effect toward zero observed for predictions. Predicted random effects typically show less variability than actually present in the population. Apart from the problem of prediction arising from inappropriate distributional assumptions, it may be quite difficult to detect deviations from the normality based on predicted random effects.

An attractive approach to relax the normality assumption is to use a mixture of normal distributions so that the true random effects distribution can be flexibly depicted, whether uni-modal or multi-modal, asymmetric or skewed. Each component of the mixture represents a certain proportion of the whole population. This type of model is thus called "heterogeneity model". Such a mixture distribution is particularly useful when the deterministic part of the model is misspecified due to omission of certain categorical variables. An estimated random effects distribution may detect such kind of misspecification and guide the practitioner. In such a case, the number of components of the estimated mixture distribution will correspond to the number of categories of the omitted covariate, see Verbeke and Lesaffre (1996), Ng et al. (2006). This implies that it can serve as a tool for exploratory cluster analysis as well as a test on the Gaussian assumption.

Verbeke and Lesaffre (1996) discussed the heterogeneity model applied to longitudinal data assuming that the random effects $\mathbf{u}_i$ are sampled from a mixture of K normal distributions (at this stage let's assume the number of components K to be

known), i.e.

$$\mathbf{u}_i \sim \sum_{k=1}^{K} p_k \, \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{D}) \ , \tag{2.5}$$

where p_k $(k = 1, 2, \ldots, K)$ is defined as such that $\sum_{k=1}^{K} p_k = 1$. In order to ensure $\mathrm{E}[\mathbf{u}_i] = \mathbf{0}$, a constraint $\mathrm{E}[\mathbf{u}_i] = \sum_{k=1}^{K} p_k \boldsymbol{\mu}_k = \mathbf{0}$ is imposed. Then $\mathrm{E}[\mathbf{y}_i] = \mathbf{X}_i\boldsymbol{\beta}$ holds as in the basic model. The marginal distribution of $\mathbf{y}_i$ is directly given by

$$\mathbf{y}_i \sim \sum_{k=1}^{K} p_k \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\mu}_k, \mathbf{V}_i), \qquad \mathbf{V}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \sigma_e\mathbf{I}_{n_i} \ . \tag{2.6}$$

The parameters $\boldsymbol{\beta}, \boldsymbol{\mu}_k, p_k$ and $\mathbf{D}$ can be estimated with the likelihood method using the expectation maximization (EM) algorithm. Verbeke and Lesaffre (1996) carried out simulations for a balanced panel with 1,000 subjects over five periods. The random effects for the individual intercept were drawn from a mixture of two normal distributions $0.5\mathcal{N}(-2, 1)+0.5\mathcal{N}(2, 1)$. This is meant to represent two heterogeneous sub-populations. They analyzed the data under the assumption of simple normality for the random effects, and also under the assumption of a two-component $(K = 2)$ mixture of normal distributions. The estimation under the former assumption yielded predictors which clearly failed to capture the true distribution. Depending on the relative size of the variance components, a uni-modal density of the predictors was obtained (when the regression error variance was large relative to the random effects variance) or an almost bi-modal density (when the regression error variance was relatively small). The structure of the design matrix $\mathbf{Z}$ can also contribute to this effect. Not surprisingly, under the second assumption one obtains predictors which reflect the true underlying distribution of the random effects. Note that in the context of cluster analysis, Mcnicholas and Murphy (2008) extended this idea to a model with additive mixture components each of which has a different mean and variance.

Ng et al. (2006) applied the idea of heterogeneity modeling to a multi-level model with repeated measurements of subjects found in groups. It was assumed that random effects specific to subjects as well as groups to which subjects belong were mutually independent, and that they possess some mixture distributions. All model parameters and variances were estimated using a special EM algorithm.

Unfortunately, there are serious difficulties for the heterogeneity model in determining the number of mixture components. Watier et al. (1999) carried out a simulation study on a normal mixture distribution without specifying the number of mixture components. They used a random intercept model with 180 clusters, $y_{ij} = x_{1ij}\beta_1 + x_{2ij}\beta_2 + u_i + e_{ij}$ $(i = 1, \ldots, 180)$ where $x_{2ij} = 0$ for all j in 90 clus-

ters and 1 for the rest. In the hierarchical Bayesian framework, they assumed prior distributions for model parameters and hyper-parameters including the number of components K. They compared two model specifications for the random intercept; model (i) with the normal distribution and model (ii) with a normal mixture distribution. The models were fit to simulation data generated with a random intercept of two-component normal mixture distribution. In spite of misspecification, model (i) did not show much difference in the fixed effects estimates. However, model (ii) resulted in a 24% decrease of the posterior standard deviation for $\hat{\beta}_2$ and 33% decrease of its MSE. This kind of decrease was not observed for $\hat{\beta}_1$. On the other hand, when the simulated random effects were normally distributed, model (ii), which was over-parametrized, did not lead to poorer performance than model (i). For both simulation data sets, model (ii) yielded the mode of the posterior distribution of K converging in the true number of the components.

Finally it should be mentioned that neither Verbeke and Lesaffre (1996) nor Verbeke and Molenberghs (2000) included discussion about the MSPE. Watier et al. (1999) studied only the MSE of model parameters but not the MSPE. No discussion about the MSPE can be found in Ng et al. (2006) either.

2.3.2 Non- and semiparametric estimation of the random effects distribution

So far we only considered fully specified random effect and error distributions. Even though the normal mixtures are a clear improvement compared to the simple normal assumption, the simulation study of Watier et al. (1999) showed some limitations due to the necessity of knowing the number of mixture components. Note first that allowing for an arbitrary number of mixture components is indeed a safe remedy for misspecification and that having m mixture components actually corresponds to a Gauss-kernel density with local bandwidths. However, we then run into identification problems yielding huge (finally infinite) variances for the estimates. In other words, we face the typical dilemma in statistics to find an appropriate trade-off between variance and bias. In the following we will summarize different proposals from the literature to attack this problem.

Assuming a normal mixture distribution with an unknown number of mixture components for the random effects distribution of a longitudinal model, Magder and Zeger (1996) proposed an ML estimation subject to a constraint that the variance of the mixture components is greater or equal to some minimum value v. The idea is to set a lower boundary for the within-cluster variability in order to avoid an undersmoothed mixture distribution estimate resulting from an inappropriate number

of components. The parameter v plays a role similar to the bandwidth of the kernel density estimation. The ML is found when the variances of the Gaussian components are all equal to the specified value of v. The model does not require strong distributional assumptions for random effects, but flexible shape and smoothness of its distribution estimate can be attained by controlling the parameter v. This model can be extended to the case of multivariate random effects. Magder and Zeger (1996) included simulation results about the MSE of the model parameter estimators as well as the random effects predictions, but not the MSPE of the response.

Tao et al. (1999) looked at a longitudinal random intercept model $y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta}+u_i+e_{ij}$ with model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma_e)^T$. They suggested a nonparametric estimation method of the random effects distribution based on what is called the predictive recursion algorithm which goes back to Newton and Zhang (1999). The idea is as follows. The objective function to be maximized is the marginal profile likelihood

$$L(\boldsymbol{\theta}, f|\mathbf{y}) = \prod_{i=1}^{m} \int_a^b l_i(\boldsymbol{\theta}|u, \mathbf{y}_i) f(u) du \ , \tag{2.7}$$

where l_i is the likelihood for the ith subject, and the interval (a, b) approximately supports all the probability mass of the random effects distribution. To estimate the density of the random effects $f(u)$, consider the posterior distribution

$$f(u|\mathbf{y}_i, \boldsymbol{\theta}) = \frac{f_{\boldsymbol{\theta}}^{i-1}(u) l_i(\boldsymbol{\theta}|u, \mathbf{y}_i)}{c_i(\boldsymbol{\theta})} \ , \tag{2.8}$$

where $c_i(\boldsymbol{\theta})$ is a normalizing constant. For a given $\boldsymbol{\theta}$, the function $f(u)$ is estimated recursively from $i = 1$ to m (randomly reordered) by

$$f^i(u) = (1 - w_i) f_{\boldsymbol{\theta}}^{i-1}(u) + w_i \frac{f_{\boldsymbol{\theta}}^{i-1}(u) l_i(\boldsymbol{\theta}|u, \mathbf{y}_i)}{c_i(\boldsymbol{\theta})} \ , \tag{2.9}$$

where $f_{\boldsymbol{\theta}}^{i-1}(u)$ is the density estimate from the previous recursion; $w_i = (i + 1)^{-\rho}$ is a user-specified weight which is a function of constant parameter $\rho \in (0, 1]$ and decreases as i increases. The initial density function $f_{\boldsymbol{\theta}}^0(u)$ can be chosen as, for example, uniform or normal. Parameter ρ affects the smoothness of the density estimate. When it is set close to zero, the current observations in $l_i(\boldsymbol{\theta}|u, \mathbf{y}_i)$ receive smaller weights and $f^i(u)$, which is a weighted average of the prior $f_{\boldsymbol{\theta}}^{i-1}$ and the Bayesian posterior density obtained from the current observations, will be closer to its prior and therefore smoother. When $f_{\boldsymbol{\theta}}^m$ from the final recursion is obtained, the parameters in $\boldsymbol{\theta}$ are estimated by maximizing the marginal profile likelihood $L(\boldsymbol{\theta}, f^m|\mathbf{y})$ to which Powell's conjugate direction search method was applied. The

authors did an extensive simulation study obtaining mainly expected findings: their semiparametric approach outperforms parametric specifications if and only if the chosen specification was wrong. The authors also conducted a simulation study of their method in the GLMM framework using ordinal response data. They obtained results similar to the LMM case about the MSE of fixed parameter estimates. Note that they did not study any MSPE. They also applied it to real ordinal data. They argued that the use of random effects density estimate can be effective in identifying misclassified responses (i.e. measurement error in the response) if it is unlikely that there are any omitted covariates, and that such detection is one of the benefits of relaxation of the normality assumption as argued by other researchers. In addition, the authors suggested the use of a density estimate to check the model fit since it plays a role similar to that of the distribution of residuals in ordinary regression analysis.

Zhang and Davidian (2001) proposed representing the density function of the random effects by truncated series expansion. First, formulate random effects $\mathbf{u}_i$ as

$$\mathbf{u}_i = \boldsymbol{L}\mathbf{v}_i \ , \tag{2.10}$$

where $\boldsymbol{L}$ is an unknown $(q \times q)$ lower triangular matrix and $\mathbf{v}_i$ is a random vector of dimension q. It is assumed that $\mathbf{v}_i$ has a sufficiently differentiable smooth density function such that it can be approximated by

$$P_K^2(\mathbf{v}_i)\phi(\mathbf{v}_i) = (\sum_{|\lambda| \leq K} a_\lambda v_i^\lambda)^2 \phi(\mathbf{v}_i) \ , \tag{2.11}$$

where $\lambda = (\lambda_1, \ldots, \lambda_q)$ is a vector of non-negative integers, $v_i^\lambda = v_{i1}^{\lambda_1} \cdots v_{iq}^{\lambda_q}$, which is a monomial of order $|\lambda| := \sum_{k=1}^q \lambda_k$, and $\phi(\mathbf{v}_i)$ is a q-dimensional standard normal density function. K is the order of the polynomial P_K. To ensure that we obtain a density function, the normalization $\int P_K^2(\mathbf{v}_i)\phi(\mathbf{v}_i)d\mathbf{v}_i = 1$ is imposed. Then, in a case with $K = 2$ and $q = 2$, $(\lambda_1, \lambda_2) = \{0, 1, 2\}$ and $P_K(\mathbf{v}_i) = a_{00} + a_{10}v_{i1} + a_{01}v_{i2} + a_{20}v_{i1}^2 + a_{11}v_{i1}v_{i2} + a_{02}v_{i2}^2$. If $K = 0$, then $P_K(\mathbf{v}_i) = a_{00} = 1$ such that $\mathbf{u}_i$ is $\mathcal{N}(\mathbf{0}, \mathbf{L}\mathbf{L}^T)$ (i.e. basic model). Let $\boldsymbol{\theta}$ be a vector of parameters to be estimated. $\boldsymbol{\theta}$ contains a vector $\mathbf{a}$, which is $(a_{00}, a_{10}, \ldots, a_{02})^T$ in the above example, and the elements of $\mathbf{L}$ in addition to the parameters in the basic model. An advantage of their methods is that the marginal log-likelihood function $l(\boldsymbol{\theta}; \mathbf{y})$ can be expressed in a closed form so that the standard optimization routine can be used for estimation. Note that K plays the role of a parameter controlling the flexibility of the shape of the density function estimator. K is selected based on model selection criterion such as AIC, BIC and Hannan-Quinn criterion. The authors applied their method to a

longitudinal cholesterol levels data set. With the baseline effects of fixed covariates being controlled, the model with $K = 1$ resulted in a bi-modal random effects density function estimate suggesting the possibility of sub-populations among subjects. The authors also investigated some properties of the model parameter estimators and the random effects predictions by simulation. However, the MSPE was not included.

A Gaussian mixture model requires determination of the components and their weights and estimation of the means and standard deviations. However, the determination of the number of components is not straightforward due to boundary problems. The maximization of the likelihood of the mixture model with varying means and standard deviations is not an easy task as was pointed out by Verbeke and Molenberghs (2000) and Ghidey et al. (2004). Zhang and Davidian (2001) proposed a method they called the penalized Gaussian mixture linear mixed model, which was developed further by Ghidey et al. (2004) as follows. Assume $\mathbf{u}_i \in \mathbb{R}^2$ for simplicity, and that $\mathbf{v}_i$ in (2.10) extends over a square of $[-b, b]$ by $[-b, b]$ with some b and practically vanishes outside of this square. Suppose a grid of equally spaced points on the interval $[-b, b]$ in both directions. The points are indexed by $j \ \ (j = 1, \ldots, J)$ in one direction and $l \ \ (l = 1, \ldots, L)$ in the other, where J and L may be different. Denote each grid point in the square by index jl. Place at each grid point jl a bivariate normal density of $\mathcal{N}\left(\boldsymbol{\mu}_{jl}, \mathbf{D}_s\right)$, where $\boldsymbol{\mu}_{jl} = (\mu_{1j} \ \mu_{2l})^T$ and $\mathbf{D}_s = diag\left(\tau_1^2, \tau_2^2\right)$. τ_1 and τ_2 are $\frac{2}{3}\left(\mu_{1j} - \mu_{1,j-1}\right)$ and $\frac{2}{3}\left(\mu_{2j} - \mu_{2,j-1}\right)$, respectively. This setting is "based on the assumption that a Gaussian density which extends over $\mu \pm 3\tau$ can be approximated by a B-spline function of degree 3 which extends over 4 equidistant sub-intervals." Then

$$f(\mathbf{u}_i) = \sum_{j=1}^{J} \sum_{l=1}^{L} c_{jl} \mathcal{N}\left(\mathbf{L}\boldsymbol{\mu}_{jl}, \mathbf{L}\mathbf{D}_s\mathbf{L}^T\right) \ , \tag{2.12}$$

where $c_{jl} = \exp(a_{jl})/(\sum_{h=1}^{J} \sum_{k=1}^{L} \exp a_{hk})$ are mixing proportions $\mathbf{a} = (a_{11}, \ldots, a_{JL})^T$ and thus $\sum_{j=1}^{J} \sum_{l=1}^{L} c_{jl} = 1$. The estimation is performed via marginal ML. In order to avoid over-fitting by using an inappropriately large number of grids, a penalty term is considered in the ML-based estimation. Note finally that while Ghidey et al. (2004) carried out simulation studies to investigate the MSE of the model parameter estimators, the MSPE was not considered.

To conclude this section, we should refer to Celeux et al. (2005) considering a large family of mixtures of mixed effects models to bring together flexible parametric methods to model the mean, the distribution, and the variance structure.

2.4 Extensions of the Covariance Structure

It is obvious to think about serial correlation in longitudinal data analysis and spatial correlation in small area statistics. In the case of having longitudinal small area data, even both may need to be addressed. Additionally, heteroscedasticity can be an issue in many applications, sometimes among the error terms, sometimes among the random effects (then often one speaks of heterogeneity in the random effects). In the following we review original articles that studied these kinds of covariance modeling.

2.4.1 Spatial correlation among random effects

In the context of SAE, Saei and Chambers (2003, 2005a) pointed out inappropriateness of ignoring spatial correlation of areas. Based on the model parameters estimated from the data of the sampled areas, the response variable is predicted for units not sampled in the sampled area as well as those in non-sampled areas. Under the assumption of no correlation between areas, the random effects predictions for areas left out of sample will be zero, which are the means of the random effects. In reality, however, bordering areas are likely to be correlated. The authors argued that estimators and predictors for areas in sample as well as out of sample can be calculated consistently with the aid of a reasonable spatial correlation model.

Suppose a random intercept model $y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i + e_{ij}$. In order to capture spatial correlation between areas i and i', the (i, i') element of the covariance matrix of the random effects can be given by

$$\sigma_u^2 \left(1 + \delta_{ii'} \exp \frac{d(i, i')}{\rho}\right)^{-1} , \tag{2.13}$$

where ρ is an unknown parameter, $d(i, i')$ is a predetermined function of the distance (not necessarily Euclidean) between the areas i and i', and $\delta_{ii'}$ is 0 for $i = i'$ and 1 otherwise.

One way of incorporating the covariance structure is the simultaneous auto-regressive model (SAR) (Salvati, 2004; Saei and Chambers, 2005a). The model is constructed starting from $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}$ with a random variable $\mathbf{v}$. It is assumed that $\mathbf{v} = \rho\mathbf{W}\mathbf{v} + \mathbf{u}$ where $\mathbf{W}$ is a matrix of proximity of neighboring areas, and ρ is the spatial dependence parameter and $\mathbf{u}$ is a vector of error terms with zero mean and unknown variance. With $\mathbf{v} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u}$ the SAR model is formulated as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u} + \mathbf{e} , \tag{2.14}$$

which is a special case of the basic LMM with the covariance matrix

$$\mathbf{D} = \sigma_u^2[(\mathbf{I} - \rho\mathbf{W}')(\mathbf{I} - \rho\mathbf{W})]^{-1} \ . \tag{2.15}$$

Model parameters can be iteratively estimated using ML or REML.

Saei and Chambers (2003, 2005a) provided simulation results assuming correlation of random effects only for directly neighboring areas. Their results showed decreases in MSE and increases in prediction efficiency. Salvati (2004) also conducted simulation studies to compare the spatial models and the basic LMM, and concluded that the larger the spatial correlation in absolute terms, the better is the accuracy of estimation measured in MSE.

Saei and Chambers (2003, 2005a) gave analytical formulas of the MSPE estimator of the EBLUP of unit level models. MSPE estimators were given under various assumptions for the covariance structure including area effects and auto-correlated time effects, time varying area effects, and spatial correlated area effects. Saei and Chambers (2005a) provided simulation studies and Saei and Chambers (2005b) studied area level models. Salvati (2004) also provided simulation studies of efficiency gains from using the spacial model in terms of the MSPE of the EBLUP.

An alternative to the SAR model is the conditional autoregressive model which models random effects distribution conditional on those of spatially neighboring areas. See e.g. Salvati (2004) and Kang et al. (2009).

2.4.2 Serial correlation between errors

It is unrealistic to assume for longitudinal data that measurements on the same subject are uncorrelated. One way of modeling the within-subject serial correlation is to modify the error term of the basic model.

Chi and Reinsel (1989) investigated the longitudinal LMM specified as $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_i$ with serial correlation among the within-subject errors $\mathbf{e}_i$. Assume the first-order auto-regression, AR(1), so that

$$e_{ij} = \phi e_{i,j-1} + r_{ij};\ r_{ij} \sim \mathcal{N}(0, \sigma^2);\ j = 1, \ldots, n_i \ . \tag{2.16}$$

The variance of $\mathbf{y}_i$ is $\mathrm{Var}[\mathbf{y}_i] = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \sigma^2\mathbf{R}_i$ where $\mathbf{R}_i$ is not an identity matrix as in the basic model. Parameters $\boldsymbol{\beta}$, the variance components and ϕ are estimated iteratively using the ML. The subject-specific random effects are predicted based on the empirical Bayes estimator. The authors applied four models to medical data. Model (i): LMM with subject specific random intercepts; model (ii): model (i) with

AR(1) for the error term; model (iii): LMM with subject specific random intercepts and slopes; model (iv): model (iii) with AR(1) for the error term. As far as this application was concerned, model (ii) turned out preferable in terms of the log-likelihood of the fitted model. The authors also mentioned an alternative model considering MA process instead of AR process. They gave the MSPE for the BLUP of the random effects and the response variable, noting that replacement of the known variance components by their estimates will lead to underestimation in the case of EBLUP.

Following Diggle et al. (1994) or Verbeke and Molenberghs (2000), the error term may be decomposed in such a way that $\mathbf{e}_i = \mathbf{e}_{1i} + \mathbf{e}_{2i}$ as follows:

$$\mathbf{y}_i \quad = \quad \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \mathbf{e}_{1i} + \mathbf{e}_{2i};\ i = 1, \dots n_i\ . \tag{2.17}$$

Here $\mathbf{e}_{1i}$ is a term capturing serial correlation distributed as $\mathcal{N}(\mathbf{0}, \tau^2\mathbf{H}_i)$ and $\mathbf{H}_i$ is a correlation matrix, and $\mathbf{e}_{2i}$ is a measurement error term with $\mathcal{N}(\mathbf{0}, \sigma_e\mathbf{I}_i)$. These terms are mutually independent, and thus the variance of e_{ij} is constant $\tau^2 + \sigma_e^2$. The variance of $\mathbf{y}_i$ is $\mathrm{Var}[\mathbf{y}_i] = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top + \tau^2\mathbf{H}_i + \mathbf{R}_i$ with $\mathbf{R}_i = \sigma_e^2\mathbf{I}_{n_i}$. The (j, k) element of $\mathbf{H}_i$ is a decreasing function of time interval between jth and kth measurements at time t_j and t_k such as $h(|t_j - t_k|)$ with $h(0) = 1$. h is often assumed to be an exponential or Gaussian function, that is, $h(u) = \exp(-\phi u)$ or $h(u) = \exp(-\phi u^2)$, respectively, with $\phi > 0$. Note that $\mathrm{Var}[\mathbf{e}_i]$ of the model discussed by Chi and Reinsel (1989) can be given in this general framework by $\mathrm{Var}[\mathbf{e}_i] = \tau^2\mathbf{H}_i + \sigma_e^2\mathbf{I}_{n_i} = \sigma_e^2(\tau^2/\sigma_e^2\mathbf{H}_i + \mathbf{I}_{n_i}) = \sigma_e^2\mathbf{R}_i$.

It is worth noting, however, that serial correlation may be confounded with the random effects and may also be explained by random effects. For example, Jones (1990) studied longitudinal data which showed increasing variances over time. While serial correlation estimated in a linear model was significant, an LMM under conditional independence assumption resulted in a better fit, which, the author argued, would often happen to relatively small data sets. Moreover, as is discussed in Verbeke and Molenberghs (2000), modeling in the above form imposes restrictions on the variance components estimates. On the other hand, modeling the serial correlation in addition to random effects can reduce the number of random effects which would otherwise be needed. The MSPE was not considered at all.

2.4.3 Heteroscedasticity in errors

The models in the previous sections assumed homogeneity for the within-cluster variance (i.e. $\mathrm{Var}[e_{ij}] = \sigma_e^2\ \forall i, j$) although this assumption may be unrealistic. Research

interests may lie in potential heterogeneity in individual variations and identifying variables related to those variations. While heterogeneity of the random effects has been studied quite a lot, less attention has been paid to this classical type of heteroscedasticity, i.e. a non-constant variance of the error term. If heteroscedasticity is taken into account, then mostly in a rather restrictive way with a fully parametric known variance function. A classical reference is Stukel and Rao (1997). In the following we review three, perhaps less known but much more flexible ways of addressing this - at least in social sciences - quite crucial issue.

Li and Stengos (1994) considered the case where the variance of the error term was simply an unknown nonparametric function of a d-dimensional covariates $\mathbf{w}_{ij}$. They proposed first estimating the variance of the random effect σ_u^2 making use of the independence of the error terms as follows. Let's denote $\mathrm{Var}[e_{ij}]$ and $y_{ij} - \mathbf{x}_{ij}^\top \boldsymbol{\beta}$ by $\sigma_e^2(\mathbf{w}_{ij})$ and U_{ij}, respectively. Since it holds that

$$\mathrm{Cov}[U_{ij}, U_{ij'}] \quad = \quad \mathrm{E}\left[U_{ij}U_{ij'}\right] = \sigma_u^2 \; \forall i,\, j \neq j' \;, \tag{2.18}$$

σ_u^2 can be estimated by

$$\hat{\sigma}_u^2 = \frac{1}{\sum_{i=1}^m n_i \left(n_i - 1\right)} \sum_i \sum_{j \neq j'} \hat{U}_{ij}\hat{U}_{ij'} \;, \tag{2.19}$$

where $\hat{U}_{ij}$ is the OLS residual $y_{ij} - \mathbf{x}_{ij}^\top \tilde{\boldsymbol{\beta}}_{OLS}$. Noting $\mathrm{Var}[U_{ij}|\mathbf{w}_{ij}] = \sigma_u^2 + \sigma_e^2(\mathbf{w}_{ij})$, denoted here by σ_{ij}^2, the variance function can be estimated by the Nadaraya-Watson (or the local polynomial regression) via

$$\hat{\sigma}_{ij}^2 = \frac{\sum_{i'=1}^m \sum_{j'=1}^{n_{i'}} K_{ij} \hat{U}_{i'j'}^2}{\sum_{i'=1}^m \sum_{j'=1}^{n_{i'}} K_{i'j'}} \;, \tag{2.20}$$

where K_{ij} is a (multivariate) kernel weight $K_{ij} = K(\mathbf{H}^{-1}(\mathbf{w}_{i'j'} - \mathbf{w}_{ij}))$ with a diagonal matrix of bandwidths $\mathbf{H} = diag(h_1, \ldots, h_d)$ and a kernel weighting function K. The variance function estimator is simply given by $\hat{\sigma}_e^2(\mathbf{w}_{ij}) = \hat{\sigma}_{ij}^2 - \hat{\sigma}_u^2$. The MSPE was not studied.

Lin et al. (1997) proposed a model that specifies within-cluster variances as random. In some cases this specification is of importance because the variance components estimates are affected by whether or not the within-cluster variance is random, while fixed effects estimates are not. The authors proposed heterogeneous within-cluster error variances $\sigma_i^2 \mathbf{R}_i$ under normality assumption, where $\mathbf{R}_i$ is a correlation matrix. Further, σ_i^2 is inverse-gamma distributed with mean σ_{0i}^2 and variance $\delta\sigma_{0i}^4$, where δ is a heterogeneity parameter. σ_{0i}^2 is determined by a vector of covariates $\mathbf{w}_i$ and

unknown parameters $\boldsymbol{\alpha}$ in the form of $\log(\sigma_{0i}^2) = \mathbf{w}_i^T\boldsymbol{\alpha}$. Note that, with $\delta = 0$ and $\mathbf{w}_i = 1$, the model is reduced to the basic LMM. The authors circumvented complication of the full marginal likelihood function due to additional specification of random within-cluster variances by using the quasi-likelihood for the estimation of the fixed effects, the pseudo-likelihood for the estimation of $\boldsymbol{\varphi}$ (parameters of the covariance matrix $\mathbf{V}$) and $\boldsymbol{\alpha}$, thereby only the first and second moments of $\mathbf{y}_i$ need to be correctly specified for the consistency of the estimators. δ is estimated by the method of moments. Under the regularity conditions the estimators obtained are consistent and asymptotically normally distributed. While consistency and asymptotic normality of estimators were considered, MSPE issues were not discussed.

In contrast to the basic model, the mixed effects model of Fay-Herriot type in the SAE context provides difficulty in estimating the variance of the error term due to lack of replications in each area. In order to account for heterogeneity in the error variances, González-Manteiga et al. (2010) proposed a non-parametric estimation method using kernel estimation. Their model is

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + u_i + \sqrt{w_i}e_i \ , \tag{2.21}$$

where $e_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. σ^2 is estimated from the sampling variance of the estimates of y_i. The heteroscedasticity weight w_i is assumed to be an unknown function of either a covariate, that is, $w_i = w(x_i)$ or the marginal expectation of the response variable, that is, $w_i = w(\mathbf{x}_i^T\boldsymbol{\beta})$, where $w(\cdot)$ is some smooth function. The error variance of the ith area is then $w_i\sigma^2$. This is estimated using observations close to the ith area weighted by the kernel function. The authors provided proof for the consistency of the estimators of the parameters $\boldsymbol{\beta}$ and σ_u^2. Note that the authors investigated the MSPE of the EBLUP for their Fey-Harriot model. They proposed two bootstrap methods to estimate the MSPE. In the first method the estimate is obtained with variance components estimates given above. The other takes into account the bias in the first method due to the estimation of the variance components. They provided a simulation study to compare their bootstrap methods and the analytical approximation given by Prasad and Rao (1990).

2.5 Relaxation of the Functional Form

Most of the inference in small area statistics is model based. Therefore, not only an adequate modeling of the random effects and error term distributions and their covariance structure should be demanded but also correct specification of the mean function is crucial. Misspecification of the mean function can easily lead to endo-

geneity of the covariates and the fundamental assumption of independence between covariates and random effects as well as error terms is violated. Moreover, the nature of response variable may require generalization of the LMM, for example, for discrete responses such as binary or Poisson.

SAE often has as its objective the prediction of the area means or totals. An LMM may need to be specified in terms of transformed variables, so that predictors need to assume appropriate forms according to the transformation. Chambers and Dorfman (2003) considered predictors for such cases.

In this section we will first review generalization of the LMM to the GLMM and then turn to nonparametric mean functions.

2.5.1 Generalized linear mixed model

The basic LMM assumes a continuous response variable. For a response variable that is not continuous but of other types such as binary, multi-categorical or count, the basic LMM needs to be extended to the GLMM. The GLMM is an extension of the generalized linear model (GLM) and inherits the main feature of the LMM that within-cluster correlations are accounted for by random effects. The GLMM can also serve as a tool to account for overdispersion in the GLM. We refer to Agresti et al. (2000) for several interesting applications of the GLMM in social science.

Assume that the jth observation of the ith cluster y_{ij} is conditionally independently distributed given the random effects $\mathbf{u}_i$, and its distribution is a member of the fully parametric exponential family. The GLMM is formulated as

$$G(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{u}_i \ , \tag{2.22}$$

where G is a known link function, and μ_{ij} is the conditional expectation $\mathrm{E}[y_{ij}|\mathbf{u}_i]$. The random effects are typically assumed to possess a known distribution. The likelihood function to be maximized is

$$L\left(\boldsymbol{\beta}, \mathbf{D}, \phi; \mathbf{y}\right) = \prod_{i=1}^{m} \int \prod_{j=i}^{n_i} f_i\left(y_{ij}|\mathbf{u}_i\right) f\left(\mathbf{u}_i\right) d\mathbf{u}_i \ , \tag{2.23}$$

where density functions f_i and f are all specified accordingly and $\mathbf{D}$ is the covariance matrix of the random effects. However, if the dimensions of the random effects, say q, are large, estimation of the unknown model parameters involves the computational difficulty of facing a q-dimensional integral. Often, this does not have an analytically closed form expression and ML-based parameter estimators are difficult

to obtain. There are mainly two approaches that circumvent the difficulty. One is numerical approximation which theoretically allows for convergence to the ML estimation, such as the quadrature method and the Monte Carlo expectation maximization (MCEM). The other is based on linearization of the GLMM such as the penalized quasi-likelihood (PQL) or the pseudo maximum likelihood (PML). These are essentially equivalent to the approach using Laplace approximation.

One way of approximation is the Gauss-Hermite quadrature technique, which can be used to approximate an integral of products of functions involving a term $\exp(-u^2)$. See, for example, Liu and Pierce (1994). A disadvantage of the Gauss-Hermite quadrature method is that it is difficult to approximate a high-dimensional integral and thus a model with crossed random factors or random factors nested in high multi-levels often cannot be handled. Another numerical approximation approach is an iterative procedure MCEM method in which unobserved random effects are considered as missing values. In the rth iteration, the expectation step calculates $\mathrm{E}[\log f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta}) f(\mathbf{u};\mathbf{D}) |\mathbf{y};\boldsymbol{\beta}^{(r-1)},\mathbf{D}^{(r-1)}]$, that is, the expectation of the conditional log-likelihood with respect to the conditional distribution of the random effects given $\mathbf{y}$ and model parameters estimated in the $(r-1)$th iteration, $\boldsymbol{\beta}^{(r-1)}$ and $\mathbf{D}^{(r-1)}$. The conditional distribution of $\mathbf{u}$ given $\mathbf{y}$, and estimates $\boldsymbol{\beta}^{(r)}$ and $\mathbf{D}^{(r)}$ is necessary in the expectation step and needs to be estimated using the Markov chain Monte Carlo (MCMC) algorithm. The expected conditional log-likelihood is maximized with respect to $\boldsymbol{\beta}$ and $\mathbf{D}$ in the maximization step. This process is iterated until convergence. For more details of the MCMC algorithm used for the MCEM, see Booth and Hobert (1999), McCulloch (1994, 1997), Zeger and Karim (1991), Malec et al. (1997) or Ghosh et al. (1998). A more recent extension can be found in Song et al. (2005), which proposed maximization by parts.

Breslow and Clayton (1993) proposed the PQL based on the maximization of quasi-likelihood with a penalty to prevent arbitrary prediction of random effects. The PQL is based on linearization and analogous to the iteratively reweighted least squares for the GLM. The model is linearized by the first order Taylor expansion around μ_{ij} and then reduced to an LMM form $t_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{u}_i + \Delta_{ij}(y_{ij} - \mu_{ij})$, where t_{ij} is the so-called working response variable or pseudo data and Δ_{ij} is the first derivative of $G(y_{ij})$ evaluated at μ_{ij}. With the working response variable obtained, $\boldsymbol{\beta}$, $\mathbf{u}$ and $\mathbf{D}$ are iteratively estimated and predicted using LMM estimation methods. This procedure is iterated until convergence. Issues about inconsistency of PQL estimators are discussed for example in Breslow and Lin (1995), Lin and Breslow (1996).

One way of extending the GLMM is to allow for non-normal distributions for the

random effects. Such a model is often called the hierarchical LMM. Motivated by the idea that the distribution of the random effects should be better determined by data or the purpose of inference, Lee and Nelder (1996) proposed models in which the random effects distribution is conjugate to the distribution of the response variable. Combinations of the distributions between the response variable and the random effects include Poisson and gamma, binomial and beta, gamma and inverse gamma, and inverse Gaussian and gamma. This approach has a technical advantage that estimation does not depend on the marginal likelihood and hence integral calculation. See also Antonio and Beirlant (2007) for application to actuarial risk data analysis.

The Bayesian framework provides another approach to the GLMM. In contrast to the frequentist methods, the fixed effects $\boldsymbol{\beta}$ and the covariance matrix of the random effects $\mathbf{D}$ are considered as random. By specifying a joint diffuse prior density $f(\boldsymbol{\beta}, \mathbf{D})$, the joint posterior distribution can be formulated as

$$f(\mathbf{u}, \boldsymbol{\beta}, \mathbf{D}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{u})\, f(\mathbf{u}|\mathbf{D})\, f(\boldsymbol{\beta}, \mathbf{D})}{f(\mathbf{y})} \ . \tag{2.24}$$

The Bayesian approach allows for not only a linear predictor involving a complicated random effects structure but also various distributions for the random effects. For conditional posterior distributions of $\mathbf{u}$, $\boldsymbol{\beta}$ and $\mathbf{D}$ that are generally intractable due to high-dimensional integrals, MCMC techniques are used to obtain the posterior or predictive distributions for estimation and prediction. See Zeger and Karim (1991), and Booth and Hobert (1998) for discussion about the MSPE in the GLMM.

For the longitudinal GLMM, Li et al. (2004) proposed an alternative approach to predict the random effects and estimate the variance components without assuming any distribution for the random effects. Their idea is based on the sufficiency score and conditional score functions. Earlier GLMM estimation methods without distributional assumptions are for example Aitkin (1999) or McCulloch (1997).

Let us come to detailed examples of discrete response models with random effects. McNeil and Wendin (2003) looked at portfolio credit risk modeling. The authors discussed a generalized random intercept model to analyze yearly defaults data of obligors and credit rating classes. Besides the response variable type, their model is different from the standard LMM for longitudinal data in that the clustering factor is the year and thus observational units are nested within year. Each level of the year-specific random intercept represents the general state of the economy of the year under consideration. Because yearly economic situations are likely to be related over some years, the model needs to account for correlation between random intercepts of neighboring years.

Let the response variable y_{tj} be the default case of obligor j $(= 1, \ldots, n_t)$ in year t $(= 1, \ldots, m)$, and let $\eta_{tj} = \mu_j + u_t$ be the linear predictor, where μ_j is a dummy variable for credit rating class to which obligor j belongs and u_t is a year-specific random intercept for year t. A binomial GLMM can be applied with u_t which is assumed to be the first order autoregressive time series as follows:

$$u_t|u_{t-1} \sim \mathcal{N}\left(\alpha u_{t-1}, \sigma_u^2\right), \; u_1 \sim \mathcal{N}\left(0, \frac{\sigma_u^2}{1-\alpha^2}\right), \; |\alpha| < 1 \; . \tag{2.25}$$

As before, the marginal joint density of $\mathbf{y}_t$ is

$$f\left(\mathbf{y}_t\right) = \int \prod_{j=1}^{n_t} f\left(y_{tj}|u_t\right) f\left(u_t\right) du_t \; . \tag{2.26}$$

However, since $f\left(u_t\right)$ is not i.i.d., the marginal likelihood for the whole sample is

$$f\left(\mathbf{y}\right) = \int \cdots \int \prod_{t=1}^{m} \prod_{j=1}^{n_t} f\left(y_{tj}|u_t\right) f\left(u_1, \ldots, u_m\right) du_1 \cdots du_m \; . \tag{2.27}$$

In order to deal with the difficulty in maximizing the likelihood, the authors suggested estimation using MCMC techniques. They fit the model to real defaults data by specifying a uniform prior distribution over $(0, 1)$ for α, a Gaussian prior with a large variance for the fixed effects, and an inverse-gamma prior for σ_u^2. Similarly, the number of defaults y_{tk} of credit rating class k in year t can be modeled with the Poisson GLMM. With the model parameters estimated, random effects can be predicted for year $m+1$ and then used for the prediction of the response conditional on the past random effects. These models can be generalized with additional variables such as obligor-specific covariates, global covariates and other random effects so that the linear predictor is given by $\eta_{tj} = \mathbf{x}_{tj}^T \boldsymbol{\beta} + \mathbf{z}_{tj}^T \mathbf{u}_t$.

When considering counting data, the possibility of extending the model by zero-inflation is an important issue in many applications. Hall (2000) studied extensions of the zero-inflated Poisson model (ZIP) and zero-inflated binomial model (ZIB). The ZIP is a mixture of distributions for two states, i.e. a Poisson distribution (Poisson distribution state) and a degenerate distribution with a point mass of zeros (zero state). Analogously to the ZIP extension, the ZIB consists of binomial and zero states. More specifically, the ZIP extended with a random intercept is given by

$$y_{ij}|u_i \sim \begin{cases} 0 & \text{with probability } p_{ij} \\ Poisson(\lambda_{ij}) & \text{with probability } 1 - p_{ij} \end{cases} , \tag{2.28}$$

where $u_i \sim \mathcal{N}\left(0, \sigma_u^2\right)$, $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{in_i})^T$ and mixing probabilities $\mathbf{p}_i = (p_{i1}, \ldots, p_{in_i})^T$. The Poisson distribution state and the mixing probabilities are modeled by $\log(\boldsymbol{\lambda}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{1}_{n_i}u_i$ ($\mathbf{1}_{n_i}$ is an n_i-dimensional vector of one's) and $\text{logit}(\mathbf{p}_i) = \mathbf{W}_i\boldsymbol{\gamma}$, respectively, with design matrices $\mathbf{X}_i$ and $\mathbf{W}_i$ and regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The marginal log-likelihood is

$$l\left(\boldsymbol{\psi}; \mathbf{y}\right) = \sum_{i=1}^{m} \log \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} f(y_{ij}|u_i; \boldsymbol{\psi}) f\left(u_i; \boldsymbol{\psi}\right) du_i \ , \tag{2.29}$$

where $\boldsymbol{\psi} = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top, \sigma_u^2)^T$ and

$$f(y_{ij}|u_i; \boldsymbol{\psi}) = \left[p_{ij} + (1 - p_{ij}) \exp\left(-\lambda_{ij}\right)\right]^{s_{ij}} \left[(1 - p_{ij}) \frac{\exp\left(-\lambda_{ij}\right) \lambda_{ij}^{y_{ij}}}{y_{ij}!}\right]^{1-s_{ij}} \tag{2.30}$$

with $s_{ij} = 1$ if $y_{ij} = 0$ and $s_{ij} = 0$ otherwise.

Similarly, the ZIB random intercept model is formulated as

$$y_{ij}|u_i \quad \sim \quad \begin{cases} 0, & \text{with probability } p_{ij} \\ binomial\left(n_{ij}, \pi_{ij}\right) & \text{with probability } 1 - p_{ij} \end{cases} \tag{2.31}$$

with $\log(\boldsymbol{\pi}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{1}_{n_i}u_i$ and $\text{logit}(\mathbf{p}_i) = \mathbf{W}_i\boldsymbol{\gamma}$. The marginal log-likelihood is

$$f(y_{ij}|u_i; \boldsymbol{\psi}) = \left[p_{ij} + (1 - p_{ij}) \left(1 - \pi_{ij}\right)^{n_{ij}}\right]^{s_{ij}} \left[(1 - p_{ij}) \begin{pmatrix} n_{ij} \\ y_{ij} \end{pmatrix} \pi_{ij}^{y_{ij}} \left(1 - \pi_{ij}\right)^{n_{ij} - y_{ij}}\right]^{1-s_{ij}} . \tag{2.32}$$

The model parameters are estimated by the EM algorithm with Gaussian quadrature, which is used to approximate the marginal distribution $f(\mathbf{y}_i)$. The author applied the ZIP and ZIB mixed models to longitudinal count data and obtained better fits than a GLMM and a ZIP/ZIB without random effects terms.

Certainly, zero-inflation can be combined with most of the other extensions we are discussing in this article. For a quite recent contribution, the combination of zero-inflation with different extensions in the GLMM context, see Alfò and Maruotti (2010) and references therein.

Olsen and Schafer (2001) applied a strategy similar to Hall (2000) to semi-continuous longitudinal data, that is, data containing a random variable whose distribution combines a continuous distribution with point masses at one or more locations. However, a mass of zeros are valid self-representing data, not proxies for negative or missing data typically handled with censored or truncated variable models. Similarly to the ZIB and ZIP models, semi-continuous response variable is assumed to be a result

of two processes: one process determining whether the response is zero or non-zero and the other determining the non-zero values. The authors modeled the former process using the logit with random effects and the latter using the LMM. The random effects included in the two processes were assumed to be jointly normal and possibly correlated. The authors used Laplace's approximation to deal with integral calculation in the likelihood maximization.

Hall and Wang (2005) considered a two-component mixture of GLMMs, which can be understood as an extension of a two-component mixture of GLMs or an extension of a GLMM to which a second component is added. This type of model is useful when there is heterogeneity in the population so that the data represent a small number of sub-populations that cannot be directly identified. Assuming two latent sub-populations underlying the data, the distribution of the conditional response is given by

$$y_{ij}|\mathbf{u}_i \sim \begin{cases} F_1(y_{ij}|\mathbf{u}_i;\theta_{1ij},\phi_1) & \text{with probability } p_{ij} \\ F_2(y_{ij}|\mathbf{u}_i;\theta_{2ij},\phi_2) & \text{with probablity } 1-p_{ij} \end{cases}, \tag{2.33}$$

where F_1 and F_2 are distributions from the exponential family with the density function conditional on random effects $\mathbf{u}_i$:

$$f_k(y_{ij}|\mathbf{u}_i;\theta_{kij},\phi_k) = \exp\left\{\frac{y_{ij}\theta_{kij} - b_k(\theta_{kij})}{\phi_k} + c_k(y_{ij},\phi_k)\right\}, \tag{2.34}$$

where $k = \{1,2\}$, b and c are known functions; θ is the canonical parameter; and ϕ is a dispersion parameter. The linear predictors of the two GLMM components may be correlated through random effects. Mixing probabilities $\mathbf{p}_i = (p_{i1},\ldots,p_{in_i})^\top$ are modeled in the form $G_p(\mathbf{p}_i) = \mathbf{W}_i\boldsymbol{\gamma}$, where G_p is a known link function such as logit; $\mathbf{W}$ is a design matrix; and $\boldsymbol{\gamma}$ is a vector of regression coefficients. Model parameters involved in the two GLMM components and $\boldsymbol{\gamma}$ are estimated using the EM algorithm with quadrature techniques.

Finally it should be mentioned that almost none of the cited articles considered the problem specific to the small area statistics of estimating the MSPE.

2.5.2 Semiparametric linear mixed models

Semiparametric regression is useful to capture complicated relationship between the response variable and covariates for which parametric models fit poorly. An LMM combined with the semiparametric regression is capable of capturing such relationship nonparametrically while accounting for the correlation structure of the response variable. A semiparametric LMM corresponding to the random intercept model can

be formulated as

$$y_{ij} = \xi(x_{ij}) + u_i + e_{ij} \ , \tag{2.35}$$

where ξ is a smooth function and u_i is the ith cluster-specific random intercept and e_{ij} is the error term. One semiparametric way of modeling ξ is the use of the piecewise polynomial spline regression of order p:

$$\xi(x_{ij}; \boldsymbol{\beta}, \mathbf{v}) = \beta_0 + \beta_1 x_{ij} + \cdots + \beta_p x_{ij}^p + \sum_{k=1}^{K} v_k (x_{ij} - \kappa_k)_+^p \ , \tag{2.36}$$

where

$$(x - \kappa_k)_+ = \begin{cases} 0 & , \text{for } x \leq \kappa_k \\ x - \kappa_k & , \text{for } x > \kappa_k \end{cases} \tag{2.37}$$

and κ_k $(k = 1, \ldots, K)$ are knots. $\mathbf{v} = (v_1, \ldots, v_K)^T$ is assumed to be a random vector distributed as $\mathbf{v} \sim \mathcal{N}(0, \sigma_v^2 \mathbf{I})$. This assumption serves to provide smoothness of the fitted spline regression. The covariance structure of $(\mathbf{u}^T \, \mathbf{v}^T \, \mathbf{e}^T)^T$ is then

$$\mathrm{Cov} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \sigma_u^2 \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_v^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2 \mathbf{I} \end{pmatrix} \ . \tag{2.38}$$

The BLUE of $\boldsymbol{\beta}$ and the BLUP of $\mathbf{u}$ are given by

$$\begin{pmatrix} \tilde{\boldsymbol{\beta}} \\ \tilde{\mathbf{u}} \end{pmatrix} = \underset{\boldsymbol{\beta}, \mathbf{u}}{\mathrm{argmin}} \left(||\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u} - \mathbf{Z}^* \mathbf{v}||^2 + \alpha_1 \left\|\mathbf{u}\right\|^2 + \alpha_2 \left\|\mathbf{v}\right\|^2 \right) \ , \tag{2.39}$$

where $\mathbf{Z}^*$ is a matrix whose (i, k) element is $(x_{ij} - \kappa_k)_+^p$; $\alpha_1 = \sigma_e^2 / \sigma_u^2$ and $\alpha_2 = \sigma_e^2 / \sigma_v^2$. Note that α_2 is a smoothing parameter. When σ_v^2 and σ_e^2 are unknown, α_2 is determined with (RE)ML estimates of σ_v^2 and σ_e^2. Then, with an estimate of σ_u^2, the EBLUE and EBLUP $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are obtained.

The semiparametric modeling in the mixed model framework has an advantage that standard LMM estimation and prediction methods can be used. This means that the (E)BLUE of $\boldsymbol{\beta}$ and (E)BLUP of $\mathbf{u}$ are obtained for a mixed effects model with a regression spline. See Gu and Ma (2005) for the generalized cross-validation method to search for a smoothing parameter.

