# Partially Linear Reduced-rank Regression

K. S. Chan
Department of Statistics & Actuarial Science,
University of Iowa, Iowa City, IA 52242
email: kchan@stat.uiowa.edu

Ming-Chung Li
EMMES Corporation,
Rockville, MD 20850, USA
email: mli@emmes.com

H. Tong
Department of Statistics, London School of Economics, UK &
University of Hong Kong, China
email: htong@hkustas2.hku.hk

Feb, 2004

**Abstract**

We introduce a new dimension-reduction technique, the Partially Linear Reduced-rank Regression (PLRR) model, for exploring possible nonlinear structure in a regression involving both multivariate response and covariate. The PLRR model specifies that the response vector loads linearly on some linear indices of the covariate, and nonlinearly on some other indices of the covariate. We give a set of sufficient conditions for the identifiability of the PLRR model. We propose a method for estimating a PLRR model, and derive the large-sample properties of the estimator. Simulation and real data analysis are used to illustrate the new approach.

Key Words: dimension-reduction technique, identifiability, large-sample properties, pollution, semiparametric methods.

## 1 Introduction

Bona fide nonparametric regression with high-dimensional covariate is seldom feasible owing to the curse of dimensionality. Much of the recent literature concerns the development of effective dimension reduction techniques for facilitating the use of nonparametric regression in exploring

possible nonlinear regression relationship; see Xia et al. (2002) and the references therein, and also §2 below. Extension of the available dimension-reduction techniques designed for the case of univariate response and multivariate covariate to the case of nonparametric regression with both multivariate response and covariate is not completely straightforward because of the need for efficiently reducing the covariate dimension simultaneously for all components of the response. Li and Chan (2001) extended the technique of linear reduced-rank regression to a semiparametric dimension-reduction technique useful for exploring the nonlinear relationship between the multivariate response and the multivariate covariate. Essentially, their so-called SemiPArametric Reduced-rank Regression (SPARR) model assumes that the response vector loads linearly on a set of nonlinear factors each of which depends on a set of linear indices of the covariate. In practice, some of the "nonlinear" factors in a SPARR model may be fairly linear over the observable data range. Here, we consider the case that the response vector loads linearly on some linear indices of the covariate, and nonlinearly on some other indices of the covariate that are orthogonal to those linear indices that linearly affect the response; the orthogonality condition is one approach for guaranteeing model identifiability; see below. The new model studied here will be referred to as the Partially Linear Reduced Rank (PLRR) model; this new model bears some resemblance to the partially linear single-index regression (Xia et al., 1999), except that the number of indices can be other than 1 and that the response is a vector. When the PLRR model is applicable, it reduces the dimension of the indices for the genuine nonlinear factors as compared to the general SPARR model so that the nonlinear part can be estimated with greater resolution.

This paper is organized as follows. In §2, we elaborate on the PLRR model, and give a set of sufficient conditions for model identifiability. An iterative estimation procedure is then proposed. The large-sample properties of the proposed estimation method are derived in §3, with proofs relegated to an appendix. A small-scale simulation study on the empirical performance of the proposed estimation scheme is reported in §4. In §5, we illustrate the PLRR model with a real dataset collected in Hong Kong. The fitted PLRR model cast new insights on the temporal dynamical structure of 4 pollution variables and their interaction with weather. We briefly conclude in §6.

## 2 The Model and an Estimation Method

Let $Y_t$ and $X_t$ be $m$ and $n$-dimensional random vectors, both of which are assumed to have

been standardized. The Partially Linear Reduced-rank Regression (PLRR) model is defined as follows:

$$Y_t = C\{DX_t + f(BX_t)\} + \epsilon_t, \quad t = 1, \cdots, T \tag{1}$$

where $\epsilon_t$ are iid, independent of $X_t$ and of zero mean and finite variance matrix, $C, D$ and $B$ are unknown matrices of respectively dimensions $m \times r_1$, $r_1 \times n$ and $r_2 \times n$ matrix; $r_1$ and $r_2$ will be referred to as the ranks of the model. The unknown function $f$ maps from $R^{r_2}$ to $R^{r_1}$. The model is identifiable if (i) the (integer) ranks $r_1 > 0$ and $r_2 > 0$ are minimal, (ii) $D$ and $B$ are orthogonal to each other, i.e. $DB^T = 0$ where the superscript $T$ denotes transpose, and (iii) after suitable permutation of the components of $Y$ and those of $X$ and re-labeling the components if necessary, the leading sub-square matrix of $C$ is the identity matrix of dimension $r_1$, and similarly that of $B$ is the $r_2 \times r_2$ identity matrix. Condition (i) implies that $f(BX)$ cannot be further decomposed as $D_1X + f_1(B_1X)$ such that $D_1B_1^T = 0$, $D_1 \neq 0$ and the row space of $B$ is the direct sum of the row spaces of $D_1$ and $B_1$. Conditions (i) and (ii) entail that the conditional mean function decomposes into a linear part and a "minimal" nonlinear part, the arguments of each part being distinct sets of orthogonal co-ordinates of $X$. To see the need of condition (ii), suppose that (1) holds. We can then subtract from the linear part a term $HBX_t$ where $H$ is any $r_1 \times r_2$ constant matrix, and add the same term to the nonlinear part, resulting in a different decomposition. Condition (ii) eliminates this source of model non-identifiability. Indeed, if condition (ii) is not satisfied, performing the aforementioned procedure with $H = B^T(BB^T)^{-1}$ will result in a decomposition satisfying (ii). Finally, condition (iii) removes model ambiguity due to rotation of $C$ and $B$. Specifically, without the constraints set by (iii), we can use any two invertible matrices $P$ and $Q$ of appropriate dimensions and effect the changes of $C$ to $CP$, $D$ to $P^{-1}D$, $B$ to $QB$ and $f(\cdot)$ to $P^{-1}f(Q^{-1}\cdot)$ without altering the conditional mean of $Y_t$ given $X_t$, causing model non-identifiability. See the appendix for a proof of the model identifiability under conditions (i)-(iii). Note that the covariate $X_t$ may contain lagged values of $Y_t$.

The PLRR model is an interesting special case of the SemiPArametric Reduced-rank Regression (SPARR) model introduced by Li and Chan (2001). The SPARR model is defined as follows:

$$Y_t = Cg(B'X_t) + \epsilon_t, \quad t = 1, \cdots, T, \tag{2}$$

where the components of $g$ may be linear or nonlinear; $B'$ is an unknown $r_2' \times n$ matrix. Here, $B'X$ is said to consist of $r_2'$ indices of $X$. The vector function $g(B'X_t)$ may be regarded as consisting of $r_1$ "nonlinear" factors on which the response loads linearly through the matrix $C$. The SPARR model provides a useful framework for dimension reduction with multivariate response and covariate. In some cases, the function $g$ may split into a linear part and another nonlinear part, which becomes the PLRR model with the rank $r_2 < r_2'$ and hence provides a more parsimonious model. Indeed, a function with a lower-dimensional argument will generally lead to more efficient estimation.

