Journal of Statistical Planning and Inference 178 (2016) 99–111

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A semiparametric multivariate partially linear model: A difference approach

Lawrence D. Brown [a], Michael Levine [b,*], Lie Wang [c]

[a] Department of Statistics, University of Pennsylvania, United States
[b] Department of Statistics, Purdue University, United States
[c] Department of Mathematics, MIT, United States

ARTICLE INFO

Article history:
Received 7 January 2015
Received in revised form 18 June 2016
Accepted 19 June 2016
Available online 1 July 2016

Keywords:
Multivariate semiparametric model
Difference-based method
Asymptotic efficiency
Partial linear model
Random field

ABSTRACT

A multivariate semiparametric partial linear model for both fixed and random design cases is considered. In either case, the model is analyzed using a difference sequence approach. The linear component is estimated based on the differences of observations and the functional component is estimated using a multivariate Nadaraya–Watson kernel smoother of the residuals of the linear fit. We show that both components can be asymptotically estimated as well as if the other component were known. The estimator of the linear component is shown to be asymptotically normal and efficient in the fixed design case if the length of the difference sequence used goes to infinity at a certain rate. The functional component estimator is shown to be rate optimal if the Lipschitz smoothness index exceeds half the dimensionality of the functional component argument. We also develop a test for linear combinations of regression coefficients whose asymptotic power does not depend on the functional component. All of the proposed procedures are easy to implement. Finally, numerical performance of all the procedures is studied using simulated data.

1. Introduction

Semiparametric models have a long history in statistics and have received considerable attention in the last 30–40 years. They have also been a subject of continuing investigation in subject areas such as econometrics. The main reason they are considered is that sometimes the relationships between the response and predictors are very heterogeneous in the same model. Some of the relationships are clearly linear whereas other ones are much harder to categorize. In many situations, a small subset of variables is presumed to have an unknown relationship with the response that is modeled nonparametrically while the rest are assumed to have a linear relationship with it. As an example, Engle et al. (1986) studied the nonlinear relationship between temperature and electricity usage where other related factors, such as income and price, are parameterized linearly. Shiller (1984) considered an earlier cost curve study in the utility industry using a partial linear model.

The model we consider in this paper is a semiparametric partial linear multivariate model

$$Y_i = a + X_i'\beta + f(U_i) + \varepsilon_i \tag{1.1}$$

* Corresponding author.
E-mail addresses: lbrown@wharton.upenn.edu (L.D. Brown), mlevins@purdue.edu (M. Levine), liewang@math.mit.edu (L. Wang).

http://dx.doi.org/10.1016/j.jspi.2016.06.005
0378-3758/

where $X_i \in \mathbb{R}^p$ and $U_i \in \mathbb{R}^q$, β is an unknown $p \times 1$ vector of parameters, a is an unknown intercept term, $f(\cdot)$ is an unknown function and ε_i are independent and identically distributed random variables with mean 0 and constant variance σ^2. We consider two cases with respect to U: a random design case whereby U is a q-dimensional random variable and a fixed design case with U_i being a q-dimensional vector where each coordinate is defined on an equispaced grid on [0, 1]. In the fixed design case the errors are independent of X_i while in the random design case they are independent of (X_i', U_i). To obtain meaningful results, the function f is assumed to belong in the Lipschitz ball class $\Lambda^\alpha(M)$ where α is the Lipschitz exponent. The version with $q = 1$ was earlier considered in Wang et al. (2011) and we only consider here the case of $q > 1$.

The bibliography concerning the case of $q = 1$ is very extensive and we refer readers to Wang et al. (2011) for details. The case where $q > 1$ has received much less attention in the past. Some of the papers that discussed that model are He and Shi (1996), Samarov et al. (2005), Schick (1996) and Müller et al. (2012). All of them considered random design case only.

In this paper, we consider the estimation of both parametric and nonparametric components. The difference sequence approach utilized in Wang et al. (2011) is generalized so that it can be used when $q > 1$. In the fixed design case, the model is best enumerated using multivariate indices. Such a model is, effectively, a semiparametric random field model. Let n be the sample size; then, using differences of observations, a $\sqrt{n}$-consistent estimator of the parametric component and a $\sqrt{n}$-consistent estimator of the intercept are constructed; to obtain $\sqrt{n}$ rate of convergence for the intercept a, the smoothness of a nonparametric component must exceed $q/2$. As is the case in Wang et al. (2011), the correlation between differences has to be ignored and the ordinary least squares approach must be used instead of the generalized least squares to obtain an optimal estimator. The reason for that is that the use of weighted least squares to estimate β in case of the difference model is conceptually analogous to use of ordinary least squares for the original model (1.1). In either case, the estimate $\hat{\beta}$ will be seriously biased due to the presence of the nonparametric component f. A similar remark is also made in Wang et al. (2011) on p. 5. These estimators can be made asymptotically efficient if the order of the difference sequence is allowed to go to infinity. The estimator of the nonparametric component is defined by using a kernel regression on the residuals and is found to be $n^{-\alpha/(2\alpha+q)}$ consistent. The hypotheses testing problem for the linear coefficients is also considered and an F-statistic is constructed. The asymptotic power of the F-test is found to be the same as if the nonparametric component is known.