In the context of SAE, Opsomer et al. (2008) applied the LMM with spline smoothing to survey data. In addition to area random effects, they adopted two-dimensional geographical coordinates as covariates using splines with radial basis functions. Their model assumption of no correlation among area effects implies that the area effect is

predicted to be zero for an area where no sample is available. However, use of spatial splines allowed to utilize information from neighboring areas to improve small are predictions. Ugarte et al. (2009) proposed a mixed effects model using B-spline bases. They investigated the behavior of the MSPE analytically as well as by using bootstrap methods.

Zhang et al. (1998) investigated a semiparametric model for longitudinal data

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \xi(t_{ij}) + \mathbf{z}_{ij}^T\mathbf{u}_i + e_1\left(t_{ij}\right) + e_{2,ij} \ , \tag{2.40}$$

where t_{ij} is a time point and ξ is a fixed smooth function; potential within-subject correlation is accounted for by an error term $e_1(t_{ij})$ for which nonstationarity is allowed. This can be further extended by allowing a nonparametric smooth function for the random effects. For longitudinal data, Wu and Zhang (2002) suggested a local polynomial mixed model

$$y_{ij} = \eta(t_{ij}) + v_i(t_{ij}) + e_{ij} \ , \tag{2.41}$$

where η is the population mean curve and v_i is a subject-specific random effects curve. v_i is assumed to have mean zero and a covariance function $\gamma\left(t, t'\right) = \mathrm{E}\left[v_i\left(t\right) v_i\left(t'\right)\right]$. $\mathbf{e}_i = \left(e_{i1}\, e_{i2} \ldots e_{in_i}\right)^T$ is assumed to have mean zero and a covariance structure $\mathbf{R}_i = \mathrm{diag}(\sigma^2(t_{i1}), \ldots, \sigma^2(t_{in_i}))$ with a variance function $\sigma^2(t_{ij})$, which means that in contrast to the methods described in Section 2.4.1 the within-subject correlation structure is modeled in the random effects specification, not in the error term. In contrast to the algorithm of Wu and Zhang (2002), which estimated the population mean function and subject-specific random effects curves simultaneously using the same bandwidth, Park and Wu (2006) used different bandwidths for each to improve efficiency of the algorithm.

Let us turn to semiparametric extension of the GLMM. The linear predictor η_{ij} of the GLMM can include semiparametric terms in the form called the generalized partially linear model

$$G(\mathrm{E}[y_{ij} \mid \mathbf{x}_{ij}, s_{ij}, u_i]) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + m(s_{ij}) + u_i \ , \tag{2.42}$$

where G is again a known link function, and m is a nonparametric smooth function of covariate s_{ij}. As is the case for the GLMM, the conditional response $y_{ij}|\mathbf{u}_i$ belongs to the exponential family. The linear predictor above can further be extended to an additive model form

$$\eta_{ij} = \sum_{h=1}^{l} f_h(x_{hij}) + u_i \ , \tag{2.43}$$

where f_h is smooth nonparametric function of covariate x_{hij}.

Gurrin et al. (2005) applied the GLMM with a predictor including additional terms of a linear spline function. Even though a parametric model fit turned out appropriate, the authors concluded that the use of spline smoothing in a mixed effects model framework allowed them to discern the shape of the relationship between variables. Lombardía and Sperlich (2008) used a generalized partially linear model and estimated the nonparametric function m using the kernel smoothing technique. They further suggested other extensions of the basic LMM to semiparametric models such as the generalized partially linear single-index model, the generalized additive partially linear model and the semiparametric separable model. See González-Manteiga et al. (2012) for further references to different nonparametric extensions of the LMM with a comparison between them.

We finally turn once again to the specific problem of analyzing the MSPE. Gu and Ma (2005) did not discuss the issue. Opsomer et al. (2008) gave an analytic formula of the MSPE of the EBLUP as well as its estimator. They also discussed the use of bootstrap procedure to obtain the MSPE. Zhang et al. (1998) investigated the model parameter estimates and their standard errors by simulation. However, they did not refer to the MSPE. Wu and Zhang (2002) also did not mention the MSPE while giving a thorough argument to the (asymptotic) MSE of the nonparametric fixed effect function, which is also found in Park and Wu (2006). Lin and Carroll (2006) also considered the semiparametric GLMM providing asymptotic discussion of different estimation approaches.

2.6 Concluding Remarks

This brief and selective review intended to bring together the literature on mixed effects models from different areas where statisticians have developed models and methods to relax various limitations of the classical simple LMM. Depending on disciplines, one speaks of random effects or mixed effects models, models for multilevels or small areas, of longitudinal or panel data studies or simply the analysis of data with repeated measurements. Certainly, at the moment of application, each model may be special and particular depending on disciplines. This is not only because of the nature of data analyzed in biometrics and medicine, econometrics or social sciences, official (administrative) statistics -in all the fields mixed effects models are still an uprising popular topic- but also or even more because of the focus of interest. Nevertheless, they all refer to the same type of model and pose rather similar questions from the statistical point of view.

With the literature being too abundant, no paper, maybe not even a single book

could give a comprehensive review of even the most important contributions. Bear in mind that we completely excluded a whole section about the Bayesian contributions, which are particularly plentiful in this field. Neither have we discussed the large literature of computational statistics dealing with the utmost complex problem of implementing estimation for multi-level (we refer here to models with several random effects) GLMM. We mentioned that LMMs are a popular computational tool for penalized spline smoothing in nonparametric statistics but are interpreted as fully deterministic although quite flexible functions.

The idea of including random effects was originally to improve efficiency by estimating the model through more adequate covariance structure modeling. Even if there is no place for specific examples, we should emphasize the usefulness of mixed effects models for prediction, in particular data matching and data mapping. Here the objectives of small area statistics should be mentioned, and we could equally well speak of the prediction of macro-parameters. We conclude this brief review with a remark that, for most of the model extensions and methods introduced, still very little is known about the estimation of the mean prediction error, i.e. the error we make when predicting level parameters (probably the main challenge of small area statistics), data matching or data mapping. Datta and Lahiri (2000) provided the second-order accurate MSPE of the EBLUP obtained with maximum-likelihood-oriented variance components estimators. In contrast to estimation methods based on analytical approximation, Hall and Maiti (2006b) proposed bias-corrected MSPE estimators and prediction intervals using parametric bootstrap methods. In Hall and Maiti (2006a), nonparametric bootstrap methods were proposed that require no distributional assumptions for the random terms to obtain bias-corrected estimators. Datta et al. (2002) investigated from the Bayesian perspective the prediction interval of the response and proposed bias-corrected intervals. However, it still seems to be an open field requiring future research. Finally we should also mention that, as mean level prediction is often of interest in small area statistics, the use of robust methods is pretty well motivated. For a recent work see Sinha and Rao (2009).

2.7 References

Adebayo, S. and L. Fahrmeir (2005). Analysing child mortality in nigeria with geoadditive discrete-time survival models. *Statistics in medicine 24*(5), 709–728.

Agresti, A., J. Booth, J. Hobert, and B. Caffo (2000). Random-effects modeling of categorical response data. *Sociological Methodology 30*(1), 27–80.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics 55*(1), 117–128.

Alfò, M. and A. Maruotti (2010). Two-part regression models for longitudinal zero-inflated count data. *Canadian Journal of Statistics 38*(2), 197–216.

Antonio, K. and J. Beirlant (2007). Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics 40*(1), 58–76.

Battese, G., R. Harter, and W. Fuller (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association 83*(401), 28–36.

Booth, J. and J. Hobert (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association 93*(441), 262–272.

Booth, J. and J. Hobert (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(1), 265–285.

Breslow, N. and D. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*(421), 9–25.

Breslow, N. and X. Lin (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika 82*(1), 81–91.

Butler, S. and T. Louis (1992). Random effects models with non-parametric priors. *Statistics in Medicine 11*(14-15), 1981–2000.

Celeux, G., O. Martin, and C. Lavergne (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling 5*(3), 243–267.

Chambers, R. and A. Dorfman (2003). Transformed variables in survey sampling. Soughampton Statistical Sciences Research Institute Methodology Working Paper M03/21, University of Soughampton.

Chambers, R. and N. Tzavidis (2006). M-quantile models for small area estimation. *Biometrika 93*(2), 255–268.

Chi, E. and G. Reinsel (1989). Models for longitudinal data with random effects and ar (1) errors. *Journal of the American Statistical Association 84*(406), 452–459.

Claeskens, G. and J. Hart (2009). Goodness-of-fit tests in mixed models. *Test 18*(2), 213–239.

Datta, G., M. Ghosh, D. Smith, and P. Lahiri (2002). On an asymptotic theory of conditional and unconditional coverage probabilities of empirical bayes confidence intervals. *Scandinavian journal of statistics 29*(1), 139–152.

Datta, G. and P. Lahiri (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica 10*(2), 613–628.

Datta, G. S. (2009). Model-based approach to small area estimation. In C. Rao (Ed.), *Handbook of Statistics Sample Surveys: Inference and Analysis*, Volume 29, Part B of *Handbook of Statistics*, pp. 251 – 288. Elsevier.

Diggle, P., P. Heagerty, K. Liang, and S. Zeger (2002). *Analysis of Longitudinal Data*. Oxford Statistical Science Series. Oxford University Press.

Diggle, P., K. Liang, and S. Zeger (1994). *Analysis of longitudinal data*. Oxford statistical science series. Clarendon Press.

Fahrmeir, L. and S. Lang (2001). Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 50*(2), 201–220.

Falorsi, P., S. Falorsi, and A. Russo (2004). Linear mixed model with correlated area time effects in small area estimation. *Atti Della XLII Riunione Scientifica SIS, Bari*, 9–11.

Fay, R. and R. Herriot (1979). Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association 74*(366), 269–277.

Fuller, W. and G. Battese (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical association 68*(343), 626–632.

Ghidey, W., E. Lesaffre, and P. Eilers (2004). Smooth random effects distribution in a linear mixed model. *Biometrics 60*(4), 945–953.

Ghosh, M., N. Nangia, and D. Kim (1996). Estimation of median income of four-person families: a bayesian time series approach. *Journal of the American Statistical Association 91*(436), 1423–1431.

Ghosh, M., K. Natarajan, T. Stroud, and B. Carlin (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association 93*(441), 273–282.

Ghosh, M. and J. Rao (1994). Small area estimation: an appraisal. *Statistical science 9*(1), 55–76.

Ghosh, M., K. Sinha, and D. Kim (2006). Empirical and hierarchical bayesian estimation in finite population sampling under structural measurement error models. *Scandinavian journal of statistics 33*(3), 591–608.

González-Manteiga, W., M. Lombardía, M. Martinez-Miranda, and S. Sperlich (2012). Kernel smoothers and bootstrapping for nonparametric mixed-effect models. Revised and resubmitted to the Journal of Multivariate Analysis.

González-Manteiga, W., M. Lombardía, I. Molina, D. Morales, and L. Santamaría (2010). Small area estimation under fay–herriot models with non-parametric estimation of heteroscedasticity. *Statistical Modelling 10*(2), 215–239.

Gu, C. and P. Ma (2005). Optimal smoothing in nonparametric mixed-effect models. *The Annals of Statistics 33*(3), 1357–1379.

Gurrin, L., K. Scurrah, and M. Hazelton (2005). Tutorial in biostatistics: spline smoothing with linear mixed models. *Statistics in Medicine 24*(21), 3361–3381.

Hall, D. (2000). Zero-inflated poisson and binomial regression with random effects: a case study. *Biometrics 56*(4), 1030–1039.

Hall, D. and L. Wang (2005). Two-component mixtures of generalized linear mixed effects models for cluster correlated data. *Statistical Modelling 5*(1), 21–37.

Hall, P. and T. Maiti (2006a). Nonparametric estimation of mean-squared prediction error in nested-error regression models. *The Annals of Statistics 34*(4), 1733–1750.

Hall, P. and T. Maiti (2006b). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(2), 221–238.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Series in Statistics. Springer.

Jiang, J. and Z. Ge (2006). Mixed models: An overview. In P. Bickel and J. Fan (Eds.), *Frontiers in statistics: dedicated to Peter John Bickel in honor of his 65th birthday*, pp. 445 – 465. World Scientific Pub Co Inc.

Jiang, J. and P. Lahiri (2006). Mixed model prediction and small area estimation. *Test 15*(1), 1–96.

Jones, R. (1990). Serial correlation or random subject effects? *Communications in statistics: Simulation and computation 19*(3-4), 1105.

Kang, E., D. Liu, and N. Cressie (2009). Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models. *Computational Statistics & Data Analysis 53*(8), 3016–3032.

Kneib, T. and L. Fahrmeir (2006). Structured additive regression for categorical space–time data: A mixed model approach. *Biometrics 62*(1), 109–118.

Laird, N. and J. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*(4), 963–974.

Lee, Y. and J. Nelder (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(4), 619–678.

Li, E., D. Zhang, and M. Davidian (2004). Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal measurements. *Biometrics 60*(1), 1–7.

Li, Q. and T. Stengos (1994). Adaptive estimation in the panel data error component model with heteroskedasticity of unknown form. *International Economic Review 35*(4), 981–1000.

Lin, X. and N. Breslow (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association 91*(435), 1007–1016.

Lin, X. and R. Carroll (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 69–88.

Lin, X., J. Raz, and S. Harlow (1997). Linear mixed models with heterogeneous within-cluster variances. *Biometrics 53*(3), 910–923.

Liu, Q. and D. Pierce (1994). A note on gauss-hermite quadrature. *Biometrika 81*(3), 624–629.

Lombardía, M. and S. Sperlich (2008). Semiparametric inference in generalized mixed effects models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 913–930.

Lombardía, M. and S. Sperlich (2012). A new class of semi-mixed effects models and its application in small area estimation. *Computational Statistics & Data Analysis 56*(10).

Longford, N. (2005). *Missing data and small-area estimation: modern analytical equipment for the survey statistician*. Springer Verlag.

Longford, N. (2010). Small area estimation with spatial similarity. *Computational Statistics & Data Analysis 54*(4), 1151–1166.

Magder, L. and S. Zeger (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of gaussians. *Journal of the American Statistical Association 91*(435), 1141–1151.

Malec, D., J. Sedransk, C. Moriarity, and F. LeClere (1997). Small area inference for binary variables in the national health interview survey. *Journal of the American Statistical Association 92*(439), 815–826.

McCulloch, C. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association 89*(425), 330–335.

McCulloch, C. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association 92*(437), 162–170.

McCulloch, C. and S. Searle (2001). *Generalized, Linear, and Mixed Models*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. John Wiley & Sons.

McNeil, A. and J. Wendin (2003). Generalized linear mixed models in portfolio credit risk modelling. ETH, Eidgenössische Technische Hochschule Zürich, Department Mathematik.

Mcnicholas, P. and T. Murphy (2008). Parsimonious gaussian mixture models. *Statistics and Computing 18*(3), 285–296.

Molenberghs, G. and G. Verbeke (2005). *Models for Discrete Longitudinal Data*. Springer Series in Statistics. Springer.

Neuhaus, J., W. Hauck, and J. Kalbfleisch (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika 79*(4), 755–762.

Newton, M. and Y. Zhang (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika 86*(1), 15–26.

Ng, S., G. McLachlan, K. Wang, L. Jones, and S. Ng (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics 22*(14), 1745–1752.

Olsen, M. and J. Schafer (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association 96*(454), 730–745.

Opsomer, J., G. Claeskens, M. Ranalli, G. Kauermann, and F. Breidt (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 265–286.

Park, J. and H. Wu (2006). Backfitting and local likelihood methods for nonparametric mixed-effects models with longitudinal data. *Journal of statistical planning and inference 136*(11), 3760–3782.

Pfeffermann, D. (2002). Small area estimation-new developments and directions. *International Statistical Review 70*(1), 125–143.

Pfeffermann, D. and C. Barnard (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics 9*(1), 73–84.

Prasad, N. and J. Rao (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association 85*(409), 163–171.

Rao, J. (1999). Some recent advances in model-based small area estimation. *Survey Methodology 25*(2), 175–186.

Rao, J. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology. John Wiley.

Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press.

Saei, A. and R. Chambers (2003). Small area estimation under linear and generalized linear mixed models with time and area effects. Soughampton Statistical Sciences Research Institute Methodology Working Paper M03/15, University of Soughampton.

Saei, A. and R. Chambers (2005a). Empirical best linear unbiased prediction for out of sample areas. Soughampton Statistical Sciences Research Institute Methodology Working Paper M05/03, University of Soughampton.

Saei, A. and R. Chambers (2005b). Out of sample estimation for small areas using area level data. Soughampton Statistical Sciences Research Institute Methodology Working Paper M05/11, University of Soughampton.

Salvati, N. (2004). Small area estimation by spatial models: the spatial empirical best linear unbiased prediction (spatial eblup). Working Paper, Dipartimento di Statistica "G. Parenti", Firenze, 2004/04.

Searle, S., G. Casella, and C. McCulloch (1992). *Variance components*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Sinha, S. and J. Rao (2009). Robust small area estimation. *Canadian Journal of Statistics 37*(3), 381–399.

Song, P., Y. Fan, and J. Kalbfleisch (2005). Maximization by parts in likelihood inference. *Journal of the American Statistical Association 100*(472), 1145–1158.

Sperlich, S. and M. Lombardía (2010). Local polynomial inference for small area statistics: estimation, validation and prediction. *Journal of Nonparametric Statistics 22*(5), 633–648.

Stukel, D. and J. Rao (1997). Estimation of regression models with nested error structure and unequal error variances under two and three stage cluster sampling. *Statistics & probability letters 35*(4), 401–407.

Tao, H., M. Palta, B. Yandell, and M. Newton (1999). An estimation method for the semiparametric mixed effects model. *Biometrics 55*(1), 102–110.

Torabi, M., G. Datta, and J. Rao (2009). Empirical bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates. *Scandinavian Journal of Statistics 36*(2), 355–369.

Ugarte, M., T. Goicoa, A. Militino, and M. Durbán (2009). Spline smoothing in small area trend estimation and forecasting. *Computational Statistics & Data Analysis 53*(10), 3616–3629.

Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association 91*(433), 217–221.

Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics 18*(2), 223–250.

Watier, L., S. Richardson, and P. Green (1999). Using gaussian mixtures with unknown number of components for mixed model estimation. In *14th International Workshop on Statistical Modelling, Graz, Austria*. Citeseer.

Wu, H. and J. Zhang (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association 97*(459), 883–897.

Wu, H. and J. Zhang (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience.

Zeger, S. and M. Karim (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American statistical association 86*(413), 79–86.

Zhang, D. and M. Davidian (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics 57*(3), 795–802.

Zhang, D., X. Lin, J. Raz, and M. Sowers (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association 93*(442), 710–719.

3 Partially Linear Mixed Effects Model

Partially Linear Mixed Effects Model without Distributional Assumptions

Ren Ohinata

Institut für Statistik und Ökonometrie, Georg-August Universität Göttingen

Abstract

An iterative estimation procedure is introduced for partially linear mixed effects models. In contrast to existing likelihood-based methods, no distributional assumptions are made for the random terms. The variance of the regression error can be heteroskedastic and is modeled either parametrically or nonparametrically. Efficient estimation of the parametric component is achieved along the Speckman (1988) approach. Estimation of the nonparametric component is kernel based. For bandwidth selection, cross validation methods are provided one of which should be more appropriate for correlated data. Bootstrap is used for subsequent inference. The derivation of different variance components estimators and fast implementation are discussed in detail. The numerical performance is studied by simulation, and the usefulness of the model and estimation procedure proposed is illustrated in two practical applications.

Key words: Mixed effects models; Partial linear models; Small area statistics; Longitudinal data; Repeated Measurement; Semiparametric regression

3.1 Introduction

In this essay we consider a partially linear mixed effects model (PLMM), specifically the partially linear random intercept (also called nested error) model. The model incorporates two classes of regression models: one is the parametric linear mixed effects model (LMM), typically used for analysis of hierarchical, clustered data; the other is the nonparametric regression model. As Stone (1985) stated, statistical models have three fundamental aspects: flexibility, dimensionality and interpretability. The LMM, specified parametrically, offers clear and practical interpretability. Modeling with random effects serves to capture heterogeneity in population which is not explained by regressors. There have been developments in statistical testing to support interpretation of the LMM. On the other hand, as is the case for any parametric regression model, the LMM requires prior knowledge about the functional form of the regression. However, the precise functional form is unlikely to be known a priori. Parametric functions with a finite number of parameters often lack in flexibility to depict the true functional relationship. A misspecified regression will suffer loss of efficiency, and coefficient estimators may incur inconsistency due to omitted-variable bias unless the regressors are uncorrelated. In reality, except for experimental data, a data design with uncorrelated regressors will seldom be available to empirical researchers. As for the functional form, Yatchew (1998) pointed out, "... most implications of economic theory are nonparametric. Typically, theoretical arguments exclude or include variables, they imply monotonicity, concavity, or homogeneity of various sorts, or they embody more complex structure such as the implications of the maximization hypothesis. They almost *never* imply a specific functional form (the pure quantity theory of money equation being one exception)".[1]

Interpretability of the nonparametric regression may be limited in comparison with the parametric regression. Nonetheless, it has an advantage of flexibility well beyond that of a parametric regression. Being free of a prior assumption about the functional form, nonparametric regression circumvents functional misspecification. However, nonparametric estimation is prone to incur imprecision of estimation which is well-known in the nonparametric estimation literature as the "curse of dimensionality". Suppose a nonparametric regression $y_i = \xi(\mathbf{w}_i) + e_i$ $(i = 1, \dots N)$, where ξ is an unknown smooth function and the error e is independent and identically distributed (i.i.d.) with zero conditional mean and finite variance. As Stone (1980) showed, the asymptotic mean squared error of a nonparametric estimator for $\xi(\mathbf{w}_i)$ is of order $O(N^{-4/(4+d)})$ if ξ is twice continuously differentiable with d denoting the dimension of the regressors $\mathbf{w}_i$. This implies firstly that the convergence of the nonparametric

[1] Italic shape is as in the original text.

regression estimator is slower than that of the parametric regression of order $O(N^{-1})$; and secondly that the higher the dimension of $\mathbf{w}_i$, the larger sample size is required to maintain the same precision of estimation. In practice, the number of variables of interest is often too large for nonparametric regression to yield estimation of reasonable precision.

One way of alleviating this problem is the use of a semiparametric model, for example, partially linear model. The nonparametric regression above can be re-specified semiparametrically as $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \gamma(\mathbf{t}_i) + e_i$ where $(\mathbf{x}_i^\top \ \mathbf{t}_i^\top) = \mathbf{w}_i^\top$ so that the dimension of $\mathbf{t}_i$ is low enough to evade the curse of dimensionality. Such a model makes sense from a practical point of view as well. For example, the researcher is interested in not all regressors but only some of them and interpretability is important only for those variables of interest. In addition, prior knowledge of the functional form may be available for some variables. It will be natural to model such variables parametrically. To the contrary, nonparametric specification will suit variables of little interest or with insufficient prior knowledge about their functional form. Semiparametric models can also be used as a means of preliminary, exploratory analysis before a fully parametric model is constructed.

Extensions of the LMM to nonparametric modeling has already been studied by many authors. For recent reviews, see for example Su and Ullah (2010) and Ohinata and Sperlich (2012). Among others, a computationally attractive alternative approach is implementation using spline smoothing techniques based on distributional assumptions; see Ruppert et al. (2003) and Opsomer et al. (2008). What all the semiparametric approaches have in common is a concern about possible misspecification of the functional form with consequently invalid inference; mixed effects model inference is always too heavily model based to claim model unbiasedness.

In this essay we discuss an extension of the partially linear model proposed by Speckman (1988) and Robinson (1988), so that the error term assumes a structure typical for the random intercept model. For the standard LMM, the random terms are often assumed to be jointly normally distributed, which provides technical convenience for maximum likelihood estimation. In contrast, we estimate the PLMM relying on the method of moments. In estimation we encounter three complications inherent in a model combining the nonparametric and mixed effects regressions. The first complication arises from the estimation of the variance components (VCs) in the semiparametric regression framework. While Speckman (1988) and Robinson (1988) discussed the partially linear model given independent data, the LMM, constructed for correlated data, is conventionally fit using generalized least squares (GLS) with the VCs. In the PLMM framework, the VC estimation needs to account for loss of

the degrees of freedom (d.f.) due to not only the estimation of the parametric function but also the nonparametric one. Neglecting the loss of the d.f. causes bias in the VC estimation in small samples. This may in turn have an influence on the GLS estimator for the parametric function in the model. To appropriately account for the d.f. in the VC estimation, we propose an iterative estimation procedure in which the VC estimator is corrected in successive rounds of iteration. Iterative procedures also serve to deal with one of the central issues discussed in Speckman (1988) and Robinson (1988), that is, key bandwidth selection required for their estimation procedures. We observed in simulation studies that without iterative estimation some estimators are quite seriously affected by bandwidth selection, and that their effects diminish through iteration process.

Secondly, while the PLMM is to analyze correlated data, conventional bandwidth selection methods for nonparametric regression estimation are constructed for independent data. They will therefore choose too small bandwidths without taking correlation structure into account, which results in undersmoothing of the nonparametric function estimate. We approach this problem with the proposal by Carmack et al. (2011), which extends the generalized cross validation.

Thirdly, asymptotic distributions of the regression coefficient estimators are difficult to derive analytically. Even in the standard LMM framework, the asymptotic distribution of the coefficient estimators relies on complicated approximations when the VCs need to be estimated. In addition, difficulty also arises from the fact that the effective d.f. of correlated data are not clearly defined. We turn to the bootstrap resampling method to provide approximate sampling distributions of coefficient estimators.

A practical concern about the use of a non-/semiparametric estimation is the computability of estimation algorithm and the availability of tailored software. With today's increasing computing power, the former is in general a relatively small concern. Nonetheless, given a large data set, some calculation in nonparametric estimation turns out prohibitively computer-intensive. For the purpose of practical implementation, we employ binning techniques to clear computational hurdles. As for the latter concern, we provide a program package `plmm` in the statistical software R for the statistical inference discussed in this essay.

The structure of the essay is as follows. Section 3.2 provides the model specification and the overview of the estimation procedure. Section 3.3 describes the estimation of the parametric and nonparametric functions of the model. The VC estimators are presented in Section 3.4 where we consider homoskedastic as well as heteroskedastic regression errors. Since the functional form of the conditional heteroskedastic

variance is unlikely to be known in practice, its nonparametric estimation is also discussed. Section 3.5 provides the random intercept predictor, followed by Section 3.6 where we discuss the testing of regression coefficients based on bootstrap resampling. In Section 3.7 we present some simulation results. In Section 3.8 the PLMM is illustrated with two empirical examples. Conclusions and further perspectives are provided in Section 3.9. Derivations omitted from the main text are collected in Appendix A.

3.2 Model Specification and Estimation Procedure

3.2.1 Model specification

We consider the random intercept model for data structured hierarchically over two levels: cluster level (denoted by subscript i) and individual observational level (subscript j). Allowing for heteroskedasticity, the model is generally specified as

$$y_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \gamma(\mathbf{t}_{ij}) + v_{ij} \ (i = 1, \ldots, m;\ j = 1, \ldots, n_i;\ N = \sum_{i=1}^{m} n_i) \quad (3.1)$$

$$v_{ij} = u_i + \alpha_{ij} e_{ij} \ , \quad (3.2)$$

where y_{ij} is a continuous response variable; $\mathbf{x}_{ij}$ is a p-dimensional vector of continuous or discrete regressors; $\mathbf{t}_{ij}$ is a d-dimensional vector of continuous regressors; u_i is the random intercept of cluster i and e_{ij} is the regression error. The number of observations in the ith cluster is denoted by n_i. Data can be unbalanced, that is, the number of observations in a cluster may vary from cluster to cluster. The fixed component consists of two subcomponents: parametric $\mathbf{x}_{ij}^\top \boldsymbol{\beta}$ and nonparametric $\gamma(\mathbf{t}_{ij})$.[2] For identification, $\mathbf{x}_{ij}$ does not contain one for the intercept. $\boldsymbol{\beta}$ is a vector of regression coefficients and γ is a smooth, at least twice continuously differentiable function. α_{ij} is a positive nonzero constant that determines heteroskedasticity in the variance of the regression error. The model with homoskedastic regression errors is a special case of (3.1) with $\alpha_{ij} = 1$. Another special case is the between-cluster heteroskedastic model which is specified with $\alpha_{ij} = \alpha_i$.

Model (3.1) can be equivalently given in stacked forms:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\gamma}(\mathbf{T}_i) + u_i \mathbf{1}_{n_i} + (\oplus_j \alpha_{ij}) \mathbf{e}_i \quad (3.3)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}(\mathbf{T}) + \mathbf{Z}\mathbf{u} + (\oplus_{ij} \alpha_{ij}) \mathbf{e} \ , \quad (3.4)$$

[2] In this essay the term "fixed" is used as in the traditional mixed effects model literature, not as in the panel data analysis literature. Thus, in equation (3.1) there are "fixed" components $\mathbf{x}_{ij}^\top \boldsymbol{\beta}$ and $\gamma(\mathbf{t}_{ij})$ and "random" components u_i and $\alpha_{ij} e_{ij}$ in v_{ij}.

where $\mathbf{y}_i = (y_{i1} \dots y_{in_i})^\top$ is stacked in $\mathbf{y} = (\mathbf{y}_1^\top \dots \mathbf{y}_m^\top)^\top$; regressors $\mathbf{X}_i = (\mathbf{x}_{i1} \dots \mathbf{x}_{in_i})^\top$ packed in matrix $\mathbf{X} = (\mathbf{X}_1^\top \dots \mathbf{X}_m^\top)^\top$ and also $\mathbf{T}_i = (\mathbf{t}_1 \dots \mathbf{t}_{n_i})^\top$ in matrix $\mathbf{T} = (\mathbf{T}_1^\top \dots \mathbf{T}_m^\top)^\top$; random effects $\mathbf{u} = (u_1 \dots u_m)^\top$; and random errors $\mathbf{e}_i = (e_{i1} \dots e_{in_i})^\top$ stacked in $\mathbf{e} = (\mathbf{e}_1^\top \dots \mathbf{e}_m^\top)^\top$; $\mathbf{1}_{n_i}$ is a n_i-dimensional vector of ones, and $\mathbf{Z} = \mathbf{I}_m \otimes \mathbf{1}_{n_i}$. Here, $\otimes$ denotes the Kronecker product[3] and $\oplus$ the Kronecker sum. Throughout the essay, when appropriate, we present models using these three forms of equations at discretion.

It is assumed that $\mathbf{X}_i$ and $\mathbf{T}_i$ have full column ranks of p and d, respectively, and that no column of $\mathbf{X}$ is a linear combination of the columns of $\mathbf{Z}$. Variables in $\mathbf{x}_{ij}$ are not necessarily independent of variables in $\mathbf{t}_{ij}$, which implies that a misspecified parametric function of $\mathbf{t}_{ij}$ may cause serious bias in the estimator of $\boldsymbol{\beta}$. For the random terms u_i and e_{ij}, we only assume zero conditional means $\mathrm{E}[u_i|\mathbf{X}_i, \mathbf{T}_i] = 0$ and $\mathrm{E}[e_{ij}|\mathbf{X}_i, \mathbf{T}_i] = 0$; and finite variances $(\sigma_u^2, \sigma_e^2) = \{\mathbb{R}^2 | \sigma_u^2 \geq 0, \sigma_e^2 > 0\}$ where $\sigma_u^2 = \mathrm{Var}[u_i]$ and $\sigma_e^2 = \mathrm{Var}[e_{ij}]$ (conventionally called variance components).[4] It is further assumed that the random intercept u_i is independent of those of other clusters (between-cluster independence) and also independent of regression errors e_{ij} $(\forall i, j)$; and that the regression errors are independent.[5] For the homoskedastic model these assumptions are summarized in matrix form as follows:

$$\mathrm{Var}\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{pmatrix} \sigma_u^2 \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 \mathbf{I}_N \end{pmatrix}, \tag{3.5}$$

where $\mathbf{I}$ is an identity matrix. Dependence between y_{ij} and $y_{ij'}$ for $j \neq j'$ is caused through the random intercept u_i they share. Let $\mathbf{v}_i = u_i \mathbf{1}_{n_i} + \mathbf{e}_i$ be stacked in $\mathbf{v} = (\mathbf{v}_1^\top \dots \mathbf{v}_m^\top)^\top$. Then $\mathrm{Var}[\mathbf{v}_i]$ denoted by $\mathbf{V}_i$ and $\mathrm{Var}[\mathbf{v}]$ by $\mathbf{V}$ are given by

$$\mathbf{V}_i = \sigma_u^2 \mathbf{J}_{n_i} + \sigma_e^2 \mathbf{I}_{n_i} \tag{3.6}$$

$$\mathbf{V} = \oplus_{i=1}^m \mathbf{V}_i , \tag{3.7}$$

where $\mathbf{J}_{n_i}$ is a matrix of ones with dimension $n_i \times n_i$. Let $\mathbf{D}$ denote $\mathrm{Var}[\mathbf{u}]$. $\mathbf{V}$ and $\mathbf{D}$ are block diagonal matrices and assumed to be positive definite and positive

[3] By abuse of the notation for the Kronecker product, we denote a block diagonal matrix whose block elements are $\{\mathbf{B}_1, \dots, \mathbf{B}_m\}$ by $\mathbf{I}_m \otimes_i \mathbf{B}_i$ where $\mathbf{B}_i$ is a matrix of an arbitrary dimension provided matrix operation permits.

[4] The estimation presented in this essay does not rely on a distributional assumption such as normality of the random terms. Nor is assumed that the random terms are symmetrically distributed, which is required for an LMM to yield "unbiased" response predictions in small samples as shown by Kackar and Harville (1981). "unbiased" is in the sense that the expectation of the response prediction is equal to the expectation of the response.

[5] Note that consistency of the generalized least squares (GLS) estimator for $\boldsymbol{\beta}$ requires $\mathbf{X}_i$ and $\mathbf{T}_i$ being strictly exogenous in the sense of $\mathrm{E}[e_{ij}|\mathbf{X}_i, \mathbf{T}_i] = 0$. This implies that $\mathrm{E}[e_{ij}|\mathbf{X}, \mathbf{T}] = 0$ under between-cluster independence assumption.

semi-definite, respectively. Extensions to heteroskedasticity are discussed in Section 3.4.

3.2.2 General estimation procedure

Our estimation procedure is motivated by a method proposed by Speckman (1988) and Robinson (1988) for the partially linear model $y_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \gamma(\mathbf{t}_{ij}) + e_{ij}$ where e_{ij} are independent homoskedastic errors. We apply their proposal to (3.1) with $\alpha_{ij} = 1$. The first step of estimation is conceptually similar to the partial correlation coefficient estimation in the classical regression. Suppose there exist finite conditional expectation functions $\mathrm{E}[y_{ij}|\mathbf{T}_i]$ and $\mathrm{E}[\mathbf{x}_{ij}^\top|\mathbf{T}_i]$. The expectation of (3.1) conditional on $\mathbf{T}_i$ is

$$\mathrm{E}[y_{ij}|\mathbf{T}_i] = \mathrm{E}[\mathbf{x}_{ij}^\top|\mathbf{T}_i]\boldsymbol{\beta} + \gamma(\mathbf{t}_{ij}) \tag{3.8}$$

since it holds that

$$\begin{aligned} \mathrm{E}[\gamma(\mathbf{t}_{ij})|\mathbf{T}_i] &= \gamma(\mathbf{t}_{ij}) && (3.9)\\ \mathrm{E}[u_i|\mathbf{T}_i] &= \mathrm{E}_{\mathbf{X}_i}\mathrm{E}[u_i|\mathbf{X}_i, \mathbf{T}_i] = 0 && (3.10)\\ \mathrm{E}[e_{ij}|\mathbf{T}_i] &= \mathrm{E}_{\mathbf{X}_i}\mathrm{E}[e_{ij}|\mathbf{X}_i, \mathbf{T}_i] = 0 && (3.11) \end{aligned}$$

by the law of iterative expectation and model assumptions. If the conditional expectations $\mathrm{E}[y_{ij}|\mathbf{T}_i]$ and $\mathrm{E}[\mathbf{x}_{ij}^\top|\mathbf{T}_i]$ are known, subtracting (3.8) from (3.1) reduces the PLMM to an LMM

$$y_{ij} - \mathrm{E}[y_{ij}|\mathbf{T}_i] = (\mathbf{x}_{ij}^\top - \mathrm{E}[\mathbf{x}_{ij}^\top|\mathbf{T}_i])\boldsymbol{\beta} + u_i + e_{ij} \ . \tag{3.12}$$

This is a standard random intercept model and $\boldsymbol{\beta}$ can be estimated with GLS. Let $\mathbf{y}_0$ be a vector whose ijth element is $y_{ij} - \mathrm{E}[y_{ij}|\mathbf{T}_i]$ and also let $\mathbf{X}_0$ be a matrix whose ijth row is $\mathbf{x}_{ij}^\top - \mathrm{E}[\mathbf{x}_{ij}^\top|\mathbf{T}_i]$. The GLS estimator of $\boldsymbol{\beta}$ for model (3.12) is given by

$$\tilde{\boldsymbol{\beta}}_{(0)} = (\mathbf{X}_0^\top \mathbf{V}^{-1} \mathbf{X}_0)^{-1} \mathbf{X}_0^\top \mathbf{V}^{-1} \mathbf{y}_0 \ . \tag{3.13}$$

An estimator of the nonparametric component $\tilde{\gamma}_{(0)}(\mathbf{t}_{ij})$ can be obtained by applying kernel regression to $y_{ij} - \mathbf{x}_{ij}^\top \tilde{\boldsymbol{\beta}}_{(0)} = \gamma(\mathbf{t}_{ij}) + u_i + e_{ij}$ as Speckman (1988) suggested.[6]

The VCs (and hence matrix $\mathbf{V}$) as well as conditional mean functions $\mathrm{E}[y_{ij}|\mathbf{T}_i]$ and $\mathrm{E}[\mathbf{x}_{ij}^\top|\mathbf{T}_i]$ are typically unknown in practice. It is natural to replace these unknowns by their estimates. Let $\tilde{y}_{0ij} = y_{ij} - \hat{\mathrm{E}}[y_{ij}|\mathbf{T}_i]$ and $\tilde{\mathbf{x}}_{0ij} = \mathbf{x}_{ij}^\top - \hat{\mathrm{E}}[\mathbf{x}_{ij}^\top|\mathbf{T}_i]$ where $\hat{\mathrm{E}}[\cdot]$ is

[6] Alternatively, γ can also be estimated by $\tilde{\gamma}_{(0)}(\mathbf{t}_{ij}) = \mathrm{E}[y_{ij}|\mathbf{T}_i] - \mathrm{E}[\mathbf{x}_{ij}^\top|\mathbf{T}_i]\tilde{\boldsymbol{\beta}}_{(0)}$ as Robinson (1988) suggested.

an estimator of the mean function. Model (3.12) is rewritten as an LMM:

$$\tilde{y}_{0ij} = \tilde{\mathbf{x}}_{0ij}^{\top}\boldsymbol{\beta} + u_i + e_{ij} \ . \tag{3.14}$$

The feasible GLS estimator is then given by

$$\hat{\boldsymbol{\beta}}_{(0)} = (\tilde{\mathbf{X}}_0^{\top}\hat{\mathbf{V}}_{(0)}^{-1}\tilde{\mathbf{X}}_0)^{-1}\tilde{\mathbf{X}}_0^{\top}\hat{\mathbf{V}}_{(0)}^{-1}\tilde{\mathbf{y}}_0 \ , \tag{3.15}$$

where $\hat{\mathbf{V}}_{(0)}$ is an estimator of $\mathbf{V}$ with the VCs replaced by their estimators. Given $\hat{\boldsymbol{\beta}}_{(0)}$, kernel regression estimator of γ is given by

$$\hat{\boldsymbol{\gamma}}_{(0)}(\mathbf{T}) = \mathbf{S}_0(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{(0)}) \ , \tag{3.16}$$

where $\mathbf{S}_0$ is a smoother matrix which depends on a bandwidth vector $\mathbf{h}$ of length d.[7]

A problem of the GLS estimator (3.15) is that the VC estimators in $\hat{\mathbf{V}}_{(0)}$ were calculated without taking into account the loss of d.f. due to the estimation of γ by (3.16). The VC estimators are therefore biased (though not asymptotically). This bias will affect the GLS estimation of $\boldsymbol{\beta}$ and the succeeding estimation of γ involving the GLS estimator. The bias in the VC estimators can be of nontrivial size, at least in small samples. As a remedy for this problem, we propose an iterative estimation procedure to correct the VC estimators for bias. The iterative estimation consists of two stages: the initial stage symbolized by subscript (0) and an succeeding iterative stage.

In the initial stage, $\hat{\boldsymbol{\beta}}_{(0)}$ and $\hat{\boldsymbol{\gamma}}_{(0)}$ are first estimated and then residuals $\hat{v}_{(0)ij} = y_{ij} - \mathbf{x}_{ij}^{\top}\hat{\boldsymbol{\beta}}_{(0)} - \hat{\boldsymbol{\gamma}}_{(0)}(\mathbf{t}_{ij})$ are calculated. The rth iteration ($r = 1, 2 \ldots$) in the iteration stage, symbolized by subscript (r), proceeds as follows.

1. Estimate the VCs $\sigma^2_{u(r)}$ and $\sigma^2_{e(r)}$ for $\mathbf{V}_{(r)}$ from the residuals $\hat{v}_{(r-1)ij}$ whereby the d.f. of the estimator $\hat{\gamma}_{(r-1)}$ is accounted for. When VC estimation has converged, the iterative procedure ends without going through the following steps.

2. Estimate $\boldsymbol{\beta}_{(r)}$ by feasible GLS:

$$\hat{\boldsymbol{\beta}}_{(r)} = (\mathbf{X}^{\top}\hat{\mathbf{V}}_{(r)}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\top}\hat{\mathbf{V}}_{(r)}^{-1}(\mathbf{y} - \hat{\boldsymbol{\gamma}}_{(r-1)}(\mathbf{T})) \tag{3.17}$$

3. Estimate $\boldsymbol{\gamma}_{(r)}$ by $\hat{\boldsymbol{\gamma}}_{(r)}(\mathbf{T}) = \mathbf{S}_r(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{(r)})$ with a smoother matrix $\mathbf{S}_r$ described in Section 3.3.2.

[7] Smoother matrix $\mathbf{S}$ also depends on the kernel function. However, since the choice of the kernel is known to have little influence on estimation, we use the Gaussian kernel throughout this essay and confine our discussion to bandwidths.

4. Obtain residuals $\hat{v}_{(r)ij} = y_{ij} - \mathbf{x}_{ij}^{\top}\hat{\boldsymbol{\beta}}_{(r)} - \hat{\boldsymbol{\gamma}}_{(r)}(\mathbf{t}_{ij})$.

5. Iterate the steps from 1. to 4. until the convergence of the VC estimates.

In the heteroskedastic regression error case, the conditional error variance function is nonparametrically estimated from $\hat{v}_{ij}$ of the last iteration. $\boldsymbol{\beta}$ is then reestimated by semiparametric GLS with $\hat{\mathbf{V}}$ constructed from the estimated variance function. Finally, function $\boldsymbol{\gamma}$ is reestimated by $\hat{\boldsymbol{\gamma}}(\mathbf{T}) = \mathbf{S}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ with a smoother matrix $\mathbf{S}$ given in Section 3.3.2 and the semiparametric GLS estimator $\hat{\boldsymbol{\beta}}$.

We emphasize that our iterative estimation starts with initial feasible GLS $\boldsymbol{\beta}$ estimation as in (3.15), not OLS estimation. Li and Ullha (1998) studied a two-step estimation in which OLS $\boldsymbol{\beta}$ estimation is followed by GLS estimation; You et al. (2010) proposed a similar two-step approach with GLS $\boldsymbol{\beta}$ estimation in the second step using a nonparametrically estimated variance function. We demonstrate by simulation (Section 3.7) and application to real data (Section 3.8.2) that our algorithm improves the efficiency of the $\boldsymbol{\beta}$ estimator.

3.3 Estimation of the Fixed Components

3.3.1 Estimation of the parametric component

The feasible GLS estimators (3.15) and (3.17) require taking the inverse of matrix $\hat{\mathbf{V}}$. Direct calculation of the inverse involves computational difficulties when the dimension N is prohibitively large. However, since $\mathbf{V}$ is a block diagonal matrix, inversion can be taken blockwise, assuming each cluster's size n_i is adequately small. For the random intercept model with homoskedastic errors, Fuller and Battese (1973) proposed an OLS estimation on transformed data. We apply their transformation to the PLMM.

Suppose the VC and function γ are known. Premultiplying both sides of the regression $\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\gamma}_i(\mathbf{T}_i) + \mathbf{v}_i$ by

$$\sigma_e \mathbf{V}_i^{-1/2} = \mathbf{I}_{n_i} - \left(1 - \left(\frac{\sigma_e^2}{\sigma_e^2 + n_i\sigma_u^2}\right)^{\frac{1}{2}}\right)\bar{\mathbf{J}}_{n_i} \tag{3.18}$$

yields a new regression with an error covariance matrix $\sigma_e^2\mathbf{I}_{n_i}$. Stukel and Rao (1997) similarly proposed a transformation for the heteroskedastic error case. Using their transformation, regression $(\oplus\alpha_{ij}^{-1})\mathbf{y}_i = (\oplus\alpha_{ij}^{-1})(\mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\gamma}_i(\mathbf{T}_i) + \mathbf{v}_i)$ is premuliplied

by

$$\sigma_e(\text{Var}[(\oplus\alpha_{ij}^{-1})\mathbf{y}_i])^{-1/2} = \mathbf{I}_{n_i} - \left(1 - \left(\frac{\sigma_e^2}{\sigma_e^2 + \eta_i\sigma_u^2}\right)^{\frac{1}{2}}\right)\eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top \ , \qquad (3.19)$$

where $\boldsymbol{\alpha}_i^{-1} = (\alpha_{i1}^{-1} \dots \alpha_{in_i}^{-1})^\top$, $\eta_i = (\boldsymbol{\alpha}_i^{-1})^\top\boldsymbol{\alpha}_i^{-1} = \sum_{j=1}^{n_i}\alpha_{ij}^{-2}$, and heteroskedasticity parameter α_{ij} is assumed to be known. The new transformed regression also has an error covariance structure $\sigma_e^2\mathbf{I}_{n_i}$. The transformed regressions are both fit by OLS. Derivation of (3.18) and (3.19) is given in Section 3.10.1. The above OLS estimators can be equivalently expressed as a GLS estimator

$$\tilde{\boldsymbol{\beta}} \ = \ (\sum_i \mathbf{X}_i^\top\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\sum_i \mathbf{X}_i^\top\mathbf{V}_i^{-1}\left(\mathbf{y}_i - \boldsymbol{\gamma}_i(\mathbf{T}_i)\right) \qquad (3.20)$$

in the homoskedastic case and

$$\tilde{\boldsymbol{\beta}} = (\sum_i \mathbf{X}_i^\top \oplus_j \alpha_{ij}^{-1}\mathbf{V}_i^{-1} \oplus_j \alpha_{ij}^{-1}\mathbf{X}_i)^{-1}\sum_i \mathbf{X}_i^\top \oplus_j \alpha_{ij}^{-1}\mathbf{V}_i^{-1} \oplus_j \alpha_{ij}^{-1}\left(\mathbf{y}_i - \boldsymbol{\gamma}_i(\mathbf{T}_i)\right) \quad (3.21)$$

in the heteroskedastic case.