Thus, the PLRR model considers the interesting case that the function $f$ in the SPARR model splits into a linear part and a nonlinear part, the arguments of the two parts being orthogonal linear combinations of $X$. This distinct-orthogonal-argument condition resembles an assumption used by Xia et al. (1999) in the case of scaler response. The SPARR model generalizes a number of existing parametric, nonparametric and semiparametric models including the Reduced-Rank Regression Model (Reinsel and Velu, 1998), the additive model (Hastie and Tibshirani, 1990), the index model (Li, 1992), partially linear model (Carroll et al., 1997; Xia et al., 1999) and projection pursuit (Friedman and Stuetzle, 1981); see Li (2000, pp. 102-4).

We now consider the estimation of a PLRR model. Were the parameters $(B, C, D)$ known, there are several approaches for estimating $f(\cdot)$, e.g., local polynomial (Fan and Gijbels, 1996) and spline smoothing (Eubank, 1988). Here we adopt the local polynomial method (of degree 1, for simplicity), owing to its generally good performance in terms of bias and variance, its ability to adapt to various types of covariate design, and absence of boundary effects; see (Fan and Gijbels, 1996; Ruppert and Wand, 1994) for details.

First, we consider the case that the bandwidth and the ranks of $C$ and $B$ are known. Let $K_h(\cdot)$ be a kernel function with $h > 0$ as the bandwidth, e.g., $K_h(\cdot)$ equals the pdf of the normal distribution with covariance matrix equals $hI$; $I$ is the identity matrix. We propose to estimate the model by minimizing the following weighted least squares criterion function (the notation $||\cdot||$ denotes the $L^2$-norm):

$$\begin{aligned} & L(C, D, B, A_{0t}, A_{1t}, t = 1, \cdots, T; h) \\ = & \sum_t \sum_i ||Y_i - CDX_i - C[A_{0t} + A_{1t}B(X_i - X_t)]||^2 K_h[B(X_i - X_t)] \end{aligned} \tag{3}$$

where $A_{0t}$ are $r_1 \times 1$ vectors and $A_{1t}$ are $r_1 \times r_2$ matrices. The arguments minimizing $L(C, D, B, A_{0t}, A_{1t}, t = 1, \cdots, T; h)$ yield the estimators $\hat{C}, \hat{D}, \hat{B}, \hat{A}_{0t}, \hat{A}_{1t}, t = 1, \cdots, T$, where $\hat{A}_{0t}$ estimates

$f(BX_t)$ and $\hat{A}_{1t}$ estimates the first derivative matrix of $f(\cdot)$ evaluated at $BX_t$.

This objective function can be motivated as follows. For a smooth function $f(\cdot)$, it can be locally approximated by a tangent plane, the effective size of the neighborhood over which the approximation is applied is controlled by the bandwidth of the kernel. Specifically, for a given $x = X_t$, we model the data around $x$ by

$$Y_i = CDX_i + C[A_{0t} + A_{1t}B(X_i - X_t)] + \text{error} \tag{4}$$

where $A_{0t}$ and $A_{1t}$ depend on $X_t$. The $A$s are then estimated by minimizing the weighted sum of squares defined with the kernel function $K$:

$$\sum_i \|Y_i - CDX_i - C[A_{0t} + A_{1t}B(X_i - X_t)]\|^2 \times K_h[B(X_i - X_t)]. \tag{5}$$

While it is desirable to minimize the preceding local least squares simultaneously for all data cases, it is more tractable to minimize the sum of (5) over all $t$, resulting in (3). Strictly speaking, the minimization of the objective function (3) only yields estimators of $f(\cdot)$ at $BX$'s. However, given $\hat{C}, \hat{D}, \hat{B}$ and $\hat{h}$, for any $u$, $f(u)$ can be estimated by $\hat{A}_0$ which minimizes

$$\sum_i ||Y_i - \hat{C}\hat{D}X_i - \hat{C}[A_0 + A_1(\hat{B}X_i - u)]||^2 K_{\hat{h}}(\hat{B}X_i - u).$$

The bandwidth and the ranks can be estimated by minimizing the criterion function

$$L(r_1, r_2) = \sum_t \|Y_t - \hat{C}\{\hat{D}X_t - \hat{f}(\hat{B}X_t)\}\|^2 \tag{6}$$

over a finite grid of $h$, $r_1$ and $r_2$, where $\hat{f}, \hat{B}, \hat{D}$ and $\hat{C}$ are estimated by the cross-validated weighted sum of squares obtained by suppressing the terms with $i = t$ in the double sum in (3). Below, we shall normalize $L(r_1, r_2)$ by the total variance of $Y$ so that $L(\cdot, \cdot)$ can be interpreted as the fraction of unexplained total variances. See §4 for numerical evidence suggesting the consistency of this rank determination procedure.

For fixed ranks and bandwidth, we outline below an iterative procedure for minimizing the objective function defined by (3), with further elaboration including useful formulas given in Appendix B. Given a set of initial values of $B, C$ and $D$, the iterative scheme cycles through the following steps until the objective function converges.

Step 1: Find $\hat{A}_{0t}$ and $\hat{A}_{1t}$ by minimizing the inner sum of the objective function in (3) with respect to $A_{0t}$ and $A_{1t}$. Denote these estimators by $\hat{A}_{0t}^{(k)}$ and $\hat{A}_{1t}^{(k)}$, where $k$ is the iteration number.

Step 2: Update $B$ by minimizing the objective function

$$\sum_t \sum_i \quad ||Y_i - \hat{C}^{(k-1)}\hat{D}^{(k-1)}X_i - \hat{C}^{(k-1)}(\hat{A}_{0t}^{(k)} + \hat{A}_{1t}^{(k)}B(X_i - X_t))||^2 K_h(\hat{B}^{(k-1)}(X_i - X_t)).$$

Let the minimizer be $\hat{B}^{(k)}$. Then we normalize $\hat{B}^{(k)}$ by transforming $\hat{B}^{(k)}$ to the form $(I, \hat{B}^*)$ after permuting the components of $X$ if necessary, where $I$ is the $r_2 \times r_2$ identity matrix and $\hat{B}^*$ is an $r_2 \times (n - r_2)$ matrix.

Step 3: Update $C$ by minimizing the criterion

$$\sum_t \sum_i \quad ||Y_i - C(\hat{D}^{(k-1)}X_i + \hat{A}_{0t}^{(k)} + \hat{A}_{1t}^{(k)}\hat{B}^{(k)}(X_i - X_t)||^2 K_h(\hat{B}^{(k)}(X_i - X_t)).$$

Let the minimizer be $\hat{C}^{(k)}$. Then we normalize $\hat{C}^{(k)}$ by transforming $\hat{C}^{(k)}$ to the form $(I, \hat{C}^{*T})^T$ after permuting the components of $Y$ if necessary, where $I$ is an $r_1 \times r_1$ identity matrix and $C^*$ is an $(m - r_1) \times r_1$ matrix.

Step 4: Update $D$ by minimizing the criterion

$$\sum_t \sum_i \quad ||Y_i - \hat{C}^{(k)}\{\hat{A}_{0t}^{(k)} + \hat{A}_{1t}^{(k)}\hat{B}^{(k)}(X_i - X_t)\} - \hat{C}^{(k)}DX_i||^2 K_h(\hat{B}^{(k)}(X_i - X_t)).$$

In Step 2, $B$ appearing in the kernel function is fixed at the value from the preceding iterate in order to simplify the minimization problem. It can be shown by adapting the proof of (C.17) in Appendix C that the separate updating of the two occurrences of $B$ is asymptotically equivalent to simultaneously updating both occurrences of $B$ in the preceding loss function. Note that the whole $B$ ($C$) is estimated even though it contains some redundant parameters before the normalization. This is done because it is not known beforehand which sub-matrices of $B$ ($C$) are of full rank. Hence we first update the whole $B$ ($C$) followed by normalization using the pivoting technique used in Gauss-Jordan elimination method (Press et al., 1992). After a few iterates, one can fix a certain square sub-matrix of $B$ ($C$) for normalization to the identity matrix.