In the random design case, the model has univariate indices and so the approach is slightly different. An attempt to generalize the approach of Wang et al. (2011) directly is fraught with difficulties since one can hardly expect to find an ordering of multivariate observations that preserves distance relationships intact. Instead, we utilize a nearest neighbor approach whereby only observations that are within a small distance from the point of interest U_0 are used to form a difference sequence. This inevitably results in difference sequences that have varying lengths for different points of interest in the range of the nonparametric component function. In order to ensure that the length of the difference sequence does not go to infinity too fast, some assumptions on the marginal density function of U_i must be imposed. As in the fixed design case, we obtain a $\sqrt{n}$-consistent estimator of the parametric component and a rate efficient estimator of the nonparametric component.

Our approach is easy to implement in practice for both random and fixed design cases and for an arbitrary dimensionality q of the functional component. Moreover, it guarantees $\sqrt{n}$ rate of convergence for the parametric component regardless of the value of q and provides an easy way of testing standard linear hypotheses about β that have an asymptotic power that does not depend on the unknown nonparametric component.

The paper is organized as follows. Section 2 discusses the fixed design case while Section 3 covers the random design case. The testing problem is considered in Section 4. Section 5 is dedicated to a simulation study that is carried out to study the numerical performance of suggested procedures. Finally, all of the proofs are collected together in the Appendix.

2. Deterministic design

We consider the following semiparametric model

$$Y_{\mathbf{i}} = a + X_{\mathbf{i}}'\beta + f(U_{\mathbf{i}}) + \varepsilon_{\mathbf{i}} \tag{2.1}$$

where $X_{\mathbf{i}} \in \mathbb{R}^p$, $U_{\mathbf{i}} \in S = [0,1]^q \subset \mathbb{R}^q$, $\varepsilon_{\mathbf{i}}$ are i.i.d. zero mean random variables with variance σ^2 and finite absolute moments of the order $\delta+2$ for some small $\delta > 0 : E|\varepsilon_{\mathbf{i}}|^{\delta+2} < \infty$. The assumption of random X_i's is also very common when a multivariate nonparametric model is considered; generally speaking, the use of random X_i's corresponds to the assumption that X_i's are generated by observational (and not experimental) data, which is much more common in practice. In the model (2.1), $\mathbf{i} = (i_1, \ldots, i_q)'$ is a multidimensional index; throughout this article, we will use bold font for all multivariate indices and a regular font for scalar ones. Each $i_k = 0, 1, \ldots, m$ for $k = 1, \ldots, q$; thus, the total sample size is $n = (m+1)^q$. This assumption ensures that $m = o(n)$ as $n \to \infty$. We will say that two indices $\mathbf{i}^1 = (i_1^1, \ldots, i_q^1) \le \mathbf{i}_2 = (i_1^2, \ldots, i_q^2)$ if $i_k^1 \le i_k^2$ for any $k = 1, \ldots, q$; the relationship between $\mathbf{i}^1$ and $\mathbf{i}^2$ is that of partial ordering. Also, for a multivariate index $\mathbf{i}$ $|\mathbf{i}| = |i_1| + \cdots + |i_q|$. Here we assume that $U_{\mathbf{i}}$ follows a fixed equispaced design: $U_{\mathbf{i}} = (u_{i_1}, \ldots, u_{i_q})' \in \mathbb{R}^q$ where each coordinate is $u_{i_k} = \frac{i_k}{m}$. Also, β is an unknown p-dimensional vector of parameters and a is an unknown intercept term. We assume that $X_{\mathbf{i}}$'s are independent random vectors and that $X_{\mathbf{i}}$ is also independent of $\varepsilon_{\mathbf{i}}$; moreover, we denote the non-singular covariance matrix of X as Σ_X. For convenience, we also denote $N = \{0, \ldots, m\}^q$. Note that in this model the intercept a cannot be absorbed in the design matrix X due to identifiability issues; in order to ensure that the model is identifiable,