In practice, the VCs, $\boldsymbol{\gamma}$ and heteroskedasticity parameter $\boldsymbol{\alpha}$ are usually unknown; the feasible GLS estimators of (3.20) and (3.21) are obtained by replacing those unknowns with their estimators. In the heteroskedastic case, the diagonal elements of $\mathbf{V}$ are determined by the conditional variance function $\text{Var}[\alpha_{ij}e_{ij}|\mathbf{w}_{ij}]$ where $\mathbf{w}_{ij}$ is a vector of conditioning variables. This unknown variance function needs to be estimated either parametrically or nonparametrically. Section 3.4.2 presents a nonparametric estimation proposed by Li and Stengos (1994). $\hat{\mathbf{V}}_i^{-1}$ required for the feasible GLS estimator (3.21) is given by

$$\hat{\mathbf{V}}_i^{-1} \ = \ \oplus\hat{\nu}_{ij}^{-2} - \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2\sum\hat{\nu}_{ij}^{-2} + 1}(\hat{\nu}_{i1}^{-2}, \dots, \hat{\nu}_{in_i}^{-2})^\top(\hat{\nu}_{i1}^{-2}, \dots, \hat{\nu}_{in_i}^{-2}) \ , \qquad (3.22)$$

where $\hat{\nu}_{ij}^{-2}$ is an estimator of $\nu_{ij}^2 = \text{Var}[\alpha_{ij}e_{ij}|\mathbf{w}_{ij}] = \alpha_{ij}^2\sigma_e^2$ (see Section 3.10.1 for derivation).[8]

Li and Stengos (1994) showed for the random intercept model that the semiparametric feasible GLS estimator $\hat{\boldsymbol{\beta}}$ is $\sqrt{N}$-consistent, asymptotically normally distributed, and adaptive, which means that the estimator of $\boldsymbol{\beta}$ is asymptotically as efficient as the estimator that would be obtained using the parametric estimator of the correctly specified variance function.

[8] Stukel and Rao transformation (3.19) can be equivalently used by setting σ_e^2 to one and α_{ij} to the square root of estimated ν_{ij}^2 without losing generality (recall $\alpha_{ij} > 0$).

3.3.2 Estimation of the nonparametric component

The model estimation we propose requires nonparametric regression estimation before the initial stage. The original PLMM model (3.1) first needs to be reduced to an LMM by plugging estimates of conditional expectations $\mathrm{E}[y_{ij}|\mathbf{t}_{ij}]$ and $\mathrm{E}[\mathbf{x}_{ij}^\top|\mathbf{t}_{ij}]$ in (3.12). We estimate them by kernel regression with a set of bandwidths. We select these bandwidths by cross validation. While kernel regression estimators are consistent, effects of the selected bandwidths on the subsequent estimation are of great interest especially in small samples.[9] In simulation studies presented in Section 3.7, we observed that they indeed affect the subsequent estimation, and that their effects, however, diminish through iterative processing.

In both initial and iteration stages, γ is also estimated by kernel regression. The multivariate local linear estimator is briefly presented below for the case of $d = 2$. Extension for the case of $d > 2$ is analytically straightforward. For ease of presentation, let a new index k denote the original index ij; k runs through $1, 2, \ldots, N$ without altering the order of the original index. Suppose $\boldsymbol{\beta}$ is given, then γ is estimated at point $\mathbf{t} = (t_1\, t_2)^\top$ by

$$\hat{\gamma}\,(\mathbf{t}) \;=\; \mathbf{e}_1^\top (\mathbf{T}_\mathbf{t}^{*\top}\mathbf{K}_\mathbf{t}\mathbf{T}_\mathbf{t}^{*})^{-1}\mathbf{T}_\mathbf{t}^{*\top}\mathbf{K}_\mathbf{t}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\;, \tag{3.24}$$

where $\mathbf{e}_1^\top = (1, 0, 0)$; $\mathbf{T}_\mathbf{t}^*$ is an $(N \times 3)$ matrix $\{1,\, t_{1k} - t_1,\, t_{2k} - t_2\}_{k=1}^N$; and $\mathbf{K}_\mathbf{t}$ is the following multiplicative kernel weighting matrix. For a given set of bandwidths $\mathbf{h} = (h_1\, h_2)^\top$,

$$\mathbf{K}_\mathbf{t} \;=\; \oplus_k K_{1k}K_{2k} \quad \text{with } K_{dk} = K\left(\frac{t_{dk} - t_d}{h_d}\right),\; d = 1, 2\;, \tag{3.25}$$

where K is the Gaussian kernel function. Selection of the bandwidths $\mathbf{h}$ is discussed in the following section. The estimators of γ at N observation points in $\mathbf{T}$ are given by

$$\hat{\boldsymbol{\gamma}}(\mathbf{T}) = \mathbf{S}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\;, \tag{3.26}$$

[9] In general, the expectation functions estimated nonparametrically are biased in small samples. As Robinson (1988) discussed, it holds after the replacement of the unknown functions that

$$y_{ij} - \hat{\mathrm{E}}[y_{ij}|\mathbf{T}_i] = (\mathbf{x}_{ij}^\top - \hat{\mathrm{E}}[\mathbf{x}_{ij}^\top|\mathbf{T}_i])\boldsymbol{\beta} + \gamma(\mathbf{t}_{ij}) - \gamma^*(\mathbf{t}_{ij}) + v_{ij} \tag{3.23}$$

where γ^* is defined as $\hat{\mathrm{E}}[\mathbf{x}_{ij}^\top|\mathbf{T}_i]\boldsymbol{\beta} - \hat{\mathrm{E}}[y_{ij}|\mathbf{T}_i]$. This implies that the estimator of $\boldsymbol{\beta}$ will be biased due to the additional error term $\gamma(\mathbf{t}_{ij}) - \gamma^*(\mathbf{t}_{ij})$ unless consistent estimators $\hat{\mathrm{E}}[y_{ij}|\mathbf{T}_i]$ and $\hat{\mathrm{E}}[\mathbf{x}_{ij}^\top|\mathbf{T}_i]$ converge fast enough. Robinson (1988) used a higher-order kernel to enable fast convergence and showed that under regularity conditions the estimator of $\boldsymbol{\beta}$ achieves $\sqrt{N}$-consistency in spite of the nonparametric component in the model. Speckman (1988) also showed $\sqrt{N}$-consistency of the $\boldsymbol{\beta}$ estimator.

where $\mathbf{S}$ is a linear smoother

$$\mathbf{S} = \{\mathbf{s}_{\mathbf{t}_k}^\top\}_{k=1}^N \tag{3.27}$$

$$\mathbf{s}_{\mathbf{t}_k}^\top = \mathbf{e}_1^\top (\mathbf{T}_{\mathbf{t}_k}^{*\top} \mathbf{K}_{\mathbf{t}_k} \mathbf{T}_{\mathbf{t}_k}^*)^{-1} \mathbf{T}_{\mathbf{t}_k}^{*\top} \mathbf{K}_{\mathbf{t}_k} \ . \tag{3.28}$$

The Nadaraya-Watson (NW) estimator of $\gamma(\mathbf{t})$ and its smoother $\mathbf{S}$ can be obtained by replacing $\mathbf{T}_t^*$ by $\mathbf{1}_N$ in (3.28).

3.3.3 Cross validation and binning

One of the methods widely used for bandwidth selection is the leave-one-out cross validation (CV). For ease of presentation, we drop the fixed parametric component from the original model (3.1) to consider a regression $y_k = \gamma(\mathbf{t}_k) + v_k$ with index k $(k = 1, 2, \ldots N)$.[10] The CV method selects a bandwidth vector $\mathbf{h}$ that minimizes the following statistic

$$\mathrm{CV}(\mathbf{h}) = \frac{1}{N} \sum_{k=1}^N \left(y_k - \hat{\gamma}_{(-k)}(\mathbf{t}_k)\right)^2 = \frac{1}{N} \sum_{k=1}^N \left(\frac{y_k - \hat{\gamma}(\mathbf{t}_k)}{1 - s_{kk}}\right)^2 \ , \tag{3.29}$$

where $\hat{\gamma}_{(-k)}(\mathbf{t}_k)$ is the kernel regression estimate obtained from the data with the kth observation left out and s_{kk} is the kth diagonal element of the smoother matrix $\mathbf{S}$.[11] An alternative bandwidth selection method is the generalized cross validation (GCV) proposed by Craven and Wahba (1979)

$$\mathrm{GCV}(\mathbf{h}) = \frac{1}{N} \sum_{k=1}^N \left(\frac{y_k - \hat{\gamma}(\mathbf{t}_k)}{1 - \mathrm{tr}(\mathbf{S})/N}\right)^2 \ . \tag{3.30}$$

Because the minimum of the residual sum of squares $\sum_{k=1}^N (y_k - \hat{\gamma}(\mathbf{t}_k))^2$ can be achieved by interpolation with small bandwidths, dependence of estimator $\hat{\gamma}(\mathbf{t}_k)$ on y_k would lead to an undersmoothed estimate of γ. While the leave-one-out CV circumvents undersmoothing by leaving out the kth observation, the GCV deals with the dependence using a penalizing function.[12]

A challenge in bandwidth selection for the PLMM arises from correlations in data. Both CV and GCV were originally developed for independent data. Thus they tend

[10] For $d > 1$, in order to to deal with scale differences of the variables in $\mathbf{t}$, we specify one bandwidth value h and set each element of $\mathbf{h}$ by multiplying h with the standard deviation of its corresponding variable. When $d = 2$, for example, $h_1 = h \cdot \mathrm{sd}[t_1]$ and $h_2 = h \cdot \mathrm{sd}[t_2]$.

[11] The second equality holds due to the fact that $\mathbf{S}$ is a linear smoother.

[12] Since function γ is estimated by kernel regression, the GCV defined above is asymptotically equivalent to the CV. See Härdle and Müller (2000).

to select too small bandwidths resulting in undersmoothed estimates of γ. Bandwidth selection for correlated data is an ongoing research field in nonparametric regression estimation. Carmack et al. (2011) proposed an extension of the GCV for correlated data. Their extended GCV is given by

$$\mathrm{GCVc}\,(\mathbf{h}) \;=\; \frac{1}{N}\sum_{k=1}^{N}\left(\frac{y_k-\hat{\gamma}(\mathbf{t}_k)}{1-\mathrm{tr}(2\mathbf{SR}-\mathbf{SRS}^\top)/N}\right)^2 , \tag{3.31}$$

where $\mathbf{R}$ is the correlation matrix of the errors $(v_1, v_2, \ldots, v_N)^\top$ (see Appendix 3.10.5 for more details).

The d.f. of the nonparametric estimator of γ is often defined as $\mathrm{tr}(2\mathbf{S}-\mathbf{SS}^\top)$ (see Hastie and Tibshirani, 1990). Carmack et al. (2011) proposed an alternative definition of the d.f. $\mathrm{tr}(2\mathbf{SR}-\mathbf{SRS}^\top)$. Note that for independent data $\mathrm{tr}(2\mathbf{SR}-\mathbf{SRS}^\top)$ is reduced to $\mathrm{tr}(2\mathbf{S}-\mathbf{SS}^\top)$. If, in addition to independence, symmetric idempotency is assumed for $\mathbf{S}$, then $\mathrm{tr}(2\mathbf{SR}-\mathbf{SRS}^\top)$ is equal to $\mathrm{tr}(\mathbf{S})$ and GCVc is equivalent to GCV by (3.30).

When the sample size is large, CV may be prohibitively computer-intensive and time-consuming. Even in a one-dimensional case ($d = 1$), calculation of the CV statistic for one value of h requires N kernel function evaluations at each t_k. Thus the total computation will be of order $O(N^2)$. For the estimation we propose, computer-intensiveness will be aggravated due to iterative estimation. To circumvent computational difficulties, the `plmm` package offers bandwidth selection using binning techniques through function `h.select` provided in the `sm` package in R. Since our smoother $\mathbf{S}$ is not symmetric idempotent, binning is also used in `plmm` to approximate $\mathrm{tr}(\mathbf{S})$ and $\mathrm{tr}(\mathbf{SS}^\top)$ for calculation of the standard d.f. $\mathrm{tr}(2\mathbf{S}-\mathbf{SS}^\top)$ of $\hat{\boldsymbol{\gamma}}$. Binning techniques are briefly described in Appendix 3.10.6.

3.4 Variance Components Estimation

This section presents methods for estimating unknown VCs. We use quadratic estimators, which are obtained by equating the sum of squared residuals to its expected value and solving the equations for the parameters to be estimated. The estimators considered here are of Swamy-Arora (SA) type[13] and fitting-of-constant (FC) type, also called Henderson's 3rd method. We iteratively estimate the VCs and succeedingly $\boldsymbol{\beta}$ and γ. Through the iterative process, the initial VC estimates are corrected

[13] Baltagi and Chang (1994) extended the VC estimators proposed by Swamy and Arora (1972) for a balanced two-way model to the case of an unbalanced one-way model.

for the loss of d.f. due to nonparametric function estimation, and thereby $\boldsymbol{\beta}$ and γ estimators are also corrected.

For both the FC and SA VC estimators, the σ_u^2 estimate may become negative. A negative estimate is conventionally set to zero in the LMM literature.[14] While the σ_e^2 estimator is unbiased, the σ_u^2 estimator is therefore biased in the sense that it needs to be corrected to zero with some unknown probability. In a simulation study on the random intercept model, Maddala and Mount (1973) investigated various VC estimators for balanced data, including the SA and FC estimators. Baltagi and Chang (1994) conducted a simulation study for unbalanced data. These studies showed that among various estimators there is little advantage of one estimator over the others; and that a negative estimate of σ_u^2 is most likely to occur when the true σ_u^2 is very small compared to σ_e^2, and hence setting a negative estimate to zero will not lead to significant loss of efficiency in estimating regression coefficients. Prasad and Rao (1990) showed in a context of the general LMM that the quadratic VC estimators are $\sqrt{m}$-consistent, and further that the probability of obtaining a negative σ_u^2 estimate converges to zero as m increases.[15]

3.4.1 Homoskedastic and heteroskedastic case with known $\boldsymbol{\alpha}$

This section presents VC estimators for homoskedastic and heteroskedastic cases with known heteroskedasticity parameters $\boldsymbol{\alpha}$. Both of the SA and FC estimators are based on the quadratic form of certain residuals. Both estimate σ_e^2 using the residuals which are known as within-residuals or least squares dummy variable (LSDV) residuals in the literature of panel data analysis. The SA estimator of σ_u^2 is constructed as a quadratic function of the residuals called between-residuals. On the other hand, the FC estimator of σ_u^2 is based on the OLS residuals. We follow the convention of setting the estimate of σ_u^2 to zero if it turns out to be negative in the initial or iteration stage.

Initial stage: Let $\tilde{y}_{ij}$ and $\tilde{\mathbf{x}}_{ij}$ denote a working response and a working regressor vector, respectively. In the initial stage, $\tilde{y}_{ij}$ is defined as $y_{ij} - \hat{\mathrm{E}}[y_{ij}|\mathbf{t}_{ij}]$ and $\tilde{\mathbf{x}}_{ij}$ as $\mathbf{x}_{ij} - \hat{\mathrm{E}}[\mathbf{x}_{ij}|\mathbf{t}_{ij}]$. In the homoskedastic case, the initial estimator of σ_e^2 is given by

$$\hat{\sigma}_e^2 \;=\; \frac{\hat{\mathbf{e}}_w^\top \mathbf{Q} \hat{\mathbf{e}}_w}{N - m - p}\,, \tag{3.32}$$

[14] For strategies to deal with a negative VC estimate, see, for example, Searle et al. (1992) and references therein.

[15] Restricted maximum likelihood VC estimators under normality assumption were also shown to be $\sqrt{m}$-consistent by Jiang (1996), and also the maximum likelihood VC estimator by Datta and Lahiri (2000).

where $\mathbf{Q} = \mathbf{I}_m \otimes \mathbf{Q}_i$ with $\mathbf{Q}_i = \mathbf{I}_{n_i} - \bar{\mathbf{J}}_{n_i}$, $\bar{\mathbf{J}}_{n_i} = \mathbf{J}_{n_i}/n_i$, and $\hat{\mathbf{e}}_w$ is a vector of within-residuals from the reduced model (3.12). The FC and SA estimators of σ_u^2 are given by

$$\hat{\sigma}^2_{u,\mathrm{FC}} = \max\left(\frac{\hat{\mathbf{v}}^\top\hat{\mathbf{v}} - (N-p)\hat{\sigma}_e^2}{N - \mathrm{tr}\left((\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\sum_{i=1}^m n_i^2\bar{\tilde{\mathbf{x}}}_{i\cdot}\bar{\tilde{\mathbf{x}}}_{i\cdot}^\top\right)}, 0\right) \text{ and} \tag{3.33}$$

$$\hat{\sigma}^2_{u,\mathrm{SA}} = \max\left(\frac{\hat{\mathbf{v}}_b^\top\mathbf{P}\hat{\mathbf{v}}_b - (m-p)\,\hat{\sigma}_e^2}{N - \mathrm{tr}\left((\tilde{\mathbf{X}}^\top\mathbf{P}\tilde{\mathbf{X}})^{-1}\sum_{i=1}^m n_i^2\bar{\tilde{\mathbf{x}}}_{i\cdot}\bar{\tilde{\mathbf{x}}}_{i\cdot}^\top\right)}, 0\right)\ , \tag{3.34}$$

respectively. Here $\mathbf{P} = \mathbf{I}_m \otimes \bar{\mathbf{J}}_{n_i}$; $\hat{\mathbf{v}}_b$ and $\hat{\mathbf{v}}$ are the between-residual vector and the OLS residual vector from (3.12), respectively. Details of notation and derivation are given in Appendix A 3.10.2.

The SA and FC estimators for the homoskedastic case can be generalized to the heteroskedastic case. First, $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + (\oplus_{ij}\alpha_{ij})\mathbf{e}$ is transformed into an LMM

$$(\oplus_{ij}\alpha_{ij}^{-1})\tilde{\mathbf{y}} = (\oplus_{ij}\alpha_{ij}^{-1})\tilde{\mathbf{X}}\boldsymbol{\beta} + (\oplus_{j}\alpha_{ij}^{-1})\mathbf{Z}\mathbf{u} + \mathbf{e}\ . \tag{3.35}$$

The estimator of σ_e^2 is given by

$$\hat{\sigma}_e^2 = \frac{\hat{\underline{\mathbf{e}}}_w^\top\mathbf{Q}\hat{\underline{\mathbf{e}}}_w}{N-m-p}\ , \tag{3.36}$$

where $\hat{\underline{\mathbf{e}}}_w$ is the within-residual vector of (3.35). The SA and FC estimators are

$$\hat{\sigma}^2_{u,\mathrm{FC}} = \max\left(\frac{(\hat{\underline{\mathbf{v}}}^\top\hat{\underline{\mathbf{v}}} - (N-p)\hat{\sigma}_e^2)}{\sum_{i,j}\alpha_{ij}^{-2} - C_{\mathrm{FC}}}, 0\right) \text{ and} \tag{3.37}$$

$$\hat{\sigma}^2_{u,\mathrm{SA}} = \max\left(\frac{\hat{\underline{\mathbf{v}}}_b^\top\mathbf{P}\hat{\underline{\mathbf{v}}}_b - (m-p)\hat{\sigma}_e^2}{\sum_i\eta_i - C_{\mathrm{SA}}}, 0\right)\ , \tag{3.38}$$

respectively. Here $\hat{\underline{\mathbf{v}}}$ and $\hat{\underline{\mathbf{v}}}_b$ are the OLS residual vector and the between-residual vector of (3.35), respectively. Details of notation and the derivation of the estimators are given in Appendix A 3.10.3.

Iteration stage: In the iteration stage, the estimation procedure differs from that in the initial stage in construction of the working variables, denoted by $\tilde{y}_{ij}$ and $\tilde{\mathbf{x}}_{ij}$ as before. In the rth iteration ($r = 1, 2, \ldots$) the working response is in matrix form $\tilde{\mathbf{y}}_{(r)} = \mathbf{y} - \hat{\boldsymbol{\gamma}}_{(r-1)}(\mathbf{T})$ and $\tilde{\mathbf{X}}_{(r)}$ is the same as the original design $\mathbf{X}$. The estimator

of σ_e^2 in the rth iteration is given by

$$\hat{\sigma}^2_{e(r)} = \frac{\hat{\mathbf{e}}^\top_{w(r)}\mathbf{Q}\hat{\mathbf{e}}_{w(r)}}{N - m - r(\mathbf{QX}) - \text{d.f.}(\hat{\boldsymbol{\gamma}}_{(r-1)})}, \tag{3.39}$$

where d.f.$(\hat{\boldsymbol{\gamma}}_{(r-1)})$ is the d.f. of freedom of the estimator of γ in the $(r-1)$th iteration and $\hat{\mathbf{e}}_{w(r)}$ is the within-residual vector. Here $\hat{\boldsymbol{\gamma}}_{(0)}$ is the estimate of $\boldsymbol{\gamma}$ in the initial stage. Details of notation and derivation are given in Appendix A 3.10.2. The FC and SA estimators of σ_u^2 are given by the same formulas as (3.33) and (3.34) with $\hat{\mathbf{v}}$ and $\hat{\mathbf{v}}_b$ replaced by the OLS residual vector $\hat{\mathbf{v}}_{(r)}$ and the between-residual vector $\hat{\mathbf{v}}_{b(r)}$, respectively.

In the heteroskedastic case, the estimator of σ_e^2 in the rth iteration is given by

$$\hat{\sigma}^2_{e(r)} = \frac{\hat{\underline{\mathbf{e}}}^\top_{w(r)}\mathbf{Q}\hat{\underline{\mathbf{e}}}_{w(r)}}{N - m - r(\mathbf{Q}\underline{\mathbf{X}}) - \text{d.f.}(\hat{\boldsymbol{\gamma}}_{(r-1)})}, \tag{3.40}$$

where $\underline{\mathbf{X}} = (\oplus_{ij}\alpha_{ij}^{-1})\mathbf{X}$. The FC and SA estimators of σ_u^2 are given by the same formulas as (3.37) and (3.38), respectively, where $\hat{\underline{\mathbf{v}}}$ and $\hat{\underline{\mathbf{v}}}_b$ are replaced by $\hat{\underline{\mathbf{v}}}_{b(r)}$ and $\hat{\underline{\mathbf{v}}}_{(r)}$. Details of notation and the derivation of the estimators are given in Appendix A 3.10.3.

The iterative process is continued until convergence. We regard convergence of the whole estimators as achieved when the following inequality is fulfilled

$$\max\left(\frac{\hat{\sigma}^2_{u(r+1)} - \hat{\sigma}^2_{u(r)}}{\hat{\sigma}^2_{u(r)}}, \frac{\hat{\sigma}^2_{e(r+1)} - \hat{\sigma}^2_{e(r)}}{\hat{\sigma}^2_{e(r)}}\right) < \epsilon \tag{3.41}$$

with ϵ being some positive small value (the default in the `plmm` package is 0.003).

3.4.2 Heteroskedastic case with unknown $\boldsymbol{\alpha}$

In practice the vector of heteroskedasticity parameters $\boldsymbol{\alpha}$ is typically unknown and needs to be estimated either parametrically or nonparametrically. This section presents nonparametric estimation of the conditional variance function of $\alpha_{ij}e_{ij}$. The variance function is assumed to be a function of conditioning variables $\mathbf{w} \in \mathbb{R}^q$, whose elements are q continuous variables of regressors $\mathbf{x}$ and $\mathbf{t}$. To estimate the variance function $\text{Var}[\alpha_{ij}e_{ij}|\mathbf{w}_{ij}]$, we apply the method Li and Stengos (1994) proposed for the random intercept model for panel data. A test of heteroskedasticity in the PLMM framework is provided by You et al. (2010). An extension to include dummies in $\mathbf{w}$ is also possible.

Let σ_{ij}^2 and ν_{ij}^2 be respective conditional variances of v_{ij} and $\alpha_{ij}e_{ij}$, where v_{ij} is the residual obtained after convergence of the iterative estimation. σ_{ij}^2 can be estimated by Nadaraya-Watson (local constant) kernel regression

$$\hat{\sigma}_{ij}^2(\mathbf{w}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} K_{\mathbf{h}}(\mathbf{w}_{ij}-\mathbf{w})\hat{v}_{ij}^2}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} K_{\mathbf{h}}(\mathbf{w}_{ij}-\mathbf{w})} \ , \tag{3.42}$$

where $\mathbf{h}$ is a q-dimensional vector of bandwidths and $K_{\mathbf{h}}(\bullet)$ is a multiplicative kernel weights function $K(\mathbf{H}^{-1}(\bullet))$ with $\mathbf{H}$ being a diagonal matrix of the bandwidths, $\oplus_{q'=1}^{q} h_{q'}^{LS}$. Since $\sigma_{ij}^2 = \sigma_u^2 + \nu_{ij}^2$ holds, ν_{ij}^2 can be estimated by

$$\hat{\nu}_{ij}^2(\mathbf{w}) = \hat{\sigma}_{ij}^2(\mathbf{w}) - \hat{\sigma}_u^2 \ , \tag{3.43}$$

where $\hat{\sigma}_u^2$ is the FC or SA estimate obtained under homoskedastic error assumption.

Bandwidths $h_{q'}^{LS}$ can be chosen by cross validation or as in Li and Stengos (1994) by the rule of thumb specified as

$$h_{q'} = \mathrm{sd}[w_{q'}]N^{-1/(4+q)} \ (q' = 1, \dots q) \ , \tag{3.44}$$

where $h_{q'}$ is the bandwidth for variable $w_{q'} \in \mathbf{w}$. A potential problem is that $\hat{\nu}_{ij}^2$ can be negative. To circumvent the problem, we arbitrarily set $\hat{\nu}_{ij}^2$ smaller than 0.001 to 0.001, which is the default setting in the `plmm` package.

While we use the FC or SA estimator of σ_u^2, Li and Stengos (1994) originally proposed estimating σ_u^2 as follows. Since $\mathrm{E}[v_{ij}] = 0$, $\mathrm{Cov}[u_i, u_{i'}] = 0$ $(i \neq i')$, and $\mathrm{Cov}[u_i, e_{ij}] = 0$ by assumption, $\mathrm{Cov}(v_{ij}, v_{ij'}) = \mathrm{E}[v_{ij}v_{ij'}] = \sigma_u^2$ for $j \neq j'$. Using the method of moments, σ_u^2 can be estimated by

$$\hat{\sigma}_u^2 \ = \ \frac{1}{\sum_{i=1}^{m} n_i\,(n_i-1)} \sum_i \sum_{j \neq j'} v_{ij}v_{ij'} \ , \tag{3.45}$$

where $v_{ij} = y_{ij} - \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} - \gamma(\mathbf{t}_{ij})$. You et al. (2010) showed that this estimator is also $\sqrt{m}$-consistent in the PLMM framework.

3.5 Prediction of Random Effects

Random effects themselves are often of interest. For instance, random effects predictions are an integral part for the so-called small area estimation,[16] where typically the mean or total of a variable of interest is to be efficiently estimated for observed

[16] Refer to Rao (2003) for an extensive overview of the small area estimation.

or unobserved areas (clusters) despite of their "small" area data sizes.[17] For the standard LMM $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$, random effects predictors are widely known and used as the best linear unbiased predictor (BLUP).[18] The BLUP is given by $\tilde{\mathbf{u}} = \mathbf{D}\mathbf{Z}^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator. It should be noted that the BLUP does not require any distributional assumption.[19]

For the model of our interest, we analogously construct a random effects predictor: $\tilde{\mathbf{u}} = \mathbf{D}\mathbf{Z}^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\gamma}}(\mathbf{T}))$ where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are consistent estimators. The ith cluster's random intercept is predicted by

$$\tilde{u}_i = \sigma_u^2 \mathbf{1}_{n_i}^\top \mathbf{V}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\gamma}}(\mathbf{T}_i) \right) \ \forall i \ , \tag{3.46}$$

which is a consistent linear predictor. The derivation of (3.46) is given in Appendix A 3.10.4.

In practice the VCs, and hence $\mathbf{V}$, are unknown. Following the convention of the LMM literature, we simply replace the unknown VCs with their consistent estimators to obtain the estimated (empirical) version of (3.46)

$$\hat{u}_i = \hat{\sigma}_u^2 \mathbf{1}_{n_i}^\top \hat{\mathbf{V}}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\gamma}}(\mathbf{T}_i) \right) \ \forall i \ . \tag{3.47}$$

Here $\mathbf{V}_i^{-1}$ is replaced in the homoskedastic case by

$$\hat{\mathbf{V}}_i^{-1} = \frac{1}{\hat{\sigma}_e^2} \left(\mathbf{I}_{n_i} - \frac{n_i \hat{\sigma}_u^2}{n_i \hat{\sigma}_u^2 + \hat{\sigma}_e^2} \bar{\mathbf{J}}_{n_i} \right) \tag{3.48}$$

and in the heteroskedastic case by

$$\hat{\mathbf{V}}_i^{-1} = \oplus \hat{\nu}_{ij}^{-2} - \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 \sum \hat{\nu}_{ij}^{-2} + 1} \left(\hat{\nu}_{i1}^{-2}, \ldots, \hat{\nu}_{in_i}^{-2} \right)^\top \left(\hat{\nu}_{i1}^{-2}, \ldots, \hat{\nu}_{in_i}^{-2} \right) \ . \tag{3.49}$$

See Appendix A 3.10.1 for the derivation of $\mathbf{V}^{-1}$.

[17] Predictions can also be used as a tool for informal checking of whether relevant cluster-level regressors are omitted from the model. If a significant cluster-level regressor is omitted, the shape of the distribution of random intercept predictions may reflect the effects of that omitted regressor. However, this way of model checking calls for caution due to so-called "shrinkage effect". For more details of this issue, see, for example, Verbeke and Molenberghs (2000).

[18] Here "best" is in the sense that $\tilde{u}$ minimizes the mean squared prediction error $\mathrm{E}[(\tilde{u} - u)^2]$; "unbiased" in the sense that $\mathrm{E}[\tilde{u}] = \mathrm{E}[u]$; and $\tilde{u}$ is "linear" in $\mathbf{y}$ of the form $b + \mathbf{c}^\top \mathbf{y}$ for constant b and vector $\mathbf{c}$.

[19] For details of the BLUP for the LMM, see, for example, Searle et al. (1992).

3.6 Test of Regression Coefficients

We now propose a simple test for the parametric component parameters $\boldsymbol{\beta}$. For ease of presentation we consider model (3.1) with $d = 1$ and $\alpha_{ij} = 1$: $y_{ij} = \mathbf{x}_{ij}^\top \beta + \gamma(t_{ij}) + v_{ij}$ where $v_{ij} = u_i + e_{ij}$. Let's write possible dependence between $x_{q,ij}$ and t_{ij} $(q = 1, \dots p)$ in terms of a functional relationship as follows:

$$x_{q,ij} \;=\; \xi_q(t_{ij}) + \delta_{q,ij} \;, \tag{3.50}$$

where ξ_q is some nonparametric smooth function and $\delta_{q,ij}$ is an i.i.d. deviation. (3.50) can be rewritten as

$$\mathbf{x}_{ij} = \boldsymbol{\xi}(t_{ij}) + \boldsymbol{\delta}_{ij} \;, \tag{3.51}$$

where $\boldsymbol{\xi}(t_{ij}) = (\xi_1(t_{ij}), \dots \xi_p(t_{ij}))^\top$, $\mathbf{x}_{ij} = (x_{1,ij}, \dots x_{p,ij})^\top$ and $\boldsymbol{\delta}_{ij} = (\delta_{1,ij}, \dots \delta_{p,ij})^\top$. Deviations $\boldsymbol{\delta}_{ij}$ are such that $\mathrm{E}[\boldsymbol{\delta}_{ij}|t_{ij}] = \mathbf{0}$ and $\mathrm{Var}[\boldsymbol{\delta}_{ij}|t_{ij}]$ is a finite diagonal matrix. In matrix form (3.51) is expressed as

$$\mathbf{X} = \boldsymbol{\Xi}(\mathbf{t}) + \boldsymbol{\Delta} \;, \tag{3.52}$$

where $\mathbf{X}$, $\boldsymbol{\Xi}$ and $\boldsymbol{\Delta}$ are matrices whose ijth row is $\mathbf{x}_{ij}^\top$, $\boldsymbol{\xi}^\top(t_{ij})$ and $\boldsymbol{\delta}_{ij}^\top$, respectively. Inserting (3.51) into $y_{ij} = \mathbf{x}_{ij}^\top \beta + \gamma(t_{ij}) + v_{ij}$ yields

$$\begin{aligned} y_{ij} &= (\boldsymbol{\xi}(t_{ij}) + \boldsymbol{\delta}_{ij})^\top \boldsymbol{\beta} + \gamma(t_{ij}) + v_{ij} \\ &= \boldsymbol{\xi}^\top(t_{ij})\boldsymbol{\beta} + \gamma(t_{ij}) + \boldsymbol{\delta}_{ij}^\top \boldsymbol{\beta} + v_{ij} \;. \end{aligned} \tag{3.53}$$

Let $\xi_y(t_{ij})$ denote the conditional expectation $\mathrm{E}[y_{ij}|t_{ij}]$:

$$\xi_y(t_{ij}) = \boldsymbol{\xi}^\top(t_{ij})\,\boldsymbol{\beta} + \gamma(t_{ij}) \;. \tag{3.54}$$

By subtracting (3.54) from (3.53), it follows that

$$y_{ij} - \xi_y(t_{ij}) \;=\; \boldsymbol{\delta}_{ij}^\top \boldsymbol{\beta} + v_{ij} \;,$$

which can be expressed in matrix form

$$\begin{aligned} \mathbf{y} - \boldsymbol{\xi}_y &= (\mathbf{X} - \boldsymbol{\Xi}(\mathbf{t}))\boldsymbol{\beta} + \mathbf{v} \\ &= \boldsymbol{\Delta}\boldsymbol{\beta} + \mathbf{v} \;. \end{aligned} \tag{3.55}$$

Thus, if $\boldsymbol{\xi}_y$ and $\boldsymbol{\Xi}$ are given, the coefficients vector $\boldsymbol{\beta}$ can be estimated by GLS with an estimate of $\mathbf{V}^{-1}$ as follows:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{gls} &= \left((\mathbf{X}-\boldsymbol{\Xi})^\top \hat{\mathbf{V}}^{-1}(\mathbf{X}-\boldsymbol{\Xi})\right)^{-1}(\mathbf{X}-\boldsymbol{\Xi})^\top \hat{\mathbf{V}}^{-1}(\mathbf{y}-\boldsymbol{\xi}_y) \\ &= (\boldsymbol{\Delta}^\top \hat{\mathbf{V}}^{-1}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}^\top \hat{\mathbf{V}}^{-1}(\boldsymbol{\Delta}\boldsymbol{\beta}+\mathbf{v}) \\ &= \boldsymbol{\beta} + (\boldsymbol{\Delta}^\top \hat{\mathbf{V}}^{-1}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}^\top \hat{\mathbf{V}}^{-1}\mathbf{v} \end{aligned} \tag{3.56}$$

with covariance matrix

$$\hat{\mathrm{Var}}[\hat{\boldsymbol{\beta}}_{gls}] = (\boldsymbol{\Delta}^\top \hat{\mathbf{V}}^{-1}\boldsymbol{\Delta})^{-1} \ . \tag{3.57}$$

Alternatively, OLS estimator can be obtained by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{ols} &= \left((\mathbf{X}-\boldsymbol{\Xi})^\top (\mathbf{X}-\boldsymbol{\Xi})\right)^{-1}(\mathbf{X}-\boldsymbol{\Xi})^\top (\mathbf{y}-\boldsymbol{\xi}_y) \\ &= (\boldsymbol{\Delta}^\top \boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}^\top (\boldsymbol{\Delta}\boldsymbol{\beta}+\mathbf{v}) \\ &= \boldsymbol{\beta} + (\boldsymbol{\Delta}^\top \boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}^\top \mathbf{v} \ . \end{aligned} \tag{3.58}$$

For this OLS estimator, a robust covariance matrix estimator is given by

$$\hat{\mathrm{Var}}_{\mathrm{robust}}[\hat{\boldsymbol{\beta}}_{ols}] = (\boldsymbol{\Delta}^\top \boldsymbol{\Delta})^{-1}\sum_{i=1}^{m}(\boldsymbol{\Delta}_i^\top \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^\top \boldsymbol{\Delta}_i)(\boldsymbol{\Delta}^\top \boldsymbol{\Delta})^{-1} \ . \tag{3.59}$$

In practice, however, $\boldsymbol{\Delta}$ needs to be estimated through nonparametric estimation of smooth functions $\xi_y, \xi_1, \ldots \xi_p$. Given consistent estimators of $\boldsymbol{\xi}_y$ and $\boldsymbol{\Xi}$, equality of (3.55) holds asymptotically. Thus, asymptotic properties of $\hat{\boldsymbol{\beta}}$ depend on the asymptotic properties of $\hat{\boldsymbol{\xi}}_y$ and $\hat{\boldsymbol{\Xi}}$ (and thus $\hat{\boldsymbol{\Delta}}$) that in turn depend on bandwidth selection, kernel function selection including the order of kernel, and distribution of variable $\mathbf{t}$. The distribution of $\mathbf{x}$ also plays a role. If a regressor x is not a continuous but, for example, binary or count variable, its ξ function needs to be estimated accordingly.

Moreover, even if $\boldsymbol{\Delta}$ is known, the covariance matrix estimator (3.57) needs to be adjusted. It is a well-known problem in the mixed effects model literature that the variance of the estimator of $\boldsymbol{\beta}$, which depends on the covariance matrix $\mathbf{V}$, will be biased when unknown $\mathbf{V}$ is simply replaced by its estimator. Kenward and Roger (1997) proposed an adjustment to obtain an approximately unbiased estimator. Also well-known in literature is the difficulty in determining the effective number of independent observations in correlated data and hence the d.f. for constructing test statistics.

Given the difficulties in determining appropriate asymptotic distributions, we employ a residual resampling method, the wild bootstrap. For the wild bootstrap, see Wu

(1986) and Shao and Tu (1995) among others. In the resampling process, predicted random effects and regression residuals are each separately multiplied by a realization of independent standard normally distributed variable. This procedure retains the first three moments of the random intercepts and the regression error as well as the independence of the random intercept and the regression error. Wild bootstrap resampling of size B is conducted as follows:

1. Obtain estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ and $(\hat{\sigma}_u^2, \hat{\sigma}_e^2)$ in a homoskedastic error case or $(\hat{\sigma}_u^2, \hat{\nu}_{ij}^2)$ in a heteroskedastic case.

2. Predict the random effect $\hat{u}_i$ and calculate the residuals by $y_{ij} - \mathbf{x}_{ij}^\top \hat{\boldsymbol{\beta}} - \hat{\gamma}(\mathbf{t}_{ij}) - \hat{u}_i$ $\forall i$.

3. For the bth bootstrap sampling ($b = 1, \dots B$), obtain an N-dimensional vector of bootstrap response by $y_{ij}^{(b)} = \mathbf{x}_{ij}^\top \hat{\boldsymbol{\beta}} + \hat{\gamma}(\mathbf{t}_{ij}) + z_{1,i}^{(b)} \hat{u}_i + z_{2,ij}^{(b)} \hat{e}_{ij}^*$ where $z_{1,i}^{(b)} \sim \mathcal{N}(0,1)$ and $z_{2,ij}^{(b)} \sim \mathcal{N}(0,1)$; $z_{1,i}^{(b)}$ and $z_{2,ij}^{(b)}$ are mutually independent; $\hat{e}_{ij}^*$ is the residual obtained in step 2.

4. Update the initial bandwidth for the estimation of $\mathrm{E}[y_{ij}^{(b)} | \mathbf{t}_{ij}]$ and obtain the bth bootstrap estimate $\hat{\boldsymbol{\beta}}^{(b)}$ from the bth bootstrap sample $\{y_{ij}^{(b)}, \mathbf{x}_{ij}, \mathbf{t}_{ij}\}$

5. Repeat step 3. and 4. for B times.

The bootstrap procedure provides an approximation to the sampling distribution of $\hat{\boldsymbol{\beta}}$ from which a bootstrap confidence interval can be constructed. An illustration of bootstrap confidence intervals is given in Section 3.8.

3.7 Finite Sample Performance: Simulation Studies

This section presents simulation studies conducted to investigate the following issues: (1) effects of the bandwidths selected for model reduction on the estimation of the VC and $\boldsymbol{\beta}$; (2) comparison of $\boldsymbol{\beta}$ estimators between GLS and OLS in the initial stage; (3) consistency of VC and $\boldsymbol{\beta}$ estimators; and (4) efficiency gains through the Li and Stengos procedure when regression errors are heteroskedastic.

The data generating process was designed for a partially linear random intercept

model with a one-dimensional nonparametric function:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{ij,2} + \beta_3 x_{3,ij} + \gamma(t_{ij}) + u_i + e_{ij} \\ \boldsymbol{\beta} &= (1,\, 1,\, 0.5,\, -1)^\top \\ x_1 &\sim \text{binom}\,(0.5) \\ x_2 &\sim \mathcal{N}\left(15,\, 3^2\right) \\ x_3 &\sim \text{unif}(1,\, 4) \\ t &\sim \text{unif}\,(0.3\pi,\, 2\pi) \\ \gamma\,(t) &= \sin\,(t) \\ u_i &\sim \mathcal{N}\,(0,\, 4) \\ e_{ij} &\sim \mathcal{N}\,(0,\, 1)\ . \end{aligned} \tag{3.60}$$

x_2 and t were generated with a correlation of about 0.7. Unbalanced data were generated with n_i chosen randomly from uniform distribution $U(3, 37)$. The number of clusters m is set to 20 for small samples while $m = 100$ for large samples. Simulation size was 500. In selecting bandwidths by CV, binning techniques were used with the number of bins being the rounded number of $8\log(N)$.[20]

3.7.1 Influence of the model-reduction bandwidths and effects of the iteration

To investigate the influence of the bandwidths selected for model reduction on the succeeding estimation of the VCs and $\boldsymbol{\beta}$, we experimented with three bandwidths: bandwidth selected by CV denoted by $\mathbf{h}_0$, and its scaled bandwidths $0.3\mathbf{h}_0$ and $3\mathbf{h}_0$. Table 3.1 and 3.2 show the average and standard deviation of 500 estimates of each model parameter using the SA VC estimator. Table 3.1 illustrates influence of different bandwidths on the estimation in the initial stage, especially on VC estimation. $\boldsymbol{\beta}$ estimators were robust to bandwidth selection in spite of variations in the VC estimates. Continuing with the estimation from the initial stage, we obtained 500 after-iteration estimates, summarized in Table 3.2. The iteration process ended with almost the same estimation values regardless of differences in the bandwidths selected. For this simulation, all the estimators resulted essentially unbiased by the end of iteration process.

[20] This number of bins is the default setting of R function `h.select`, which selects a bandwidth by CV using binning techniques. For $N > 100$, its default is the rounded number of $8\log(N)/d$.

	$\hat{\sigma}_u$	$\hat{\sigma}_e$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
$0.3\mathbf{h}_0$	3.507 (1.364)	1.083 (0.106)	0.989 (0.107)	0.494 (0.021)	-0.988 (0.065)
$\mathbf{h}_0$	3.644 (1.401)	1.054 (0.084)	0.997 (0.104)	0.498 (0.018)	-0.996 (0.061)
$3\mathbf{h}_0$	3.697 (1.417)	1.107 (0.088)	0.998 (0.108)	0.500 (0.018)	-0.998 (0.063)

Table 3.1: Influence of the model-reduction bandwidths on the initial stage parameter estimation using the SA VC estimator.

	$\hat{\sigma}_u$	$\hat{\sigma}_e$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
$0.3\mathbf{h}_0$	3.938 (1.481)	1.063 (0.088)	1.000 (0.108)	0.488 (0.036)	-1.000 (0.065)
$\mathbf{h}_0$	3.939 (1.481)	1.063 (0.088)	1.000 (0.108)	0.489 (0.036)	-1.000 (0.065)
$3\mathbf{h}_0$	3.939 (1.481)	1.063 (0.088)	1.000 (0.108)	0.489 (0.036)	-1.000 (0.065)

Table 3.2: Influence of the model-reduction bandwidths on the parameter estimation after iteration using the SA VC estimator.

We estimated the model parameters using the FC VC estimator as well. The effects of different bandwidths on the following estimation (Table 3.3) almost disappeared by the end of iteration as seen in Table 3.4. For this simulation, estimators with the FC VC estimators can also be regarded as unbiased.

The average squared error of the σ_u^2 estimates was 1.956 for the FC VC estimator while 2.194 for the SA VC estimator. It is not clear whether one estimator should be preferred to the other for the iterative PLMM estimation. As Maddala and Mount (1973) and Baltagi and Chang (1994) argued, it will be sensible to estimate the PLMM with both VC estimation methods to see if they yield largely different estimates. Robustness of the $\boldsymbol{\beta}$ estimator against the biased VC estimators in the initial stage is in agreement with the bootstrap simulation studies on the random intercept model by Bellmann et al. (1989).

	$\hat{\sigma}_u$	$\hat{\sigma}_e$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
$0.3\mathbf{h}_0$	3.563 (1.278)	1.083 (0.106)	0.989 (0.107)	0.494 (0.021)	-0.988 (0.065)
$\mathbf{h}_0$	3.697 (1.314)	1.054 (0.084)	0.997 (0.104)	0.498 (0.018)	-0.996 (0.061)
$3\mathbf{h}_0$	3.752 (1.333)	1.107 (0.088)	0.998 (0.108)	0.500 (0.018)	-0.998 (0.063)

Table 3.3: Influence of the model-reduction bandwidths on the initial stage parameter estimation using the FC VC estimator.

	$\hat{\sigma}_u$	$\hat{\sigma}_e$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
$0.3\mathbf{h}_0$	3.909 (1.397)	1.063 (0.088)	1.000 (0.108)	0.488 (0.036)	-1.000 (0.065)
$\mathbf{h}_0$	3.909 (1.397)	1.063 (0.088)	1.000 (0.108)	0.489 (0.036)	-1.000 (0.065)
$3\mathbf{h}_0$	3.909 (1.397)	1.063 (0.088)	1.000 (0.108)	0.489 (0.036)	-1.000 (0.065)

Table 3.4: Influence of the model-reduction bandwidths on the parameter estimation after iteration using the FC VC estimator.

3.7.2 Comparison between OLS and GLS estimators

The estimation procedure we propose relies on GLS $\boldsymbol{\beta}$ estimation. On the other hand, the procedure proposed by You et al. (2010) estimates $\boldsymbol{\beta}$ by OLS after reducing the PLMM to an LMM.[21] This section compares the performance of the GLS and OLS $\boldsymbol{\beta}$ estimators in the initial stage. As in the previous section, we experimented with three model-reduction bandwidths ($\mathbf{h}_0$, $0.3\mathbf{h}_0$, $3\mathbf{h}_0$). Table 3.5 shows the mean and standard deviation of 500 OLS estimates for each model parameter. Results about the σ_e^2 estimator are not given because the variance function of e_{ij} was nonparametrically estimated in You et al. (2010). In terms of the average of the estimates, Table 3.5 shows patterns similar to Table 3.1 and 3.3: the σ_u^2 estimator (3.45) was affected by bandwidth selection; the OLS $\boldsymbol{\beta}$ estimator turned out unbiased being insensitive to the bandwidths selected. However, all the standard deviations of the OLS estimators resulted larger than those of the GLS estimators. These results imply that our estimation procedure provides more efficient $\boldsymbol{\beta}$ estimator (and smaller bias in the iterative σ_u^2 estimator in a small sample) than the one discussed in You et al. (2010).

	$\hat{\sigma}_u^2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
$0.3\mathbf{h}_0$	3.472 (1.446)	0.989 (0.224)	0.493 (0.041)	-0.987 (0.130)
$\mathbf{h}_0$	3.606 (1.494)	0.997 (0.224)	0.498 (0.039)	-0.995 (0.126)
$3\mathbf{h}_0$	3.657 (1.517)	0.999 (0.226)	0.500 (0.039)	-0.998 (0.126)

Table 3.5: Influence of the model-reduction bandwidths on the OLS parameter estimation. The VC were estimated by (3.45).

3.7.3 Convergence of the parameter estimators

Here we examine the consistency of the model parameter estimators by plotting sampling distributions (the mean and standard deviation are given in each figure).

[21] After OLS $\boldsymbol{\beta}$ estimation, γ is nonparametrically estimated, followed by VC estimation by the formula given in (3.45). See Section 3.8.2 for more details of You et al. (2010).