To start the algorithm, we determine the initial values with the following procedure: First, we run a reduced-rank linear regression of $Y$ on $X$ to find the initial values of $C$ and $D$. Then, we run a projection pursuit regression (Venables and Ripley, 2002) of $Y - \bar{Y} - CDX$ on $HX$ where $H = (I - D^T(DD^T)^{-1}D)$, via the ppr function in R, with the options nterms= $r_2$, max.term= 2. Next, we set $B = \alpha H$ where $\alpha$ is the coefficient matrix from the preceding projection pursuit. The idea is to approximate the nonlinear function $f(\cdot)$ by projection-pursuit regression. And $HX$ is used as the regressor in order to ensure that $B$ is orthogonal to $D$, as required by model identifiability of the PLRR model.

# 3 Asymptotic Properties of the Estimator

We first derive the asymptotic distribution for $\hat{f}$ (and $\hat{f}'$) with the parameters $B, C$ and $D$ assumed known. Indeed, the proof shows that the same result applies if these parameters are known up to an error of order $o_p\{(Th^{r_2})^{-1/2}+h^2\}$. The latter convergence rate holds if $\hat{B}$, $\hat{C}$ and $\hat{D}$ differ from the corresponding true values by an error of $O_P(1/\sqrt{T})$ and $Th^{r_2+4} = O(1)$. It is shown that $\hat{f}$ is asymptotically normal with a bias of order $h^2$ with the rate of convergence being $O_P\{(Th^{r_2})^{-1/2}\}$. Hence the optimal bandwidth according to the mean integrated squared error (MISE) criterion is of the order $O(T^{-1/(r_2+4)})$ where $r_2$ is the rank of $B$. Then we show that if $\hat{B}$, $\hat{C}$ and $\hat{D}$ have convergence rate of $O_P(T^{-1/2})$, then under suitable conditions, these estimators are asymptotically normal. In summary, under some suitable conditions, the bandwidth can be chosen to ensure both the asymptotic normality of $\hat{B}$, $\hat{C}$ and $\hat{D}$, as well as the $(Th^{r_2})^{-1/2}$ convergence rate of $\hat{f}$, at the expense of under-smoothing $\hat{f}$. That is, the bandwidth is of smaller order compared to the rate $O(T^{-1/(r_2+4)})$, the optimal order for estimating $f$ according to the MISE criterion.

Initially, we consider the independent case for ease of exposition, and show at the end of the section how to extend the results to the case of dependent variables with suitable mixing rates. The proofs in Appendix C make use of some techniques in Carroll et al. (1997).

## 3.1 Asymptotic Distribution of the Nonparametric part

Let $g(\cdot) = g(\cdot; B)$ be the marginal density of $U = BX$. Denote by $C_0, B_0$ and $f_0$ the true parameters and the true function, respectively. Also, let $U_0 = B_0X$ and $g_0(\cdot) = g(\cdot; B_0)$ be the pdf of $U_0$. Define the $r_2 \times r_2$ matrices $k_2$, $\nu_2$, scalar $\nu_0$, $r_1 \times 1$ vector $k_{2,f_0,h}$ and $m \times m$ matrix $\Sigma(u)$ by the following formulas:

$$
\begin{aligned}
k_2 &= \int ww^T K(w)dw; && (7)\\
\nu_0 &= \int K^2(w)dw; && (8)\\
\nu_2 &= \int ww^T K^2(w)dw; && (9)\\
k_{2,f_0,h}(u) &= h^2 \int (I_{r_1} \otimes w^T) f_0''(u) wK(w)dw; && (10)\\
\Sigma(u) &= Cov(Y|U_0 = u). && (11)
\end{aligned}
$$

where $w$ denotes an $r_2$-dimensional vector and the integrals are over $R^{r_2}$. The $(r_1r_2) \times r_2$ matrix $f_0''(u)$ consists of the second derivatives of $f_0$ (see (C.7) for the definition). Because of the

identification conditions, we require $C$ and $B$, up to permutations of $X$ and $Y$, to be of the form

$$C = \begin{pmatrix} I \\ C^* \end{pmatrix}, \text{ and } B = (I, B^*). \tag{12}$$

**Condition 1:**

(i) The matrix $C_0^T C_0$ is positive definite.

(ii) The marginal density of $B_0 X$ is positive and continuous at the point $u$.

(iii) The function $f_0(\cdot)$ and its second derivatives are bounded and uniformly Lipschitz continuous; i.e., for some $D$, $||f_0''(u) - f_0''(v)|| \leq D||u - v||$ for all $u$ and $v$, where $D$ is a positive number.

(iv) The matrices $\nu_0 C_0^T \Sigma(u) C_0$ and $\nu_2 \otimes C_0^T \Sigma(u) C_0$ are finite and positive definite at $u$. Denote $\bar{f}_i = \bar{f}_i(u) = f_0(u) + f_0'(u)(U_i - u)$ and $V_1 = \sqrt{h} X_1^* q_1(\bar{f}_1, Y_1) K_h(U_1 - u)$, where

$$X_1^* = \begin{pmatrix} I_{r_1} \\ (\frac{U_1 - u}{h}) \otimes I_{r_1} \end{pmatrix},$$

and $q_1(x, y) = 2C^T(y - Cx)$. Assume $E(V_{i1} V_{j1} V_{l1} V_{m1}) < \infty$ for all $i, j, l,$ and $m$, where $V_{i1}$ is the $i$th element of the $V_1$.

(v) The kernel $K$ is a non-degenerate symmetric density function with bounded first derivative and bounded support.

Condition 1(i) ensures the validity of (12) and Condition 1(v) can be relaxed at the expense of more complex conditions.

**Theorem 3.1** *Assume that $\{Y_i, X_i, i = 1, 2, \cdots, T\}$ are i.i.d. random vectors, and the bandwidth $h$ satisfies the condition that as $T \to \infty, h \to 0, Th^{r_2} \to \infty, Th^{r_2+4} = O(1)$. Under Condition 1, as $T \to \infty$,*

$$(Th^{r_2})^{1/2} \left( \begin{bmatrix} \hat{f}(u) - f_0(u) \\ h\{vec[\hat{f}'(u) - f_0'(u)]\} \end{bmatrix} - \frac{1}{2} \begin{bmatrix} k_{2,f_0,h}(u) \\ 0 \end{bmatrix} \right) \tag{13}$$

*is asymptotically normal with mean zero and the block diagonal covariance matrix*

$$\Sigma_{g_0}(u) \equiv \begin{pmatrix} \Sigma^{11} & 0 \\ 0 & \Sigma^{22} \end{pmatrix} \tag{14}$$

*where*

$$\begin{aligned}\Sigma^{11} &= \nu_0(C_0^TC_0)^{-1}C_0^T\Sigma(u)C_0(C_0^TC_0)^{-1}/g_0(u)\\ \Sigma^{22} &= (k_2^{-1}\nu_2k_2^{-1})\otimes((C_0^TC_0)^{-1}(C_0^T\Sigma(u)C_0)(C_0^TC_0)^{-1})/g_0(u).\end{aligned}$$

Followings are several remarks which aim to clarify the use of the preceding theorem.

1. Note that if $h = O(T^{-1/r})$ with $r_2 < r \le r_2+4$, then the bandwidth condition of Theorem 3.1 are satisfied.

2. Theorem 3.1 indicates that the local polynomial fit for the $j$th component of $f_0(u)$ has the squared asymptotic bias and covariance matrix respectively as:

$$\text{squared bias} \approx k_{2,f_0,h,j}^2(u)/4, \tag{15}$$

$$\text{covariance matrix} \approx \frac{1}{Th^{r_2}}\Sigma_{f_0,j,j}(u) \tag{16}$$

The optimal bandwidth for estimating the $f_{j,0}(u)$ can be determined by minimizing the asymptotic mean integrated square error (AMISE), to be defined below. For a given function $\omega(\cdot)$ with compact support, the AMISE with weight $g_0(\cdot)w(\cdot)$ equals, up to a negligible term,

$$\begin{aligned}AMISE &= \int E[\sum_{j=1}^m(\hat{f}_j(u)-f_{j,0}(u))^2]g_0(u)w(u)du\\ &\approx \frac{1}{4}\sum_{j=1}^m\int k_{2,f_0,h,j}^2g_0(u)w(u)du+\sum_{j=1}^m\frac{1}{Th^{r_2}}\int\Sigma_{f_0,j,j}(u)g_0(u)w(u)du\\ &= \frac{h^4}{4}\int\sum_{j=1}^m\left[\int(e_j^T\otimes w^T)f_0''(u)wk(w)dw\right]^2g(u)w(u)du\\ &\quad+\frac{1}{Th^{r_2}}\nu_0\sum_{j=1}^me_j^T(C_0^TC_0)^{-1}C_0^T\int\Sigma(u)w(u)duC_0(C_0^TC_0)^{-1}e_j.\end{aligned}$$

where $e_j$ denote the unit column vector with 1 in the $j$th position. Consequently, the optimal bandwidth minimizing the AMISE is $O_P(T^{-1/(r_2+4)})$; specifically

$$h_{\text{opt}} = T^{-1/(r_2+4)}\left[\frac{r_2\nu_0\sum_j e_j^T(C_0^TC_0)^{-1}C_0^T\int\Sigma(u)w(u)duC_0(C_0^TC_0)^{-1}e_j}{\int\sum_j\left[\int(e_j^T\otimes w^T)f_0''(u)wk(w)dw\right]^2g_0(u)w(u)du}\right]^{1/(r_2+4)}.$$

This result suggests that when we do a search for the optimal bandwidth over a grid, the grid for higher ranks $r_2$ may be set by rescaling the grid for $r_2 = 1$ by a factor of the order $T^{1/5-1/(r_2+4)}$, which is done in the numerical analysis below. This is motivated by the following

consideration. Let $h(r)$ be the optimal bandwidth with $r_2 = r$. Then $h(r) = O_p(T^{-1/(r+4)}) = O_p(T^{-1/5} * T^{1/5-1/(r+4)}) = T^{1/5-1/(r+4)}O_p(h(1))$, were the true $r_2 = 1$.

### 3.2 Asymptotic Distribution of the Parametric Part

We will assume that $\text{vec}(\hat{B}^*)$, $\text{vec}(\hat{C}^*)$ and $\text{vec}(\hat{D})$ are within some $T^{-1/2}$-neighborhood of their corresponding true values, i.e., $\text{vec}(\hat{B}^* - B_0^*) = O_p(T^{-1/2})$, etc. Let $\epsilon_t = Y_t - C_0\{D_0X + f_0(B_0X_t)\}$, $R = D_0X$ and $U = B_0X$. Denote by $A^{-1}$ the inverse of a square matrix $A$. The following conditions will be needed below.

**Condition 2:**

(i) The function $f_0''(\cdot)$ is continuous in $u \in \mathcal{D}$, a compact set, which is the support of the random variable $U_0$.

(ii) The density of $U_0$ has continuous second derivatives on the set $\mathcal{D}$.

(iii) The conditional density of $U_t = B_0X_t$ given $Y_t$ exists and is uniformly bounded.

(iv) All moments of the error $\epsilon_t$ exist, i.e., $E(|\epsilon|^k) < \infty$ for $k \geq 0$.

(v) The matrix $Q$ defined in Theorem 3.2 is invertible.

Again, these conditions can be relaxed at the expense of more complex conditions.

**Theorem 3.2** *Let the coefficient matrices $\hat{B}^*$ and $\hat{C}^*$ be the estimators satisfying the normalization conditions (12), and $\hat{B}$ and $\hat{D}$ satisfy the constraints $\hat{D}\hat{B}^T = 0$. Assume Conditions 1 and 2 hold and $Th^4 \to 0, \ln T/(Th^{r_2}) \to 0$ and $T^{1-\delta}h^{r_2} \to \infty$ for some arbitrary but fixed $\delta > 0$. Then, as $T \to \infty$,*

$$T^{1/2}\begin{pmatrix} vec(\hat{B}^* - B_0^*) \\ vec(\hat{C}^* - C_0^*) \\ vec(\hat{D} - D_0) \end{pmatrix} \xrightarrow{D} N(0, \mathcal{P}P\mathcal{P}^T) \tag{17}$$

*where, by an abuse of notation, $X$ is partitioned as $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ with $X_1$ being of dimensional $r_2$ and corresponding to the components of $X$ whose coefficients in the indices are fixed according to the constraint (12); similarly partitioned are $X_t = \begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix}$; $D = (D_1, D_2)$ and $D_0 = (D_{10}, D_{20})$;*

$P = Var\{g_0(U)[\Lambda - E(\Lambda|U)C_0(C_0^TC_0)^{-1}C_0^T]\epsilon\}$;

$$\Lambda = \begin{pmatrix} (X_2 \otimes I_{r_2}) f_0'^T (B_0 X) C_0^T \\ 0_{r_1(m-r_1)\times r_1}, f_0(B_0 X) \otimes I_{m-r_1} \\ X \otimes C_0^T \end{pmatrix};$$

$$Q_1 = E[g(U)\Lambda C_0 f_0'(U)(X_2^T \otimes I_{r_2})] - E[g(U)\Lambda C_0 f_0'(U) E(X_2^T \otimes I_{r_2}|U)];$$

$$Q_2 = E\left[g(U)\Lambda \begin{pmatrix} 0_{r_1\times r_1(m-r_1)} \\ (R^T + f_0^T(U)) \otimes I_{m-r_1} \end{pmatrix}\right] - E\left[g(U)\Lambda C_0 (C_0^T C_0)^{-1} C_0^T \begin{pmatrix} 0_{r_1\times r_1(m-r_1)} \\ (E(R^T|U) + f_0^T(U)) \otimes I_{m-r_1} \end{pmatrix}\right];$$

$$Q_3 = E[g(U)\Lambda(X^T \otimes C_0)] - E[g(U)\Lambda C_0 (C_0^T C_0)^{-1} C_0^T E(X^T \otimes C|U)];$$

$$Q = (Q_1, Q_2, Q_3);$$

$\mathcal{P} = Q^{-1}(I - H^T(HQ^{-1}H^T)^{-1}HQ^{-1})$, *where* $H = [(I_{r_2} \otimes D_{20})K_{r_2,n-r_2}, 0_{r_2\times r_1,(m-r_1)\times r_1}, B_0 \otimes I_{r_1}]$*; the commutation matrix* $K_{p,q}$ *is a matrix consisting of ones and zeroes such that, for any* $p \times q$ *matrix* $M$, $K_{p,q}vec(M) = vec(M^T)$*; see (Turkington, 2002, p. 30). Specifically,* $K_{p,q} = [I_p \otimes e_1^q, I_p \otimes e_2^q, \cdots, I_p \otimes e_q^q]$ *where* $e_i^q$ *is the i-th column vector of* $I_q$*, the* $q \times q$ *identity matrix.*