Figure 3.1 displays histograms of the SA VC estimates for small samples. Histograms of SA VC estimates obtained from large samples are shown in Figure 3.2. For both of σ_u^2 and σ_e^2, the average of the estimates became closer to the true parameter value with a smaller standard deviation. Also the histograms are less skewed and more bell-shaped for large samples. As for the FC VC estimator, the simulation study produced results similar to the SA VC estimator.

The GLS $\boldsymbol{\beta}$ estimator (using the SA VC estimator) turned out unbiased for small samples (Figure 3.3) as well as large samples (Figure 3.4). As expected, the standard deviation of the estimates became smaller as the number of clusters (or more precisely, the number of observations N) increased. The GLS estimator with the FC VC estimator showed similar results.

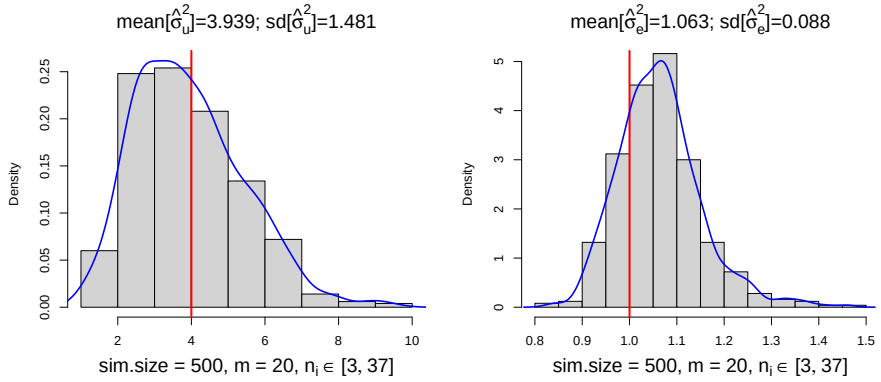


Figure 3.1: Histograms of SA VC estimates from small samples.

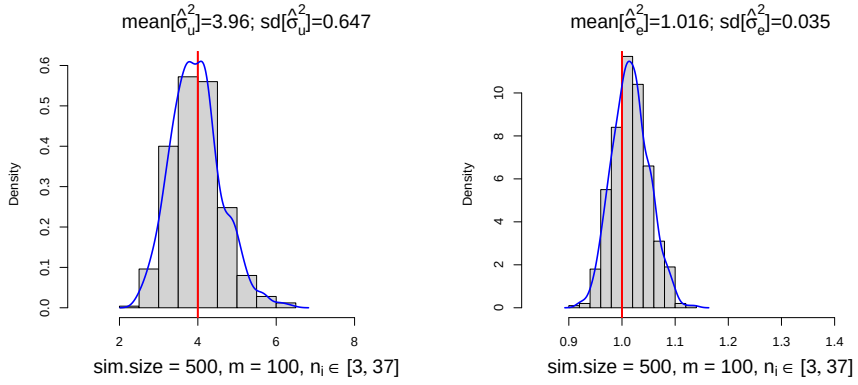


Figure 3.2: Histograms of SA VC estimates from large samples.

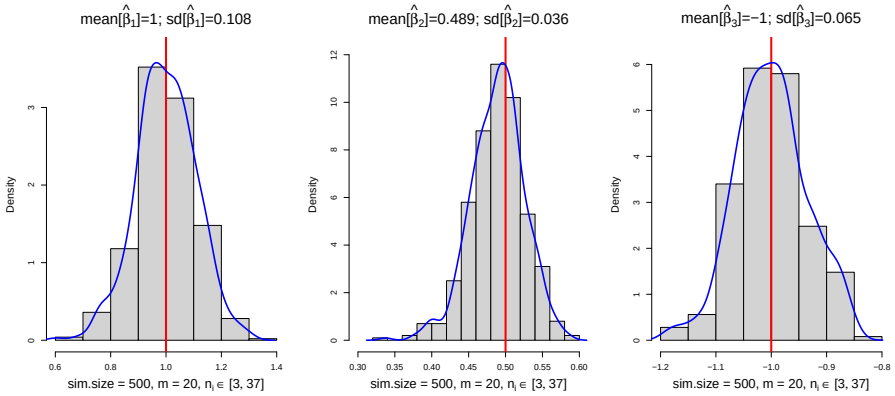


Figure 3.3: Histograms of $\boldsymbol{\beta}$ estimates from small samples.

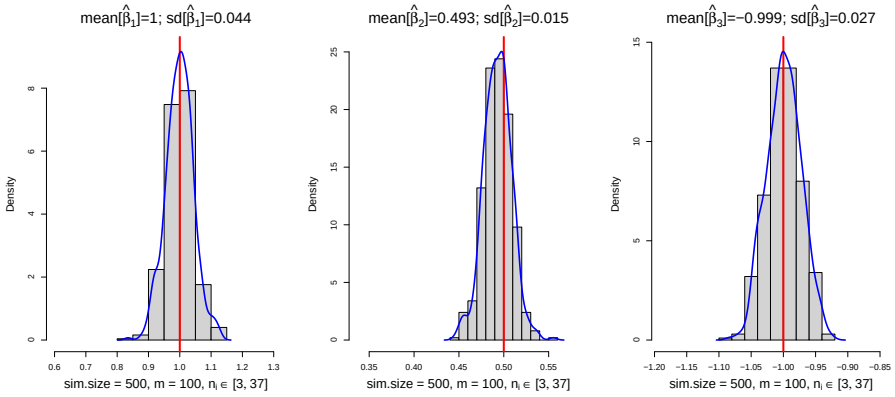


Figure 3.4: Histograms of $\boldsymbol{\beta}$ estimates from large samples.

3.7.4 Efficiency gain by GLS using nonparametric weight function

This section demonstrates the efficiency performance of the GLS estimator using a weight function obtained nonparametrically through the Li and Stengos method. We simulated heteroskedastic data by generating regression errors with heteroskedasticity parameter $\alpha_{ij} = x_{3,ij}^2$. γ and the heteroskedastic variance function were estimated nonparametrically by local linear and Nadaraya-Watson (NW) kernel regression, respectively. The SA VC estimator was used for all the results. The FC VC estimator yielded similar results. Figure 3.5 shows histograms of the GLS $\boldsymbol{\beta}$ estimates obtained under homoskedasticity assumption. As expected, the figures indicate unbiasedness of the estimators. Figure 3.6 and 3.7 display histograms of GLS $\boldsymbol{\beta}$ estimates obtained by using a weight function. The variance function was estimated using the rule-of-thumb type bandwidth selection for Figure 3.6 and CV for Figure 3.7. The sampling distributions show unbiasedness of the GLS estimator using weights and also confirm

efficiency gain, given the heteroskedasticity conditioning variable specified correctly.

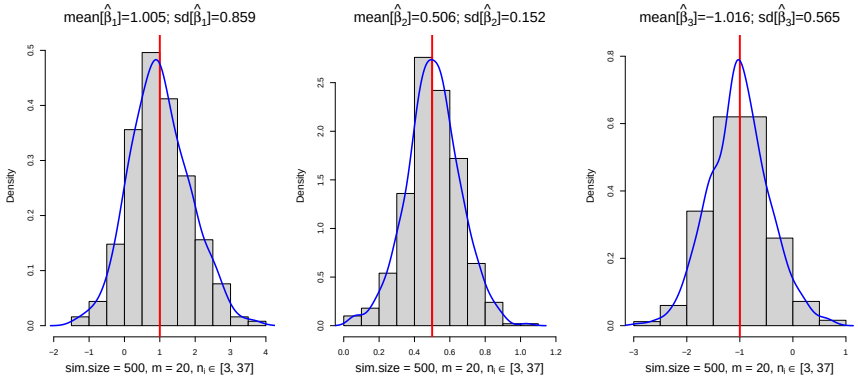


Figure 3.5: GLS $\boldsymbol{\beta}$ estimates under homoskedasticity assumption.

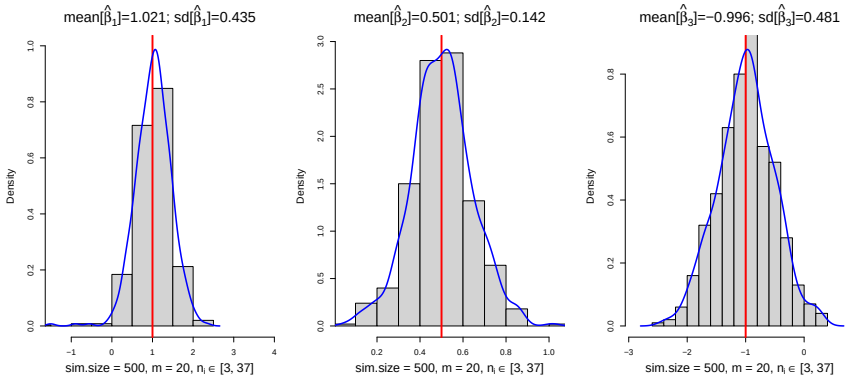


Figure 3.6: Histograms of GLS $\boldsymbol{\beta}$ estimates using a weight function. The rule of thumb type bandwidth selection was used.

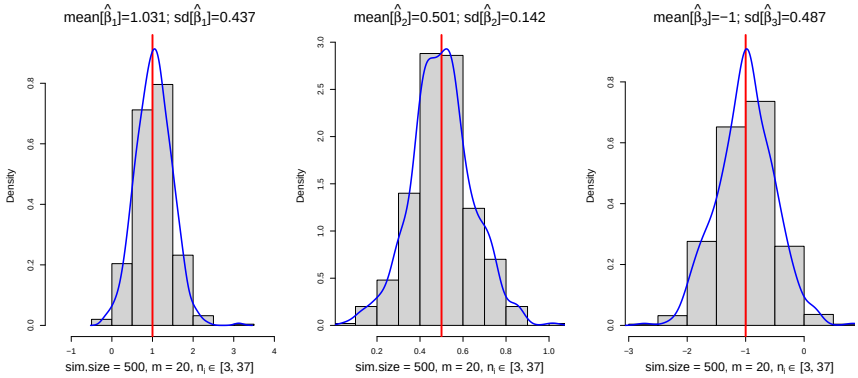


Figure 3.7: Histograms of GLS $\boldsymbol{\beta}$ estimates using a weight function. CV was used for bandwidth selection.

3.8 Estimator's Performance in Practice

3.8.1 Application 1: Panel wage equation

This section illustrates the PLMM by comparing analyses by PLMM and LMM. We analyze a National Longitudinal Survey (NLSY) data set on 4,697 women at the age of 14-26 in 1968, which is unbalanced panel data over 21 years.[22] The data consists of 28,091 observations, where each woman forms a cluster of time-series observations. The variables used in the analysis are described in Table 3.6. Correlations between variables are given in Table 3.7. There are relatively large correlations between *ttl_exp* and *age* (0.76) as well as *ttl_exp* and *tenure* (0.67). This implies that a misspecified function of *age* and *tenure* will cause correlation between *ttl_exp* and the error terms, i.e. endogeneity of the regressors.

ln_wage	logarithm of wage	min. 0, max. 5.263916
grade	completed years of schooling	min. 0, max. 18
age	current age	min. 14, max. 46
ttl_exp	total work experience	min. 0, max. 28.88461
tenure	job tenure, in years	min. 0, max. 25.91667
race	whether black or not	1 if black, 0 otherwise
not_smsa	whether residing in SMSA or not	1 if not SMSA, 0 otherwise
south	whether south or not	1 if south, 0 otherwise

Table 3.6: Description of the variables. Source: Stata Longitudinal-Data/Panel-Data Reference Manual Release 11, pp.6.

	grade	*age*	*ttl_exp*	*tenure*	*race*	*not_smsa*	*south*
ln_wage	.44	.28	.42	.37	-.14	-.22	-.19
grade		.19	.24	.15	-.17	-.12	-.13
age			.76	.44	-.02	.02	.03
ttl_exp				.67	-.03	-.01	-.00
tenure					.01	.00	-.02
race						-.07	.27
not_smsa							.18

Table 3.7: Correlations between variables.

Given a large number of women (number of clusters), a mixed effects model is appealing for construction of a parsimonious parametric model. Here we consider a

[22] Data are from StataCorp (2009) in Longitudinal/Panel-Data Reference Manual Release 11.

parametric random intercept model:

$$\begin{aligned} ln\_wage_{ij} &= \beta_0 + \beta_1 grade + \beta_2 age_{ij} + \beta_3 age_{ij}^2 + \beta_4 ttl\_exp + \beta_5 ttl\_exp^2 + \\ &\quad \beta_6 tenure + \beta_7 tenure^2 + \beta_8 race + \beta_9 not\_smsa + \\ &\quad \beta_{10} south + u_i + e_{ij} \ , \end{aligned} \tag{3.61}$$

where the regression error e_{ij} is assumed to be homoskedastic. Note that the effects of continuous variables *age*, *ttl_exp* and *tenure* are each modeled as a quadratic function. The estimated parameters of model (3.61) are displayed in the first column (LMM) in Table 3.8.[23] The VCs were estimated using the standard SA estimator for the LMM.

The coefficient estimate of ttl_exp^2 appears puzzling. Since the marginal effect of work experience on wage typically decreases, one would expect a negative coefficient for its squared term. However, the estimated coefficient (0.000312) was significantly positive with a bootstrap standard error of 0.000160; robust estimate of the standard error was 0.000163. The Hausman specification test on this LMM rejected the null hypothesis. This implies either that the specification of the fixed components of the model is correct but the random effects are correlated with regressors; or that functional forms of regressors are misspecified and hence correlations with random terms resulted through correlations between *age*, *ttl_exp* and *tenure*.

To examine the latter implication, we constructed three PLMMs: *age* was nonparametrically modeled in plmm (1), *tenure* in plmm (2), and both *age* and *tenure* in plmm (3). We used local linear kernel regression to estimate the nonparametric functions. Estimation results are given in the 2nd, 3rd and 4th columns of Table 3.8. We obtained standard errors given in parentheses using the wild bootstrap described in Section 3.6 with a simulation size of 500. While the estimated coefficient of ttl_exp^2 was significantly positive for plmm (1), it was positive but insignificant for plmm (2) and negative and insignificant for plmm (3). Figure 3.8 displays bootstrap sampling distributions of the estimator of ttl_exp^2 for the three PLMMs. Figure (a) and (b) in Figure 3.9 plot the estimated nonparametric functions of plmm (1) and (2). Figure 3.10 (a) is the estimated nonparametric function of plmm (3), where the curve lacks smoothness, capturing too much noise. We reestimated plmm (3) by increasing the bandwidths selected for plmm (3) by 100% in each direction (*age* and *tenure*) in the iterations and obtained Figure 3.10 (b). The estimated coefficient of ttl_exp^2 was -0.000125.[24] From the results of plmm (3) and its reestimate, we conclude that

[23] StataCorp (2009), pp.448-456 and 471-473 provides more estimates using various panel data models and discussion.

[24] The other coefficient reestimates were *grade* 0.0676, *ttl_exp* 0.0315, *race* -0.0466, *not_smsa* $-$

there is no statistical evidence for increasing marginal effects of work experience. A histogram of random intercept predictions and summary of bootstrap samples are provided in Appendix B.

	LMM	plmm (1)	plmm (2)	plmm (3)
$grade$	.0646 (.0021)	.0644 (.0017)	.0647 (.0017)	.0669 (.0017)
age	.0369 (.0045)		.0375 (.0034)	
age^2	-.0007 (.0001)		-.0007 (.0001)	
ttl_exp	.0287 (.0031)	.0278 (.0027)	.0258 (.0027)	.0299 (.0028)
ttl_exp^2	.000312 (.000160)	.000270 (.000127)	.000162 (.000132)	-.000034 (.000141)
$tenure$	.0396 (.0024)	.0398 (.0017)		
$tenure^2$	-.0020 (.0002)	-.0020 (.0001)		
$race$	-.0536 (.0088)	-.0507 (.0084)	-.0535 (.0084)	-.0466 (.0084)
not_smsa	-.1336 (.0095)	-.1319 (.0070)	-.1326 (.0070)	-.1291 (.0070)
$south$	-.0881 (.0094)	-.0889 (.0065)	-.0859 (.0065)	-.0847 (.0064)
σ_u	.2395	.2388	.2393	.2382
σ_e	.2907	.2903	.2909	.2907

Table 3.8: Estimates of LMM and three PLMMs. Standard deviations were estimated from 500 bootstrap samples.

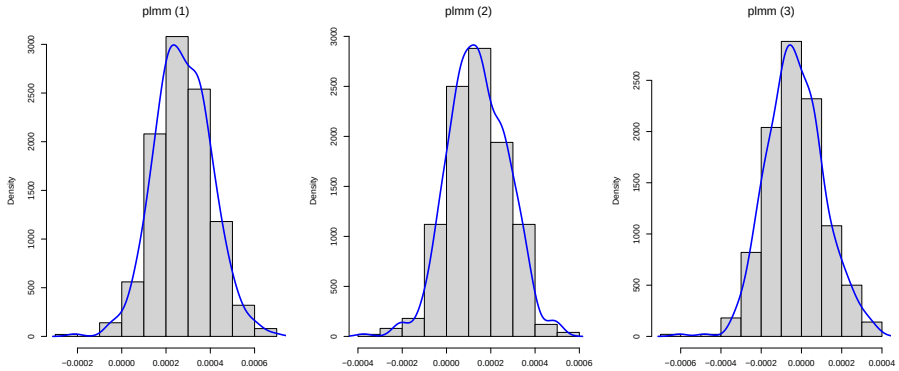
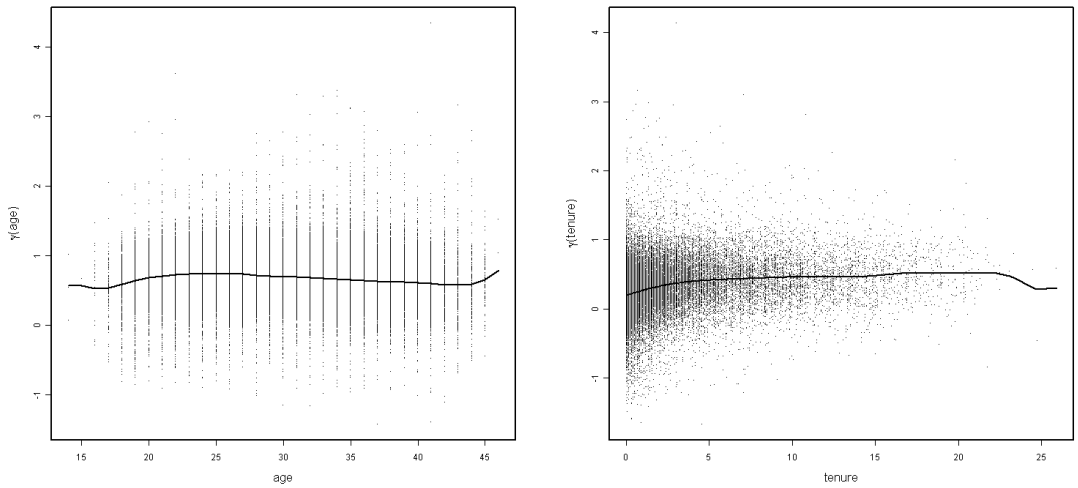


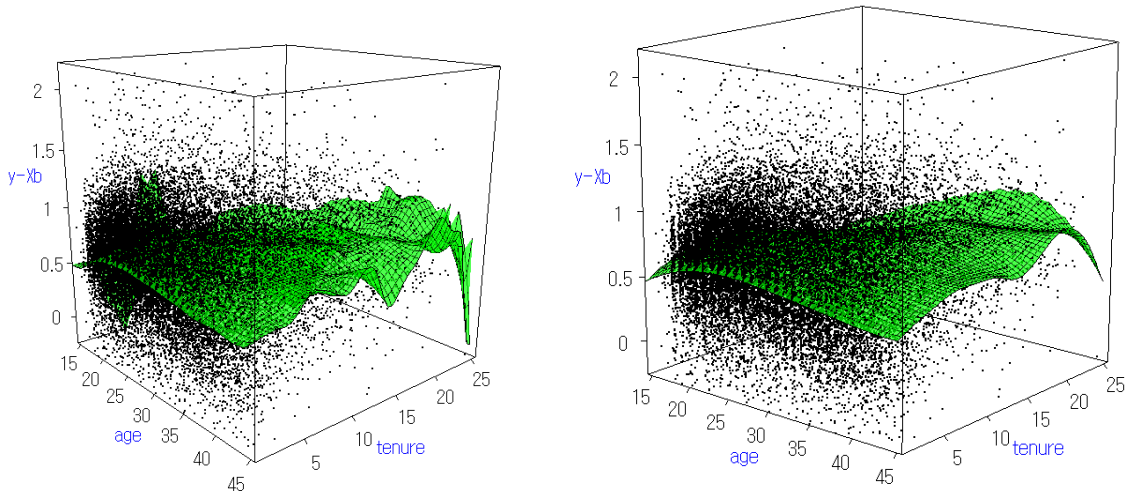
Figure 3.8: Bootstrap sampling distributions of the estimated coefficient of ttl_exp^2 by plmm (1), (2) and (3).

0.1294, $south$ -0.0847.



(a) $\hat{\gamma}(age)$ (b) $\hat{\gamma}(tenure)$

Figure 3.9: Plots of the estimated nonparametric functions.



(a) $\mathbf{h} = (0.880, 1.572)$ (b) $\mathbf{h} = (1.764, 3.150)$

Figure 3.10: Plots of the estimated nonparametric function $\hat{\gamma}(age, tenure)$.

The results suggest that the parametric specification of *age* and *tenure* in (3.61) were possibly inappropriate. On rejection of the null hypothesis by the Hausman

test, a mixed effects model is often replaced by an alternative fixed effects model with cluster-specific dummy variables; a fixed effects model will be used all the more when it is difficult to find an appropriate functional specification to avoid endogeneity. However, inference based on fixed effects models is limited to the sample. In addition, they incur the incidental parameter problem, that is, inconsistency of dummy coefficient estimators unless the cluster sizes increase. Even though a Hausman type statistical test needs to be conducted on a PLMM, the results above indicate an alternative approach to a simple fixed effects model "solution" when correlation between regressors and random effects are suspected.

3.8.2 Application 2: Health expenditure

This section compares two partially linear random intercept models: our proposal and the one proposed by You et al. (2010). They provided not only estimation procedures but also testing methods for random effects specification as well as heteroskedasticity of the regression error. While their model specification is the same as (3.1), which is reduced to an LMM given by (3.12), their succeeding estimation procedures differ from ours in certain ways. Below is a summary of notable differences.

- Fixed component parameters $\boldsymbol{\beta}$ are estimated from a reduced model by the "semiparametric least squares estimation" (SLSE), which is an OLS estimator $\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}_0^\top \tilde{\mathbf{X}}_0)^{-1} \tilde{\mathbf{X}}_0^\top \tilde{\mathbf{y}}_0$ as apposed to the feasible GLS estimator (3.15). Following the SLSE, nonparametric component γ is estimated by the local linear kernel regression $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \boldsymbol{\gamma}(\mathbf{T}) + \mathbf{v}$.
- After the fixed components estimation, the VCs are estimated nonparametrically. Whether regression errors are homoskedastic or heteroskedastic, the variance function is estimated with the method proposed by Li and Stengos (1994) from the residuals $\hat{\mathbf{v}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\gamma}}(\mathbf{T})$ and the estimate of σ_u^2 by (3.45).
- The bandwidths used for model reduction as well as nonparametric γ function estimation are selected by the so-called "leave one block out cross-validation". The bandwidths for the conditional variance function estimation were obtained by the rule-of-thumb method proposed by Yu and Jones (2004).

It should be emphasized that their SLSE relies on the OLS estimator, which can have serious consequences as we will shortly see. You et al. (2010) illustrated their estimation procedures by analyzing Australian medical expenditure data. The data set is a balanced random sample of 200 individuals collected annually over 5 years;

each individual forms a cluster with $n_i = 5$.[25] Table 3.9 describes the variables used in the analysis. Correlations between regressors (*linc*, *age*, *insur*) displayed in Table 3.10 suggest little linear dependence between them.

medexp	response variable: annual medical expenditure in hundreds of dollars
linc	logarithm of annual income in thousands of dollars
age	age in years
insur	binary variable: 1 if the individual has private health insurance

Table 3.9: Description of the variables. Source: Hill et al. (2008), p.414.

	linc	*age*	*insur*
medexp	.089	.665	.284
linc		.080	-.091
age			.073

Table 3.10: Correlations between variables.

A parametric random intercept model is specified follows:

$$medexp_{ij} = \beta_0 + \beta_1 linc_{ij} + \beta_2 age_{ij} + \beta_3 age_{ij}^2 + \beta_4 insur_{ij} + u_i + e_{ij} \ . \qquad (3.62)$$

The estimated model parameters are displayed in the first column (LMM) of Table 3.11. The effect of *linc* was statistically insignificant; the effect of *age* was statistically significant taking a quadratic convex form; and *insur* was also significant.

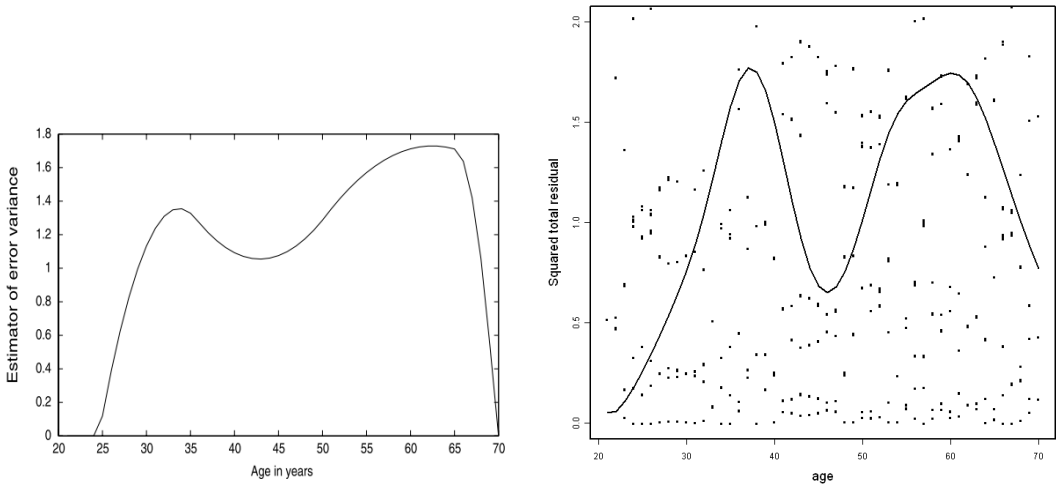
You et al. (2010) analyzed the data using the PLMM with the effect of *age* modeled as a nonparametric function. The second column of Table 3.11 shows the estimates by SLSE. In contrast to the insignificant effect obtained by the LMM, the coefficient of *linc* turned out significantly positive. The authors further tested the hypothesis of homoskedastic variance of the regression error and obtained rejection of the hypothesis. Figure 3.11 (a) is their estimate of the regression error variance function. They reestimated the model by "weighed semiparametric least squares estimation" (WSLSE), which is based on an estimated variance function through the Li and Stengos method. The results are given in the third column (WSLSE). The sign and significance of the coefficients remained approximately the same as the SLSE.[26]

[25] Data are from Hill et al. (2008).

[26] The strongly positive and highly significant effect is little credible for the Australian health (insurance) system. One would rather expect no significance since the difference in expenditures due to income does not make much sense once insurants have decided on either compulsory insurance only or private insurance additionally.

	LMM	SLSE	WSLSE	plmm (1)	plmm (2)
linc	-.1494 (.2802)	.5624 (.0297)	.6137 (.0193)	-.0102 (.2811)	-.0137 (.2856)
age	-.0903 (.0458)				
age^2	.0024 (.0005)				
insur	1.3621 (.0960)	1.5830 (.1619)	1.2129 (.0934)	1.4161 (.0948)	1.4743 (.1315)
σ_u	1.615	1.511	1.511	1.500	1.500
σ_e	1.031	NA	NA	1.078	NA

Table 3.11: Estimates of the coefficients and VCs. Standard error in parentheses were estimated by bootstrap resampling for LMM, plmm (1) and plmm (2). Analytic asymptotic standard errors provided by You et al. (2010) are given for SLSE and WSLSE.

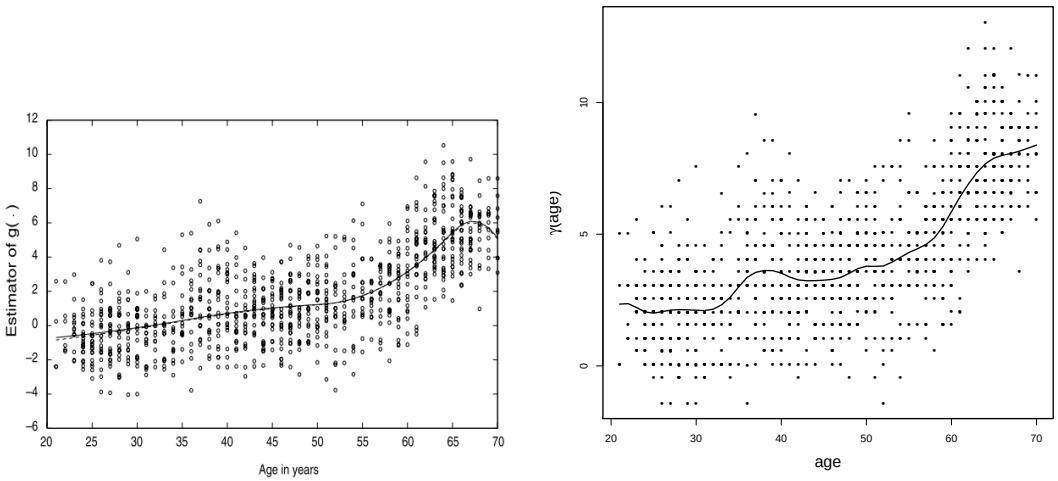


(a) Source: You et al. (2010) p.1100 (b) plmm(1)

Figure 3.11: Estimated conditional variance function.

Figure 3.12 (a) is a plot of the nonparametric function of *age* estimated by You et al. (2010). Difference in significance of *linc* between the LMM and WSLSE (or SLSE) would suggest that the LMM (3.62) was misspecified with respect to the functional form of *age*. However, the shape of the estimated function appears perplexing. Judging from the figure, the quadratic functional form of *age* in the LMM does not appear inappropriate. A quadratic function is almost what the nonparametric estimate suggests as a parametric form, and thus the nonparametric estimation of the function of *age* could not explain the difference in the estimated coefficient of *linc* between the LMM and the SLSE (or WSLSE). In addition, given the relatively

low correlations between regressors, estimation of the coefficient of *linc* is unlikely to be affected so much by misspecification as implied by the results of SLSE and WSLSE.



(a) Source: You et al. (2010) p. 1100 (b) By plmm (2) with bandwidth 1.456

Figure 3.12: Estimate of nonparametric function $\gamma(age)$.

We applied our PLMM proposal to the same semiparametric model specification as in You et al. (2010). We obtained the estimates for the homoskedastic error model shown in the fourth column (plmm (1)) of Table 3.11. Note that the estimates coincided with those of the LMM in sign and insignificance. Using the residuals obtained from plmm (1), we also estimated the heteroskedastic variance function by the NW kernel regression. The estimate is displayed in Figure 3.11 (b). Both figures in Figure 3.11 capture heteroskedasticity as a function of *age* and detect a bimodal shape with a valley between the modes around the middle of age 40 and 50.[27] We reestimated the model using the variance function estimate. The coefficient estimates in the fifth column (plmm (2)) remained similar to those of plmm (1) in terms of sign and significance. Figure 3.12 (b) displays the local linear kernel regression estimate of $\gamma(age)$ with the bandwidth selected using binning techniques. The shape of the estimated function would also suggest a quadratic function.

To investigate the cause of the difference between our results and those of You et al. (2010), we estimated the fixed component coefficients by OLS (with $\hat{\mathbf{V}}_{(0)}^{-1} = \mathbf{I}_N$ in

[27] The range above 2 along the vertical axis is cut off. The entire figure is given in Appendix B.

(3.15)) without subsequent iteration. We obtained estimates of 0.6050 and 1.5910 for *linc* and *insur*, respectively, and 1.477 for the estimate of σ_u by (3.45). These estimates are all close to the results of the SLSE. We further examined the updating process of the estimates by plmm (1). Table 3.12 clearly shows that the estimates were far apart from the above OLS estimates already at the initial stage. This disagreement of the estimates is plausible as is suggested by the simulation study in Section 3.7, where we observed a larger MSE of OLS estimators.

	Initial	1. iteration	2. iteration	3. iteration
linc	-.00006	-.00628	-.00830	-.01019
insur	1.37668	1.41503	1.41606	1.41605

Table 3.12: Updating process of the plmm (1) estimates.

A histogram of the random effects predictions and a summary of bootstrap estimates of the sampling distributions of the $\boldsymbol{\beta}$ estimators are given in Appendix B.

To complete this section, Table 3.13 presents the bandwidths selected by different CV methods for the estimation of the nonparametric component γ. The function γ is estimated by local linear kernel regression (and the variance function by NW kernel regression). For both plmm (1) and plmm (2), the extended GCV (GCVc) by Carmack et al. (2011) selected larger bandwidths than those by other CV methods. Figure 3.13 shows the estimate of $\gamma(age)$ for plmm (2) obtained using GCVc, which appears smoother and more appropriate than the estimate in Figure 3.12 (b).

	CV(1)	CV(2)	GCV	GCVc
plmm (1)	1.454	2.049	2.031	2.561
plmm (2)	1.456	2.049	2.032	2.495

Table 3.13: Bandwidths selected by different CV methods. CV(1) is CV with binning and CV(2) is the leave-one-out CV.

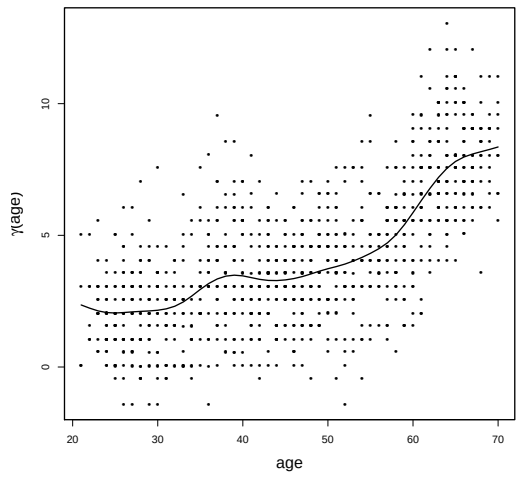


Figure 3.13: Estimate of nonparametric function $\gamma(age)$ by plmm(2) using the bandwidth 2.495 obtained by GCVc.

3.9 Concluding Remarks

In this essay we have considered the PLMM, specifically, the partially linear random intercept model without distributional assumptions for the random terms of the model and introduced an iterative estimation procedure. We showed by simulation that our iterative procedure worked well even for relatively small samples and yielded consistent estimators. Simulation studies also showed that iterative procedures alleviate concerns about the potential effects that the bandwidth selection for model reduction of the PLMM to the LMM may have on the succeeding estimation. It was also demonstrated that, if regression errors are heteroskedastic, the PLMM estimation using a nonparametrically estimated variance function improves the mean squared error. We illustrated the PLMM by analyzing two real data sets. We first investigated an empirical case in which coefficients estimated by the LMM were not amenable to economic interpretation. On rejection of the null hypothesis by the Hausman test, we fit a PLMM as an alternative approach to model misspecification and obtained interpretable coefficient estimates. In the other data analysis, we compared our PLMM estimation with the one proposed by You et al. (2010). The result indicated that our iterative GLS estimation yields more reasonable coefficient estimates than their OLS-type semiparametric estimation. For this application, we obtained more sensible bandwidths using the extended GCV to prevent undersmoothing due to correlations in data.

We have provided detailed derivations of estimators under different sets of assump-

tions on the variance structure, data adaptive choice of bandwidths, efficient implementation through binning, matrix decomposition and model transformation to reduce computational burdens. We also offered a resampling scheme for subsequent statistical inference. All the procedures are implemented in the statistical software R.

Our empirical applications point to directions of further research, in particular, two types of specification test for the PLMM. One is an extension of the Hausman test applied to the LMM. In the NLSY data example, even though the PLMM yielded an interpretable model estimate as opposed to the LMM, its specification with random effects still needs to be tested. The other specification test is with respect to the appropriateness of a parametric function instead of nonparametric one. In practice, it will be of interest whether a parametrically specified function is acceptable as an alternative to nonparametric function. In the medical data example, it was left untested whether the parametric function specified in the LMM suffices in place of the nonparametric function. In the partially linear model framework, Härdle and Mammen (1993) proposed a statistic based on the wild bootstrap to test whether the parametric function estimate is significantly different from the nonparametric estimate. The idea has been extended in Sperlich and Lombardía (2010). This type of specification test in the PLMM framework is another issue of future research.

One immediate extension of the partially linear random intercept model is a partially linear random coefficient model that models not only the intercept but also slope coefficients as cluster-specific random variables. This will be an semiparametric extension of Swamy (1970). Moreover, researchers are often interested in whether the form of an unknown smooth function differs between some fixed groups. Another natural extension will thus be to model an interaction between a discrete factor variable and the nonparametric function.

Concerning the use of kernel regression, we used a global bandwidth. As one sees in Figure 3.10 (a), the bandwidth selected is too small in boundary regions of the support. An extension that allows the use of local bandwidths will help to avoid local undersmoothing. In practice, one could also look for adequate prior transformations of the regressors that enter the model nonparametrically, and then use a global bandwidth.

Another extension possibility is application of Vilar Fernández and Francisco Fernández (2002). They proposed a nonparametric kernel regression to improve asymptotic efficiency based on the standard GLS concept. Recall that, while we use the feasible GLS estimation for $\boldsymbol{\beta}$, we estimated nonparametric function γ with the standard kernel regression ignoring the correlation structure of the errors. Application of their

proposal will produce an estimator given by

$$\hat{\gamma}(\mathbf{t}) = \mathbf{e}_1^\top \left(\mathbf{T}_\mathbf{t}^{*\top}\boldsymbol{\Omega}^\top\mathbf{K}\boldsymbol{\Omega}\mathbf{T}_\mathbf{t}^*\right)^{-1}\mathbf{T}_\mathbf{t}^{*\top}\boldsymbol{\Omega}^\top\mathbf{K}\boldsymbol{\Omega}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \ , \tag{3.63}$$

where $\boldsymbol{\Omega}$ is such that $\boldsymbol{\Omega}^\top\boldsymbol{\Omega} = \mathbf{V}^{-1}$. In fact, (3.24) can be expressed as a special case of (3.63) with $\boldsymbol{\Omega} = \mathbf{I}$. The estimator was originally proposed for the nonparametric kernel regression with an auto-regressive error structure.[28] The authors reported improvement in the mean integrated squared error through simulation. We conducted simulation studies in which we experimented with different decompositions of $\mathbf{V}^{-1}$ to construct Ω (obviously, Ω is not unique). However, we haven't observed promising results yet. Regarding the use of a GLS type estimator, Ruckstuhl et al. (2000) and Lin and Carroll (2000) argued that incorporation of the correlation structure would not lead to asymptotically more efficient estimation. On the other hand, Su and Ullah (2007) proposed a different way of incorporating the correlation structure to achieve higher efficiency over the standard kernel regression.

Finally, our future research will address the small area estimation, where the estimation of the mean squared prediction error of the response is of paramount importance. In particular, application of the PLMM to offer reliable, easy-to-handle estimators for area specific parameters and their confidence/prediction intervals will be one of the focal points of our research.

3.10 Appendix A

3.10.1 Calculation of $\mathbf{V}^{-1}$ by spectral decomposition

Homoskedastic case $(u_i + e_{ij})$:

$$\mathbf{V}_i^{-1} = (n_i\sigma_u^2 + \sigma_e^2)^{-1}\bar{\mathbf{J}}_{n_i} + \sigma_e^{-2}(\mathbf{I}_{n_i} - \bar{\mathbf{J}}_{n_i}) \tag{3.64}$$

$$\mathbf{V}^{-1} = (\oplus(n_i\sigma_u^2 + \sigma_e^2)^{-1}) \otimes \bar{\mathbf{J}}_{n_i} + \sigma_e^{-2}\mathbf{I}_m \otimes (\mathbf{I}_{n_i} - \bar{\mathbf{J}}_{n_i}) \ . \tag{3.65}$$

Between-cluster heteroskedastic case $(u_i + \alpha_i e_{ij})$:

$$\mathbf{V}_i^{-1} = (n_i\sigma_u^2 + \alpha_i^2\sigma_e^2)^{-1}\bar{\mathbf{J}}_{n_i} + (\alpha_i^2\sigma_e^2)^{-1}(\mathbf{I}_{n_i} - \bar{\mathbf{J}}_{n_i}) \tag{3.66}$$

$$\mathbf{V}^{-1} = (\oplus(n_i\sigma_u^2 + \alpha_i^2\sigma_e^2)^{-1}) \otimes \bar{\mathbf{J}}_{n_i} + (\oplus(\alpha_i^2\sigma_e^2)^{-1}) \otimes (\mathbf{I}_{n_i} - \bar{\mathbf{J}}_{n_i}) \ . \tag{3.67}$$

[28] For a review, see González-Manteiga et al., 2012 among others.

Within-cluster heteroskedastic case ($u_i + \alpha_{ij}e_{ij}$): because of the structure of $\mathbf{V}_i$

$$\begin{aligned}\mathbf{V}_i &= n_i\sigma_u^2\bar{\mathbf{J}}_{n_i} + \oplus_j\alpha_{ij}^2\sigma_e^2 \\ &= \oplus_j(n_i\sigma_u^2 + \alpha_{ij}^2\sigma_e^2)\bar{\mathbf{J}}_{n_i} + \oplus_j\alpha_{ij}^2\sigma_e^2(\mathbf{I}_{n_i} - \bar{\mathbf{J}}_{n_i}) \ , \end{aligned} \tag{3.68}$$

such a spectral representation as the other cases is not available except for the variance of transformed data $(\oplus\alpha_{ij}^{-1})\mathbf{y}_i$. Due to the idempotency of $\eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top$ and $\mathbf{I}_{n_i} - \eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top$,

$$\begin{aligned}\mathrm{Var}[(\oplus\alpha_{ij}^{-1})\mathbf{y}_i] &= (\eta_i\sigma_u^2 + \sigma_e^2)\eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top + \\ &\quad \sigma_e^2(\mathbf{I}_{n_i} - \eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top) \end{aligned} \tag{3.69}$$

$$\begin{aligned}(\mathrm{Var}[(\oplus\alpha_{ij}^{-1})\mathbf{y}_i])^{-1} &= (\eta_i\sigma_u^2 + \sigma_e^2)^{-1}\eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top + \\ &\quad \sigma_e^{-2}(\mathbf{I}_{n_i} - \eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top) \ , \end{aligned} \tag{3.70}$$

where $\boldsymbol{\alpha}_i^{-1} = (\alpha_{i1}^{-1}\dots\alpha_{in_i}^{-1})^\top$ and $\eta_i = (\boldsymbol{\alpha}_i^{-1})^\top\boldsymbol{\alpha}_i^{-1} = \sum_{j=1}^{n_i}\alpha_{ij}^{-2}$. Since $(\mathrm{Var}[(\oplus\alpha_{ij}^{-1})\mathbf{y}_i])^{-1}$ $= \oplus\alpha_{ij}\mathbf{V}_i^{-1}\oplus\alpha_{ij}$,

$$\begin{aligned}\mathbf{V}_i^{-1} &= \oplus\alpha_{ij}^{-1}\left((\eta_i\sigma_u^2 + \sigma_e^2)^{-1}\eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top + \sigma_e^{-2}\left(\mathbf{I}_{n_i} - \eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top\right)\right)\oplus\alpha_{ij}^{-1} \\ &= \oplus\alpha_{ij}^{-1}\left(\frac{1}{\sigma_e^2}\left(\mathbf{I}_{n_i} - \frac{\eta_i\sigma_u^2}{\eta_i\sigma_u^2 + \sigma_e^2}\eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top\right)\right)\oplus\alpha_{ij}^{-1} \\ &= \frac{1}{\sigma_e^2}\left(\oplus\alpha_{ij}^{-2} - \frac{\sigma_u^2}{\eta_i\sigma_u^2 + \sigma_e^2}(\alpha_{i1}^{-2},\dots,\alpha_{in_i}^{-2})^\top(\alpha_{i1}^{-2},\dots,\alpha_{in_i}^{-2})\right) \\ &= \oplus\nu_{ij}^{-2} - \frac{\sigma_u^2}{\sigma_e^{-2}\eta_i\sigma_u^2 + 1}\frac{1}{\sigma_e^2}(\alpha_{i1}^{-2},\dots,\alpha_{in_i}^{-2})^\top\frac{1}{\sigma_e^2}(\alpha_{i1}^{-2},\dots,\alpha_{in_i}^{-2}) \\ &= \oplus\nu_{ij}^{-2} - \frac{\sigma_u^2}{\sigma_u^2\sum\nu_{ij}^{-2} + 1}(\nu_{i1}^{-2},\dots,\nu_{in_i}^{-2})^\top(\nu_{i1}^{-2},\dots,\nu_{in_i}^{-2}) \ , \end{aligned} \tag{3.71}$$

where $\nu_{ij}^2 = \alpha_{ij}^2\sigma_e^2$.

3.10.2 Derivation of VC estimators in the homoskedastic case

This section presents the derivation of the homoskedastic VC estimators used in the initial stage and the iterative process. The VCs are estimated for an LMM specified with working variables $\tilde{y}$ and $\tilde{\mathbf{x}}$ as

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \ . \tag{3.72}$$

Note that $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$ change according to the estimation stage: in the initial stage $\tilde{\mathbf{y}}$ is a vector whose ijth element is $y_{ij} - \hat{\mathrm{E}}[y_{ij}|\mathbf{t}_{ij}]$ and $\tilde{\mathbf{X}}$ is a matrix whose ijth row

is $\mathbf{x}_{ij}^{\top} - \hat{\mathrm{E}}[\mathbf{x}_{ij}^{\top}|\mathbf{t}_{ij}]$. It holds that $\mathbf{Q}_i\mathbf{1}_{n_i} = \mathbf{0}$ and $\mathbf{QZ} = \mathbf{0}$ (recall $\mathbf{Q}_i = \mathbf{I}_{n_i} - \bar{\mathbf{J}}_{n_i}$, $\mathbf{Q} = \mathbf{I}_m \otimes \mathbf{Q}_i$ and $\mathbf{P} = \mathbf{I}_m \otimes \bar{\mathbf{J}}_{n_i}$). Transforming (3.72) with $\mathbf{Q}$ yields

$$\begin{aligned} \mathbf{Q}\tilde{\mathbf{y}} &= \mathbf{Q}\tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{QZu} + \mathbf{Qe} \\ &= \mathbf{Q}\tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{Qe} \ . \end{aligned} \tag{3.73}$$

Let $\hat{\boldsymbol{\beta}}_w$ be the within-estimator (LSDV estimator) for regression (3.73). The within-residual vector denoted here by $\hat{\mathbf{e}}_w$ is given by

$$\begin{aligned} \hat{\mathbf{e}}_w &= \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_w \\ &= (\mathbf{I}_N - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^{\top}\mathbf{Q}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\top}\mathbf{Q})\tilde{\mathbf{y}} \\ &= (\mathbf{I}_N - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^{\top}\mathbf{Q}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\top}\mathbf{Q})\mathbf{e} \ . \end{aligned} \tag{3.74}$$

Since the trace of idempotent matrix $(\mathbf{Q} - \mathbf{Q}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^{\top}\mathbf{Q}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\top}\mathbf{Q})$ is $N - m - r(\mathbf{Q}\tilde{\mathbf{X}})$ and $\mathrm{E}[\mathbf{ee}^{\top}] = \sigma_e^2\mathbf{I}_N$, the quadratic estimator of σ_e^2 is obtained as follows:

$$\begin{aligned} \mathrm{E}[\hat{\mathbf{e}}_w^{\top}\mathbf{Q}\hat{\mathbf{e}}_w] &= \mathrm{E}[\mathbf{e}^{\top}(\mathbf{Q} - \mathbf{Q}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^{\top}\mathbf{Q}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\top}\mathbf{Q})\mathbf{e}] \\ &= \mathrm{tr}\left((\mathbf{Q} - \mathbf{Q}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^{\top}\mathbf{Q}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\top}\mathbf{Q})\mathrm{E}[\mathbf{ee}^{\top}]\right) \\ &= \left(N - m - r(\mathbf{Q}\tilde{\mathbf{X}})\right)\sigma_e^2 \end{aligned} \tag{3.75}$$

$$\therefore\ \hat{\sigma}_e^2 = \frac{\hat{\mathbf{e}}_w^{\top}\mathbf{Q}\hat{\mathbf{e}}_w}{N - m - r(\mathbf{Q}\tilde{\mathbf{X}})} \ , \tag{3.76}$$

where $r(\bullet)$ denotes the rank of $\bullet$.