**Remark**: The condition $T^{1-\delta}h^{r_2} \to \infty$ for some arbitrary but fixed $\delta > 0$ implies the validity of (4.5) in Masry (1996) which is required by Lemma 1 of Li and Chan (2001).

Note that if $h = O(T^{-1/r})$ with $4 > r \geq r_2$, then the bandwidth condition in Theorem 3.2 holds. In particular, the asymptotic normality result for the parameter estimates obtains only for $r_2 \leq 3$. It is of interest to further investigate the limiting distribution for dimensions higher than 3.

We now consider how to relax the i.i.d. assumption. Let $\mathcal{F}_a^b$ be the $\sigma$-algebra of events generated by the random variables $\{Y_t, X_t, a \leq t \leq b\}$ and $L_2(\mathcal{F}_a^b)$ denote the collection of all second-order stationary random variables which are $\mathcal{F}_a^b$-measurable. The stationary process $\{Y_t, X_t\}$ is *strongly mixing* (Rosenblatt, 1956) if

$$\sup_{\substack{A\in\mathcal{F}^0 \\ B\in\mathcal{F}_k^{-\infty}}} |P(A\cap B) - P(A)P(B)| = \alpha(k) \to 0 \text{ as } k \to \infty.$$

The coefficients $\alpha(k)$ are known as the strong mixing coefficients.

**Condition 3:**

(i) $|g_{X_1,X_{l+1}}(u,v;l) - g_{X_1}(u)g_{X_{l+1}}(v)| < A_1 < \infty$ for all $l \geq 1$ where $g_{X_1}(u)$ and $g_{X_1,X_{l+1}}(u,v;l)$ denote, respectively, the probability density of $B_0X_1$ and of $(B_0X_1, B_0X_{l+1})$.

(ii) The process $\{Y_i, X_i\}$ is strongly mixing with $\sum_{j=1}^{\infty} j^a[\alpha(j)]^{1-2/v} < \infty$ for some $v > 2$ and $a > 1 - 2/v$.

(iii) The conditional density $f_{U_t|Y_t}(u|y)$ of $U_t$ given $Y_t$ exists and is bounded, i.e., $f_{U_t|Y_t}(u|y) \leq C_1 < \infty$ for some $C_1$.

(iv) The conditional density $f_{(U_t,U_{t+l})|(Y_t,Y_{t+l})}$ of $(U_t, U_{t+l})$ given $(Y_t, Y_{t+l})$ exists and is bounded, i.e., there exists $C_2$ such that, for all $l \geq 1$,

$$f_{(U_t,U_{t+l})|(Y_t,Y_{t+l})}((u,v)|(y_1,y_2)) \leq C_2 < \infty.$$

Theorem 3.1 continues to hold in the dependent case if we assume Condition 3 in addition to the conditions in Theorem 3.1. The proof of Theorem 3.1 has to be modified as follows. Replace $E(W_T)$ by

$$E_T = \frac{h^{r_2}}{\sqrt{Th^{r_2}}} \sum_{i=1}^{T} X_i^* E[q_1(\bar{f}_i, Y_i - CDX_i)|U_i]$$

so that $W_T - E_T$ is the sum of a martingale difference sequence, and that (C.13) continues to hold under Condition 3. Similarly Theorem 3.2 continues to hold if we assume that in addition to the conditions in Theorem 3.2, Condition 3 holds.

# 4 Simulation Study

We report some simulation results checking the empirical performance of the estimation scheme proposed in § 2. The simulation model has the following specification: $m = n = 4$ and $r_1 = r_2 = 2$, with

$$C = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \\ 1.0 & 1.0 \\ 1.0 & -1.0 \end{pmatrix}; D = \begin{pmatrix} 5.0 & -5.0 & 0.0 & 5.0 \\ -5.0 & -5.0 & 5.0 & 0.0 \end{pmatrix}; B = \begin{pmatrix} 1.0 & 0.0 & 1.0 & -1.0 \\ 0.0 & 1.0 & 1.0 & 1.0 \end{pmatrix}.$$

The normalization constraints on $C$ and $B$ fix 8 of the 24 parameters. The $X$'s are independent and identically distributed with the four components of $X$ independent and marginally from the uniform $U(0,1)/sqrt3$ distribution. Recall that $U = BX$. The two nonlinear functions $f_1(u_1, u_2)$ and $f_2(u_1, u_2)$ are defined by the formulas $f_1 = \sin(\pi(u_1 - a)/(b - a)) - 0.6\sin(\pi\sqrt{3}u_2)$ and $f_2 = 2\cos(\pi\sqrt{3}u_2)$ where $a = \sqrt{3}/2 - 1.645/\sqrt{12}$ and $b = \sqrt{3}/2 + 1.645/\sqrt{12}$ are chosen to ensure that the design is relatively thick in the tails (Carroll et al., 1997). The errors are assumed to be normally distributed, uncorrelated over time and also contemporaneously uncorrelated, and with identical noise variance equal to 0.01 (standard deviation being approximately 0.1).

While the consistency of the rank estimators has not been established, we conjecture that the rank estimators are consistent under suitable regularity conditions, which seems to be supported by the simulation results reported below. We simulated data with sample size $T = 50, 100$ and 200 from the above model, and estimated the ranks by minimizing the criterion function defined by (6) with $r_1$ and $r_2$ between 1 to 4, and the bandwidth parameter being selected from 0.05, 0.1 to 0.5 with increment equal to 0.1. Each experiment was replicated 100 times. Table 1 reports the frequencies of each combination of ranks that were selected. Note that for $T = 50$, the ranks were correctly selected 70 times out of 100, and for $T = 100$, the ranks were correctly estimated 95 times out of 100 times, and then 99 times out of 100 times when $T = 200$. Hence, the results provide some empirical evidence that the ranks can be consistently estimated by the proposed method. In practice, semiparametric analysis of a high-dimensional multivariate system generally requires larger sample size than 50 so the case of $T = 50$ is included mainly for checking the performance of the estimation scheme for the case of low data-per-parameter ratio.

We have also checked the accuracy of the asymptotic normality distributional result for the estimators of $\hat{B}, \hat{C}$ and $\hat{D}$. Specifically, we construct 95% confidence intervals for each free parameter according to Theorem 3.2 and compute their empirical coverage rates of the intervals. For $T = 200$, the average empirical coverage rates for the coefficients of $B, C$ and $D$ were 93.5%, 94.8% and 91.7% respectively, and these coverage rates became 98.0%, 95.5% and 97% respectively when $T = 300$, with all experiments replicated 200 times. Thus, the empirical coverage rates were close to their nominal values.