FC estimator of σ_u^2: Let $\hat{\mathbf{v}}$ be a vector of OLS residuals from $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \mathbf{v}$, where $\mathbf{v} = \mathbf{Zu} + \mathbf{e}$. It holds that

$$\begin{aligned} \hat{\mathbf{v}} &= \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_{ols} \\ &= (\mathbf{I}_N - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\top})\tilde{\mathbf{y}} \\ &= (\mathbf{I}_N - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\top})\mathbf{v} \ . \end{aligned} \tag{3.77}$$

Then it follows that

$$\begin{aligned}
\mathrm{E}[\hat{\mathbf{v}}^\top\hat{\mathbf{v}}] &= \mathrm{E}[\mathbf{v}^\top(\mathbf{I}_N - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top)\mathbf{v}] \\
&= \mathrm{tr}\left((\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top)\mathrm{E}[\mathbf{v}\mathbf{v}^\top]\right) \\
&= \mathrm{tr}\left((\mathbf{I}-\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top)(\sigma_u^2\mathbf{Z}\mathbf{Z}^\top + \sigma_e^2\mathbf{I}_N)\right) \\
&= \sigma_u^2\mathrm{tr}(\mathbf{Z}\mathbf{Z}^\top) - \sigma_u^2\mathrm{tr}\left((\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{Z}\mathbf{Z}^\top\tilde{\mathbf{X}}\right) \\
&\quad -\sigma_e^2\mathrm{tr}(\mathbf{I}_N-\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top)\ . \qquad (3.78)
\end{aligned}$$

The first, second and third terms of (3.78) have respective traces of

$$\begin{aligned}
\mathrm{tr}(\mathbf{Z}\mathbf{Z}^\top) &= \mathrm{tr}(\mathbf{I}_m \otimes_i \mathbf{J}_{n_i}) \\
&= N \qquad (3.79) \\
\mathrm{tr}\left((\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{Z}\mathbf{Z}^\top\tilde{\mathbf{X}}\right) &= \mathrm{tr}\left((\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top(\mathbf{I}_m \otimes_i \mathbf{1}_{n_i})(\mathbf{I}_m \otimes_i \mathbf{1}_{n_i}^\top)\tilde{\mathbf{X}}\right) \\
&= \mathrm{tr}((\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\sum_{i=1}^m n_i^2\bar{\tilde{\mathbf{x}}}_i\bar{\tilde{\mathbf{x}}}_i^\top) \qquad (3.80) \\
\mathrm{tr}(\mathbf{I}_N-\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top) &= N-p\ , \qquad (3.81)
\end{aligned}$$

where $\bar{\tilde{\mathbf{x}}}_i$ is a p-dimensional vector of the cluster means of the regressors $\tilde{\mathbf{x}}_i$. Inserting (3.79), (3.80) and (3.81) in (3.78) yields

$$\mathrm{E}[\hat{\mathbf{v}}^\top\hat{\mathbf{v}}] = N\sigma_u^2 - \sigma_u^2\mathrm{tr}((\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\sum_{i=1}^m n_i^2\bar{\tilde{\mathbf{x}}}_i\bar{\tilde{\mathbf{x}}}_i^\top) - (N-p)\sigma_e^2\ . \qquad (3.82)$$

The FC estimator of σ_u^2 is thus given by

$$\hat{\sigma}_{u,\mathrm{FC}} = \max\left(\frac{\hat{\mathbf{v}}^\top\hat{\mathbf{v}} - (N-p)\hat{\sigma}_e^2}{N - \mathrm{tr}\left((\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\sum_{i=1}^m n_i^2\bar{\tilde{\mathbf{x}}}_i\bar{\tilde{\mathbf{x}}}_i^\top\right)}, 0\right)\ . \qquad (3.83)$$

SA estimator of σ_u^2: With the between-estimator denoted by $\hat{\boldsymbol{\beta}}_b$ for regression (3.72), the vector of between-residuals denoted here by $\hat{\mathbf{v}}_b$ is given by

$$\begin{aligned}
\hat{\mathbf{v}}_b &= \tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}_b \\
&= (\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\mathbf{P}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{P})\tilde{\mathbf{y}} \\
&= (\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\mathbf{P}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\mathbf{P})\,(\mathbf{Z}\mathbf{u} + \mathbf{e})\ .
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathrm{E}[\hat{\mathbf{v}}_b^\top \mathbf{P} \hat{\mathbf{v}}_b] &= \mathrm{E}[(\mathbf{Z}\mathbf{u}+\mathbf{e})^\top (\mathbf{P} - \mathbf{P}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top \mathbf{P}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \mathbf{P})(\mathbf{Z}\mathbf{u}+\mathbf{e})] \\
&= \mathrm{tr}\left((\mathbf{P} - \mathbf{P}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top \mathbf{P}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \mathbf{P})\mathrm{E}[(\mathbf{Z}\mathbf{u}+\mathbf{e})(\mathbf{Z}\mathbf{u}+\mathbf{e})^\top]\right) \\
&= \mathrm{tr}\left((\mathbf{P} - \mathbf{P}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top \mathbf{P}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \mathbf{P})(\sigma_u^2 \mathbf{Z}\mathbf{Z}^\top + \sigma_e^2 \mathbf{I}_n)\right) \\
&= \sigma_u^2 \mathrm{tr}(\mathbf{P}\mathbf{Z}\mathbf{Z}^\top) - \sigma_u^2 \mathrm{tr}(\mathbf{P}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top \mathbf{P}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \mathbf{P}\mathbf{Z}\mathbf{Z}^\top) \\
&\quad -\sigma_e^2 \mathrm{tr}(\mathbf{P} - \mathbf{P}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top \mathbf{P}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \mathbf{P}) \; . \qquad (3.84)
\end{aligned}$$

The first, second and third terms have respective traces of

$$\begin{aligned}
\mathrm{tr}(\mathbf{P}\mathbf{Z}\mathbf{Z}^\top) &= \mathrm{tr}(\mathbf{Z}\mathbf{Z}^\top) \\
&= N \qquad (3.85) \\
\mathrm{tr}(\mathbf{P}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top \mathbf{P}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \mathbf{P}\mathbf{Z}\mathbf{Z}^\top) &= \mathrm{tr}((\tilde{\mathbf{X}}^\top \mathbf{P}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \mathbf{Z}\mathbf{Z}^\top \tilde{\mathbf{X}}) \\
&= \mathrm{tr}((\tilde{\mathbf{X}}^\top \mathbf{P}\tilde{\mathbf{X}})^{-1} \sum_{i=1}^{m} n_i^2 \bar{\tilde{\mathbf{x}}}_i \bar{\tilde{\mathbf{x}}}_i^\top) \qquad (3.86) \\
\mathrm{tr}(\mathbf{P} - \mathbf{P}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top \mathbf{P}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \mathbf{P}) &= m - p \; . \qquad (3.87)
\end{aligned}$$

Inserting (3.85), (3.86) and (3.87) in (3.84) gives

$$\mathrm{E}[\hat{\mathbf{v}}_b^\top \mathbf{P} \hat{\mathbf{v}}_b] = N\sigma_u^2 - \sigma_u^2 \mathrm{tr}((\tilde{\mathbf{X}}^\top \mathbf{P}\tilde{\mathbf{X}})^{-1} \sum_{i=1}^{m} n_i^2 \bar{\tilde{\mathbf{x}}}_i \bar{\tilde{\mathbf{x}}}_i^\top) + (m-p)\,\sigma_e^2 \; . \qquad (3.88)$$

Thus the SA estimator of σ_u^2 is given by

$$\hat{\sigma}_{u,\mathrm{SA}}^2 = \max\left(\frac{\hat{\mathbf{v}}_b^\top \mathbf{P} \hat{\mathbf{v}}_b - (m-p)\hat{\sigma}_e^2}{N - \mathrm{tr}\left((\tilde{\mathbf{X}}^\top \mathbf{P}\tilde{\mathbf{X}})^{-1} \sum_{i=1}^{m} n_i^2 \bar{\tilde{\mathbf{x}}}_i \bar{\tilde{\mathbf{x}}}_i^\top\right)}, 0 \right) \; . \qquad (3.89)$$

Note that, in the iteration stage, estimators are obtained with the working variables $\tilde{y}$ and $\tilde{\mathbf{x}}$ in (3.72) modified as follows: $\tilde{\mathbf{y}}$ is a vector whose ijth element is $y_{ij} - \hat{\gamma}(\mathbf{t}_{ij})$ and $\tilde{\mathbf{X}}$ is the same as the original design $\mathbf{X}$.

By taking into account the loss of d.f. due to the estimation of γ, the estimator of σ_e^2 is given by

$$\hat{\sigma}_{e(r)}^2 = \frac{\hat{\mathbf{e}}_{w(r)}^\top \mathbf{Q} \hat{\mathbf{e}}_{w(r)}}{N - m - r(\mathbf{Q}\mathbf{X}) - \mathrm{d.f.}(\hat{\boldsymbol{\gamma}}_{(r-1)})} \; , \qquad (3.90)$$

where the subscript (r) indicates the rth iteration ($\hat{\boldsymbol{\gamma}}_{(0)}$ is the estimator of γ in the initial stage). $r(\mathbf{Q}\mathbf{X})$ is p if none of the regressors in $\mathbf{x}$ is invariant within the cluster.

The d.f. of the nonparametric estimator of γ is defined here as

$$\text{d.f.}(\hat{\boldsymbol{\gamma}}_{(r)}) = 2\text{tr}(\mathbf{S}_{(r)}) - \text{tr}(\mathbf{S}_{(r)}\mathbf{S}_{(r)}^\top) \ , \tag{3.91}$$

where $\mathbf{S}_{(r)}$ is the smoother matrix for γ estimation. Alternatively, the d.f. can be defined following Carmack et al. (2011) by

$$\text{d.f.}(\hat{\boldsymbol{\gamma}}_{(r)}) = 2\text{tr}(\mathbf{S}_{(r)}\mathbf{R}_{(r)}) - \text{tr}(\mathbf{S}_{(r)}\mathbf{R}_{(r)}\mathbf{S}_{(r)}^\top) \ , \tag{3.92}$$

where $\mathbf{R}_{(r)}$ is the correlation matrix of the random terms. See Appendix A 3.10.5 for more details.

Replacing the estimator of σ_e^2 in (3.83) and (3.89) by (3.90) yields the estimators in the rth iteration:

$$\hat{\sigma}_{u(r),\text{FC}} = \max\left(\frac{\hat{\mathbf{v}}_{(r)}^\top\hat{\mathbf{v}}_{(r)} - (N-p)\hat{\sigma}_{e(r)}^2}{N - \text{tr}\left((\mathbf{X}^\top\mathbf{X})^{-1}\sum_{i=1}^m n_i^2\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top\right)}, 0\right) \tag{3.93}$$

$$\hat{\sigma}_{u(r),\text{SA}}^2 = \max\left(\frac{\hat{\mathbf{v}}_{b(r)}^\top\mathbf{P}\hat{\mathbf{v}}_{b(r)} - (m-p)\hat{\sigma}_{e(r)}^2}{N - \text{tr}\left((\mathbf{X}^\top\mathbf{P}\mathbf{X})^{-1}\sum_{i=1}^m n_i^2\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i^\top\right)}, 0\right) \ . \tag{3.94}$$

3.10.3 Derivation of VC estimators in the heteroskedastic case

This section presents the derivation of the heteroskedastic VC estimators used in the initial and iteration stages. Note that working variables $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{X}}$ change according to the estimation stage. In the initial stage $\tilde{\mathbf{y}}$ is a vector whose ijth element is $y_{ij} - \hat{\text{E}}[y_{ij}|\mathbf{t}_{ij}]$ and $\tilde{\mathbf{X}}$ is a matrix whose ijth row is $\mathbf{x}_{ij}^\top - \hat{\text{E}}[\mathbf{x}_{ij}^\top|\mathbf{t}_{ij}]$. Estimation starts with premultiplying the heteroskedastic model $\tilde{y}_{ij} = \tilde{\mathbf{x}}_{ij}^\top\boldsymbol{\beta} + u_i + \alpha_{ij}e_{ij}$ by the inverse of the heteroskedastic parameter α_{ij} to obtain $\alpha_{ij}^{-1}\tilde{y}_{ij} = \alpha_{ij}^{-1}\tilde{\mathbf{x}}_{ij}^\top\boldsymbol{\beta} + \alpha_{ij}^{-1}u_i + e_{ij}$ or in matrix form

$$\begin{aligned}(\oplus_j\alpha_{ij}^{-1})\tilde{\mathbf{y}}_i &= (\oplus_j\alpha_{ij}^{-1})\tilde{\mathbf{X}}_i\boldsymbol{\beta} + u_i(\oplus_j\alpha_{ij}^{-1})\mathbf{1}_{n_i} + \mathbf{e}_i \\ &= (\oplus_j\alpha_{ij}^{-1})\tilde{\mathbf{X}}_i\boldsymbol{\beta} + u_i\boldsymbol{\alpha}_i^{-1} + \mathbf{e}_i \ ,\end{aligned} \tag{3.95}$$

where $\boldsymbol{\alpha}_i^{-1} = (\alpha_{i1}^{-1}\,\alpha_{i2}^{-1}\ldots\alpha_{in_i}^{-1})^\top$. The spectral decomposition representation of $\text{Var}[(\oplus_j\alpha_{ij}^{-1})\mathbf{y}_i]$ is

$$\begin{aligned}\text{Var}[(\oplus_j\alpha_{ij}^{-1})\tilde{\mathbf{y}}_i] &= \sigma_u^2\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top + \sigma_e^2\mathbf{I}_{n_i} \\ &= (\eta_i\sigma_u^2 + \sigma_e^2)\eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top + \sigma_e^2(\mathbf{I}_{n_i} - \eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top) \ ,\end{aligned} \tag{3.96}$$

where $\eta_i = (\boldsymbol{\alpha}_i^{-1})^\top \boldsymbol{\alpha}_i^{-1} = \sum_{j=1}^{n_i} \alpha_{ij}^{-2}$. Rewrite (3.95) as

$$\begin{aligned} \underline{\mathbf{y}}_i &= \underline{\mathbf{X}}_i\boldsymbol{\beta} + u_i\boldsymbol{\alpha}_i^{-1} + \mathbf{e}_i \\ &= \underline{\mathbf{X}}_i\boldsymbol{\beta} + \underline{\mathbf{v}}_i \ , \end{aligned} \tag{3.97}$$

where $\underline{\mathbf{y}}_i = (\oplus_j \alpha_{ij}^{-1})\tilde{\mathbf{y}}_i$, $\underline{\mathbf{X}}_i = (\oplus_j \alpha_{ij}^{-1})\tilde{\mathbf{X}}_i$ and $\underline{\mathbf{v}}_i = u_i\boldsymbol{\alpha}_i^{-1} + \mathbf{e}_i$. (3.97) is further rewritten as

$$\begin{aligned} \underline{\mathbf{y}} &= \underline{\mathbf{X}}\boldsymbol{\beta} + \mathbf{1}_m \otimes_i (u_i\boldsymbol{\alpha}_i^{-1} + \mathbf{e}_i) \\ &= \underline{\mathbf{X}}\boldsymbol{\beta} + \underline{\mathbf{v}} \ . \end{aligned} \tag{3.98}$$

Let $\mathbf{Q}_i$ be $\mathbf{I}_{n_i} - \eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top$, which appears in the second term of (3.96). ($\mathbf{Q}_i$ defined here is generally different from the one in the previous section. The $\mathbf{Q}_i$ of the homoskedastic case is a special case with $\alpha_{ij} = 1 \ \forall ij$.) Let's define $\mathbf{P}_i$ as $\eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top$. Since $\mathbf{P}_i$ is symmetric idempotent, $\mathbf{Q}_i = \mathbf{I}_{n_i} - \mathbf{P}_i$, $\mathbf{P} = \mathbf{I}_m \otimes \mathbf{P}_i$ and $\mathbf{Q} = \mathbf{I}_m \otimes \mathbf{Q}_i$ are all symmetric idempotent.

In order to obtain the within-estimator of $\boldsymbol{\beta}$, we apply the following heteroskedastic within-transformation to (3.97) using $\mathbf{Q}_i$:

$$\begin{aligned} \mathbf{Q}_i\underline{\mathbf{y}}_i &= \mathbf{Q}_i(\underline{\mathbf{X}}_i\boldsymbol{\beta} + u_i\boldsymbol{\alpha}_i^{-1} + \mathbf{e}_i) \\ &= \mathbf{Q}_i\underline{\mathbf{X}}_i\boldsymbol{\beta} + \mathbf{Q}_i\mathbf{e}_i \ . \end{aligned} \tag{3.99}$$

The second equality is due to $u_i\mathbf{Q}_i\mathbf{a}_i^{-1} = u_i(\boldsymbol{\alpha}_i^{-1} - \boldsymbol{\alpha}_i^{-1}) = \mathbf{0}$. This transformation yields analogs to the homoskedastic within-transformation as follows:

$$\begin{aligned} \mathbf{Q}_i(\oplus_j \alpha_{ij}^{-1})\tilde{\mathbf{y}}_i &= (\oplus_j \alpha_{ij}^{-1})\tilde{\mathbf{y}}_i - \eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top(\oplus_j \alpha_{ij}^{-1})\tilde{\mathbf{y}}_i \\ &= (\oplus_j \alpha_{ij}^{-1})\tilde{\mathbf{y}}_i - \bar{y}_i\boldsymbol{\alpha}_i^{-1} \ , \end{aligned} \tag{3.100}$$

where $\bar{y}_i = \frac{\sum_j \alpha_{ij}^{-2}\tilde{y}_{ij}}{\sum_j \alpha_{ij}^{-2}}$;

$$\begin{aligned} \mathbf{Q}_i(\oplus_j \alpha_{ij}^{-1})\tilde{\mathbf{X}}_i &= (\oplus_j \alpha_{ij}^{-1})\tilde{\mathbf{X}}_i - \eta_i^{-1}\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top(\oplus_j \alpha_{ij}^{-1})\tilde{\mathbf{X}}_i \\ &= (\oplus_j \alpha_{ij}^{-1})\tilde{\mathbf{X}}_i - \boldsymbol{\alpha}_i^{-1}\bar{\mathbf{x}}_i^\top \ , \end{aligned} \tag{3.101}$$

where $\bar{\mathbf{x}}_i^\top = \left(\frac{\sum_j \alpha_{ij}^{-2}\tilde{x}_{ij,1}}{\sum_j \alpha_{ij}^{-2}} \ \frac{\sum_j \alpha_{ij}^{-2}\tilde{x}_{ij,2}}{\sum_j \alpha_{ij}^{-2}} \ \dots \ \frac{\sum_j \alpha_{ij}^{-2}\tilde{x}_{ij,p}}{\sum_j \alpha_{ij}^{-2}} \right)$; and

$$\mathbf{Q}_i\mathbf{e}_i = \mathbf{e}_i - \bar{e}_i\boldsymbol{\alpha}_i^{-1} \ , \tag{3.102}$$

where $\bar{e}_i = \frac{\sum_j \alpha_{ij}^{-2}e_{ij}}{\sum_j \alpha_{ij}^{-2}}$.

The heteroskedastic within-estimator of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}}_w = (\underline{\mathbf{X}}^\top \mathbf{Q} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{Q} \underline{\mathbf{y}}$. The heteroskedastic within-residual vector is obtained by

$$\begin{aligned} \hat{\underline{\mathbf{e}}}_w &= \underline{\mathbf{y}} - \underline{\mathbf{X}} \hat{\boldsymbol{\beta}}_w \\ &= (\mathbf{I} - \underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \mathbf{Q} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{Q}) \underline{\mathbf{y}} \\ &= (\mathbf{I} - \underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \mathbf{Q} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{Q}) \mathbf{e} \; . \end{aligned} \tag{3.103}$$

Since the trace of idempotent matrix $\mathbf{Q} - \underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \mathbf{Q} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{Q}$ is $N - m - r(\mathbf{Q}\underline{\mathbf{X}})$ and $\mathrm{E}[\mathbf{e}\mathbf{e}^\top] = \sigma_e^2 \mathbf{I}_N$, the quadratic estimator of σ_e^2 is obtained as follows:

$$\begin{aligned} \mathrm{E}[\hat{\underline{\mathbf{e}}}_w^\top \mathbf{Q} \hat{\underline{\mathbf{e}}}_w] &= \mathrm{E}[\mathbf{e}^\top (\mathbf{Q} - \mathbf{Q}\underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \mathbf{Q} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{Q}) \mathbf{e}] \\ &= \mathrm{tr}\left((\mathbf{Q} - \underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \mathbf{Q} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{Q}) \mathrm{E}[\mathbf{e}\mathbf{e}^\top] \right) \\ &= (N - m - r(\mathbf{Q}\underline{\mathbf{X}}))\, \sigma_e^2 \end{aligned} \tag{3.104}$$

$$\therefore \hat{\sigma}_e^2 = \frac{\hat{\underline{\mathbf{e}}}_w^\top \mathbf{Q} \hat{\underline{\mathbf{e}}}_w}{N - m - r(\mathbf{Q}\underline{\mathbf{X}})} \; , \tag{3.105}$$

where $r(\bullet)$ denotes the rank of $\bullet$.

FC estimator of σ_u^2 The FC estimator of σ_u^2 is obtained using the OLS residuals of regression (3.98). Given the OLS estimator $\hat{\boldsymbol{\beta}}_{ols} = (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \underline{\mathbf{y}}$, the vector of residuals $\widehat{\underline{\mathbf{v}}}$ is expressed as

$$\begin{aligned} \widehat{\underline{\mathbf{v}}} &= \underline{\mathbf{y}} - \underline{\mathbf{X}} \hat{\boldsymbol{\beta}} \\ &= (\mathbf{I} - \underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top) \underline{\mathbf{y}} \\ &= (\mathbf{I} - \underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top) \underline{\mathbf{v}} \; . \end{aligned} \tag{3.106}$$

Since $\mathbf{I} - \underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top$ is a symmetric idempotent matrix and $\mathrm{Var}[\underline{\mathbf{v}}] = \sigma_u^2 \mathbf{I}_m \otimes \boldsymbol{\alpha}_i^{-1} (\boldsymbol{\alpha}_i^{-1})^\top + \sigma_e^2 \mathbf{I}_N$, the quadratic estimator of σ_u^2 is obtained as follows.

$$\begin{aligned} \mathrm{E}[\widehat{\underline{\mathbf{v}}}^\top \widehat{\underline{\mathbf{v}}}] &= \mathrm{E}[\underline{\mathbf{v}}^\top (\mathbf{I} - \underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top) \underline{\mathbf{v}}] \\ &= \mathrm{tr}\left((\mathbf{I} - \underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top) \mathrm{E}[\underline{\mathbf{v}}^\top \underline{\mathbf{v}}] \right) \\ &= \mathrm{tr}\left((\mathbf{I} - \underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top) (\sigma_u^2 \mathbf{I}_m \otimes \boldsymbol{\alpha}_i^{-1} (\boldsymbol{\alpha}_i^{-1})^\top + \sigma_e^2 \mathbf{I}_N) \right) \\ &= \sigma_u^2 \mathrm{tr}(\mathbf{I}_m \otimes \boldsymbol{\alpha}_i^{-1} (\boldsymbol{\alpha}_i^{-1})^\top) - \sigma_u^2 \mathrm{tr}(\underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \left(\mathbf{I}_m \otimes \boldsymbol{\alpha}_i^{-1} (\boldsymbol{\alpha}_i^{-1})^\top \right)) \\ &\quad + \sigma_e^2 \mathrm{tr}(\mathbf{I} - \underline{\mathbf{X}} (\underline{\mathbf{X}}^\top \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top) \; . \end{aligned} \tag{3.107}$$

The first, second and third terms have the respective traces of

$$\begin{aligned}\mathrm{tr}(\mathbf{I}_m \otimes \boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top) &= \sum_i \mathrm{tr}(\boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top) \\ &= \sum_{i,j} \alpha_{ij}^{-2} \ , \end{aligned} \tag{3.108}$$

$$\begin{aligned}& \mathrm{tr}\left(\underline{\mathbf{X}}(\underline{\mathbf{X}}^\top\underline{\mathbf{X}})^{-1}\underline{\mathbf{X}}^\top(\mathbf{I}_m \otimes_i \boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top)\right) \\ = \ & \mathrm{tr}\left((\mathbf{I}_m \otimes_i \boldsymbol{\alpha}_i^{-1})^\top\underline{\mathbf{X}}(\underline{\mathbf{X}}^\top\underline{\mathbf{X}})^{-1}\underline{\mathbf{X}}^\top(\mathbf{I}_m \otimes_i \boldsymbol{\alpha}_i^{-1})\right) \\ = \ & \mathrm{tr}\left((\mathbf{1}_m \otimes_i (\boldsymbol{\alpha}_i^{-1})^\top\underline{\mathbf{X}}_i)(1 \otimes \underline{\mathbf{X}}^\top\underline{\mathbf{X}})^{-1}(\mathbf{1}_m^\top \otimes_i \underline{\mathbf{X}}_i^\top\boldsymbol{\alpha}_i^{-1})\right) \\ = \ & \mathrm{tr}\left(\mathbf{J}_m \otimes_i \left((\boldsymbol{\alpha}_i^{-1})^\top(\oplus_j\alpha_{ij}^{-1})\tilde{\mathbf{X}}_i(\underline{\mathbf{X}}^\top\underline{\mathbf{X}})^{-1}\tilde{\mathbf{X}}_i^\top(\oplus_j\alpha_{ij}^{-1})\boldsymbol{\alpha}_i^{-1}\right)\right) \\ = \ & \mathrm{tr}\{\mathbf{J}_m \otimes_i ((\sum_j \alpha_{ij}^{-2}\tilde{\mathbf{x}}_{ij}^\top)(\sum_{k,j} \alpha_{kj}^{-2}\tilde{\mathbf{x}}_{kj}\tilde{\mathbf{x}}_{kj}^\top)^{-1}(\sum_j \alpha_{ij}^{-2}\tilde{\mathbf{x}}_{ij}))\} \\ = \ & \sum_{i=1}^{m}((\sum_j \alpha_{ij}^{-2}\tilde{\mathbf{x}}_{ij}^\top)(\sum_{k,j} \alpha_{kj}^{-2}\tilde{\mathbf{x}}_{kj}\tilde{\mathbf{x}}_{kj}^\top)^{-1}(\sum_j \alpha_{ij}^{-2}\tilde{\mathbf{x}}_{ij})) \ , \end{aligned} \tag{3.109}$$

$$\mathrm{tr}(\mathbf{I} - \underline{\mathbf{X}}(\underline{\mathbf{X}}^\top\underline{\mathbf{X}})^{-1}\underline{\mathbf{X}}^\top) = N - p \ . \tag{3.110}$$

By inserting (3.108), (3.109) and (3.110) in (3.107), it follows that

$$\mathrm{E}[\widehat{\underline{\mathbf{v}}}^\top\widehat{\underline{\mathbf{v}}}] = \sigma_u^2(\sum_{i,j} \alpha_{ij}^{-2} - C_{\mathrm{FC}}) + \sigma_e^2(N-p) \tag{3.111}$$

$$\therefore \ \hat{\sigma}_{u,\mathrm{FC}}^2 = \max\left(\frac{\widehat{\underline{\mathbf{v}}}^\top\widehat{\underline{\mathbf{v}}} - (N-p)\hat{\sigma}_e^2}{\sum_{i,j}\alpha_{ij}^{-2} - C_{\mathrm{FC}}}, 0\right) \ , \tag{3.112}$$

where $C_{\mathrm{FC}} = \sum_{i=1}^{m}\left((\sum_j \alpha_{ij}^{-2}\tilde{\mathbf{x}}_{ij}^\top)(\sum_{k,j} \alpha_{kj}^{-2}\tilde{\mathbf{x}}_{kj}\tilde{\mathbf{x}}_{kj}^\top)^{-1}(\sum_j \alpha_{ij}^{-2}\tilde{\mathbf{x}}_{ij})\right)$.

SA estimator of σ_u^2 Estimation starts by transforming regression (3.98) with the heteroskedastic between-transformation matrix $\mathbf{P}$ to obtain the between-estimator $\hat{\boldsymbol{\beta}}_b$, which is given by $\hat{\boldsymbol{\beta}}_b = (\underline{\mathbf{X}}^\top\mathbf{P}\underline{\mathbf{X}})^{-1}\underline{\mathbf{X}}^\top\mathbf{P}\underline{\mathbf{y}}$. The between-residual vector, here denoted by $\hat{\underline{\mathbf{v}}}_b$, is expressed as

$$\begin{aligned}\hat{\underline{\mathbf{v}}}_b &= \underline{\mathbf{y}} - \underline{\mathbf{X}}\hat{\boldsymbol{\beta}}_b \\ &= (\mathbf{I} - \underline{\mathbf{X}}(\underline{\mathbf{X}}^\top\mathbf{P}\underline{\mathbf{X}})^{-1}\underline{\mathbf{X}}^\top\mathbf{P})\underline{\mathbf{y}} \\ &= (\mathbf{I} - \underline{\mathbf{X}}(\underline{\mathbf{X}}^\top\mathbf{P}\underline{\mathbf{X}})^{-1}\underline{\mathbf{X}}^\top\mathbf{P})\underline{\mathbf{v}} \ . \end{aligned} \tag{3.113}$$

Since $\mathrm{Var}[\underline{\mathbf{v}}] = \sigma_u^2 \mathbf{I}_m \otimes \boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top + \sigma_e^2 \mathbf{I}_N$, the quadratic estimator of $\sigma^2_{u,\mathrm{SA}}$ is obtained as follows. It holds that

$$\begin{aligned}
\mathrm{E}[\hat{\underline{\mathbf{v}}}_b^\top \mathbf{P} \hat{\underline{\mathbf{v}}}_b] &= \mathrm{E}[\underline{\mathbf{v}}^\top (\mathbf{P} - \mathbf{P}\underline{\mathbf{X}}(\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{P}) \underline{\mathbf{v}}] \\
&= \mathrm{tr}\left((\mathbf{P} - \mathbf{P}\underline{\mathbf{X}}(\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{P}) \mathrm{E}[\underline{\mathbf{v}}\underline{\mathbf{v}}^\top] \right) \\
&= \mathrm{tr}\left((\mathbf{P} - \mathbf{P}\underline{\mathbf{X}}(\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{P}) (\sigma_u^2 \mathbf{I}_m \otimes \boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top + \sigma_e^2 \mathbf{I}_n) \right) \\
&= \sigma_u^2 \mathrm{tr}\left(\mathbf{P}(\mathbf{I}_m \otimes \boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top) \right) \\
&\quad - \sigma_u^2 \mathrm{tr}\left(\mathbf{P}\underline{\mathbf{X}}(\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{P} (\mathbf{I}_m \otimes \boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top) \right) \\
&\quad + \sigma_e^2 \mathrm{tr}\left(\mathbf{P} - \mathbf{P}\underline{\mathbf{X}}(\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{P} \right) . \qquad (3.114)
\end{aligned}$$

Since $\mathbf{P}_i \boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top = \boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top$, the traces of the first and third terms of (3.114) are

$$\begin{aligned}
\mathrm{tr}\left(\mathbf{P}(\mathbf{I}_m \otimes \boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top) \right) &= \mathrm{tr}\left(\mathbf{I}_m \otimes \boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top \right) \\
&= \sum_{i,j} \alpha_{ij}^{-2} = \sum_i \eta_i \text{ and} \qquad (3.115)
\end{aligned}$$

$$\mathrm{tr}(\mathbf{P} - \mathbf{P}\underline{\mathbf{X}}(\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{P}) = m - p \ , \qquad (3.116)$$

respectively. Calculation similar to (3.109) yields the trace of the second term of (3.114):

$$\begin{aligned}
& \mathrm{tr}\left(\mathbf{P}\underline{\mathbf{X}}(\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top \mathbf{P} (\mathbf{I}_m \otimes \boldsymbol{\alpha}_i^{-1}(\boldsymbol{\alpha}_i^{-1})^\top) \right) \\
= \ & \mathrm{tr}\left((\mathbf{I}_m \otimes_i \boldsymbol{\alpha}_i^{-1})^\top \underline{\mathbf{X}}(\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}})^{-1} \underline{\mathbf{X}}^\top (\mathbf{I}_m \otimes_i \boldsymbol{\alpha}_i^{-1}) \right) \\
= \ & \mathrm{tr}\left(\mathbf{J}_m \otimes_i \left((\sum_j \alpha_{ij}^{-2} \tilde{\mathbf{x}}_{ij}^\top)(\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}})^{-1} \sum_j \alpha_{ij}^{-2} \tilde{\mathbf{x}}_{ij} \right) \right) \\
= \ & \sum_{i=1}^m \left((\sum_j \alpha_{ij}^{-2} \tilde{\mathbf{x}}_{ij}^\top)(\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}})^{-1} \sum_j \alpha_{ij}^{-2} \tilde{\mathbf{x}}_{ij} \right) . \qquad (3.117)
\end{aligned}$$

Inserting (3.115), (3.117) and (3.116) into (3.114) yields

$$\mathrm{E}[\hat{\underline{\mathbf{v}}}_b^\top \mathbf{P} \hat{\underline{\mathbf{v}}}_b] = \sigma_u^2 (\sum_i \eta_i - C_{\mathrm{SA}}) + \sigma_e^2 (m - p) \ . \qquad (3.118)$$

Therefore,

$$\hat{\sigma}^2_{u,\mathrm{SA}} = \max\left(\frac{\hat{\underline{\mathbf{v}}}_b^\top \mathbf{P} \hat{\underline{\mathbf{v}}}_b - (m-p)\hat{\sigma}_e^2}{\sum_i \eta_i - C_{\mathrm{SA}}}, 0 \right) \ , \qquad (3.119)$$

where $C_{\mathrm{SA}} = \sum_{i=1}^{m} \left((\sum_j \alpha_{ij}^{-2} \tilde{\mathbf{x}}_{ij}^\top)(\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}})^{-1} (\sum_j \alpha_{ij}^{-2} \tilde{\mathbf{x}}_{ij}) \right)$. Here $\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}}$ can be expressed as

$$\begin{aligned}
\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}} &= (\mathbf{1}_m^\top \otimes_k \underline{\mathbf{X}}_k^\top)(\mathbf{I}_m \otimes_k \eta_k \boldsymbol{\alpha}_k^{-1} (\boldsymbol{\alpha}_k^{-1})^\top)(\mathbf{1}_m \otimes_k \underline{\mathbf{X}}_k) \\
&= m \otimes_k (\underline{\mathbf{X}}_k^\top \eta_k \boldsymbol{\alpha}_k^{-1} (\boldsymbol{\alpha}_k^{-1})^\top \underline{\mathbf{X}}_k) \\
&= \sum_{k=1}^{m} \eta_k \tilde{\mathbf{X}}_k \boldsymbol{\alpha}_k^{-2} (\boldsymbol{\alpha}_k^{-2})^\top \tilde{\mathbf{X}}_k \\
&= \sum_{k=1}^{m} \eta_k (\sum_j \alpha_{kj}^{-2} \tilde{\mathbf{x}}_{kj})(\sum_j \alpha_{kj}^{-2} \tilde{\mathbf{x}}_{kj}^\top) \ . \qquad (3.120)
\end{aligned}$$

In the iteration stage, estimators are obtained with the working variables $\tilde{y}$ and $\tilde{\mathbf{x}}$ in (3.95) modified as follows: $\tilde{\mathbf{y}}$ is a vector whose ijth element is $y_{ij} - \hat{\gamma}(\mathbf{t}_{ij})$ and $\tilde{\mathbf{X}}$ is the same as the original design $\mathbf{X}$.

By taking into account the loss of d.f. due to the estimation of γ, the estimator of σ_e^2 is given by

$$\hat{\sigma}_{e(r)}^2 = \frac{\hat{\underline{\mathbf{e}}}_{w(r)}^\top \mathbf{Q} \hat{\underline{\mathbf{e}}}_{w(r)}}{N - m - r(\mathbf{Q}\underline{\mathbf{X}}) - \mathrm{d.f.}(\hat{\boldsymbol{\gamma}}_{(r-1)})} \ , \qquad (3.121)$$

where the subscript (r) indicates the rth iteration ($\hat{\boldsymbol{\gamma}}_{(0)}$ is the estimator of γ in the initial stage). $r(\mathbf{Q}\underline{\mathbf{X}})$ is p if none of the columns of $\underline{\mathbf{X}}$ is cluster-wise invariant. The d.f. of the nonparametric estimate of γ is defined as

$$\mathrm{d.f.}(\hat{\boldsymbol{\gamma}}_{(r)}) = 2\mathrm{tr}(\mathbf{S}_{(r)}) - \mathrm{tr}(\mathbf{S}_{(r)} \mathbf{S}_{(r)}^\top) \ , \qquad (3.122)$$

where $\mathbf{S}_{(r)}$ is the smoother matrix of γ estimation. An alternative definition of the d.f. following Carmack et al. (2011) is given by

$$\mathrm{d.f.}(\hat{\boldsymbol{\gamma}}_{(r)}) = 2\mathrm{tr}(\mathbf{S}_{(r)} \mathbf{R}_{(r)}) - \mathrm{tr}(\mathbf{S}_{(r)} \mathbf{R}_{(r)} \mathbf{S}_{(r)}^\top) \ , \qquad (3.123)$$

where $\mathbf{R}_{(r)}$ is the correlation matrix of the random terms. See Appendix A 3.10.5 for more details.

Replacing the estimator of σ_e^2 in (3.112) and (3.119) by (3.121) yields the estimators in the rth iteration:

$$\hat{\sigma}_{u(r),\mathrm{FC}}^2 = \max \left(\frac{\hat{\underline{\mathbf{v}}}_{(r)}^\top \hat{\underline{\mathbf{v}}}_{(r)} - (N-p)\hat{\sigma}_{e(r)}^2}{\sum_{i,j} \alpha_{ij}^{-2} - C_{\mathrm{FC}}}, 0 \right) \ , \qquad (3.124)$$

where $C_{\mathrm{FC}} = \sum_{i=1}^{m} \left((\sum_j \alpha_{ij}^{-2} \mathbf{x}_{ij}^\top)(\sum_{k,j} \alpha_{kj}^{-2} \mathbf{x}_{kj} \mathbf{x}_{kj}^\top)^{-1} (\sum_j \alpha_{ij}^{-2} \mathbf{x}_{ij}) \right)$; and

$$\hat{\sigma}^2_{u(r),\mathrm{SA}} = \max \left(\frac{\hat{\underline{\mathbf{v}}}_{b(r)}^\top \mathbf{P} \hat{\underline{\mathbf{v}}}_{b(r)} - (m-p)\hat{\sigma}^2_{e(r)}}{\sum_i \eta_i - C_{\mathrm{SA}}}, 0 \right) , \tag{3.125}$$

where $C_{\mathrm{SA}} = \sum_{i=1}^{m} \left((\sum_j \alpha_{ij}^{-2} \mathbf{x}_{ij}^\top)(\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}})^{-1} (\sum_j \alpha_{ij}^{-2} \mathbf{x}_{ij}) \right)$ with
$\underline{\mathbf{X}}^\top \mathbf{P} \underline{\mathbf{X}} = \sum_{k=1}^{m} \eta_k (\sum_j \alpha_{kj}^{-2} \mathbf{x}_{kj})(\sum_j \alpha_{kj}^{-2} \mathbf{x}_{kj}^\top)$.

3.10.4 Derivation of the random intercept predictor

This section presents the derivation of the random intercept predictors $\tilde{u}_i$. The predictors are consistent and linear in the form of

$$\tilde{u} = b + \mathbf{c}^\top \mathbf{y} \tag{3.126}$$

with some scalar b and vector $\mathbf{c}$. $\tilde{u}$ is such that minimizes the following mean squared prediction error

$$\begin{aligned} \mathrm{E}[(\tilde{u} - u)^2] &= \mathrm{E}[b^2 + 2b(\mathbf{c}^\top \mathbf{y} - u) + (\mathbf{c}^\top \mathbf{y} - u)^2] \\ &= b^2 + 2b\mathrm{E}[\mathbf{c}^\top \mathbf{y} - u] + \mathrm{E}[(\mathbf{c}^\top \mathbf{y} - u)^2] \, . \end{aligned} \tag{3.127}$$

b and $\mathbf{c}$ required for the best predictor are obtained as follows. Differentiating (3.127) with respect to b and setting the derivative $\partial \mathrm{E}[(\tilde{u} - u)^2] / \partial b$ to zero yields

$$\begin{aligned} 2(b + \mathrm{E}[\mathbf{c}^\top \mathbf{y} - u]) &= 0 \\ \therefore b &= -\mathrm{E}[\mathbf{c}^\top \mathbf{y} - u] \, . \end{aligned} \tag{3.128}$$

By inserting (3.128) into (3.127),

$$\begin{aligned} \mathrm{E}[(\tilde{u} - u)^2] &= \mathrm{E}[(\mathbf{c}^\top \mathbf{y} - u)^2] - (\mathrm{E}[\mathbf{c}^\top \mathbf{y} - u])^2 \\ &= \mathrm{Var}[\mathbf{c}^\top \mathbf{y} - u] \\ &= \mathbf{c}^\top \mathbf{V} \mathbf{c} - 2\mathbf{c}^\top \mathrm{Cov}[\mathbf{y}, u] + \sigma_u^2 \, . \end{aligned} \tag{3.129}$$

Differentiating (3.129) with respect to $\mathbf{c}$ and setting the derivative $\partial \mathrm{E}[(\tilde{u} - u)^2] / \partial \mathbf{c}$ to zero gives

$$\begin{aligned} 2\mathbf{V}\mathbf{c} - 2\mathrm{Cov}[\mathbf{y}, u] &= 0 \\ \therefore \mathbf{c} &= \mathbf{V}^{-1} \mathrm{Cov}[\mathbf{y}, u] \, . \end{aligned} \tag{3.130}$$

Thus inserting (3.128) and (3.130) into (3.126) yields

$$\begin{aligned} \tilde{u} &= \mu_u + \mathrm{Cov}[\mathbf{y}^\top, u]\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y}) \\ &= \mathrm{Cov}[\mathbf{u}^\top, u]\mathbf{Z}^\top\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y}) \ , \end{aligned} \tag{3.131}$$

where $\boldsymbol{\mu}_\mathbf{y} = \mathrm{E}[\mathbf{y}|\mathbf{X}, \mathbf{T}]$ and $\mu_u = \mathrm{E}[u] = 0$ by assumption. The last equality holds due to the assumption of between-cluster independence. Taking the expectation of (3.131) shows the unbiasedness of the predictor in the sense that $\mathrm{E}[\tilde{u}] = \mathrm{E}[u]$. Note that predictor (3.131) requires knowledge of only the first two moments of $\mathbf{y}$ and $\mathbf{u}$, but not a distributional assumption such as normality.

From (3.131), the ith random intercept predictor is given by

$$\begin{aligned} \tilde{u}_i &= \mathrm{Cov}[\mathbf{u}^\top, u_i](\mathbf{I}_m \otimes \mathbf{Z}_i^\top)(\mathbf{I}_m \otimes \mathbf{V}_i^{-1}) \left(\mathbf{1}_m \otimes (\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i})\right) \\ &= \sigma_u^2 \mathbf{e}_i^\top \left(\mathbf{1}_m \otimes \mathbf{1}_i^\top \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i})\right) \\ &= \sigma_u^2 \mathbf{1}_i^\top \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_{\mathbf{y}_i}) \ , \end{aligned} \tag{3.132}$$

where $\mathbf{e}_i \in \mathbb{R}^m$ is a vector of zero except for the ith element being one. Replacing $\boldsymbol{\mu}_{\mathbf{y}_i}$ with $\mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\gamma}(\mathbf{T}_i)$ gives

$$\tilde{u}_i = \sigma_u^2 \mathbf{1}_{n_i}^\top \mathbf{V}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\gamma}(\mathbf{T}_i)\right) \ \forall i = 1, \ldots, m \ . \tag{3.133}$$

Replacing unknown VC, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ with their consistent estimators yields (3.47).

3.10.5 Extended generalized cross validation (GCVc)

Suppose a one-dimensional kernel regression estimation of $y_k = \gamma(t_k) + v_k$ where $\mathrm{E}[v_k|t_k] = 0$ holds ($k = 1, 2, \ldots, N$). Here we consider the mean squared prediction error (MSPE), $\mathrm{E}[(y_k - \hat{\gamma}(t_k))^2]$. Let $\hat{\gamma}_{\mathrm{cv}}$ denote an estimate of γ with the bandwidth h selected by either CV ($\hat{\gamma}_{\mathrm{cv}}(t_k) = \hat{\gamma}_{(-k)}(t_k)$) or GCV ($\hat{\gamma}_{\mathrm{cv}}(t_k) = \hat{\gamma}(t_k)$). Here $\hat{\gamma}(t_k)$ is an estimator obtained using the whole sample. The MSPE of $\hat{\gamma}_{\mathrm{cv}}(t_k)$ is

$$\begin{aligned} \mathrm{E}[(y_k - \hat{\gamma}_{\mathrm{cv}}(t_k))^2] &= \mathrm{E}[(y_k - \gamma(t_k))^2] + \mathrm{E}[(\gamma(t_k) - \hat{\gamma}_{\mathrm{cv}}(t_k))^2] \\ &\quad +2\mathrm{E}[(y_k - \gamma(t_k))\,(\gamma(t_k) - \hat{\gamma}_{\mathrm{cv}}(t_k))] \\ &= \sigma_v^2 + \mathrm{E}[(\gamma(t_k) - \hat{\gamma}_{\mathrm{cv}}(t_k))^2] + 2\mathrm{E}[v_k\hat{\gamma}_{\mathrm{cv}}(t_k)] \ . \end{aligned} \tag{3.134}$$

Provided the last term of (3.134) is zero, minimization of the MSPE of $\hat{\gamma}_{\mathrm{cv}}$ is equivalent to minimization of the MSE $\mathrm{E}[(\gamma(t_k) - \hat{\gamma}(t_k))^2]$. This implies that the average squared prediction errors (ASPE), $N^{-1}\sum_{k=1}^N (y_k - \hat{\gamma}(t_k))^2$ is expected to minimize

the average MSE, $N^{-1}\sum_{k=1}^{N}\mathbf{E}\left(\gamma(t_k)-\hat{\gamma}(t_k)\right)^2$. If the errors are independent, the last term will be (asymptotically) zero for the CV due to $\mathrm{E}[v_k\hat{\gamma}_{(-k)}(t_k)]=0$ and for the GCV due to the penalization of the term $\mathrm{E}[\sum_{k=1}^{N}s_{kk}v_k^2]$ involved in $\mathrm{E}[v_k\hat{\gamma}_{\mathrm{cv}}(t_k)]$, where s_{kk} is the kth diagonal element of the smoother matrix $\mathbf{S}$.

For correlated data, however, $\mathrm{E}[v_kv_l]$ for $(k\neq l)$ is not necessarily zero and therefore neither the CV nor the GCV has the last term of (3.134) being zero. Therefore, if correlations are positive, CV and GCV tend to select a smaller bandwidth than the optimal that minimizes the average MSE, $N^{-1}\sum_{k=1}^{N}\mathbf{E}[(\gamma(t_k)-\hat{\gamma}(t_k))^2]$.

In order to correct the GCV for correlations between errors, Carmack et al. (2011) proposed an alternative definition of the residual degrees of freedom. First, note that for the GCV, the second term of (3.134) is

$$\begin{aligned}
\mathrm{E}[(\gamma(t_k)-\hat{\gamma}(t_k))^2] &= \mathrm{Var}[\gamma(t_k)-\hat{\gamma}(t_k)]+(\mathrm{E}[\gamma(t_k)-\hat{\gamma}(t_k)])^2\\
&= \mathrm{Var}[\sum_j s_{kj}y_j]+(\mathrm{E}[\gamma(t_k)-\sum_j s_{kj}y_j])^2\\
&= \mathrm{Var}[\sum_j s_{kj}y_j]+(\mathrm{E}[\gamma(t_k)-\sum_j s_{kj}\gamma(x_j)-\sum_j s_{kj}v_j])^2\\
&= \sum_j\sum_l s_{kj}s_{kl}\mathrm{Cov}[v_j,v_l]+(\gamma(t_k)-\sum_j s_{kj}\gamma(x_j))^2 \qquad (3.135)
\end{aligned}$$

and that the third term is

$$\begin{aligned}
\mathrm{E}\left[v_k\hat{\gamma}(t_k)\right] &= \mathrm{Cov}[v_k,\hat{\gamma}(t_k)]\\
&= \mathrm{Cov}[v_k,\sum_j s_{kj}y_j]\\
&= \sum_j s_{kj}\mathrm{Cov}[v_k,v_j]\ , \qquad (3.136)
\end{aligned}$$

where s_{kj} is the (k,j) element of smoother $\mathbf{S}$. Then (3.134) is rewritten as

$$\begin{aligned}
\mathrm{E}[(y_k-\hat{\gamma}(t_k))^2] &= \sigma_v^2+\sum_j\sum_l s_{kj}s_{kl}\mathrm{Cov}[v_j,v_l]+(\gamma(t_k)-\sum_j s_{kj}\gamma(x_j))^2\\
&\quad -2\sum_j s_{kj}\mathrm{Cov}[v_k,v_j]\ . \qquad (3.137)
\end{aligned}$$

Thus the expectation of the APSE is given by

$$\begin{aligned}
\mathbf{E}\left[\text{ASPE}(h)\right] &= \frac{1}{N}\sum_{k=1}^{N}\mathbf{E}\left(y_k-\hat{\gamma}(t_k)\right)^2 \\
&= \frac{1}{N}\sum_{k}(\gamma(t_k)-\sum_{j}s_{kj}\gamma(x_j))^2 + \\
&\quad \frac{1}{N}(N\sigma_v^2+\sum_{k}\sum_{j}\sum_{l}s_{kj}s_{kl}\text{Cov}[v_j,v_l]-2\sum_{k}\sum_{j}s_{kj}\text{Cov}[v_k,v_j]) \\
&= \frac{1}{N}\sum_{k}C_k + \\
&\quad \frac{\sigma_v^2}{N}(N+\sum_{k}\sum_{j}\sum_{l}s_{kj}s_{kl}\text{Corr}[v_j,v_l]-2\sum_{k}\sum_{j}s_{kj}\text{Corr}[v_k,v_j])\ ,
\end{aligned} \tag{3.138}$$

where C_k denotes $(\gamma(t_k)-\sum_j s_{kj}\gamma(x_j))^2$. Let $\mathbf{R}$ denote the correlation matrix of the errors. The second term in the parentheses in (3.138) is $\text{tr}(\mathbf{SRS}^\top)$ and the third term $-2\text{tr}(\mathbf{SR})$. Therefore it holds that

$$\begin{aligned}
\mathbf{E}\left[\text{ASPE}(h)\right] &= \frac{1}{N}\sum_{k}C_k+\frac{\sigma_v^2}{N}\left(N+\text{tr}(\mathbf{SRS}^\top-2\mathbf{SR})\right) \\
\mathbf{E}[\sum_{k=1}^{N}(y_k-\hat{\gamma}(t_k))^2] &= \sum_{k}C_k+\sigma_v^2\left(N-\text{tr}(2\mathbf{SR}-\mathbf{SRS}^\top)\right)\ .
\end{aligned} \tag{3.139}$$

From this observation, Carmack et al. (2011) proposed an alternative definition of the residual degrees of freedom given by $N-\text{tr}(2\mathbf{SR}-\mathbf{SRS}^\top)$. This is reduced to $N-\text{tr}(2\mathbf{S}-\mathbf{SS}^\top)$ when data are uncorrelated ($\mathbf{R}=\mathbf{I}$), and further to $N-\text{tr}\left(\mathbf{S}\right)$ when data are uncorrelated and $\mathbf{S}$ is symmetric idempotent ($\mathbf{SS}^\top=\mathbf{S}$). As an alternative to the GCV statistic (3.30), they proposed the following extended GCV statistic:

$$\text{GCVc}\left(h\right) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{y_k-\hat{\gamma}(t_k)}{1-\text{tr}(2\mathbf{SR}-\mathbf{SRS}^\top)/N}\right)^2\ . \tag{3.140}$$

In practical implementation, if the dimension of $\mathbf{S}$ and $\mathbf{R}$ is large, calculation of $\text{tr}(\mathbf{SR})$ and $\text{tr}(\mathbf{SRS}^\top)$ may encounter computational difficulties. Even though binning techniques cannot be used, calculation of trace becomes manageable through the use of the diagonal structure of $\mathbf{R}$. First, note that smoother $\mathbf{S}$ can be rewritten

as

$$\mathbf{S} = (\mathbf{s}_1\mathbf{s}_2\cdots\mathbf{s}_N)^\top \tag{3.141}$$

$$\mathbf{s}_k^\top = (\mathbf{s}_{k1}^\top\mathbf{s}_{k2}^\top\cdots\mathbf{s}_{km}^\top);\ k = 1,\dots N\ , \tag{3.142}$$

where $\mathbf{s}_k^\top$ is the kth row vector of $\mathbf{S}$, composed of m subvectors $\mathbf{s}_{ki}^\top$ whose length is each n_i. It holds that

$$\begin{aligned}\mathrm{tr}(\mathbf{SRS}^\top) &= \mathrm{tr}(\mathbf{RS}^\top\mathbf{S}) \\ &= \mathrm{tr}(\mathbf{R}\sum_{k=1}^{N}\mathbf{s}_k\mathbf{s}_k^\top) \\ &= \sum_{k=1}^{N}\mathrm{tr}(\mathbf{R}\mathbf{s}_k\mathbf{s}_k^\top) \\ &= \sum_{k=1}^{N}\mathbf{s}_k^\top\mathbf{R}\mathbf{s}_k\ .\end{aligned} \tag{3.143}$$

Since $\mathbf{R}$ is a block diagonal matrix $\oplus_i\mathbf{R}_i$,

$$\begin{aligned}\mathbf{s}_k^\top\mathbf{R}\mathbf{s}_k &= (\mathbf{s}_{k1}^\top\mathbf{s}_{k2}^\top\cdots\mathbf{s}_{km}^\top)\oplus_i\mathbf{R}_i(\mathbf{s}_{k1}^\top\mathbf{s}_{k2}^\top\cdots\mathbf{s}_{km}^\top)^\top \\ &= \sum_{i=1}^{m}\mathbf{s}_{ki}^\top\mathbf{R}_i\mathbf{s}_{ki}\ .\end{aligned} \tag{3.144}$$

$\mathrm{tr}(\mathbf{SRS}^\top)$ is therefore given by

$$\mathrm{tr}(\mathbf{SRS}^\top) = \sum_{k=1}^{N}\sum_{i=1}^{m}\mathbf{s}_{ki}^\top\mathbf{R}_i\mathbf{s}_{ki}\ . \tag{3.145}$$

The (k,k) element of $\mathbf{SR}$ can be expressed as $\mathbf{s}_k^\top\,[\mathbf{R}]_{,k} = \mathbf{s}_{ki}^\top[\mathbf{R}_i]_{,k}$, where i is the index of the cluster to which kth observation belongs ($[\mathbf{R}]_{,k}$ and $[\mathbf{R}_i]_{,k}$ are the kth column vector of $\mathbf{R}$ and $\mathbf{R}_i$, respectively). Thus $\mathrm{tr}(\mathbf{SR})$ is given by

$$\mathrm{tr}(\mathbf{SR}) = \sum_{k=1}^{N}\sum_{i=1}^{m}\mathbf{s}_{ki}^\top[\mathbf{R}_i]_{,k}I(y_k\in\mathbf{y}_i)\ , \tag{3.146}$$

where I is an indicator function.

3.10.6 Binning

For ease of description, we consider a one-dimensional kernel regression $y_k = \gamma(t_k) + e_k$ where $\mathrm{E}[e_k|t_k] = 0$ holds ($k = 1, 2, \ldots, N$). The Nadaraya-Watson and local linear estimators of $\gamma(t)$ are given by

$$\hat{\gamma}(t) = \frac{T_0(t)}{S_0(t)} \qquad \text{(Nadaraya-Watson)} \tag{3.147}$$

$$\hat{\gamma}(t) = \frac{S_2(t)T_0(t) - S_1(t)T_1(t)}{S_2(t)S_0(t) - S_1(t)^2} \qquad \text{(Local linear)} , \tag{3.148}$$

respectively, where

$$T_l(t) = \sum_{k=1}^{N} K_h(t_k - t)(t_k - t)^l y_k \quad l = 0, 1 \tag{3.149}$$

$$S_l(t) = \sum_{k=1}^{N} K_h(t_k - t)(t_k - t)^l y_k \quad l = 0, 1, 2 \tag{3.150}$$

with kernel function K_h which depends on a bandwidth h. N kernel function values need to be evaluated for the estimate of γ at point t. Thus computation of order $O(N^2)$ is required for the estimation of γ at N points of t. The same argument is also true for the calculation of $\mathrm{tr}(\mathbf{S})$ and $\mathrm{tr}(\mathbf{S}\mathbf{S}^\top)$ where $\mathbf{S}$ is a smoother matrix for the estimation of γ. CV repeats this order of calculation for each of the candidate bandwidths.

The basic idea of binning is to reduce the number of kernel evaluations by using only a set of summary data created by binning. Here we consider the so-called "simple binning".[29] Suppose G grid points τ_g ($g = 1, \ldots G$) equi-spaced with bin width λ, and G bins $B_g = (\tau_g - \lambda/2,\ \tau_g + \lambda/2)$ over the support of t. Let I_g denote an index set such that $I_g = \{k : y_k \in B_g\}$. The binned data set is then $\{\bar{y}_g, \tau_g, c_g\}_{g=1}^{G}$ where $\bar{y}_g$ is the simple average of $\{y_k : y_k \in B_g\}$ and c_g is the number of indeces in I_g.

Let $\tau_{g(k)}$ denote the grid point of the bin into which t_k is binned. $T_l(t)$ in (3.149) is

[29] Function `h.select` of R package `sm`, which is used in `plmm`, also employs simple binning for CV. Calculation of $\mathrm{tr}(\mathbf{S})$ and $\mathrm{tr}(\mathbf{S}\mathbf{S}^\top)$ in the `plmm` package is also implemented using simple binning if the sample size is large ($N > 100$ in the default).

approximated by $\bar{T}_l(\tau)$ defined as follows:

$$\begin{aligned}\bar{T}_l(\tau) &= \sum_{k=1}^{N} K_h(\tau_{g(k)}-\tau)(\tau_{g(k)}-\tau)^l y_k \\ &= \sum_{g=1}^{G}\sum_{k\in I_g} K_h(\tau_g-\tau)(\tau_g-\tau)^l y_k \\ &= \sum_{g=1}^{G} K_h(\tau_g-\tau)(\tau_g-\tau)^l c_g \bar{y}_g \ . \end{aligned} \tag{3.151}$$

Likewise, $S_l(\tau)$ in (3.150) is approximated by $\bar{S}_l(\tau)$:

$$\begin{aligned}\bar{S}_l(\tau) &= \sum_{g=1}^{G}\sum_{k\in I_g} K_h(\tau_{g(k)}-\tau)(\tau_{g(k)}-\tau)^l \\ &= \sum_{g=1}^{G} K_h(\tau_g-\tau)(\tau_g-\tau)^l c_g \ . \end{aligned} \tag{3.152}$$

γ estimators (3.147) and (3.148) are obtained by replacing T_l and S_l with their binning analogs $\bar{T}_l$ and $\bar{S}_l$. It follows that the leave-one-out CV statistic (3.29) is approximated by

$$\mathrm{CV}\,(h) = \frac{1}{G}\sum_{g=1}^{G}\left(\bar{y}_g - \hat{\gamma}_{(-g)}(\tau_g)\right)^2 \ , \tag{3.153}$$

where $\hat{\gamma}_{(-g)}$ is the estimator obtained with the gth grid point being left out. Note that, since the grid points are equally spaced, the estimation of γ at G grid points requires computation of order $O(G)$ instead of $O(G^2)$. One dimensional binning techniques can be extended to a multi-dimensional case ($d > 1$).

$\mathrm{tr}(\mathbf{S})$ and $\mathrm{tr}(\mathbf{S}\mathbf{S}^\top)$, which are required for computation of the d.f. of $\hat{\gamma}$, can also be approximated using binning techniques (see Turlach and Wand, 1996). $\mathbf{T}^*(t)$, $\mathbf{K}$ given in Section 3.3.2 are approximated by their binning analogs. Continuing with a one-dimensional case, define a $(G \times 2)$ matrix $\bar{\mathbf{T}}^*(t)$ and a weight matrix $\bar{\mathbf{K}}(h)$ as follows:

$$\bar{\mathbf{T}}^*_\tau = \{1,\, \tau_g - \tau\}_{g=1}^{G} \tag{3.154}$$

$$\bar{\mathbf{K}}_\tau = \oplus_g K\left(\frac{\tau_g-\tau}{h}\right) \ . \tag{3.155}$$

At point τ, the local weighted least squares estimator is obtained by minimizing the

following objective function:

$$\sum_{g=1}^{G} \left(\bar{y}_g - \zeta_0 - \zeta_1(\tau_g - \tau)\right)^2 c_g K\left(h^{-1}(\tau_g - \tau)\right) \ . \tag{3.156}$$

This can be rewritten in matrix form as

$$(\mathbf{C}^{-1}\mathbf{y}_\cdot - \bar{\mathbf{T}}_\tau^{*}\boldsymbol{\zeta})^\top \bar{\mathbf{K}}_\tau \mathbf{C} (\mathbf{C}^{-1}\mathbf{y}_\cdot - \bar{\mathbf{T}}_\tau^{*}\boldsymbol{\zeta}) \ , \tag{3.157}$$

where $\mathbf{y}_\cdot$ is a G-dimensional vector $\{c_g \bar{y}_g\}_{g=1}^G$ and $\mathbf{C} = \oplus_g c_g$. Minimizing (3.157) with respect to $\boldsymbol{\zeta}$ yields the estimator of ζ_0:

$$\hat{\zeta}_0 \ = \ \mathbf{e}_1^\top (\bar{\mathbf{T}}_\tau^{*\top} \bar{\mathbf{K}}_\tau \mathbf{C} \bar{\mathbf{T}}_\tau^{*})^{-1} \bar{\mathbf{T}}_\tau^{*\top} \bar{\mathbf{K}}_\tau \mathbf{y}_\cdot \ . \tag{3.158}$$

The binning analogs to (3.27) and (3.28) are therefore given by

$$\bar{\mathbf{S}} \ = \ \{\bar{\mathbf{s}}_{\tau_g}^\top\}_{g=1}^G \tag{3.159}$$

$$\bar{\mathbf{s}}_{\tau_g}^\top \ = \ \mathbf{e}_1^\top (\bar{\mathbf{T}}_{\tau_g}^{*\top} \bar{\mathbf{K}}_{\tau_g} \mathbf{C} \bar{\mathbf{T}}_{\tau_g}^{*})^{-1} \bar{\mathbf{T}}_{\tau_g}^{*\top} \bar{\mathbf{K}}_{\tau_g} \ . \tag{3.160}$$

Since the kth diagonal element s_{kk} of $\mathbf{S}$ is obtained by

$$\begin{aligned} s_{kk} \ &= \ \mathbf{e}_1^\top (\mathbf{T}_{t_k}^{*\top} \mathbf{K}_{t_k} \mathbf{T}_{t_k}^{*})^{-1} \mathbf{T}_{t_k}^{*\top} \mathbf{K}_{t_k} \mathbf{e}_k \\ &= \ K(0) \mathbf{e}_1^\top (\mathbf{T}_{t_k}^{*\top} \mathbf{K}_{t_k} \mathbf{T}_{t_k}^{*})^{-1} \mathbf{e}_1 \ , \end{aligned} \tag{3.161}$$

$\mathrm{tr}(\mathbf{S}) = \sum_{k=1}^N s_{kk}$ can be approximated as follows:

$$\mathrm{tr}(\mathbf{S}) \approx \sum_{g=1}^{G} \bar{s}_{gg} c_g \ , \tag{3.162}$$

where $\bar{s}_{gg}$, binning analog to (3.161), is given by

$$\bar{s}_{gg} \ = \ K(0) \mathbf{e}_1^\top (\bar{\mathbf{T}}_{\tau_g}^{*\top} \bar{\mathbf{K}}_{\tau_g} \mathbf{C} \bar{\mathbf{T}}_{\tau_g}^{*})^{-1} \mathbf{e}_1 \ . \tag{3.163}$$

Similarly, the kth diagonal element of $\mathbf{S}\mathbf{S}^\top$, denoted here by $[\mathbf{S}\mathbf{S}^\top]_{kk}$, can be approx-

imated as follows:

$$\begin{aligned}[\mathbf{S}\mathbf{S}^\top]_{kk} &= \sum_{k'=1}^{N} s_{kk'}^2 \\ &\approx \sum_{g'=1}^{G} \bar{s}_{gg'}^2 c_g \\ &= [\bar{\mathbf{S}}\mathbf{C}\bar{\mathbf{S}}^\top]_{gg} \ , \end{aligned} \tag{3.164}$$

where $s_{kk'}$ is (k,k')th element of $\mathbf{S}$ and $\bar{s}_{gg'}$ is the (g,g')th element of $\bar{\mathbf{S}}$. Then it follows that

$$\mathrm{tr}[\mathbf{S}\mathbf{S}^\top] \approx \sum_{g=1}^{G} [\bar{\mathbf{S}}\mathbf{C}\bar{\mathbf{S}}^\top]_{gg} \ . \tag{3.165}$$

A high dimensional extension ($d > 1$) of the trace calculations above can be similarly obtained and they are implemented in the `plmm` package.

The accuracy of binning techniques depends on the number of bins G. In the `plmm` package, we set the number of bins to the rounded number of $8\log(N)/d$ if $N > 100$ and otherwise binning is not employed. This number is the default setting of the R function `h.select` for the CV using binning. For accuracy of binning and more details, see Turlach and Wand (1996), Fan and Marron (1994), Hurvich et al. (1998), and references therein.

3.11 Appendix B

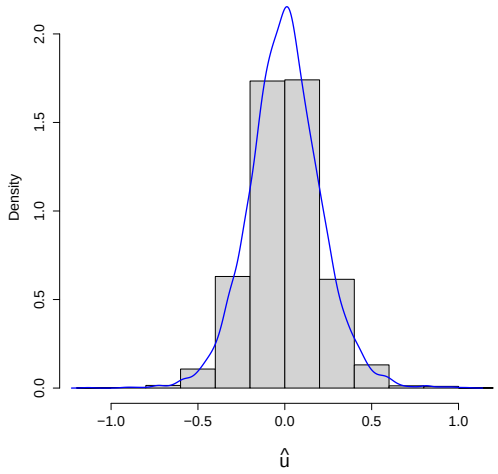


Figure 3.14: Histogram of the predicted random intercepts by plmm (3) (Section 3.8.1).

plmm		*grade*	*age*	age^2	*ttl_exp*	ttl_exp^2	*tenure*	$tenure^2$	*race*	*not_smsa*	*south*
(1)	mean	.0650			.0271	.000276	.0398	-.0020	-.0501	-.1316	-.0890
	sd	.0017			.0027	.000127	.0017	.0001	.0084	.0070	.0065
(2)	mean	.0649	.0379	-.0007	.0258	.000142			-.0538	-.1326	-.0860
	sd	.0017	.0034	.0001	.0027	.000132			.0084	.0070	.0065
(3)	mean	.0691			.0284	-.0000296			-.0444	-.1281	-.0839
	sd	.0017			.0028	.000141			.0084	.0070	.0064

Table 3.14: Summary of the bootstrap sampling distributions (Section 3.8.1).

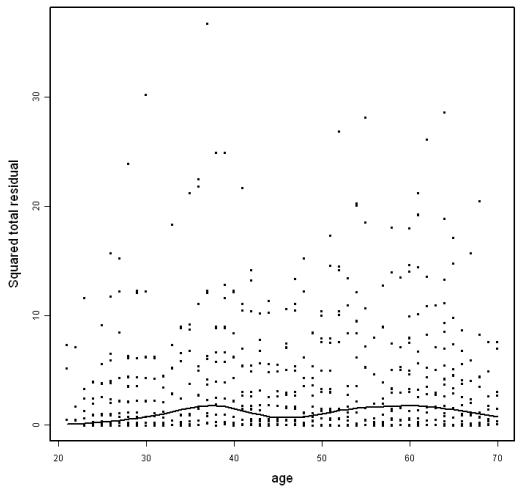


Figure 3.15: Estimated conditional variance function (Section 3.8.2).

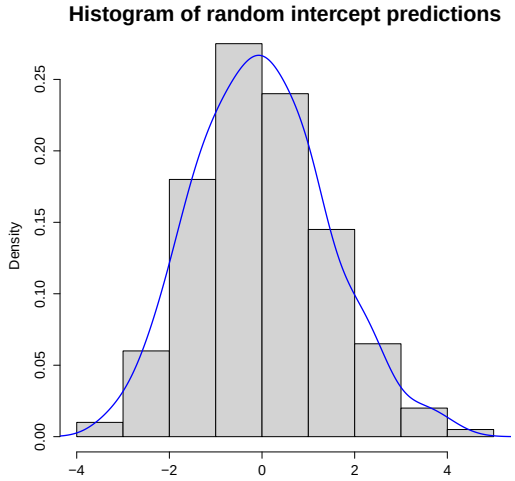


Figure 3.16: Histogram of the predicted random intercepts by plmm (2) (Section 3.8.2).

plmm		linc	insur
(1)	mean	-.0137	1.417
	sd	.2811	.0948
(2)	mean	-.0191	1.470
	sd	.2856	.1315

Table 3.15: Summary of the bootstrap sampling distributions (Section 3.8.2).

3.12 References

Baltagi, B. (2005). *Econometric analysis of panel data*, Volume 13. Wiley.

Baltagi, B. and Y. Chang (1994). Incomplete panels:: A comparative study of alternative estimators for the unbalanced one-way error component regression model. *Journal of Econometrics 62*(2), 67–89.

Bellmann, L., J. Breitung, and J. Wagner (1989). Bias correction and bootstrapping of error component models for panel data: Theory and applications. *Empirical Economics 14*(4), 329–342.

Butar, F. and P. Lahiri (2003). On measures of uncertainty of empirical bayes small-area estimators. *Journal of Statistical Planning and Inference 112*(1-2), 63–76.

Carmack, P., J. Spence, and W. Schucany (2011). Generalized correlated cross-validation (gccv). Technical report, Department of Mathematics, University of Central Arkansas.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik 31*, 377–403.

Datta, G. and P. Lahiri (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica 10*(2), 613–628.

Fan, J. and J. Marron (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics 3*(1), 35–56.

Fuller, W. and G. Battese (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical association 68*(343), 626–632.

González-Manteiga, W., M. Lombardía, M. Martinez-Miranda, and S. Sperlich (2012). Kernel smoothers and bootstrapping for nonparametric mixed-effect models. Revised and resubmitted to the Journal of Multivariate Analysis.

González-Manteiga, W., M. Lombardía, I. Molina, D. Morales, and L. Santamaría (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational statistics & data analysis 51*(5), 2720–2733.

Härdle, W. and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics 21*(4), 1926–1947.

Härdle, W. and M. Müller (2000). Multivariate and semiparametric kernel regression. In M. Schimek (Ed.), *Smoothing and Regression*, pp. 357–391. John Wiley & Sons, Inc.

Härdle, W., M. Müller, S. Sperlich, and A. Werwatz (2004). *Nonparametric and Semiparametric Models*. Heidelberg: Springer Verlag.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Monographs on Statistics and Applied Probability. Chapman and Hall.

Hill, R., W. Griffiths, and G. Lim (2008). *Principles of Econometrics*. John Wiley & Sons.

Hurvich, C., J. Simonoff, and C. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(2), 271–293.

Jiang, J. (1996). Reml estimation: Asymptotic behavior and related topics. *The Annals of Statistics 24*(1), 255–286.

Kackar, R. and D. Harville (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in statistics-theory and methods 10*(13), 1249–1261.

Kenward, M. and J. Roger (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics 53*(7), 983–997.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science 18*(2), 199–210.

Li, Q. and T. Stengos (1994). Adaptive estimation in the panel data error component model with heteroskedasticity of unknown form. *International Economic Review 35*(4), 981–1000.

Li, Q. and A. Ullha (1998). Estimating partially linear panel data models with one-way error components. *Econometric Reviews 17*(2), 145–166.

Lin, X. and R. Carroll (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association 95*(450), 520–534.

Lombardía, M. and S. Sperlich (2008). Semiparametric inference in generalized mixed effects models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 913–930.

Maddala, G. and T. Mount (1973). A comparative study of alternative estimators for variance components models used in econometric applications. *Journal of the American Statistical Association 68*(342), 324–328.

Ohinata, R. and S. Sperlich (2012). Some recent advances in modeling with mixed effects for small areas, multi-level and panel models. Discussion Paper, University of Göttingen.

Opsomer, J., G. Claeskens, M. Ranalli, G. Kauermann, and F. Breidt (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(1), 265–286.

Prasad, N. and J. Rao (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association 85*(409), 163–171.

Rao, J. (2003). *Small Area Estimation*. Wiley Series in Survey Methodology. John Wiley.

Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society 56*(4), 931–954.

Ruckstuhl, A., A. Welsh, and R. Carroll (2000). Nonparametric function estimation of the relationship between two repeatedly measured variables. *Statistica Sinica 10*(1), 51–72.

Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge Series on Statistical and Probabilistic Mathematics. Cambridge University Press.

Searle, S., G. Casella, and C. McCulloch (1992). *Variance components*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley.

Shao, J. and D. Tu (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics. Springer Verlag.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B (Methodological) 50*(3), 413–436.

Sperlich, S. and M. Lombardía (2010). Local polynomial inference for small area statistics: estimation, validation and prediction. *Journal of Nonparametric Statistics 22*(5), 633–648.

StataCorp (2009). *Stata Longitudinal-Data/Panel-Data Reference Manual: Release 11*. StataCorp LP.

Stone, C. (1980). Optimal rates of convergence for nonparametric estimators. *The annals of Statistics 8*(6), 1348–1360.

Stone, C. (1985). Additive regression and other nonparametric models. *The annals of Statistics 13*(2), 689–705.

Stukel, D. and J. Rao (1997). Estimation of regression models with nested error structure and unequal error variances under two and three stage cluster sampling. *Statistics & probability letters 35*(4), 401–407.

Su, L. and A. Ullah (2007). More efficient estimation of nonparametric panel data models with random effects. *Economics Letters 96*(3), 375–380.

Su, L. and A. Ullah (2010). Nonparametric and semiparametric panel econometric models: Estimation and testing. In A. Ullah and D. E. A. Giles (Eds.), *Handbook of Empirical Economics and Finance*. Chapman and Hall/CRC.

Swamy, P. (1970). Efficient inference in a random coefficient regression model. *Econometrica: Journal of the Econometric Society 38*(2), 311–323.

Swamy, P. and S. Arora (1972). The exact finite sample properties of the estimators of coefficients in the error components regression models. *Econometrica: Journal of the Econometric Society 40*(2), 261–275.

Turlach, B. and M. Wand (1996). Fast computation of auxiliary quantities in local polynomial regression. *Journal of Computational and Graphical Statistics 5*(4), 337–350.

Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer.

Vilar Fernández, J. and M. Francisco Fernández (2002). Local polynomial regression smoothers with ar-error structure. *Test 11*(2), 439–464.

Wand, M. (1994). Fast computation of multivariate kernel estimators. *Journal of Computational and Graphical Statistics 3*(4), 433–445.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics 14*(4), 1261–1295.

Yatchew, A. (1998). Nonparametric regression techniques in economics. *Journal of Economic Literature 36*(2), 669–721.

You, J., X. Zhou, and Y. Zhou (2010). Statistical inference for panel data semiparametric partially linear regression models with heteroscedastic errors. *Journal of Multivariate Analysis 101*(5), 1079–1101.

Yu, K. and M. Jones (2004). Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association 99*(465), 139–144.

4 Index Model

Comparison of the Principal Component and Directed Principal Component Index Models: An Empirical Study

Ren Ohinata

Institut für Statistik und Ökonometrie, Georg-August Universität Göttingen

Abstract

Principal component regression is widely used to construct a household's welfare indicator from a number of variables. This approach has an intrinsic weakness that principal components, thus weights crucial to building an indicator, are determined outside the regression. An alternative approach, use of directed principal components, provides optimal weights with respect to the response. In this essay these approaches are used in semiparametric index model estimation and compared through an analysis of household survey data. Bootstrap inference is also illustrated for the alternative approach. The data analysis demonstrates the potential of the alternative approach as a tool of exploratory data analysis. The directed principal component index model requires continuous data. Category means data approach proposed by Kolenikov and Angeles (2009) is applied to deal with categorical variables.

Key words: Index model; Principal component analysis; Semiparametric regression; Welfare indicator

4.1 Introduction

In a study of micro-economic data, it is often of great interest to construct a welfare "indicator" of the socio-economic status (SES) of a household. Income or expenditure data can be used, alone or combined with other variables, as a measure of a household's welfare. When those data are unavailable or unreliable due to reporting errors, researchers turn to variables that serve as proxies. Because one single proxy variable does not fully capture SES, it is a conventional practice to construct a welfare indicator composed of many proxy variables. As Kolenikov and Angeles (2009) and references therein pointed out, a number of categorical variables are used as proxies in empirical studies. While categorical data are typically easier to collect and less prone to reporting errors than income or expenditure data, their use as a measure of the SES tends to involve large measurement errors. This problem is also alleviated by incorporating into the indicator as many proxy categorical variables as possible. This approach reduces measurement errors and improves reliability and stability of the indicator.

Given a number of regressors, one way to construct a welfare indicator is to estimate a regression model

$$y = m(\mathbf{x}^\top \boldsymbol{\beta}) + u \ , \tag{4.1}$$

where $\mathbf{x}$ is a d-dimensional vector of regressors; $\boldsymbol{\beta}$ is a vector of weights, called "index vector"; and the regression error u has zero conditional mean $\mathrm{E}[u|\mathbf{x}] = 0$. Here m is some unknown smooth link function which links the response $y \in \mathbb{R}$ and "index" $\mathbf{x}^\top \boldsymbol{\beta} \in \mathbb{R}$. Once model (4.1) is estimated with a given set of weights, SES is predicted by the indicator $\hat{m}(\mathbf{x}^\top \boldsymbol{\beta})$. Prediction quality mainly depends on the appropriateness of two components: link function m and index vector $\boldsymbol{\beta}$.

Even though a parametric form can be assumed for m, it is reasonable on many occasions to assume no functional form for m and estimate it nonparametrically. This is because, without reliable knowledge about the functional form, a restrictive parametric specification of m may result in a severely biased indicator. In nonparametrics, model (4.1) with a nonparametric function m is called a single index model (SIM). Model (4.1) is a useful alternative to a fully nonparametric model such as $y = m^*(\mathbf{x}) + u$. The estimation of m^* faces a challenge of the so-called "curse of dimensionality": with an increasing number of the design space dimension d, the estimation of m^* loses precision and becomes even infeasible. Use of an index helps to avoid the curse by reducing the number of function arguments from d to one. The SIM can be extended to a multi-index model (MIM) which has M function

arguments:

$$y = m(\mathbf{x}^\top \boldsymbol{\beta}_1, \mathbf{x}^\top \boldsymbol{\beta}_2, \ldots, \mathbf{x}^\top \boldsymbol{\beta}_M) + u \ , \tag{4.2}$$

where $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_M$ are orthogonal index vectors and M is assumed to be less than d and reasonably small. For the construction of a welfare indicator from a large number of regressors, index models serve as a "semiparametric" compromise between restrictive parametric and flexible but hardly estimable fully nonparametric model.

Creating M indices from d regressors means reducing the dimension of the design space of regression from d to M. Dimension reduction is meaningful and useful when there are relatively strong correlations between regressors, in other words, when information in a d-dimensional design space is concentrated effectively in a smaller M-dimensional subspace. As is seen in model (4.1) and (4.2), dimension reduction is an intrinsic process of estimating an index model. Dimension reduction methods to determine the optimal index vectors are therefore of great concern.

In order to obtain appropriate index vectors, one may turn to, for example, some weighting scheme, expert opinions, or monetary values. Among others, principal component analysis (PCA) is a widely used statistical method. For model (4.2) PCA provides M eigenvectors (also called principal components, PC) corresponding to the largest eigenvalues of the covariance or correlation matrix of $\mathbf{x}$; these vectors are used as index vectors. We call this type of model "PC MIM", and likewise we say "PC SIM" when $M = 1$.

Use of PCA appears reasonable because such eigenvectors can identify the subspace which concentrates most of the information contained in the regressors. However, an important potential weakness of the PCA in the determination of index vectors is that eigenvectors are calculated without any reference to the response variable. Even though the subspace spanned by those M eigenvectors retain as much information contained in the design space as possible, it may have lost information relevant to the functional relationship with the response. While the performance of an index model depends on the choice of index vectors, it is not known a priori which PC is the most associated, or even whether PCs are associated, with the response. Even if each regressor relates enough with the response, an inappropriate index vector may distort the estimation of their functional relationship. This problem may be mitigated by increasing the number of indices in the index model. Nonetheless the essential weakness remains the same.

Let's call an index vector "directed principal component" (DPC) that relates the most to the response variable through a certain functional relationship and consider an alternative SIM

$$y = g(\mathbf{x}^\top \boldsymbol{\gamma}) + e \ , \tag{4.3}$$

which we will call "DPC SIM". Use of a DPC as the index vector $\boldsymbol{\gamma}$ overcomes the weakness of the PC SIM. In estimation of model (4.3), the DPC is determined simultaneously with the link function g as a vector that is "directed" to the gradient of g. The DPC SIM can also be extended to a "DPC MIM" with M DPCs and hence M indices. DPC index models are expected to provide a better welfare indicator in comparison with PC index models. Another advantage of the DPC index model is that it enables a statistical test of the effects of regressors. This is impossible for the PC index model because PCs are determined independent of the response.

In this essay we apply index models to Bangladesh demographic and household survey data and compare PC and DPC models. The analysis illustrates the potential of the DPC model to capture latent data structure. We use the statistical software R, in particular, the `sm` package for kernel regression and the `EDR` package for implementation of the DPC index model estimation.

The rest of the essay is structured as follows. Section 4.2 describes some theoretical background of our data analysis. We sketch the idea and estimation procedure of the DPC SIM in Section 4.2.1 and bootstrap inference in Section 4.2.2. Use of categorical proxy variables poses a problem to the DPC index model because its estimation requires data to be continuous. Section 4.2.3 presents techniques to deal with categorical regressors, followed by Section 4.2.4 where we describe a technical problem of bandwidth selection for regression involving categorical regressors. Applications of index models are presented in Section 4.3. Concluding remarks follow in Section 4.4. More technical details of the DPC index model estimation are given in Appendix A.

4.2 Method

4.2.1 Estimation of single index model

In this section we sketch the basic ideas and estimation procedure for the DPC SIM. More details of the estimation procedure are given for the MIM in Appendix A.

Suppose a SIM takes the form

$$\begin{aligned} y &= f(\mathbf{x}) + e \\ f(\mathbf{x}) &= g(\mathbf{x}^\top \boldsymbol{\gamma}) \ , \end{aligned} \tag{4.4}$$

where f and g are smooth functions $f : \mathbb{R}^d \to \mathbb{R}$, $g : \mathbb{R} \to \mathbb{R}$; $\boldsymbol{\gamma}$ is a d-dimensional index vector. $\boldsymbol{\gamma}$ is normalized for model identification, i.e. $\|\boldsymbol{\gamma}\| = 1$ ($\|\cdot\|$ is the

Euclidean length). A real-valued d-dimensional vector $\mathbf{x}$ is assumed to have a support of $[-1,1]^d$ (if necessary, data are transformed accordingly).

Several estimation approaches are known in semiparametrics literature. An M-estimation-type approach estimates $\boldsymbol{\gamma}$ by solving a minimization problem

$$\hat{\boldsymbol{\gamma}} \quad = \quad \arg\min_{\boldsymbol{\gamma}} \sum_{i=1}^{n} \phi(y_i, \hat{g}(\mathbf{x}^\top \boldsymbol{\gamma})) \ , \tag{4.5}$$

where n is the sample size; $\hat{g}$ is some estimate of the link function g; and the contrast function ϕ is, for example, $-\log L(\hat{g}, \boldsymbol{\gamma}; y_i, \mathbf{x}_i)$ for the semiparametric maximum likelihood estimation or $(y_i - \hat{g}(\mathbf{x}_i^\top \boldsymbol{\gamma}))^2$ for the semiparametric least squares estimation.

The analysis of this essay is based on an alternative approach called the average derivative estimation (Stoker 1986, Powell et al. 1989), which we briefly present below. The gradient vector of (4.4) at point $\mathbf{x}_i$ is given by

$$\nabla_f(\mathbf{x}_i) \quad = \quad g'(\mathbf{x}_i^\top \boldsymbol{\gamma})\boldsymbol{\gamma} \ , \tag{4.6}$$

where $g'(\mathbf{x}_i^\top \boldsymbol{\gamma}) = dg/d(\mathbf{x}_i^\top \boldsymbol{\gamma})$. This implies that the gradient is proportional to the index vector $\boldsymbol{\gamma}$, directed in the same direction as $\boldsymbol{\gamma}$ at each point of $\mathbf{x}$. Since $\mathrm{E}[\nabla_f(\mathbf{x}_i)] = \mathrm{E}[g'(\mathbf{x}_i^\top \boldsymbol{\gamma})]\boldsymbol{\gamma}$, a natural idea is to estimate the index vector from a linear functional of the gradient $\mathbf{b} = \frac{1}{n}\sum_{i=1}^{n} \nabla_f(\mathbf{x}_i)$, which will be estimated by

$$\hat{\mathbf{b}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\nabla}_f(\mathbf{x}_i) \ . \tag{4.7}$$

Then the estimator of the index vector is obtained by

$$\hat{\boldsymbol{\gamma}} = \hat{\mathbf{b}} / \left\| \hat{\mathbf{b}} \right\| \ . \tag{4.8}$$

The estimation of $\boldsymbol{\gamma}$ requires the estimation of $\mathbf{b}$, which in turn requires the estimation of $\nabla_f(\mathbf{x}_i)$ at each point. The following kernel regression simultaneously estimate $f(\mathbf{x}_i)$ and $\nabla_f(\mathbf{x}_i)$:

$$\begin{pmatrix} \hat{f}(\mathbf{x}_i) \\ \hat{\nabla}_f(\mathbf{x}_i) \end{pmatrix} \quad = \quad \underset{\xi_0 \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^d}{\arg\min} \sum_{j=1}^{n} \left(y_j - \xi_0 - (\mathbf{x}_j - \mathbf{x}_i)^\top \boldsymbol{\xi} \right)^2 K\left(\frac{\|\mathbf{x}_j - \mathbf{x}_i\|^2}{h_1^2} \right) \ , \tag{4.9}$$

where K is a kernel weighting function and h_1 a bandwidth of a spherical window around $\mathbf{x}_i$. Estimation by (4.9) faces the curse of dimensionality when d is large. Hristache, Juditsky, and Spokoiny (2001) proposed an iterative estimation by kernel

regression with an ellipsoidal window rather than a spherical window so that the number of observations falling in the window becomes large enough for estimation. The basic idea is as follows.

Approximation of $g(\mathbf{x})$ by the tangent hyperplane $\Delta y = \nabla_f(\mathbf{x})^\top \Delta\mathbf{x}$ in the directions orthogonal to the gradient is relatively good because $f(\mathbf{x})$ and hence $\nabla_f(\mathbf{x})$ do not change much in those directions. This implies that estimation can be improved using an ellipsoidal window $\{\mathbf{x} : \left\|(\mathbf{x}-\mathbf{x}_i)^\top\boldsymbol{\gamma}\right\| \leq h\}$ which contain enough observations inside. Such a window is obtained by expanding a spherical window in the directions orthogonal to the gradient and shrinking it in the direction of the gradient.[1]

In practice, estimation is improved by iterative estimation using an ellipsoidal window. After obtaining a pilot estimate $\hat{\boldsymbol{\gamma}}_1$ through (4.9), $\boldsymbol{\gamma}$ is reestimated with an estimator of $\nabla_f(\mathbf{x}_i)$ by

$$\begin{pmatrix} \hat{f}^{(2)}(\mathbf{x}_i) \\ \hat{\nabla}_f^{(2)}(\mathbf{x}_i) \end{pmatrix} = \underset{\xi_0\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^d}{\arg\min} \sum_{j=1}^{n} \left(y_j - \xi_0 - (\mathbf{x}_j-\mathbf{x}_i)^\top\boldsymbol{\xi}\right)^2 K\left(\frac{\|\mathbf{S}_2(\mathbf{x}_j-\mathbf{x}_i)\|^2}{h_2^2}\right), \tag{4.10}$$

where $\mathbf{S}_2 = (\mathbf{I} + \rho_2^{-2}\hat{\mathbf{b}}_1\hat{\mathbf{b}}_1^\top)^{1/2}$ creates an ellipsoidal window $\{\mathbf{x}|\, \|\mathbf{S}_2(\mathbf{x}-\mathbf{x}_i)\| < h_2\}$ with a bandwidth $h_2 > h_1$ and a parameter $\rho_2 < 1$ to control the shape of the ellipsoid.[2]

Estimation of $\boldsymbol{\gamma}$ is repeated in iteration process with an increasing h and a decreasing ρ and an iteratively updated $\mathbf{S}$. The estimator of the index vector $\boldsymbol{\gamma}$ is shown to be consistent with the parametric convergence rate $\sqrt{n}$. Finally, after the iterative process, the link function g is estimated with $\hat{\boldsymbol{\gamma}}$ by kernel regression.

4.2.2 Bootstrap inference

When the effect of regressors on the response variable is of interest, the DPC index model has an advantage over the PC model. Because PCs are determined without any reference to the relationship between the response and regressors, a test of the significance of index vector coefficients does not provide any statistical evidence on

[1] $\left\|(\mathbf{x}-\mathbf{x}_i)^\top\boldsymbol{\gamma}\right\| = 0 < h$ holds for a vector $(\mathbf{x}-\mathbf{x}_i)$ lying orthogonal to the index vector. To the contrary, the smaller the angle which $(\mathbf{x}-\mathbf{x}_i)$ and the index vector make, it is less likely that $\left\|(\mathbf{x}-\mathbf{x}_i)^\top\boldsymbol{\gamma}\right\| \leq h$ holds.

[2] Since $\mathbf{I} + \rho^{-2}\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^\top$ is symmetric and positive definite, $(\mathbf{I} + \rho^{-2}\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^\top)$ can be given by $(\mathbf{I} + \rho^{-2}\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^\top)^{1/2}(\mathbf{I}+\rho^{-2}\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^\top)^{1/2}$. $(\mathbf{I}+\rho^{-2}\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^\top)$ has the largest eigenvalue $1+\rho^{-2} > 1$ whose eigenvector $\hat{\boldsymbol{\gamma}}$ and the other eigenvalues are all 1 with eigenvectors being orthogonal to $\hat{\boldsymbol{\gamma}}$. Therefore, the minor axis of this ellipsoid, to which direction the ellipsoid is compressed, is along $\hat{\boldsymbol{\gamma}}$. In literature $\mathbf{S}_2$ is sporadically given by $\mathbf{S}_2 = (\mathbf{I}+\rho^{-2}\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^\top)^{-1/2}$. However, with $(\mathbf{I}+\rho^{-2}\hat{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}^\top)^{-1/2}$, $\hat{\boldsymbol{\gamma}}$ will be along the major axis, to which direction the window is expanded.

the effect of regressors. In contrast, DPCs are determined in relation to the response; the effect of a regressor can be statistically evaluated by testing the significance of DPC coefficients.

However, their sampling distributions are generally hard to obtain analytically. Thus we use a resampling method, the wild bootstrap, to simulate the distributions. We note that the wild bootstrap retains the first three moments of the regression errors and thus potential heteroskedasticity. For the wild bootstrap, see Wu (1986) and Shao and Tu (1995) among others. In case of the SIM, we conduct bootstrap resampling of size R as follows.

1. Obtain estimates $\hat{\boldsymbol{\gamma}}$ and $\hat{g}$ from the original sample.
2. Calculate the residuals $\hat{e}_i$ by $y_i - \hat{g}(\mathbf{x}_i^\top \hat{\boldsymbol{\gamma}})$.
3. For the rth bootstrap sampling ($r = 1, \ldots, R$), obtain an n-dimensional vector of the bootstrap responses by $y_i^{(r)} = \hat{g}(\mathbf{x}_i^\top \hat{\boldsymbol{\gamma}}) + z_i^{(r)} \hat{e}_i$ with $z_i^{(r)} \sim \mathcal{N}(0, 1)$.
4. Obtain the rth bootstrap estimates $\hat{\boldsymbol{\gamma}}^{(r)}$ and $\hat{g}^{(r)}$ from the rth bootstrap sample $\{y_i^{(r)}, \mathbf{x}_i\}_{i=1}^n$.
5. Repeat step 3. and 4. for R times.

Bootstrap resampling for the MIM is analogous to the above. The bootstrap procedure provides an approximation to the sampling distributions of DPC coefficient estimators, and a bootstrap confidence interval can be built from the R bootstrap estimates.

4.2.3 Use of categorical variables

Kolenikov and Angeles (2009) studied problems of the use of categorical data in PCA. Since indices are constructed by linear transformation of regressors, indices are not necessarily independent unless the regressors are normally distributed. Orthogonal transformation certainly guarantees zero sample correlation between the indices, but not their independence. In addition, the moment (Pearson) correlation of categorical variables is an underestimation in absolute value of the "true" correlation between their underlying joint standard normal variables. This downward bias may influence the standard PCA, which is based on sample covariances or correlations. On the other hand, recall that the DPC index model is based on the average derivative estimation and it requires continuous variables (or at least real-valued variables of interval-scale type). Lack of interpretable numerical distance between categories is a serious problem for the DPC index model.

In the PCA framework Kolenikov and Angeles (2009) investigated several approaches to those problems. Their approaches include: PCA using polychoric and polyserial correlations and PCA using moment correlations of "category means data". Horowitz and Härdle (1996) proposed an approach to handle categorical regressors in the index model framework. However, their approach is infeasible for more than one or two categorical regressors involved in the model. In our data analysis, we use category means data for regressors of index models.

Polychoric correlation: Analogously to the standard assumption made for the response variable in the ordered probit or logit regression model, it is reasonable to assume a latent multivariate standard normal distribution which underlie categorical variables.