# 5 Hong Kong Pollution Data

We now illustrate our approach with a real dataset collected for studying the effects of pollutant and weather on circulatory and respiratory diseases. The pollutant and weather data are the daily average levels of Sulphur Dioxide (S, ($\mu g/m^{-3}$), log-transformed), Nitrogen Dioxide (N, ($\mu g/m^{-3}$)), respirable suspended Particulates (P, ($\mu g/m^{-3}$), log-transformed), Ozone (O, ($\mu g/m^{-3}$), square-root transformed) Temperature (T ($^{\circ}C$)) and relative Humidity (H (%)). The data were collected daily in Hong Kong from January 1st, 1994, to December 31st, 1995, and have been analyzed before by Xia et al. (2002), Cai et al. (2000) and Fan and Zhang (1999). However, previous studies mainly focused on analyzing the effects of pollutants and weather on daily number of daily hospital admissions of patients suffering from circulatory and

respiratory problems. Here, we focus on the underlying temporal dynamics of the pollutants and its interactions with weather.

The scatter diagrams of pairs of the 6 variables are displayed in Figure 1, which indicate the contemporaneous relationships between the variables. These relationships mostly appear to be linear, but there are appreciable curvature in the scatter diagrams between Temperature and some of the pollutant variables, and less so between Humidity and the pollutant variables. The extents of temporal dependence in these variables are summarized by the correlation functions that are plotted in Figure 2. Except for Sulphur Dioxide, the other three pollutants (Nitrogen Dioxide, Particulates and Ozone) are highly seasonal and significantly auto-correlated and cross-correlated over time, and so are Temperature and Humidity. The auto-correlation function of Temperature appears to decay linearly initially and hence Temperature appears to be nonstationary. The multivariate nonlinear structure of this dataset has earlier been explored by Chan and Li (2002) who fitted a SPARR model with the response variables consisting of the four pollutant variables (S, N, P, O) and the covariates being the lags 1, 2 and 7 of the response variables, as well as current Temperature and its lag 1, and current Humidity and its lag 1. Lags 1 and 2 of the response variables are included in the regressors to model short-term memory of the system and lag 7 for possible weekly effects; weekly effects may be appreciable as Hong Kong is a very busy and densely populated city. Using a criterion function analogous to the objective function defined by (6), Chan and Li (2002) estimated a SPARR model with the ranks $r_1 = r_2 = 2$ for the Hong Kong pollution data. The fitted SPARR model suggests some piecewise linear structure in the data.

As Figure 1 suggests that the dynamics of the system is largely linear except that the pollutants may relate nonlinearly with Temperature, and possibly less so with Humidity, the PLRR model may be appropriate for this dataset. Using the same set of 16-dimensional covariates used by Chan and Li (2002), we estimated the ranks using the criterion function defined by (6). The left side of Table 2 suggests that while the objective function is strictly minimized when $r_1 = 4$ and $r_2 = 1$. The combination $r_1 = 3$ and $r_2 = 1$ is very competitive and since for $r_2 = 1$, the objective function drops precipitously from $r_2 = 1$ to $r_2 = 3$ and the decrease in the objective function from $r_2 = 3$ to $r_2 = 4$ is relatively small, we choose the ranks to be $r_1 = 3$ and $r_2 = 1$. This represents a simplification over the SPARR model fitted by Chan and Li (2002) as the nonlinear factor now depends on a 1-dimensional index as compared to the 2-dimensional

arguments for the nonlinear functions of the fitted SPARR model.

However, for the fitted PLRR model with $r_1 = 3$ and $r_2 = 1$, all the coefficients in $\hat{C}, \hat{B}$ and $\hat{D}$ (not reproduced here for saving space) related to lags 2 and 7 of the response variables were found to be insignificant, with individual 5% significance level. We subsequently trimmed the 16-dimensional covariate to an 8-dimensional covariate vector consisting of the lag 1 of the response variables, Temperature and its lag 1, and Humidity and its lag 1. The right panel of Table 2 reports the corresponding criterion function for the refined PLRR model with the 8-dimensional covariate, for $1 \leq r_1 \leq 4$ and $1 \leq r_2 \leq 4$. Notice for $r_1 = 4$ and $r_2 = 3, 4$, the objective function cannot be computed as normalization for $\hat{C}$ or $\hat{B}$ cannot be done. Again, the ranks $r_1 = 3$ and $r_2 = 1$ provide arguably the best estimates and the corresponding estimates of $C, D$ and $B$ with their standard errors enclosed in parentheses are reported below (significant estimates are bold-faced, standard errors of those estimates subject to normalization constraints are undefined and hence marked as NAs):

$$\hat{C} = \begin{pmatrix} S & 0.000 & \mathbf{1.000} & 0.000 \\ s.e. & (NA) & (NA) & (NA) \\ N & \mathbf{1.000} & 0.000 & \text{-}0.000 \\ s.e. & (NA) & (NA) & (NA) \\ P & \mathbf{0.799} & \text{-}0.005 & \mathbf{0.271} \\ s.e. & (0.036) & (0.043) & (0.041) \\ O & 0.000 & 0.000 & \mathbf{1.000} \\ s.e. & (NA) & (NA) & (NA) \end{pmatrix}$$

$$\hat{D} = \begin{pmatrix} & S-lag1 & N-lag1 & P-lag1 & O-lag1 & T & T-lag1 & H & H-lag1 \\ row1 & \text{-}0.042 & \mathbf{0.460} & 0.231 & \text{-}0.162 & \text{-}0.053 & \text{-}0.131 & 0.029 & \text{-}0.146 \\ s.e. & (0.173) & (0.224) & (0.291) & (0.190) & (0.108) & (0.127) & (0.196) & (0.229) \\ row2 & \mathbf{0.608} & \text{-}0.108 & 0.089 & \text{-}0.306 & 0.003 & \text{-}0.072 & 0.079 & \text{-}0.095 \\ s.e. & (0.327) & (0.418) & (0.533) & (0.363) & (0.133) & (0.178) & (0.362) & (0.408) \\ row3 & \text{-}0.108 & \mathbf{0.253} & \text{-}0.005 & \mathbf{0.360} & 0.035 & 0.135 & \mathbf{\text{-}0.598} & \mathbf{0.203} \\ s.e. & (0.049) & (0.063) & (0.065) & (0.062) & (0.105) & (0.105) & (0.056) & (0.063) \end{pmatrix}$$

$$\hat{B} = \begin{pmatrix} & S-lag1 & N-lag1 & P-lag1 & O-lag1 & T & T-lag1 & H & H-lag1 \\ row1 & 0.022 & -0.042 & -0.024 & 0.287 & \mathbf{1.000} & \mathbf{-0.951} & -0.010 & -0.011 \\ s.e. & (0.289) & (0.370) & (0.478) & (0.340) & (NA) & (0.198) & (0.318) & (0.354) \end{pmatrix}$$

The nonlinear function depends on a single index $BX_t$, but only the coefficients of temperature and its lag 1 appear to be significant in $B$. Interestingly, upon suppressing the insignificant coefficients, $BX_t$ essentially becomes the first difference of Temperature. Figure 3 displays the scatter plots of $f(U)$ versus $U = BX_t$, suggesting minor curvature in the first component that resembles saturation effects at both tails; the second and third components of $f$ appear to be piecewise linear functions. For the linear factors $DX_t$, its first component essentially depends on lag 1 of Nitrogen Dioxide, the second component on the lag 1 of Sulphur Dioxide and the third component depends on lags 1 of Nitrogen Dioxide and Ozone, as well as Humidity and its lag 1. Figure 4 displays the cross- and auto-correlation of the residuals, indicating non-zero contemporaneous correlations but the cross- and autocorrelations are generally insignificant, suggesting that the PLRR model provides a good fit to the data.