Suppose a latent variable $X_p^* \sim \mathcal{N}(0,1)$ underlying a categorical variable X_p with K_p categories $\{k_p\}_{k_p=1}^{K_p}$. Then divide the range of X_p^* into K_p subranges with a set of cutpoints: $\boldsymbol{\alpha}_p = (\alpha_{p,0}, \alpha_{p,1}, \ldots, \alpha_{p,K_p})$ with $\alpha_{p,0} = -\infty$ and $\alpha_{p,K_p} = \infty$. It is assumed that the outcome $X_p = k_p$ is observed if $\alpha_{p,k-1} < x_p^* < \alpha_{p,k}$. Likewise, suppose a latent variable $X_q^* \sim \mathcal{N}(0,1)$ for a categorical variable X_q with K_q categories and $K_q + 1$ cutpoints. Their joint distribution is given by

$$\begin{pmatrix} X_p^* \\ X_q^* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho_{pq} \\ \rho_{pq} & 1 \end{pmatrix}\right) . \tag{4.11}$$

ρ_{pq} is the polychoric correlation between X_p and X_q. The probability of observing an outcome $(x_p = k_p,\, x_q = k_q)$ is given by

$$\begin{aligned} \Pr[X_p = k_p,\, X_q = k_q] \;=\; & \Phi_2(\alpha_{p,k}, \alpha_{q,k}; \rho_{pq}) - \Phi_2(\alpha_{p,k-1}, \alpha_{q,k}; \rho_{pq}) - \\ & \Phi_2\left(\alpha_{p,k}, \alpha_{q,k-1}; \rho_{pq}\right) + \Phi_2\left(\alpha_{p,k-1}, \alpha_{q,k-1}; \rho_{pq}\right) , \end{aligned} \tag{4.12}$$

where $\Phi_2(\cdot)$ is the cumulative distribution function of two-dimensional standard normal distribution. The cutpoint vector $\boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}_q$, and correlation coefficient ρ_{pq} are simultaneously estimated by maximizing the following likelihood:

$$L(\boldsymbol{\alpha}_p, \boldsymbol{\alpha}_q, \rho_{pq}; \mathbf{x}_p, \mathbf{x}_q) = \prod_{i=1}^{n} \pi(x_{pi}, x_{qi}; \boldsymbol{\alpha}_p, \boldsymbol{\alpha}_q, \rho_{pq}) , \tag{4.13}$$

where $\pi(x_{pi}, x_{qi}; \boldsymbol{\alpha}_p, \boldsymbol{\alpha}_q, \rho_{pq})$ is the probability for the ith observation (x_{pi}, x_{qi}).

Polyserial correlation: The polyserial correlation between a categorical variable X_p and a standard normal variable X_q is given by ρ_{pq} as in the following:

$$\begin{pmatrix} X_p^* \\ X_q \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho_{pq} \\ \rho_{pq} & 1 \end{pmatrix}\right) . \tag{4.14}$$

Probability of observing an outcome $\{X_p = k_p,\, X_q = x_q\}$ is given by

$$\begin{aligned} \Pr[X_p = k_p, X_q = x_q] &= \Pr[\alpha_{p,k-1} < X_p^* < \alpha_{p,k} | X_q = x_q]\phi(x_q) \\ &= \left(\Phi(\alpha_{p,k} - \mathrm{E}[X_p^* | X_q = x_q]) - \Phi(\alpha_{p,k-1} - \mathrm{E}[X_p^* | X_q = x_q])\right)\phi(x_q) \\ &= \left(\Phi(\alpha_{p,k} - \rho_{pq}x_q) - \Phi(\alpha_{p,k-1} - \rho_{pq}x_q)\right)\phi(x_q) , \end{aligned} \tag{4.15}$$

where ϕ is the standard normal probability density function. The likelihood to be maximized is given by

$$L(\boldsymbol{\alpha}_p, \rho_{p,q}; \mathbf{x}_p, \mathbf{x}_q) = \prod_{i=1}^{n} \pi(x_{pi}, x_{qi}; \boldsymbol{\alpha}_p, \rho_{pq}) , \tag{4.16}$$

where $\pi(x_{pi}, x_{qi}; \boldsymbol{\alpha}_p, \rho_{pq})$ is the probability of observing the ith observation (x_{pi}, x_{qi}).

Since the polychoric and polyserial correlations are estimated by maximum likelihood estimation, their estimators are consistent, asymptotically normal and efficient. In practice, likelihood maximization is often computationally infeasible for a high-dimensional multivariate distribution with parameters $(\boldsymbol{\alpha}_1^\top, \boldsymbol{\alpha}_2^\top, \ldots, \boldsymbol{\rho}^\top)$. One way to circumvent computational difficulties is to use a "two-step estimation" approach. The two-step approach proceeds as follows. First, the cutpoints $\{\alpha_{p,k}\}$ are estimated by

$$\hat{\alpha}_{p,k} = \Phi^{-1}\left(\frac{-0.5 + I(x_p \leq k_p)}{N}\right) \; \forall\, p, k , \tag{4.17}$$

where I is an indicator function. Then $\boldsymbol{\rho}$ is estimated by maximizing the joint normal likelihood function (4.13) or (4.16) with respect to $\boldsymbol{\rho}$ where $\boldsymbol{\alpha}_p$ and $\boldsymbol{\alpha}_q$ are replaced with their estimates by (4.17).

A polychoric/-serial correlation matrix is constructed by replacing the off-diagonal elements of the moment correlation matrix with the corresponding estimates of $\boldsymbol{\rho}$. It has been shown that the deviation of a two-step estimate of ρ_{pq} from its ML estimate is negligible (Olsson, 1979; Maydeu-Olivares et al., 2009).

Category means data: PCA can also be conducted using the moment correlation matrix of category means data. A category mean is defined for each category of a categorical variable as the mean of its underlying standard normal distribution

conditional on the category in question. For example, the category mean of the kth category of variable X_p is given by[3]

$$\begin{aligned} \mathrm{E}\left[X_p^* | X_p = k_p\right] &= \int_{\alpha_{p,k-1}}^{\alpha_{p,k}} u\phi\left(u\right) du \\ &= \frac{\phi(\alpha_{p,k-1}) - \phi(\alpha_{p,k})}{\Phi(\alpha_k) - \Phi(\alpha_{k-1})} \ . \end{aligned} \tag{4.18}$$

For implementation, cutpoints $\{\alpha_{p,k}\}$ are estimated by (4.17). In contrast to crude category data, category means reflect the underlying normal distribution and distances between category means are more informative.

Kolenikov and Angeles (2009) compared PCA based on the polychoric/-serial correlation matrix and the moment correlation matrix of category means data. They found no major difference between these alternatives. Their study also showed that PCA based on the moment correlation matrix of crude categorical data yielded a comparable result. This means to us that the DPC index model using category means data can reasonably be compared with PC index models based on any of the three types of correlation matrix (crude, polychoric/-serial and category means).

4.2.4 Bandwidth selection for nonparametric link function estimation

Along with the index space estimation, the link function is estimated by kernel regression. In literature and practice, the ordinary leave-one-out cross validation is one of the most widely used bandwidth selection methods for kernel regression. However, link functions estimated in our studies often gave an optical impression that the ordinary cross validation chose too small a bandwidth, resulting in undersmoothing of function estimates. Categorical variables with a relatively small number of categories collect masses of observations at specific points in the data space. This fact leads to a problem that, no matter how much undersmoothing may result, the ordinary cross validation tends to select too small a bandwidth which yields a curve passing through the middle of each observation mass.

As a remedy for this problem, we turned to binning technique. This technique is usually used to deal with a large data set that makes the ordinary cross validation prohibitively computer-intensive and time-consuming. Binning-based cross validation circumvents computational problems by creating a reasonable number of bins and collecting neighboring observations into their nearest bin. The optimal bandwidth is selected by cross-validating the bin averages of binned data. For details of

[3]The category mean is defined incorrectly on p.137 in Kolenikov and Angeles (2009) and is given modified above.

binning, see, for example, Turlach and Wand (1996), Fan and Marron (1994). In our data analysis, after the estimation of the indices we estimated the link function m of (4.1) with bandwidths obtained using binning techniques.

With binning technique applied, a mass or masses of observations collected in a bin are represented by one average. A crucial aspect of binning applied to a model involving categorical regressors is that bandwidth selection becomes overly sensitive to the bin width (i.e. the number of bins). This is because a small change of bin width makes observation masses suddenly fall in or out of a bin. In our studies we used an arbitrary "rule of thumb" number of bins to stabilize bandwidth selection.[4] See Figure 4.1 for an example of large fluctuations and stabilization of the selected bandwidths in response to the number of bins. In general, when the number of bins increases, the bandwidth selected converges in general to the one selected by the ordinary cross validation. However, this is not the case when data contain masses of observations at specific points like in our studies. This implies that cross validation statistics used in our study are not approximations to the ordinary cross validation statistics.

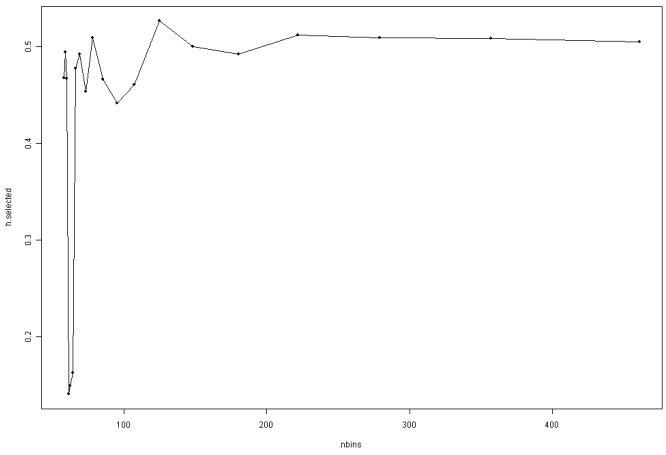


Figure 4.1: Example of bandwidth selection using binning techniques for a model involving categorical regressors. The bandwidth selected is on the vertical axis and the number of bins on the horizontal axis.

[4] We used function `h.select` of R package `sm`. The default number of the bins is set to $8 * log(n)/M$ (binning is used only for $n > 100$). In our studies we set the number of bins arbitrarily to 5 times the default.

4.3 Application

4.3.1 Data and models

This section presents an empirical comparison of PC and DPC index models. We analyze a data set of 1,451 households from the Bangladesh Demographic and Health Survey. We search for an indicator to predict the body mass index (BMI) of the mother of a household.[5] The variables used as regressors for the construction of our indicator consist of two education variables, five housing characteristics variables and two durable goods variables. BMI and the education variables are considered continuous; the housing characteristics and the durable goods variables are ordered categorical variables. Table 4.1 describes the variables.[6]

BMI	mother's body mass index	Response
$faEdu$	father's years spent in school	Education
$moEdu$	mother's years spent in school	Education
$water$	3 ordered categories of the source of water	Housing characteristics
$toilet$	3 ordered categories of the type of toilet	Housing characteristics
$cook$	3 ordered categories of the type of main cooking fuel	Housing characteristics
$floor$	2 categories of the main floor material	Housing characteristics
$roof$	3 ordered categories of the main roof material	Housing characteristics
$trans$	3 ordered categories of the ownership of transport means	Durable goods
$hhItem$	4 ordered categories of the ownership of durable goods	Durable goods

Table 4.1: Data description. Data were prepared and provided by Dr. Nils-Hendrik Klann.

Prior to analysis, we assumed positive association between BMI and each of the regressors. Categories were ordered in number according to this prior assumption. The continuous variables were standardized (BMI was only normalized). We transformed all the categorical data into category means data.[7]

[5] The BMI is used not only as a direct measurement of nutritional status. A mother's BMI can be used as a proxy measure for her child's nutritional status whose measurements are of poor quality or unavailable.

[6] The moment correlations, histograms and scatter plots of the original data as well as the polychoric/serial correlation matrix are given in Appendix B.

[7] Since category means data depend on the cutpoints estimated by (4.17), we performed a rough test on their accuracy. In general, if there are d regressors including at least one categorical variable, there will be $d-1$ bivariate likelihood functions of either (4.13) or (4.16) to estimate a cutpoint of a categorical variable, say, α^*. Consequently there will be $d-1$ ML estimates $\{\hat{\alpha}_j^*\}_{j=1}^{d-1}$. The

Table 4.2 is the correlation matrix of the category means data. BMI is positively, although generally weakly, correlated with all the regressors; positive correlations are in agreement with our prior assumption.

	$faEdu$	$moEdu$	$water$	$toilet$	$cook$	$floor$	$roof$	$trans$	$hhItem$
BMI	.23	.20	.10	.25	.27	.37	.25	.04	.31
$faEdu$		.62	.16	.37	.31	.36	.33	.17	.33
$moEdu$			.14	.30	.27	.31	.30	.13	.29
$water$				.23	.28	.29	.20	-.04	.19
$toilet$					.43	.47	.40	.08	.43
$cook$						.50	.35	.05	.41
$floor$							.47	.04	.51
$roof$								.14	.32
$trans$									.17

Table 4.2: Sample correlation matrix. Category means data are used for the categorical variables.

Figure 4.2 displays the histograms of all the variables. Effect of the transformation of categorical variables can be seen, for example, in the histogram of $hhItem$. Its categories were originally denoted by ("1", "2", "3", "4"). As a result of assigning a real category mean value to each category, bars of the histogram are not equispaced. Even though the education variables clearly deviate from normality, we only standardized them without any further transformation.

intersection of their $d-1$ 95% confidence intervals is roughly $\bigcap_{j=1}^{d-1}[\hat{\alpha}_j^* - 2\text{se}(\hat{\alpha}_j^*); \hat{\alpha}_j^* + 2\text{se}(\hat{\alpha}_j^*)]$. For the data we considered ($d = 9$), any cutpoint estimate was contained in this intersection of the confidence intervals. Thus we regard our cutpoint estimates as reasonable approximations to maximum likelihood estimates.

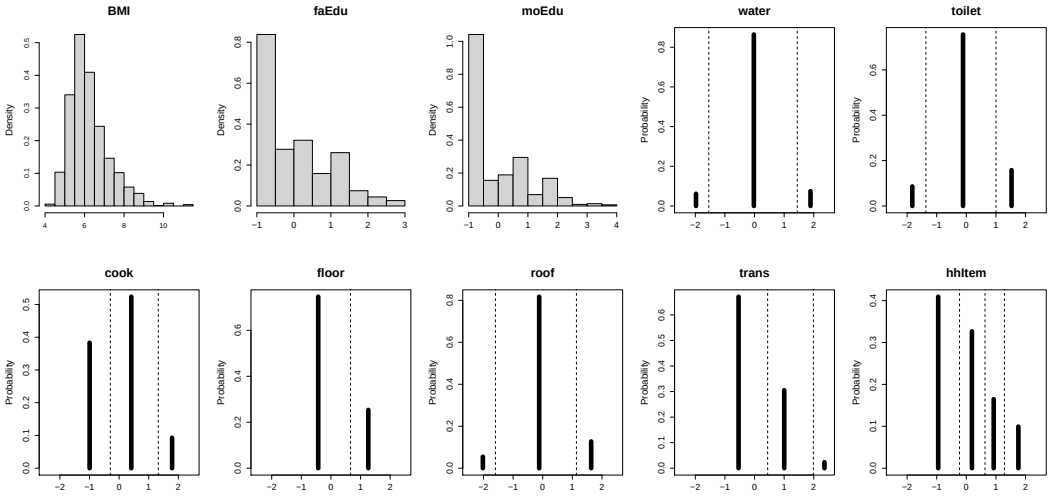


Figure 4.2: Histograms of the sample data. Dashed lines are drawn at cutpoints in the plots of the categorical variables.

In spite of statistical confirmation of the positive relationship between BMI and the regressors, a closer look into data provides somewhat complicated perspectives. Let's define "the 1st quartile group" as a group whose members are households with BMI values less than the first quartile of BMI data and let it be denoted by "Q_1". Likewise, let Q_j denote the "jth quartile group" ($j = 2, 3$), which is a group composed of the households with BMI between the jth and $(j-1)$th quartiles of BMI data; and finally let "Q_4" contain the households with BMI above the third quartile.[8] Table 4.3 presents the quartile group averages of data with respect to each variable. Figure 4.3 visualizes these quartile group averages. Heights of the bars are adjusted so that the height of the average of Q_1 is unity.

	BMI	$faEdu$	$moEdu$	$water$	$toilet$	$cook$	$floor$	$roof$	$trans$	$hhItem$
Q_4	7.625	.415	.399	.176	.362	.396	.435	.294	.070	.410
Q_3	6.327	.013	-.071	-.094	-.037	-.052	-.047	-.014	.025	-.016
Q_2	5.784	-.250	-.198	-.046	-.073	-.152	-.168	-.095	-.001	-.117
Q_1	5.209	-.177	-.130	-.035	-.252	-.191	-.219	-.184	-.094	-.277

Table 4.3: Averages of the variables for each quartile group.

[8] There are 484 households in each of the first, second and third quartile groups and 483 households in the fourth quartile group.

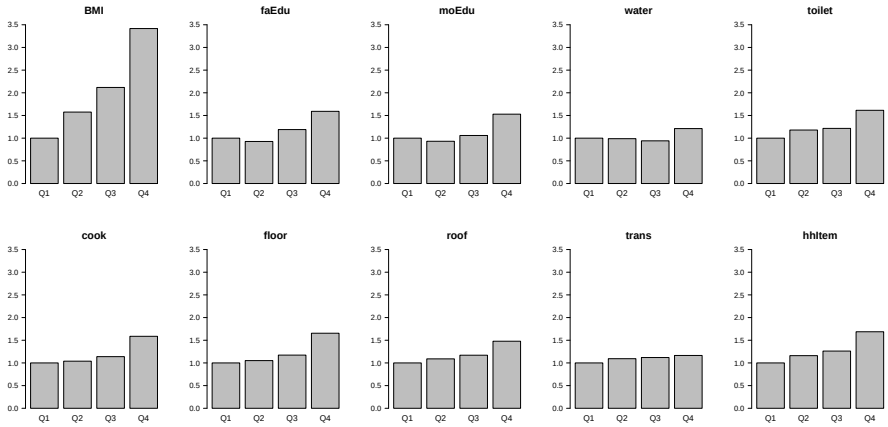


Figure 4.3: Barplots of the averages of the variables for each quartile group. Heights of the bars are adjusted so that the average of Q_1 is unity.

Quartile group averages show interesting variation patterns. For most of the regressors, the difference between Q_3 and Q_4 was disproportionately larger than the differences between the first three quartile groups. This can partly be attributed to the skewed distributions of regressors. However, relatively little skewed variables such as *water*, *toilet* and *roof* also follow this variation pattern. It will certainly be reasonable to assume for Q_4 some positive relationship between BMI and regressors. However, their relationships in the first three quartile groups seem rather ambiguous; the averages of the education variables and *water* are not even in increasing order. Figure 4.3 suggests that BMI and regressors are related partly but not over the whole sample. It may be sensible to interpret the barplot patterns as an indication of some unobserved factor on which the effects of regressors on BMI are conditional, may it be household-related, region-related or otherwise.

4.3.2 PC and DPC single index model analysis

This section presents estimates of the SIM. PC and DPC SIMs are specified as

$$\text{PC: } BMI = m(\mathbf{x}^\top \boldsymbol{\beta}) + u \tag{4.19}$$

$$\text{DPC: } BMI = g(\mathbf{x}^\top \boldsymbol{\gamma}) + e \ , \tag{4.20}$$

where a vector of regressors $\mathbf{x}$ is $(faEdu, moEdu, water, toilet, cook, floor, roof, trans, hhItem)^\top$. m and g are smooth link functions. For regression error terms u and e, we assume zero conditional mean and fulfillment of regularity conditions. In the following, all the index vector estimates are presented normalized so that the length of each vector is unity. To evaluate model fitting, we refer to a nonparametric

version of R^2 statistic which is defined as

$$R^2 = 1 - \frac{\sum_i \left(\hat{e}_i - \bar{\hat{e}}.\right)^2}{\sum_i \left(y_i - \bar{y}.\right)^2} \ . \tag{4.21}$$

We used the R package `EDR` for estimation of DPC index models.[9]

Table 4.4 shows the eigenvalues of the correlation matrix (Table 4.2) and their cumulative proportions in the total variance of 9. About 40% of the total variation in the regressors is explained by the first PC and 52% by the first two PCs. The corresponding scree plot appears in Figure 4.4. Judging from the scree plot and the fact that only the first two eigenvalues exceed one, it will be reasonable to compare DPC index models with PC models with at most the first two index vectors.

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9
Eigenvalue	3.48	1.17	.94	.80	.69	.57	.56	.42	.37
Cum.Prop.	.39	.52	.62	.71	.79	.85	.91	.96	1.00

Table 4.4: Eigenvalues and cumulative proportions of the variation explained by the PCs.

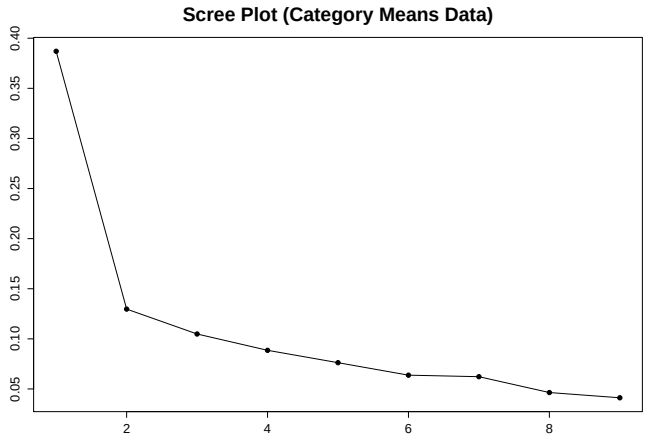


Figure 4.4: Scree plot of the eigenvalues.

[9] To deal with categorical data, we modified the function `edr`. We changed the initial bandwidth for iterative kernel regression estimation from the package specification of $0.85(d/n \prod_{j=1}^{d} IQR(x_k))^{1/d}\sqrt{d}$ to $0.85(d/n \prod_{j=1}^{d} 1.34\hat{\sigma}_j)^{1/d}\sqrt{d}$, where $IQR(x_k)$ is the sample interquartile range of x_k and $\hat{\sigma}_k = \sqrt{(n-1)^{-1}\sum_i (x_{ik} - \bar{x}_k)^2}$. The modification was done using the R function `trace`.

The first row of Table 4.5 shows the first PC estimate. Since all the variables are positively weighted, the first PC score can be interpreted as an index of the overall welfare of a household. Relative sizes of PC coefficients (weights) happen to be similar to the correlations between BMI and the regressors: for example, the pair of the variables with the largest and smallest PC weights ($floor$ and $trans$) correspond to the variables pair with the largest and smallest correlations.

The second row of Table 4.5 is the first DPC estimate. The pattern of signs stand in contrast to that of the first PC: five coefficients have positive weights and the other four have negative ones. It also differs in terms of the relative sizes of coefficients. In the first DPC, for example, *cook* and $floor$ are loaded much more heavily than the others in comparison with the first PC. In addition, a remarkable dissimilarity to the first PC lies in the education variables: they are weighted with opposite signs to each other, implying opposite effects of the education of father and mother even though they are positively correlated (correlation coefficient 0.62). As opposed to the first PC, the coefficients of the fist DPC are hard to interpret.

	$faEdu$	$moEdu$	$water$	$toilet$	$cook$	$floor$	$roof$	$trans$	$hhItem$
PC_1	.361	.330	.221	.378	.365	.409	.350	.109	.366
DPC_1	-.235	.117	-.074	-.144	.438	.792	.210	-.037	.203

Table 4.5: 1st PC and DPC.

Figure 4.5 plots the coefficients of both principal components. The dashed line is a 45 degree line. It is clearly seen that their weight structure is very different. Table 4.6 lists all the PCs. None of the PCs seems to be directed in a direction similar to that of the 1st DPC (even though some linear combination of a few PCs may be). This means that the original data of the regressors are projected onto very different spaces in model (4.19) and (4.20).

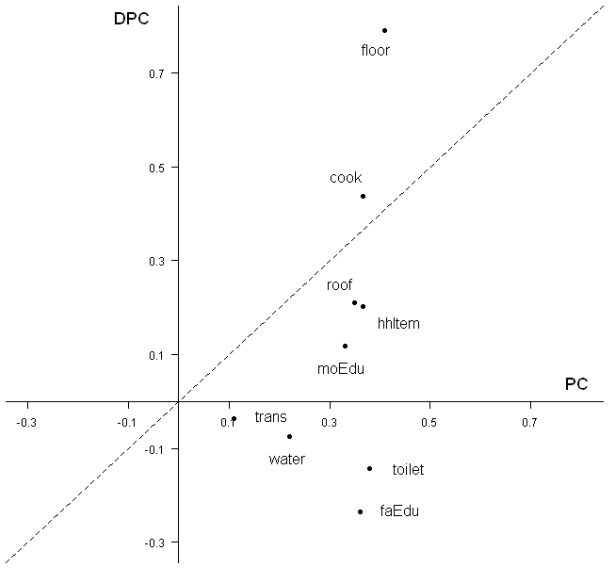


Figure 4.5: Coefficients of the PC and the DPC. The dashed line is a 45-degree line.

	faEdu	*moEdu*	*water*	*toilet*	*cook*	*floor*	*roof*	*trans*	*hhItem*
PC_1	.361	.330	.221	.378	.365	.409	.350	.109	.366
PC_2	-.394	-.418	.440	.115	.248	.220	.026	-.593	.039
PC_3	-.413	-.506	-.189	.120	.089	.109	.155	.620	.309
PC_4	-.010	-.027	-.837	.189	.082	.164	.071	-.451	.145
PC_5	.056	.050	.020	-.013	.210	-.019	-.833	.017	.506
PC_6	.015	-.039	.140	.464	-.795	-.048	-.001	-.117	.340
PC_7	-.058	.097	.019	-.747	-.258	.349	.203	-.102	.435
PC_8	-.155	.155	.035	-.048	.215	-.772	.326	-.144	.422
PC_9	.716	-.650	.015	-.135	.031	-.156	.081	-.068	.098

Table 4.6: PC coefficients.

After obtaining each household's first PC score by $\mathbf{x}^\top\hat{\boldsymbol{\beta}}$ ($\hat{\boldsymbol{\beta}} = PC_1$), we estimated the link function m nonparametrically. The estimate is plotted in Figure 4.6. Nonparametric R^2 was 0.150.[10] In the figure a dashed line is drawn at the PC score which yields the third quartile of BMI prediction, i.e. $\widehat{BMI}_{0.75} = \hat{m}(\mathbf{x}^\top\hat{\boldsymbol{\beta}})$ with $\mathbf{x}^\top\hat{\boldsymbol{\beta}}$ at the point of the dashed line.

[10]To make sure that the function estimate and the resultant R^2 are not influenced by an inappropriately large bandwidth, we experimented with a 30% smaller bandwidth. The newly estimated function was optically unchanged and its R^2 was 0.151.

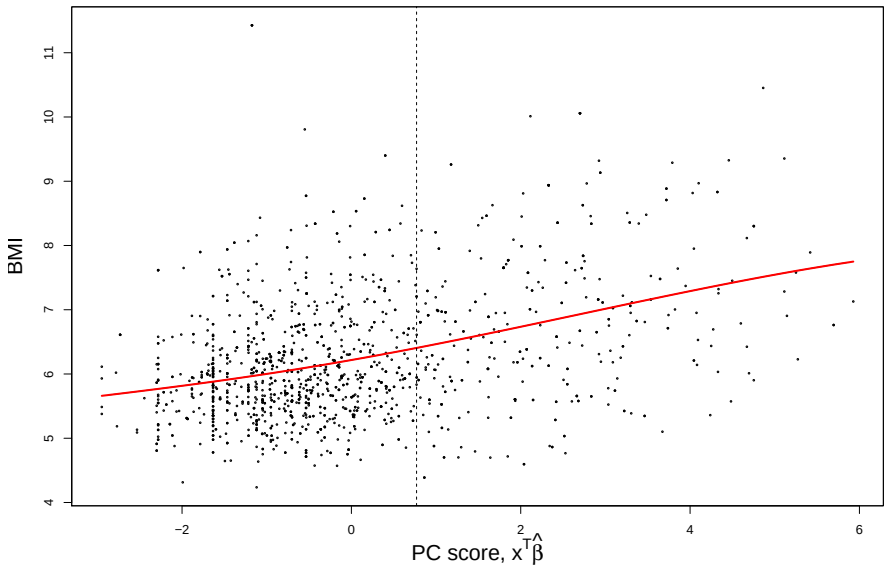


Figure 4.6: Estimated link function $\hat{m}$ for model (4.19). The dashed line is at the PC score corresponding to the third quartile of predicted BMI.

With each household's 1st DPC score obtained by $\mathbf{x}^\top\hat{\boldsymbol{\gamma}}$ ($\hat{\boldsymbol{\gamma}} = DPC_1$), we estimated the link function g nonparametrically. Figure 4.7 displays the estimated function with a dashed line at the DPC score which yields the third quartile of the BMI values predicted by $\hat{g}(\mathbf{x}^\top\hat{\boldsymbol{\gamma}})$. The nonparametric R^2 was 0.185, about 23% larger than the corresponding R^2 of the PC index model. Estimated function $\hat{g}$ is obviously not linear in the DPC score and almost constant up to a DPC score about zero.[11]

[11] To make sure that the function estimate and the resultant R^2 are not influenced by an inappropriately small bandwidth, we experimented with a 30% larger bandwidth. The estimated function was still optically very similar to Figure 4.7 and R^2 was 0.183.

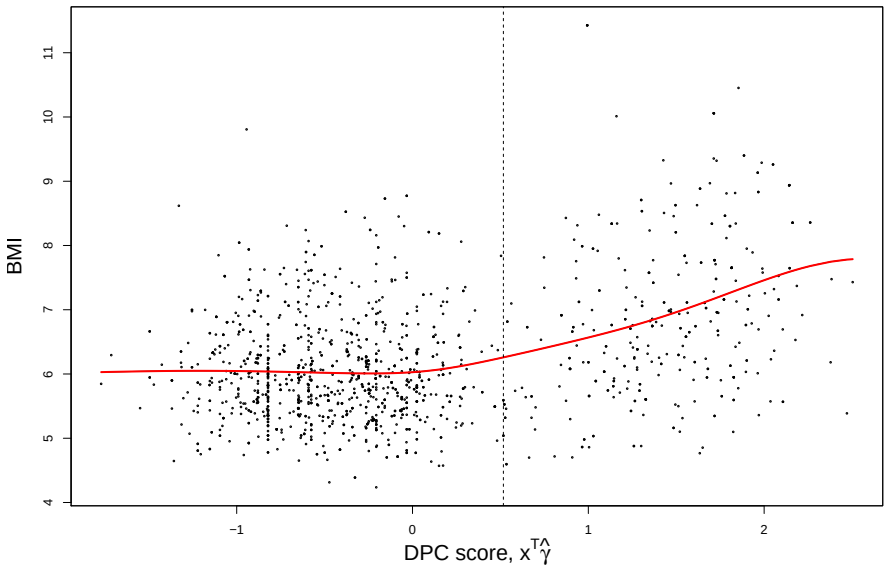


Figure 4.7: Estimated link function $\hat{g}$ for model (4.20). The dashed line is at the DPC score corresponding to the third quartile of predicted BMI.

To investigate the shape of the estimated function $\hat{g}$, we reestimated model (4.19) and (4.20) for a subsample composed of only two quartile groups Q_3 and Q_4. The reestimated coefficients of the PC and DPC are given in Table 4.7. The direction of the first PC remained almost the same: the signs were unchanged and the relative sizes of the coefficients remained very similar. This implies that the first PCs are almost the same for the whole sample and the subsample ($Q_3 \cup Q_4$). On the other hand, the first DPC has changed: the sign of *water* was reversed and the relative size of the coefficients changed. Figure 4.8 displays the reestimated link functions $\hat{m}$ and $\hat{g}$ for the subsample. There is relatively clear relationship between BMI and PC/DPC scores in both figures. In contrast to Figure 4.7, there is no structural break in the right figure of Figure 4.8. We interpret these results as follows. First, the change in the DPC suggests that function f is essentially different depending on subsamples.[12] The flat part of the estimated link function in Figure 4.7 is a reflection of little relationship between the BMI and regressors in Q_1 and Q_2 (and Q_3). Secondly, the DPC for the whole sample with four negative coefficients are not necessarily distorted by the observations in the Q_1 and Q_2; it is rather a depiction of a complicated association among variables.

[12]Recall DPC SIM: $BMI = f(\mathbf{x}) + e = g(\mathbf{x}^\top \boldsymbol{\gamma}) + e$.

	$faEdu$	$moEdu$	$water$	$toilet$	$cook$	$floor$	$roof$	$trans$	$hhItem$
PC_1	.340	.320	.227	.388	.381	.407	.352	.102	.368
DPC_1	-.145	.281	.198	-.221	.150	.850	.104	-.079	.223

Table 4.7: 1st PC and DPC for a subsample.

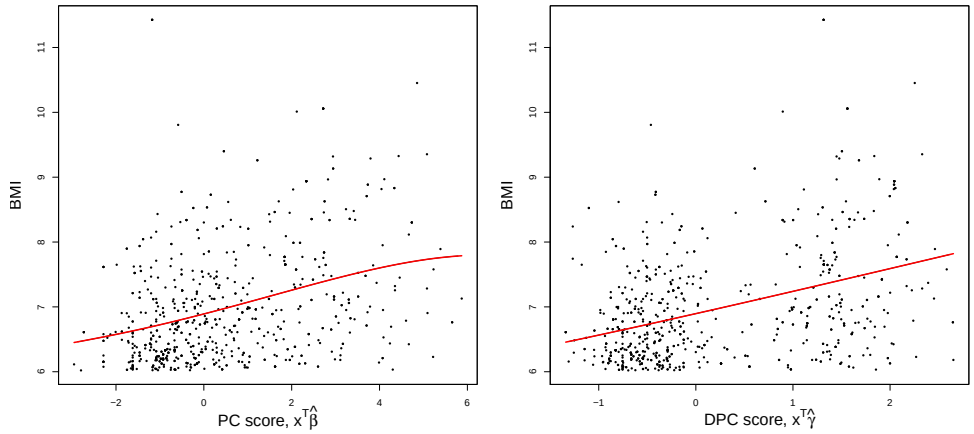


Figure 4.8: Reestimated link functions $\hat{m}$ (left) and $\hat{g}$ (right) for a subsample.

4.3.3 Bootstrap inference

We tested significance of the DPC coefficients using the wild bootstrap as described in Section 4.2.2. The bootstrap simulation size R was set to 1,000. Table 4.8 is a matrix of the estimated correlations between DPC coefficient estimators. Some estimators have relatively strong correlations (for example, $floor$ and $cook$, and $faEdu$ and $moEdu$). However, we present only marginal bootstrap sampling distributions in Figure 4.9 and the 0.025 and 0.975 quantiles of the bootstrap estimates for each variable in Table 4.9. None of the coefficients except that of $faEdu$ was significantly different from 0 at the 5% level.

	moEdu	*water*	*toilet*	*cook*	*floor*	*roof*	*trans*	*hhItem*
faEdu	-.50	.01	.01	-.08	-.25	-.18	-.00	-.18
moEdu		-.07	.07	.00	.06	.05	-.19	-.02
water			.01	-.32	-.25	-.11	.15	-.24
toilet				-.17	-.20	-.17	-.06	-.28
cook					.58	.24	-.03	.35
floor						.37	-.07	.47
roof							-.18	.35
trans								-.24

Table 4.8: Estimated correlations between the DPC coefficient estimators.

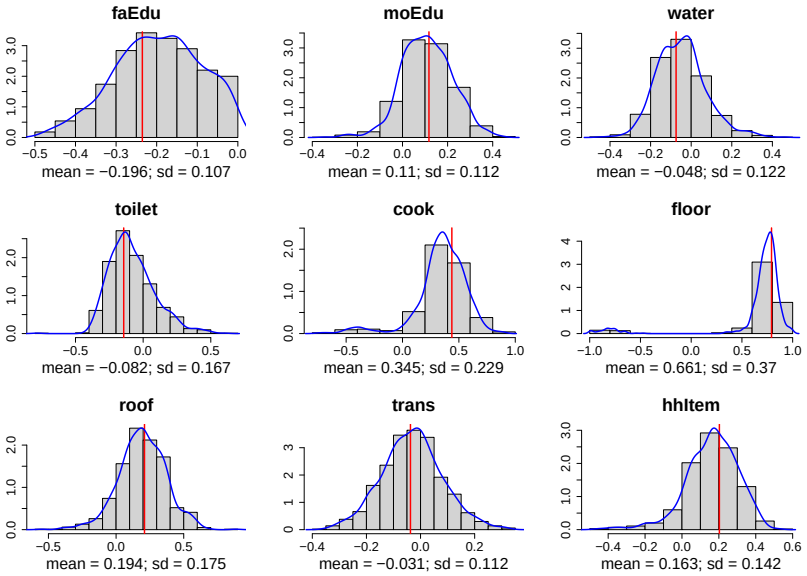


Figure 4.9: Bootstrap sampling distribution of $\hat{\gamma}$. Vertical solid lines indicate the original DPC coefficient estimates. Kernel density estimates are added to the histograms.

	faEdu	*moEdu*	*water*	*toilet*	*cook*	*floor*	*roof*	*trans*	*hhItem*
2.5%	-.417	-.114	-.262	-.338	-.385	-.824	-.179	-.257	-.178
97.5%	-.013	.319	.230	.299	.688	.915	.537	.198	.403

Table 4.9: 0.025 and 0.975 quantiles of the bootstrap sampling distributions.

4.3.4 Comparison between PC and DPC index models

In general, it will be of interest to statistically test the significance of deviation between the two index vectors, i.e. PC and DPC. However, due to their clearly different

structures and the insignificance of the DPC coefficients except for $faEdu$, here we compare only the prediction performance of the models. We refer to the following statistics: Spearman's and Kendall's rank correlation coefficients, and proportion of correct predictions. A household's BMI was predicted by the estimated conditional expectation, $\hat{m}(\mathbf{x}^\top\hat{\boldsymbol{\beta}})$ and $\hat{g}(\mathbf{x}^\top\hat{\boldsymbol{\gamma}})$. Spearman's and Kendall's rank correlation coefficients were based on the order of the original BMI observations and that of the predicted BMI values.

Similarly to the the quartile groups Q_j (created according to the original BMI observations), we classified the households into four "prediction quartile groups" according to BMI predictions. The jth ($j = 1, \ldots, 4$) prediction quartile group is denoted by "P_j". The size of P_j is the same as Q_j. We denote by p_i the predicted quartile group index $\{1, \ldots, 4\}$ of the ith household ($i = 1, ..., 1451$), and likewise q_i (index of the quartile group $\{1, \ldots, 4\}$ to which the ith household belongs). Correct prediction is defined as such that p_i is the same as q_i. We counted correct predictions for the whole sample (i.e. $\sum_i I(p_i = q_i)$ where I is an indicator function).

Table 4.10 shows the statistics including the nonparametric R^2. For a comparison purpose we additionally estimated model (4.19) using polychoric/-serial correlations. As far as R^2 is concerned, the DPC SIM produced a better fit than the PC models. However, Spearman's and Kendall's rank correlation coefficients slightly favored the PC SIM. No method showed an advantage over the others in terms of the proportion of correct predictions. It should be noted, however, that the predicted BMI values in the negative range of the DPC score are almost the same (see Figure 4.7), and that BMI predictions in this range should therefore be ascribed to randomness rather than systematic functional relationship. As a whole, the prediction performance of the DPC SIM was comparable with that of the PC models in spite of little functional relationship estimated in the negative DPC score range.

Method	R^2	Spearman	Kendall	Prediction
PC SIM (Polychor.)	.153	.327	.223	33.56 %
PC SIM (Cat.Mean)	.150	.326	.222	34.46 %
DPC SIM	.185	.294	.200	35.77 %

Table 4.10: Nonparametric R^2, Spearman's and Kendall's rank correlation coefficients and proportion of correct predictions for the whole sample. "Polychor." and "Cat.Mean" mean polychoric/-serial and category means data, respectively.

We further examined the proportions of correct predictions in each prediction quartile group (Table 4.11).[13] Patterns of prediction performance are quite similar among the different approaches: both PC and DPC models achieved relatively large proportions for P_4; and neither of them showed good performance in the rest of the quartile groups, particularly in P_2 and P_3.[14] In the case of the DPC model, the DPC scores that resulted in relatively good prediction lie in the range on the right of the dashed line where a linear relationship with the BMI is observed (see Figure 4.7). This contrasts with the PCA-based prediction: regardless of group-wise differences in prediction performance, a linear relationship is observed over the whole PC score range across the parting dashed line (see Figure 4.6).

Method	Total	P_1	P_2	P_3	P_4
PC SIM (Polychor.)	33.56	32.51	27.27	24.24	50.28
PC SIM (Cat.Mean)	34.46	34.16	28.65	24.52	50.55
DPC SIM	35.77	32.78	30.85	28.93	50.55

Table 4.11: Proportions of correct predictions (%) in each prediction quartile group.

Distributions of group averages provide a further insight into the prediction performances. Similarly to Table 4.3, we calculated the averages of the variables in each prediction quartile group (Table 4.12 and 4.13). Note the averages in P_4 resulting from the DPC model (Table 4.13): all the regressors have a positive average in spite of some negative coefficients of the DPC. The averages in P_4 are by far larger than those in the other quartile groups. This means that negative coefficients in the DPC do not contradict our prior assumption about positive effects of regressors on BMI. This implies that the coefficients of the DPC were determined in relation to BMI in such a manner that, for example, a negative effect of $faEdu$ on BMI was compensated with counter-positive effects produced by certain regressors through some linear or nonlinear relationship between $faEdu$ and those regressors.

Figure 4.10 is a plot of the PC SIM-based prediction quartile group averages given in Table 4.12. The figure provides a clear functional association between the BMI and

[13] For P_j, proportion of correct predictions is $\sum_i I(q_i = j, p_i = q_i) / \sum_i I(q_i = j)$.

[14] To evaluate this prediction performance, let t be a hyper-geometrically distributed random variable with population size 1,451 and probability 0.25. Then, 0.25% and 97.5% percentiles of t of 363 drawings are 77 and 105, respectively. This means that, when we consider just one prediction quartile group and randomly assign 363 households to this group, we would expect with a 95% probability a proportion of correct predictions from 21.21% to 28.93%. In this sense, predictions in P_2 and P_3 are hardly different from random assignment.

the regressors over the whole quartile groups: the averages of all the regressors are in the same increasing order as the predicted BMI. These figures are in agreement with a linearly increasing function estimate $\hat{m}$ (Figure 4.6).

On the other hand, in the first three quartile groups DPC SIM predicted almost no variation for the BMI group averages nor for the group averages of most of the regressors (Figure 4.11). Lack of ordering in the averages of regressors is also remarkable. This is in accordance with almost constant BMI predictions in the negative DPC score range (Figure 4.7). When we recall the indication from Table 4.3 and Figure 4.3 that there is no simple overall association between the BMI and regressors, the DPC SIM estimate seems to verify that indication.

Our interpretation of the prediction performance of the DPC model is that, in search of the optimal index vector and link function, it found a functional relationship only in a subsample consisting mostly of P_4. In this sense, poor overall prediction performance was not because of an inappropriate index vector estimate. To the contrary, poor prediction performance of the PC model was rather because the first PC, determined without regard to BMI, happened to be relevant only to subsample P_4 and irrelevant to the rest of the sample.

	BMI	$faEdu$	$moEdu$	$water$	$toilet$	$cook$	$floor$	$roof$	$trans$	$hhItem$
P_4	6.821	1.025	.878	.409	.813	.778	.926	.686	.187	.889
P_3	6.243	.141	.161	-.094	-.017	.123	-.154	-.092	.119	.174
P_2	6.051	-.397	-.403	-.152	-.201	-.190	-.354	-.133	.039	-.320
P_1	5.881	-.766	-.634	-.163	-.593	-.709	-.415	-.460	-.344	-.741

Table 4.12: Averages of prediction quartile groups by the PC SIM.

	BMI	$faEdu$	$moEdu$	$water$	$toilet$	$cook$	$floor$	$roof$	$trans$	$hhItem$
P_4	6.936	.620	.559	.361	.671	.809	1.244	.632	.058	.817
P_3	6.062	-.052	-.217	-.098	-.167	-.543	-.382	-.302	.008	-.322
P_2	6.029	-.476	-.330	-.114	-.293	-.457	-.429	-.184	-.090	-.270
P_1	6.012	-.091	-.010	-.147	-.209	.194	-.429	-.145	.024	-.222

Table 4.13: Averages of prediction quartile groups by the DPC SIM.

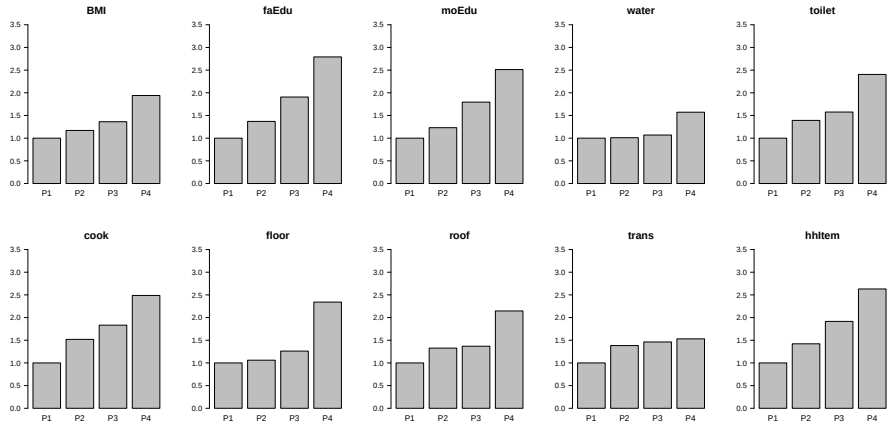


Figure 4.10: Averages of prediction quartile groups by the PC SIM. Heights of the bars are adjusted so that the average of P_1 is unity.

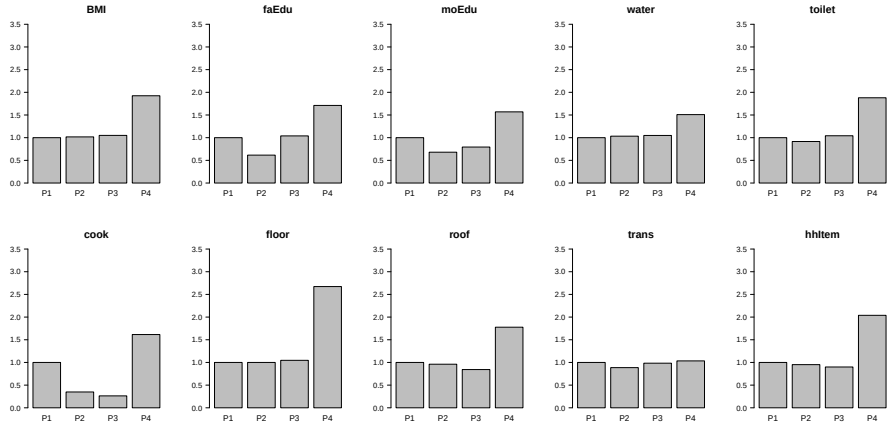


Figure 4.11: Averages of prediction quartile groups by the DPC SIM. Heights of the bars are adjusted so that the average of P_1 is unity.

As far as the data set analyzed is concerned, we have found little evidence that the DPC index model outperforms the PC counterpart in prediction. However, we note that, if the effects of regressors on the BMI are conditional on unobserved household characteristics, further investigation into household characteristics with the aid of the households' DPC scores may lead to finding of influential latent factors.

4.3.5 Analysis using cluster average data

The 1,451 households analyzed in the previous sections belong to 336 clusters in the sample. This section presents PC and DPC index model analysis using cluster-wise averaged category means data (obtained by taking the average of category means data for each cluster, and henceforth called "cluster average data"). Large residuals seen in Figure 4.6 and 4.7 and relatively low nonparametric R^2 statistics suggest a lot of noise in data. If households in the same cluster are relatively homogeneous, cluster-wise averaging will help to reduce noise without losing relevant information about the functional relationship. Moreover, if there are factors underlying the structural break in the link function g found by the DPC SIM, and if they are related with cluster characteristics, a similar structural break will be observed in the analysis of cluster average data.

Figure 4.12 is a histogram of the cluster size. The cluster size is highly unbalanced. It may be necessary to assign an appropriate weight to each cluster. However, we present only an analysis using crude cluster-wise averaged category means data.

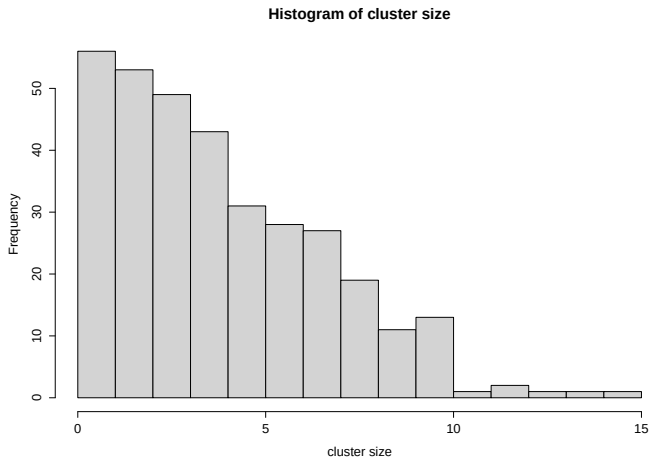


Figure 4.12: Histogram of the cluster size.

Table 4.14 shows the correlation matrix of cluster average data. Table 4.15 is a table of the eigenvalues and their cumulative proportions of the total variation explained by the PCs. 45% of the variation is explained by the first PC, which is by 5% larger than in the original data (Table 4.4). We fit the SIMs corresponding to model (4.19) and (4.20).

	$faEdu$	$moEdu$	$water$	$toilet$	$cooking$	$floor$	$roof$	$trans$	$hhItem$
BMI	.37	.35	.17	.44	.38	.48	.34	.07	.42
$faEdu$		.66	.19	.48	.32	.46	.43	.19	.44
$moEdu$			.19	.38	.32	.40	.40	.18	.35
$water$				.28	.35	.39	.25	-.04	.25
$toilet$					.54	.54	.40	.10	.49
$cooking$						.64	.43	-.02	.46
$floor$							.51	.08	.62
$roof$								.18	.42
$trans$									.14

Table 4.14: Correlation matrix of the cluster average data.

	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7	PC_8	PC_9
Eigenvalue	3.99	1.22	.83	.75	.60	.54	.46	.32	.29
Cum.Prop.	.44	.58	.67	.75	.82	.88	.93	.97	1.00

Table 4.15: Eigenvalues and cumulative proportions obtained from cluster average data.

The first PC and DPC are shown in Table 4.16. As was the case for the original category means data, the first PC can be interpreted as a measure of overall welfare of a household. The DPC contrasts with that obtained from the original category means data: most of its coefficients are positive; only those of $water$ and $trans$ are negative with relatively small size. The structures of the coefficients are visually presented in Figure 4.13. Compared with Figure 4.5, the coefficients of the PC and DPC are plotted now closer to the 45 degree line.

	$faEdu$	$moEdu$	$water$	$toilet$	$cooking$	$floor$	$roof$	$trans$	$hhItem$
PC_1	.361	.330	.233	.372	.363	.414	.346	.100	.368
DPC_1	.050	.309	-.030	.376	.378	.670	.026	-.095	.398

Table 4.16: 1st PC and DPC obtained from the cluster average data.

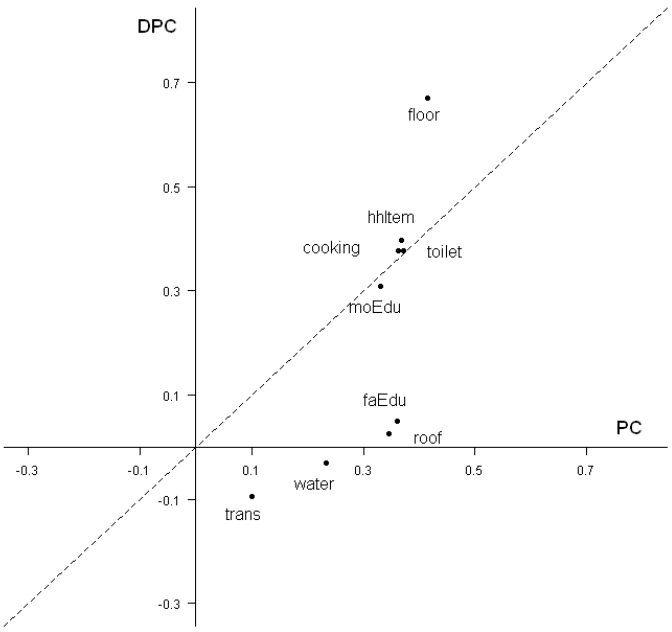


Figure 4.13: Coefficients of the PC and the DPC obtained from the cluster average data. The dashed line is a 45-degree line.

Figure 4.14 is a plot of the estimated link function corresponding to m of model (4.19). The nonparametric R^2 was 0.288.[15] Figure 4.15 shows the estimated function corresponding to g of model (4.20). The estimated function looks partly undersmoothed. Nonparametric R^2 was 0.345.[16] There seems to be no systematic break in the shape of the estimated function $\hat{g}$ and it shares an upward-sloping tendency with the estimate by the PC model. These results suggest that cluster-wise averaging has filtered out not only noise in data but also households' characteristics that yielded the DPC for the original sample; and that potential factors underlying the structural break are unlikely to be related with clusters.

[15]The estimated link function may seem oversmoothed. However, even with a 70% smaller bandwidth, the function estimate, although somewhat curvy, yielded a R^2 0.292.

[16]Undersmoothing is because of the use of a global bandwidth, which is determined without regard to the local density of the DPC score. A large R^2 is partly due to undersmoothing. However, R^2 remained 0.320 even with a 70% larger bandwidth, which produced a more smooth function.

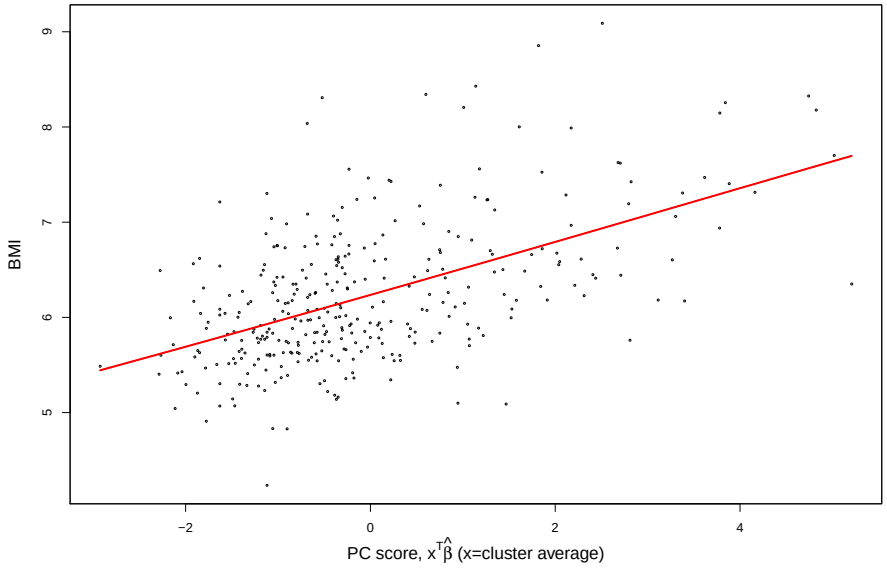


Figure 4.14: Estimated link function $\hat{m}$ for the cluster average data.

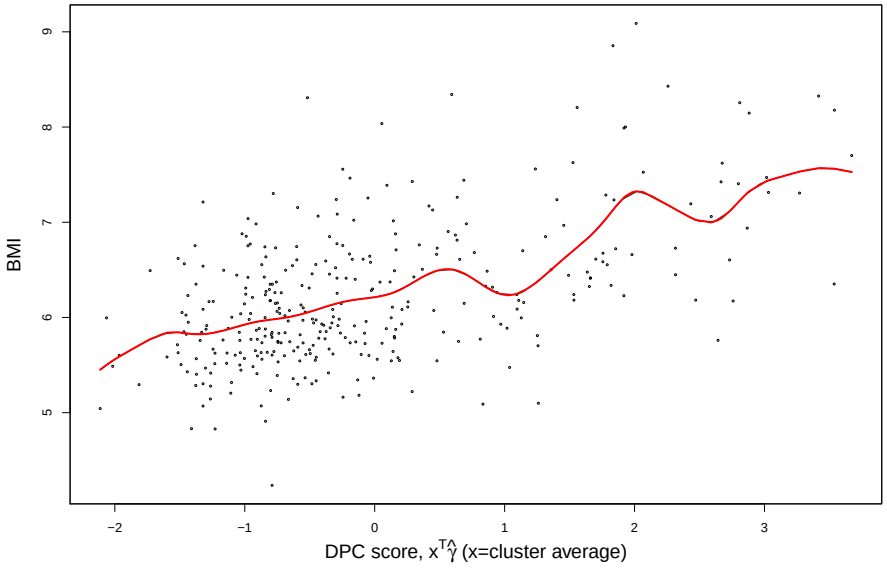


Figure 4.15: Estimated link function $\hat{g}$ for the cluster average data.

4.3.6 Multi index model

This section briefly covers application of the MIM to the same category means data as in Section 4.3.2. The PC and DPC MIMs are specified with $M = 2$ as

$$\text{PC: } BMI = m(\mathbf{x}^\top \boldsymbol{\beta}_1, \mathbf{x}^\top \boldsymbol{\beta}_2) + u \tag{4.22}$$

$$\text{DPC: } BMI = g(\mathbf{x}^\top \boldsymbol{\gamma}_1, \mathbf{x}^\top \boldsymbol{\gamma}_2) + e \ , \tag{4.23}$$

where m and g are smooth functions; and u and e are regression errors with $\mathrm{E}[u|\mathbf{x}] = \mathrm{E}[e|\mathbf{x}] = 0$. The coefficients of the first two PCs are listed in Table 4.17. In the second PC, education variables and some housing characteristics variables were loaded negatively and relatively heavily. As opposed to the first PC, interpretation of the second PC seems difficult. The third and fourth rows of Table 4.17 are the first and second DPC estimates, respectively. The first DPC retained the same pattern of signs as the DPC SIM estimate; relatively large-sized coefficients of *cook* and $floor$ are also in common. Neither of the two DPCs is amenable to interpretation with respect to the BMI. Figure 4.16 gives a visual image of the structures of the first and second PC and DPC. The index spaces spanned by those vectors seems to be quite different.

	$faEdu$	$moEdu$	$water$	$toilet$	$cook$	$floor$	$roof$	$trans$	$hhItem$
PC_1	.361	.330	.221	.378	.365	.409	.350	.109	.366
PC_2	-.394	-.418	.440	.115	.248	.220	.026	-.593	.039
DPC_1	-.148	.058	-.130	-.108	.593	.751	.035	-.067	.157
DPC_2	.261	-.244	.365	.238	-.067	.096	.491	.259	.600

Table 4.17: 1st and 2nd PC and DPC.

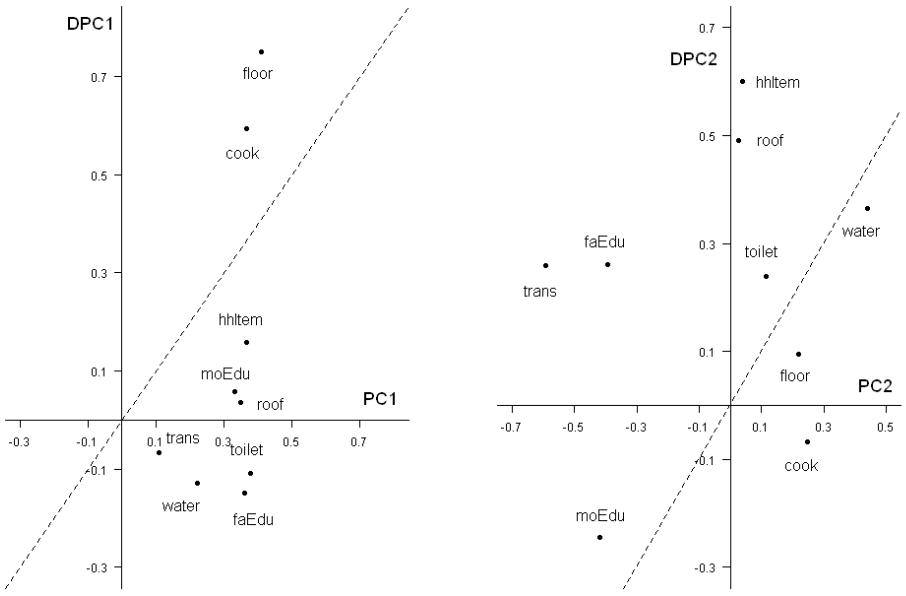


Figure 4.16: Coefficients of the 1st and 2nd PCs and DPCs. The dashed lines are 45-degree lines.

Figure 4.17 displays the estimated function $\hat{m}$ for model (4.22). The left figure shows $\hat{m}$ from the perspective of the first PC score and the right figure from the second PC score perspective. On the whole, the estimated function retained a linearly increasing tendency although it shows downward-sloping shapes near boundary areas of the the second PC score.

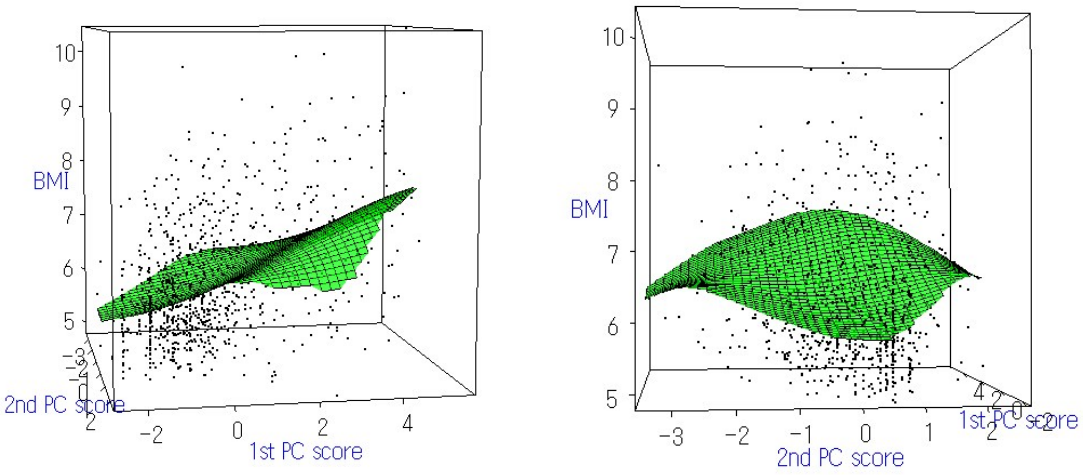


Figure 4.17: Estimated link function $\hat{m}$ for model (4.22): (left) from the first PC perspective and (right) from the second PC perspective.

The estimated function $\hat{g}$ is plotted in Figure 4.18. Potential factors underlying a structural break found in the DPC SIM may have been split between the two DPC index vectors. The structural break in $\hat{g}$ along the first DPC direction has become somewhat ambiguous even though its functional form in this direction still resembles that of the DPC SIM estimate.

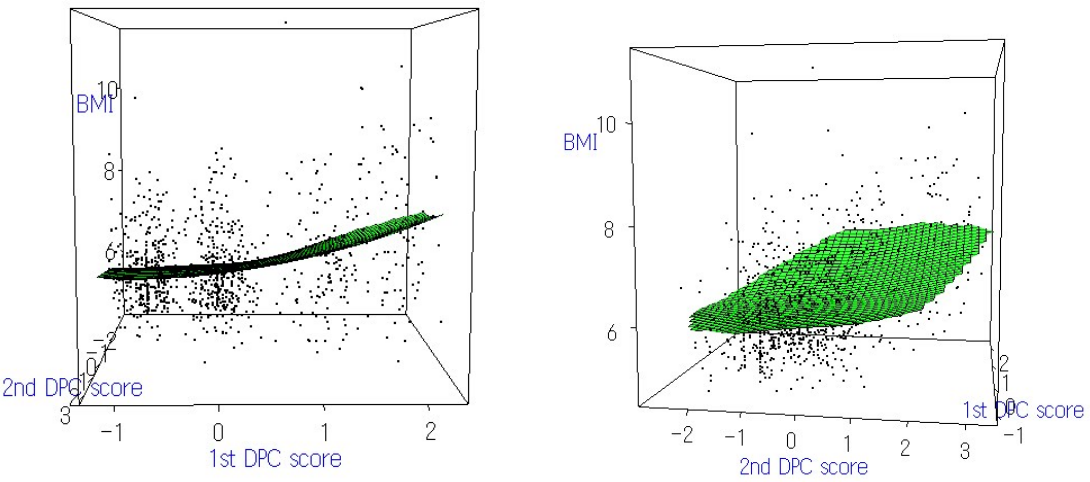


Figure 4.18: Estimated link function $\hat{g}$ for model (4.23): (left) from the first DPC perspective and (right) from the second DPC perspective.

To evaluate the MIMs, we calculated the same set of statistics as before (Table 4.18). Nonparametric R^2 of the PC model increased from 0.150 to 0.174. On the other hand, the increase in R^2 of the DPC model was marginal (from 0.185 to 0.190). This implies that information about the functional relationship is effectively concentrated in one-dimensional space given by the first SIM index vector. Rank correlation coefficients increased only slightly for all the methods. In terms of the proportion of correct predictions there was no improvement over the SIM estimation (or slight deterioration).

Method	R^2	Spearman	Kendall	Prediction
PC MIM (Polychor.)	.174	.339	.232	32.32 %
PC MIM (Cat.Mean)	.174	.342	.233	33.77 %
DPC MIM	.190	.321	.218	33.49 %

Table 4.18: Nonparametric R^2, Spearman's and Kendall's rank correlation coefficients and proportion of correct predictions.

We further examined the prediction performance in each prediction quartile group (see Table 4.19). For both PC and DPC MIMs, the pattern of proportions of correct prediction was similar to that of the SIM prediction: proportions were low in the first quartile groups, particularly P_2 and P_3; and prediction in P_4 was relatively good. Overall, there was no improvement in prediction by the MIM.

Method	Total	P_1	P_2	P_3	P_4
PC MIM (Polychor.)	32.32	33.06	25.34	20.94	50.00
PC MIM (Cat.Mean)	33.77	34.71	28.65	22.04	49.72
DPC MIM	33.49	34.99	25.34	23.69	50.00

Table 4.19: Proportions of correct predictions (%) in each prediction quartile group.

To complete this section, Table 4.20 and 4.21 show the prediction quartile group averages of the variables. Their barplots are given in Figure 4.19 and 4.20. There was little difference between the average structures between the PC SIM and PC MIM. As for the DPC MIM, the group average structure has diverted from that of the DPC SIM and become similar to that of the PC counterpart.

	BMI	$faEdu$	$moEdu$	$water$	$toilet$	$cook$	$floor$	$roof$	$trans$	$hhItem$
P_4	6.851	.889	.747	.452	.785	.836	.987	.647	.104	.877
P_3	6.260	.110	.217	-.126	-.027	.093	-.177	-.052	.036	.179
P_2	6.062	-.280	-.376	-.168	-.182	-.182	-.378	-.159	.059	-.268
P_1	5.905	-.717	-.585	-.157	-.574	-.744	-.429	-.434	-.199	-.785

Table 4.20: Averages of prediction quartile groups by the PC MIM.

	BMI	$faEdu$	$moEdu$	$water$	$toilet$	$cook$	$floor$	$roof$	$trans$	$hhItem$
P_4	6.931	.654	.569	.382	.694	.851	1.206	.632	.088	.851
P_3	6.172	-.055	-.178	-.045	.013	.124	-.345	-.013	.235	.401
P_2	6.050	-.060	-.165	-.093	-.202	-.411	-.429	-.111	.055	-.401
P_1	5.944	-.538	-.224	-.243	-.503	-.562	-.429	-.507	-.377	-.849

Table 4.21: Averages of prediction quartile groups by the DPC MIM.

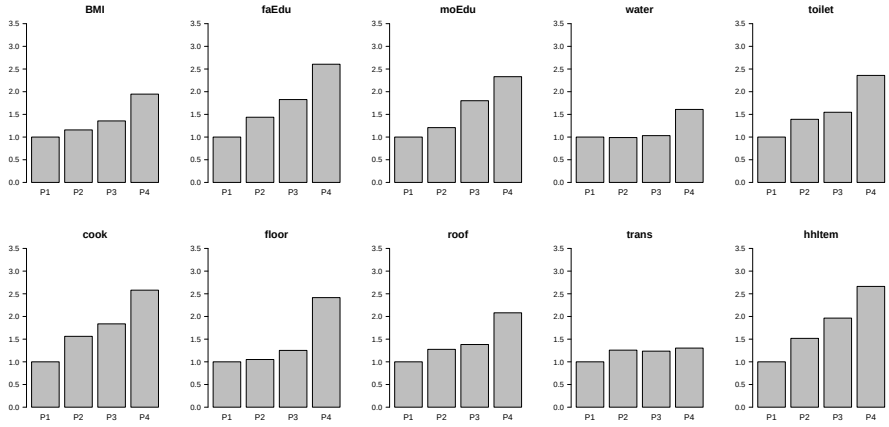


Figure 4.19: Averages of prediction quartile groups by the PC MIM.

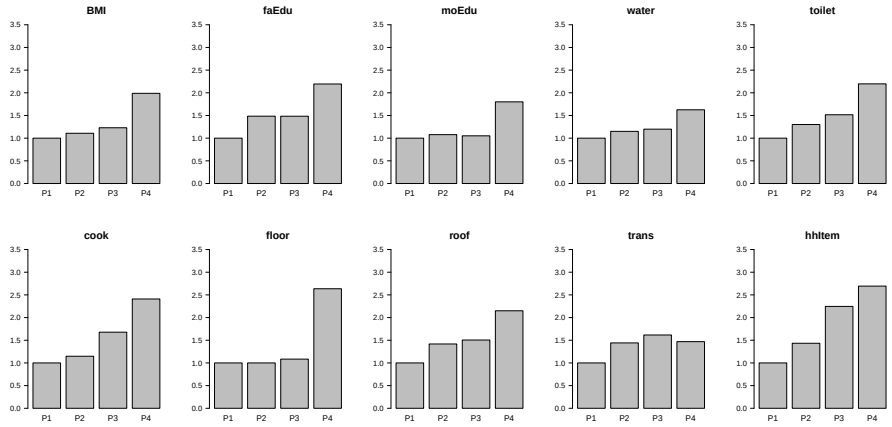


Figure 4.20: Averages of prediction quartile groups by the DPC MIM.

4.4 Concluding Remarks

Construction of a reliable welfare indicator of a household's SES is of great research interest. PCA has been a conventional tool to obtain a few indices that summarize information contained in a large number of variables. However, PCs are determined without regard to the response variable the fact of which undermines the prediction performance of PC-based regression models. Motivated by this intrinsic disadvantage of PCA approaches, we turned to DPCs, which are determined in relation to the response. The index model provides a regression framework which balances rigidness of the parametric model and flexibility of the nonparametric one. We compared the performance of PC and DPC index models by applying them to demographic and health survey data.

As Kolenikov and Angeles (2009) discussed, categorical variables require special treatment for PC-oriented methods. This is true for the DPC index model all the more because it is applicable by definition only to continuous data. We addressed this issue by using category means data proposed by Kolenikov and Angeles (2009). We consider that the results presented in this essay generally justify the use of category means data for DPC index models.

Contrary to our expectation, PC index models did not underperform their DPC counterparts. Comparable prediction performances of the PC models are partly explained by the fact that PCA is applied implicitly in accordance with prior knowledge about the association between the response and regressors. Indeed, we recoded original unordered categorical data, for example, roof material, into ordered data based on prior knowledge of correlation between the BMI and the material. Each regressor

was a priori coded so that its increases in value are positively related with the BMI. Thus, the first PC score, which represents overall availability of the determinants of a positive BMI, cannot be a bad index.

In our study we have found little evidence that the DPC index model outperforms the PC counterpart in prediction. However, our analysis illustrated the potential of the DPC index model to serve as a tool of exploratory analysis. We interpret the result obtained by the DPC SIM as an indication of some unknown structure underlying the population. If the effects of regressors on the BMI are conditional on unobserved household characteristics, further investigation of households and their DPC scores may lead to finding of latent structure.

The fact that the DPC index model did not outperform the PC counterpart may partly be due to relatively weak association between the response and regressors, which was indicated by our test using bootstrap simulation (even though it is unlikely that the regressors are jointly insignificant). We studied similar household survey data from other countries without finding remarkable difference between PC and DPC models. More empirical researches, especially using data containing more continuous variables and stronger association between the response and regressors, will shed light on the performance of the DPC index model.

4.5 Appendix A

Suppose a MIM of the form

$$\begin{aligned} y &= f(\mathbf{x}) + e \\ f(\mathbf{x}) &= g(\mathbf{x}^\top \boldsymbol{\gamma}_1, \mathbf{x}^\top \boldsymbol{\gamma}_2, \ldots, \mathbf{x}^\top \boldsymbol{\gamma}_M) = g(\boldsymbol{\Gamma}\mathbf{x}) \ , \end{aligned} \tag{4.24}$$

where $\mathbf{x} \in \mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}$, $g : \mathbb{R}^M \to \mathbb{R}$ and $\boldsymbol{\Gamma} : \mathbb{R}^d \to \mathbb{R}^M$. $\{\boldsymbol{\gamma}_j\}_{j=1}^M$ are orthogonal index vectors. $\boldsymbol{\Gamma}$ is thus a linear orthogonal mapping. The M-dimensional subspace spanned by $\{\boldsymbol{\gamma}_j\}$ is referred to as index space (also called effective dimension space). The d-vector is further assumed to have a support of $[-1, 1]^d$ (if necessary, data are transformed accordingly). M is assumed to be known and $M < d$.

The gradient vector of (4.24) at point $\mathbf{x}_i$ is given by

$$\nabla f(\mathbf{x}_i) = g_1' \boldsymbol{\gamma}_1 + g_2' \boldsymbol{\gamma}_2 + \cdots + g_M' \boldsymbol{\gamma}_M \ , \tag{4.25}$$

where g_j' is the partial derivative of g with respect to its jth argument. (4.25) implies that the gradient at any point is a linear combination of the index vectors and therefore belongs to the index space.

Samarov (1993) discussed the use of derivatives to explore regression structure by nonparametric functional estimation. The following two observations are relevant for the MIM in consideration:

$$\mathrm{E}[\nabla_f \nabla_f^\top] = \sum_{j,k=1}^{M} \mathrm{E}[g'_j g'_k] \boldsymbol{\gamma}_j \boldsymbol{\gamma}_k^\top \tag{4.26}$$

$$\mathrm{E}[\nabla_f^2] = \sum_{j,k=1}^{M} \mathrm{E}[g''_{jk}] \boldsymbol{\gamma}_j \boldsymbol{\gamma}_k^\top \ , \tag{4.27}$$

where ∇_f^2 is the Hessian of f; g''_{jk} is the second-order (cross) partial derivative of g with respect to its jth and kth arguments; and expectation is with respect to $\mathbf{x}$. (4.26) and (4.27) are both a matrix of dimension $d \times d$ constructed by a linear combination of outer products $\boldsymbol{\gamma}_j \boldsymbol{\gamma}_k^\top$. Since $M < d$, (4.26) and (4.27) imply that the index vectors $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_M$ lie in an M-dimensional subspace spanned by the eigenvectors of $\mathrm{E}[\nabla_f \nabla_f^\top]$ or $\mathrm{E}[\nabla_f^2]$ corresponding to their nonzero eigenvalues. It is, however, more convenient to base the estimation of the index space on (4.26) from the following observation. Let's denote $\mathrm{E}[\nabla_f \nabla_f^\top]$ by $\mathbf{M}^*$. The directional derivative of f in the direction of $\mathbf{a}$ ($\mathbf{a} \neq \mathbf{0}$) is given by

$$\frac{df(\mathbf{x})}{d\mathbf{a}} = \nabla_f^\top(\mathbf{x})\mathbf{a} \tag{4.28}$$

and it follows that

$$\mathrm{E}\left[\frac{df(\mathbf{x})}{d\mathbf{a}}^\top \frac{df(\mathbf{x})}{d\mathbf{a}}\right] = \mathbf{a}^\top \mathbf{M}^* \mathbf{a} \ . \tag{4.29}$$

$\mathbf{M}^*$ can be spectral-decomposed into $\mathbf{C}\boldsymbol{\Lambda}\mathbf{C}^\top$, where diagonal matrix $\boldsymbol{\Lambda}$ has M nonzero eigenvalues sorted in decreasing order ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$). Thus, if $\mathbf{a}$ is the eigenvector of $\mathbf{M}^*$ corresponding to λ_1, $\mathbf{a}^\top \mathbf{M}^* \mathbf{a}$ takes the maximum value, which means that $\mathbf{a}$ is directed to the same direction as the gradient vector in expectation. On the other hand, if $\mathbf{a}$ is an eigenvector corresponding to zero eigenvalue, $\mathbf{a}$ is orthogonal to that gradient vector. Note also, that, since $\mathbf{a}_j^\top \mathbf{M}^* \mathbf{a}_j = \lambda_j$ for jth eigenvector $\mathbf{a}_j$, nonzero eigenvalue λ_j reflects the extent to which f is expected to vary due to $d\mathbf{x}$ along the direction of jth index vector.

The observation above leads to the basic idea of the average derivative estimation, that is, estimating the index space from the first M eigenvectors of a sample analog matrix of $\mathbf{M}^*$ given by

$$\mathbf{M} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\nabla}_f(\mathbf{x}_i) \boldsymbol{\nabla}_f^\top(\mathbf{x}_i) \ . \tag{4.30}$$

Several approaches to the calculation of $\mathbf{M}$ have been proposed in literature. $\mathbf{M}$ is a quadratic functional of the gradient of f and is harder to calculate than the following linear functional suggested by Ibragimov et al. (1986). Hristache, Juditsky, Polzehl, and Spokoiny (2001) applied the idea of estimating linear functional as follows.

Consider a set of orthonormal basis functions for finite sum $\{\psi_l(\mathbf{x}_i)\}$ $(l = 1, 2, \ldots, L)$, that is,

$$\sum_{i=1}^{n} \psi_l(\mathbf{x}_i)\psi_{l'}(\mathbf{x}_i) \quad = \quad \delta_{ll'} \ , \tag{4.31}$$

where $\delta_{ll'}$ is the Kronecker delta. Suppose that the gradient function $\nabla_f(\mathbf{x}_i)$ can be approximated by an orthogonal series expansion using $\{\psi_l(\mathbf{x}_i)\}$:

$$\nabla_f(\mathbf{x}_i) \quad = \quad \mathbf{b}_1\psi_1(\mathbf{x}_i) + \mathbf{b}_2\psi_2(\mathbf{x}_i) + \cdots + \mathbf{b}_L\psi_L(\mathbf{x}_i) \ , \tag{4.32}$$

where $\mathbf{b}_1, \mathbf{b}_2, \ldots$ are d-dimensional vectors. From (4.31) and (4.32), $\mathbf{b}_l$ is obtained by

$$\mathbf{b}_l \quad = \quad \sum_{i=1}^{n} \nabla_f(\mathbf{x}_i)\psi_l(\mathbf{x}_i) \ . \tag{4.33}$$

Since $\mathbf{b}_l$ is a linear functional of gradients of f and all the gradients of f belong to the index space, $\mathbf{b}_l$ also belongs to the index space.

Let $\mathbf{B}$ denote a $(d \times L)$ matrix $(\mathbf{b}_1\, \mathbf{b}_2 \cdots \mathbf{b}_L)$. Then it holds $r(\mathbf{B}) \leq M$ ($r(\bullet)$ is the rank of $\bullet$). Suppose further that $\{\psi_l(\mathbf{x}_i)\}$ is such that $r(\mathbf{B}) = M$. Then $(\mathbf{b}_1\, \mathbf{b}_2 \cdots \mathbf{b}_L)$ spans the index space. Let $\mathbf{M}_L$ be a $(d \times d)$ matrix such that

$$\mathbf{M}_L \quad = \quad \sum_{l=1}^{L} \mathbf{b}_l\mathbf{b}_l^\top = \mathbf{B}\mathbf{B}^\top \ . \tag{4.34}$$

Since $r(\mathbf{M}_L) = M$, the first M eigenvectors of $\mathbf{M}_L$ corresponding nonzero eigenvalues estimate the index space.

Let $\mathbf{C}_M$ be a $(d \times M)$ matrix of the first M eigenvectors of $\mathbf{B}\mathbf{B}^\top (= \mathbf{M}_L)$. Singular value decomposition of $\mathbf{B}$ yields

$$\begin{aligned} \mathbf{B} \quad &= \quad \mathbf{C}_M\mathbf{\Lambda}_M^{1/2}\mathbf{O}_M^\top \\ \mathbf{B}\mathbf{O}_M \quad &= \quad \mathbf{C}_M\mathbf{\Lambda}_M^{1/2} \ , \end{aligned} \tag{4.35}$$

where $\mathbf{O}_M$ is a $(L \times M)$ matrix of the first M eigenvectors of $\mathbf{B}^\top\mathbf{B}$, and $\mathbf{\Lambda}_M^{1/2}$ is a diagonal matrix of elements $(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_M})$. Note that M nonzero eigenvalues

are the same for both $\mathbf{B}\mathbf{B}^\top$ and $\mathbf{B}^\top\mathbf{B}$. Rewrite $\mathbf{C}_M\mathbf{\Lambda}_M^{1/2}$ as

$$\mathbf{C}_M\mathbf{\Lambda}_M^{1/2} = (\sqrt{\lambda_1}\boldsymbol{\theta}_1\ \sqrt{\lambda_2}\boldsymbol{\theta}_2\ \cdots\ \sqrt{\lambda_m}\boldsymbol{\theta}_M)\ . \tag{4.36}$$

Then a linear orthogonal mapping $\mathbf{\Gamma}$ can be expressed so that (4.24) is uniquely identified by

$$\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_M)^\top = (\sqrt{\lambda_1}\boldsymbol{\theta}_1\ \sqrt{\lambda_2}\boldsymbol{\theta}_2\ \cdots\ \sqrt{\lambda_m}\boldsymbol{\theta}_M)^\top = (\mathbf{C}_M\mathbf{\Lambda}_M^{1/2})^\top\ . \tag{4.37}$$

$\mathbf{\Gamma} = (\mathbf{C}_M\mathbf{\Lambda}_M^{1/2})^\top$ can be obtained equivalently as in Hristache, Juditsky, Polzehl, and Spokoiny (2001) from the spectral decomposition of the $L \times L$ matrix $\mathbf{B}^\top\mathbf{B}$:

$$\mathbf{B}^\top\mathbf{B} \ = \ \mathbf{O}_L\mathbf{\Lambda}_L\mathbf{O}_L^\top\ . \tag{4.38}$$

Thus, using the first M eigenvectors of $\mathbf{O}_L$, $\mathbf{\Gamma} = (\mathbf{B}\mathbf{O}_M)^\top = (\mathbf{C}_M\mathbf{\Lambda}_M^{1/2})^\top$ is obtained. $\mathbf{O}_M^\top\mathbf{B}^\top\mathbf{B}\mathbf{O}_M = \mathbf{\Lambda}_M$ follows from (4.38), which implies that $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_M$ are orthogonal. Use of the spectral decomposition (4.38), instead of the calculation of $\mathbf{C}_M\mathbf{\Lambda}_M^{1/2}$ from the spectral decomposition of $\mathbf{B}\mathbf{B}^\top$, has an advantage. In practice, the estimate of $\mathbf{B}$ has a rank d, not M. The optimal estimate of $\mathbf{B}$ is obtained from spectral decomposition of $\hat{\mathbf{B}}^\top\hat{\mathbf{B}}$ where $\hat{\mathbf{B}}$ is an estimate of $\mathbf{B}$.

In order to estimate the index space, unknown $\mathbf{B}$ needs to be estimated by $\hat{\mathbf{B}} = (\hat{\mathbf{b}}_1\, \hat{\mathbf{b}}_2 \cdots \hat{\mathbf{b}}_L)$ with

$$\hat{\mathbf{b}}_l \ = \ \sum_{i=1}^{n} \hat{\nabla} f(\mathbf{x}_i)\psi_l(\mathbf{x}_i)\ , \tag{4.39}$$

which requires estimation of the gradient $\nabla f(\mathbf{x}_i)$. The following kernel regression simultaneously estimates $f(\mathbf{x}_i)$ and $\nabla f(\mathbf{x}_i)$:

$$\begin{pmatrix} \hat{f}(\mathbf{x}_i) \\ \hat{\nabla} f(\mathbf{x}_i) \end{pmatrix} \ = \ \underset{\xi_0 \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^d}{\arg\min} \sum_{j=1}^{n} \left(y_j - \xi_0 - (\mathbf{x}_j - \mathbf{x}_i)^\top \boldsymbol{\xi} \right)^2 K\left(\frac{\|\mathbf{x}_j - \mathbf{x}_i\|^2}{h_1^2} \right) \tag{4.40}$$

$$= \ (\mathbf{X}_i^\top \mathbf{K} \mathbf{X}_i)^{-1} \mathbf{X}_i^\top \mathbf{K} \mathbf{y} \tag{4.41}$$

with

$$\mathbf{X}_i \ = \ \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \mathbf{x}_1 - \mathbf{x}_i & \mathbf{x}_2 - \mathbf{x}_i & \cdots & \mathbf{x}_n - \mathbf{x}_i \end{pmatrix}^\top \tag{4.42}$$

$$\mathbf{K} \ = \ \oplus_{j=1}^{n} K(h_1^{-2} \|\mathbf{x}_j - \mathbf{x}_i\|^2)\ , \tag{4.43}$$

where K is a kernel weighting function and h_1 is a bandwidth of a spherical window

around $\mathbf{x}_i$. This estimation will face the curse of dimensionality when d is large because there will not be enough observations within a spherical window with radius h_1. Nonetheless, $\hat{\mathbf{B}}$ obtained from $\hat{\nabla}_f(\mathbf{x}_i)$ in (4.40) contains information about the index space. Hristache, Juditsky, Polzehl, and Spokoiny (2001) proposed an iterative estimation of $\mathbf{B}$ by exploiting the assumed orthogonal structure of index vectors.

Approximation of $g(\mathbf{x})$ by the tangent hyperplane $\Delta y = \nabla_f(\mathbf{x})^\top \Delta\mathbf{x}$ in the directions orthogonal to the index space is relatively good (in those directions $f(\mathbf{x})$ and hence $\nabla_f(\mathbf{x})$ do not change much). This leads to an idea of improving the estimation by using an ellipsoidal window which will contain enough observations inside. Such an ellipsoidal window is obtained by expanding the spherical window in the direction of the space orthogonal to the index space and shrinking it in the directions of the index space. Using the orthogonal structure of index vectors, an ellipsoid given by $\{\mathbf{x} : \|\mathbf{\Gamma}(\mathbf{x} - \mathbf{x}_i)\| \leq h\}$ serves as an appropriate window.

After obtaining a pilot estimate by (4.40), $f(\mathbf{x})$ and $\nabla_f(\mathbf{x})$ are reestimated by kernel regression with an ellipsoidal window. Since $\mathbf{\Gamma}$ is unknown, the window is replaced by $\{\mathbf{x}|\, \|\mathbf{S}_2(\mathbf{x}_j - \mathbf{x}_i)\| < h_2\}$ with

$$\begin{aligned} \mathbf{S}_2 &= (\mathbf{I} + \rho_2^{-2}\hat{\mathbf{\Gamma}}_1^\top \hat{\mathbf{\Gamma}}_1)^{1/2} \\ &= (\mathbf{I} + \rho_2^{-2}\hat{\mathbf{B}}_1\mathbf{O}_1(\hat{\mathbf{B}}_1\mathbf{O}_1)^\top)^{1/2} \\ &= (\mathbf{I} + \rho_2^{-2}\hat{\mathbf{B}}_1\hat{\mathbf{B}}_1^\top)^{1/2}\ , \end{aligned} \tag{4.44}$$

where $\rho_2 < 1$ and $h_2 > h_1$. Reestimation by kernel regression is given by

$$\begin{pmatrix} \hat{f}^{(2)}(\mathbf{x}_i) \\ \hat{\nabla}_f^{(2)}(\mathbf{x}_i) \end{pmatrix} = \underset{\xi_0\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^d}{\arg\min} \sum_{j=1}^{n} \left(y_j - \xi_0 - (\mathbf{x}_j - \mathbf{x}_i)^\top \boldsymbol{\xi}\right)^2 K\left(\frac{\|\mathbf{S}_2(\mathbf{x}_j - \mathbf{x}_i)\|^2}{h_2^2}\right) . \tag{4.45}$$

In the iterative estimation process, the expansion and compression of a kernel window takes place through an increasing parameter h and a decreasing parameter ρ so that there remain enough observations in the window. The iterative estimation proceeds as follows. Let k denote the iteration round ($k = 1, 2, \ldots$).

1. Initialize parameters: ρ_1, ρ_{min} (minimum value of ρ), $a_\rho < 1$, h_1, $a_h > 1$ and $\{\psi_l\}$. Set $k = 1$ and $\hat{\mathbf{B}}_0 = \mathbf{0}$.

2. Compute $\mathbf{S}_k = (\mathbf{I} + \rho_k^{-2}\hat{\mathbf{B}}_{k-1}\hat{\mathbf{B}}_{k-1}^\top)^{1/2}$.

3. Estimate $f(\mathbf{x}_i)$ and $\nabla_f(\mathbf{x}_i)$ by

$$\begin{pmatrix} \hat{f}^{(k)}(\mathbf{x}_i) \\ \hat{\nabla}_f^{(k)}(\mathbf{x}_i) \end{pmatrix} = \underset{\xi_0 \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^d}{\arg\min} \sum_{j=1}^{n} \left(y_j - \xi_0 - (\mathbf{x}_j - \mathbf{x}_i)^\top \boldsymbol{\xi} \right)^2 K \left(\frac{\|\mathbf{S}_k(\mathbf{x}_j - \mathbf{x}_i)\|^2}{h_k^2} \right)$$
$$= (\mathbf{X}_i^\top \mathbf{K}_k \mathbf{X}_i)^{-1} \mathbf{X}_i^\top \mathbf{K}_k \mathbf{y} \ , \tag{4.46}$$

where $\mathbf{K}_k = \oplus_{j=1}^{n} K(h_k^{-2} \|\mathbf{S}_k(\mathbf{x}_j - \mathbf{x}_i)\|^2)$.

4. Compute $\hat{\mathbf{b}}_l^{(k)} = \sum_{i=1}^{n} \hat{\nabla}_f^{(k)}(\mathbf{x}_i)\psi_l(\mathbf{x}_i)$ to obtain $\hat{\mathbf{B}}_k = (\hat{\mathbf{b}}_1^{(k)} \hat{\mathbf{b}}_2^{(k)} \cdots \hat{\mathbf{b}}_L^{(k)})$.

5. Update the parameters: $h_{k+1} = a_h h_k$, $\rho_{k+1} = a_\rho \rho_k$. Set k to $k+1$ and repeat step 2. to 5. until $\rho_{k+1} < \rho_{min}$.

Finally, after the iterative estimation process, link function g is estimated by kernel regression for the estimated index space $\hat{\boldsymbol{\Gamma}}$.

Since $\hat{\mathbf{b}}_l = \sum_{i=1}^{n} \hat{\nabla}_g(\mathbf{x}_i)\psi_l(\mathbf{x}_i)$, estimation of $\mathbf{b}$ requires well-defined estimator of $\hat{\nabla}_g(\mathbf{x}_i)$. In order to prevent the variance of $\hat{\nabla}_f(\mathbf{x}_i)$ from becoming too large, data points around the neighborhood of $\mathbf{x}_i$ need to satisfy some local regularity. Hristache, Juditsky, and Spokoiny (2001) and Hristache, Juditsky, Polzehl, and Spokoiny (2001) proposed a modified estimation using a weighting scheme applied to the terms of $\hat{\mathbf{b}}_l$ depending on the local design. Polzehl and Sperlich (2009) proposed a further modification, assuming the prior knowledge of the dimension of index space M. Their algorithm penalizes a search of an index space outside the presumed space of dimension M or somewhat larger than M. It has been shown that these modifications, especially the penalized algorithm, reduces the error in estimation of the link function.

4.6 Appendix B

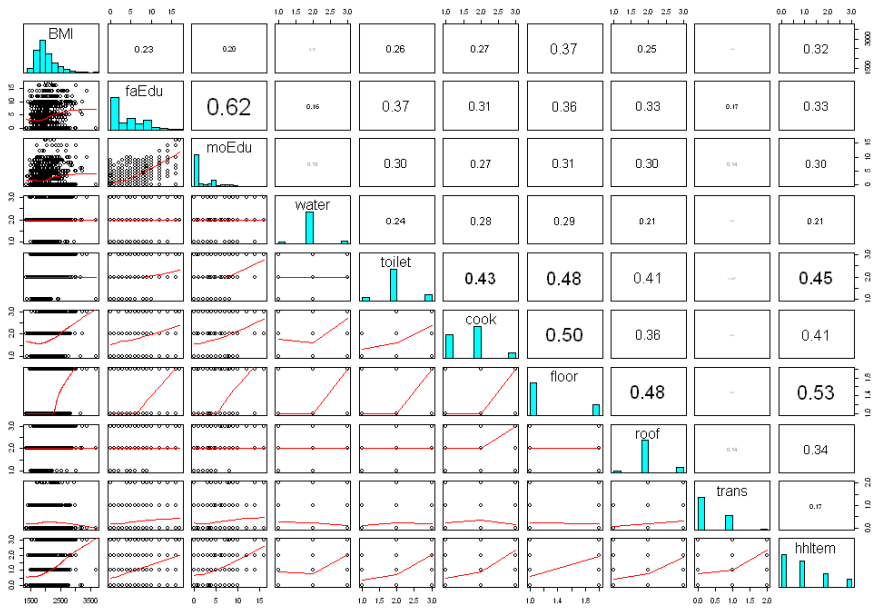


Figure 4.21: Correlation matrix, histograms and scatter plots of the original data.

	faEdu	*moEdu*	*water*	*toilet*	*cook*	*floor*	*roof*	*trans*	*hhItem*
BMI	.23	.20	.15	.31	.31	.51	.33	.05	.36
faEdu		.62	.23	.46	.36	.50	.44	.21	.37
moEdu			.21	.36	.31	.43	.40	.17	.33
water				.38	.40	.47	.33	-.08	.28
toilet					.58	.71	.63	.13	.57
cook						.73	.50	.07	.50
floor							.77	.07	.69
roof								.23	.45
trans									.24

Table 4.22: Polychoric/-serial correlation matrix.

4.7 References

Fan, J. and J. Marron (1994). Fast implementations of nonparametric curve estimators. *Journal of Computational and Graphical Statistics*, 35–56.

Horowitz, J. and W. Härdle (1996). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 1632–1640.

Hristache, M., A. Juditsky, J. Polzehl, and V. Spokoiny (2001). Structure adaptive approach for dimension reduction. *The Annals of Statistics 29*(6), 1537–1566.

Hristache, M., A. Juditsky, and V. Spokoiny (2001). Direct estimation of the index coefficient in a single-index model. *The Annals of Statistics 29*(3), 593–623.

Ibragimov, I. A., A. S. Nemirovskii, and R. Z. Khas'minskii (1986). Some problems on nonparametric estimation in gaussian white noise. *Theory of Probability and its Applications 31*, 391–406.

Kolenikov, S. and G. Angeles (2009). Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer? *Review of Income and Wealth 55*(1), 128–165.

Maydeu-Olivares, A., C. García-Forero, D. Gallardo-Pujol, and J. Renom (2009). Testing categorized bivariate normality with two-stage polychoric correlation estimates. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences 5*(4), 131.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika 44*(4), 443–460.

Polzehl, J. and S. Sperlich (2009). A note on structural adaptive dimension reduction. *Journal of Statistical Computation and Simulation 79*(6), 805–818.

Powell, J., J. Stock, and T. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica: Journal of the Econometric Society*, 1403–1430.

Samarov, A. (1993). Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association 88*(423), 836–847.

Shao, J. and D. Tu (1995). *The Jackknife and Bootstrap*. Springer Series in Statistics. Springer Verlag.

Stoker, T. (1986). Consistent estimation of scaled coefficients. *Econometrica 54*(6), 1461–1481.

Turlach, B. and M. Wand (1996). Fast computation of auxiliary quantities in local polynomial regression. *Journal of Computational and Graphical Statistics 5*(4), 337–350.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics 14*(4), 1261–1295.