In summary, the fitted PLRR model suggests that over the study period (i) Sulphur Dioxide depended linearly on its lag 1 and nonlinearly on the first difference of Temperature, (ii) Nitrogen Dioxide depended linearly on its lag 1 and somewhat nonlinearly on the first difference of Temperature, (iii) Particulate depended linearly on lag 1 of Sulphur Dioxide, lag 1 of Nitrogen Dioxide, lag 1 of Ozone, Humidity and its lag 1, and nonlinearly on the first difference of Temperature, and (iv) Ozone depended linearly on the lag 1 of Nitrogen Dioxide, lag 1 of Ozone, Humidity and its lag 1, and nonlinearly on the first difference of Temperature. The nonlinear effect of Temperature was strongest on Ozone, with Ozone reducing sharply when the weather cooled down substantially from the previous day.

## 6 Conclusion

The above analysis with the Hong Kong pollution data illustrates the potential usefulness of PLRR model in eliciting the linear and nonlinear regression structure. The estimation of the PLRR models requires computing an objective function that comprises many local sum of squares, and hence is computer-intensive. Further research is required to expedite the computation efforts to facilitate this approach for very large datasets. Another interesting issue concerns the exploration of adopting different approaches for decomposing a nonlinear regression function into a sum of linear and nonlinear parts. Here, we impose the orthogonality condition between

the indices in the nonlinear part and those in the linear part. But the question remains whether other approaches may be applied to effect the decomposition of the nonlinear function into a linear part and another nonlinear part.

# 7 Acknowledgements

KSC and HT thank the RGC (Hong Kong, grant no HKU 7111/02P) for partial support.

# A Identifiability of the PLRR model

Assume conditions (i) to (iii) hold. It follows from the proof of of the identifiability of the SPARR model given in Li and Chan (2001) that both $C$ and the sum $DX + f(BX)$ are unique. Thus, it remains to prove that the decomposition of the sum into linear and nonlinear parts satisfying conditions (i) to (iii) is unique. Suppose that we have two such decompositions so that $DX + f(BX) = \tilde{D} + \tilde{f}(\tilde{B}X)$, $\forall X$. We shall prove that the two decompositions are identical. Let $\mathcal{D} \subset R^n$ be a vector subspace of the row space of $D$, the vector subspace spanned by the row vectors of $D$, and $\tilde{\mathcal{D}}$ be that of $\tilde{D}$. Let $(\mathcal{D} + \tilde{\mathcal{D}})^{\perp}$ be the space orthogonal to the direct sum $\mathcal{D} + \tilde{\mathcal{D}}$. For any $X \in R^n$, there exist $X_1 \in \tilde{\mathcal{D}}, X_2 \in \mathcal{D}$ and $X_3 \in (\mathcal{D} + \tilde{\mathcal{D}})^{\perp}$ such that $X = X_1 + X_2 + X_3$. Consequently, $\tilde{B}X_1 = 0$ and $BX_2 = 0$. Then,

$$\begin{aligned}
\tilde{f}(\tilde{B}X) + \tilde{D}X &= \tilde{f}(\tilde{B}(X_2 + X_3)) + \tilde{D}X \\
&= f(B(X_2 + X_3)) + D(X_2 + X_3) - \tilde{D}(X_2 + X_3) + \tilde{D}X \\
&= f(BX_3) + D(X_2 + X_3) + \tilde{D}X_1 \\
&= \tilde{f}(\tilde{B}X_3) + \tilde{D}(X_1 + X_3) + DX_2. \qquad \text{(A.1)}
\end{aligned}$$

If $\mathcal{D} + \tilde{\mathcal{D}}$ contains $\tilde{\mathcal{D}}$ as a proper subset, then $(\mathcal{D} + \tilde{\mathcal{D}})^{\perp}$ is a proper subspace of $\mathcal{B}$, in which case there exist a positive integer $k < r_2$, a $r_2 \times k$ matrix $B_1$ and a $k \times n$ matrix $B_2$ such that $\tilde{B}X_3 = B_1B_2X$. Letting $\check{f}$ be the function defined by $\check{f}(\cdot) = f(B_1\cdot)$, we have $\tilde{f}(\tilde{B}X_3) = \check{f}(B_2X)$, resulting in a decomposition with a nonlinear part whose argument is of lower dimension than that of $\tilde{B}$, contradicting condition (i). Hence, $\mathcal{D} + \tilde{\mathcal{D}} = \tilde{\mathcal{D}}$. Similarly, $\mathcal{D} + \tilde{\mathcal{D}} = \mathcal{D}$ hence $\mathcal{D} = \tilde{\mathcal{D}}$. Consequently, $\mathcal{B} = \tilde{\mathcal{B}}$, hence $B = \tilde{B}$ by condition (iii).

Recall that $B = (I, B^*)$. Partition $D$, $\tilde{D}$ and $X$ accordingly as $D = (D_1, D_2), \tilde{D} = (\tilde{D}_1, \tilde{D}_2)$ and $X = (X_1, X_2)$. Letting $X_2 = 0$, the equality $DX + f(BX) = \tilde{D}X + \tilde{f}(BX)$ entails that

$$D_1X_1 + f(X_1) = \tilde{D}_1X_1 + \tilde{f}(X_1), \qquad \text{(A.2)}$$

for all $X_1 \in R^{r_2}$. In other words, $f$ and $\tilde{f}$ differ additively by some linear function. Hence,

$$\begin{aligned} & DX + f(BX) = \tilde{D}X + \tilde{f}(BX) \\ \Rightarrow \quad & DX + f(BX) = \tilde{D}X + (D_1 - \tilde{D}_1)BX + f(BX) \\ \Rightarrow \quad & DX = \tilde{D}X + (D_1 - \tilde{D}_1)(X_1 + B^*X_2) \\ \Rightarrow \quad & (D_2 - \tilde{D}_2)X_2 = (D_1 - \tilde{D}_1)B^*X_2. \end{aligned}$$

Since this is true for all $X_2$, $D_2 - \tilde{D}_2 = (D_1 - \tilde{D}_1)B^*$. Condition (ii) implies that $0 = DB^T = D_1 + D_2B^{*T}$ so that $D_1 = -D_2B^{*T}$. Therefore, $D_2 - \tilde{D}_2 = (D_1 - \tilde{D}_1)B^* = -(D_2 - \tilde{D}_2)B^{*T}B^*$, so $(D_2 - \tilde{D}_2)(I + B^{*T}B^*) = 0$ implying that $D_2 = \tilde{D}_2$ because $I + B^{*T}B^*$ is invertible. Thus, $D = \tilde{D}$. This completes the proof of the model identifiability under conditions (i) to (iii).

# B Some Formulas Useful for Implementing Steps 1 to 3.

We recall some well-known results that will be needed later; see (Reinsel and Velu, 1998, p. 4–6) and Wand (2002). For any matrices $A, B, C$ of appropriate dimensions that make the operations below well-defined, we have

$$\begin{aligned} \mathrm{vec}(ABC) &= (C^T \otimes A)\mathrm{vec}(B) \\ &= (I \otimes AB)\mathrm{vec}(C) \\ &= (C^TB^T \otimes I)\mathrm{vec}(A) \\ \mathrm{tr}(ABCB^T) &= (\mathrm{vec}(B))^T(C^T \otimes A)\mathrm{vec}(B) \\ \frac{\partial \mathrm{tr}(CZ)}{\partial C} &= Z^T \\ \frac{\partial \mathrm{tr}(CXX^TC^T)}{\partial C} &= 2CXX^T, \quad \frac{\partial \mathrm{tr}(CXX^TC^T)}{\partial C^T} = 2XX^TC^T \end{aligned}$$

**Theorem B.1** *Let $B^*$ minimize $\sum_i tr((Y_i - A_iBX_i)^TW_i(Y_i - A_iBX_i))$. Then the normal equation is given by*

$$\sum_i (A_i^TW_iY_iX_i^T - A_i^TW_iA_iBX_iX_i^T) = 0.$$

*Alternatively,*

$$\sum_i (X_iX_i^T \otimes (A_i^TW_iA_i))vec(B) = \sum_i vec(A_i^TW_iY_iX_i^T).$$

Proof:

$$\begin{aligned} g(B) &= \mathrm{tr}((Y_i - A_iBX_i)^T W_i (Y_i - A_iBX_i)) \\ &= \mathrm{tr}(Y_i^T W_i Y_i - 2Y_i^T W_i A_i BX_i + (A_iBX_i)^T W_i (A_iBX_i)) \\ &= \text{constant} - 2\mathrm{tr}(BX_iY_i^T W_i A_i) + \mathrm{tr}(X_i^T B^T A_i^T W_i A_i BX_i) \\ &= \text{constant} - g_1(B) - g_2(B). \end{aligned}$$

Observe that

$$\partial g_1(B)/\partial B = -2A_i^T W_i Y_i X_i^T;$$

$$\begin{aligned} g_2(B) &= \mathrm{tr}(X_i^T B^T A_i^T W_i A_i BX_i) \\ &= \mathrm{tr}(A_i^T W_i A_i BX_i X_i^T B^T) \\ &= (\mathrm{vec}(B))^T (X_iX_i^T \otimes A_i^T W_i A_i)\mathrm{vec}(B), \\ \partial g_2(B)/\partial \mathrm{vec}(B) &= 2(X_iX_i^T \otimes A_i^T W_i A_i)\mathrm{vec}(B), \\ \partial g_2(B)/\partial B &= 2A_i^T W_i A_i BX_i X_i^T. \end{aligned}$$

Therefore,

$$\partial g(B)/\partial B = -2A_i^T W_i Y_i X_i^T + 2A_i^T W_i A_i BX_i X_i^T.$$

**Theorem B.2** *Let $B$ minimize $\sum_i tr((Y_i - ABX_i)^T(Y_i - ABX_i))$ subject to the linear constraints $BC^T = 0$, where $A$ is $m \times r, B$ is $r \times n$ and $C$ is $s \times n$. Then the solution satisfies the following equation:*

$$\begin{aligned} [\sum_i (X_iX_i^T) \otimes (A^TA)]vec(B) &= \sum_i vec(A^T Y_i X_i^T), \\ (C \otimes I_r)vec(B) &= 0. \end{aligned}$$

We now state the updating formulas for the iterative estimation scheme, one component at a time, as solutions of some linear least squares problem (the components that not being updated are set as their most recent values in the iteration):

- Update $\{A_{0t}, A_{1t}\}$ by minimizing

$$\begin{aligned}
& \sum_t \sum_{i \neq t} ||Y_i - C[DX_i + A_{0t} + A_{1t}B(X_i - X_t)]||^2 K_h(B(X_i - X_t)) \\
= & \sum_t \sum_{i \neq t} ||Y_i - CDX_i - C[A_{0t} + A_{1t}B(X_i - X_t)]||^2 K_h(B(X_i - X_t)) \\
= & \sum_t \sum_{i \neq t} ||Y_i^* - C[A_{0t} + A_{1t}B(X_i - X_t)]||^2 K_h(B(X_i - X_t)) \\
& \text{where } Y_i^* = Y_i - CDX_i \\
= & \sum_{t, i \neq t} ||Y_{it}^* - C[A_{0t}K_h^{1/2}(B(X_i - X_t)) + A_{1t}(BX)_{it}^*]||^2 \\
& \text{where } Y_{it}^* = Y_i^* K_h^{1/2}(B(X_i - X_t)), (BX)_{it}^* = B(X_i - X_t)K_h^{1/2}(B(X_i - X_t)) \\
= & \sum_{t, i \neq t} ||Y_{it}^* - C \left[ (A_{0t}, A_{1t}) \begin{pmatrix} K_h^{1/2}(B(X_i - X_t)) \\ (BX)_{it}^* \end{pmatrix} \right] ||^2.
\end{aligned}$$

- Update $B$ by minimizing

$$\begin{aligned}
& \sum_t \sum_{i \neq t} ||Y_i - C[DX_i + A_{0t} + A_{1t}B(X_i - X_t)]||^2 K_h(B(X_i - X_t)) \\
= & \sum_t \sum_{i \neq t} ||Y_i^* - CA_{0t} - CA_{1t}B(X_i - X_t)||^2 K_h(B(X_i - X_t)) \\
& \text{where } Y_i^* = Y_i - CDX_i \\
= & \sum_{t, i \neq t} ||Y_{it}^* - CA_{1t}BX_{it}^*||^2 \\
& \text{where } Y_{it}^* = Y_i^* K_h^{1/2}(), X_{it}^* = (X_i - X_t)K_h^{1/2}() \\
= & \sum_{t, i \neq t} ||Y_{it}^* - C_{1t}BX_{it}^*||^2 \\
& \text{where } C_{1t} = CA_{1t}.
\end{aligned}$$

- Update $C$ by minimizing

$$\begin{aligned}
& \sum_t \sum_{i \neq t} ||Y_i - C[DX_i + A_{0t} + A_{1t}B(X_i - X_t)]||^2 K_h(B(X_i - X_t)) \\
= & \sum_{t, i \neq t} ||Y_i^* - CX_{it}^*||^2 \\
& \text{where } Y_{it}^* = Y_i K_h^{1/2}() \\
& X_{it}^* = [DX_i + A_{0t} + A_{1t}B(X_i - X_t)]K_h^{1/2}().
\end{aligned}$$

- Update $D$ by minimizing

$$\begin{aligned} & \sum_t \sum_{i \neq t} ||Y_i - C[A_{0t} + A_{1t}B(X_i - X_t)] - CDX_i||^2 K_h(B(X_i - X_t)) \\ = & \sum_{t, i \neq t} ||Y_{it}^* - CDX_{it}^*||^2 \end{aligned}$$

subject to $DB^T = 0$ where

$$\begin{aligned} Y_{it}^* &= (Y_i - C[A_{0t} + A_{1t}B(X_i - X_t)]) * K_h^{1/2}(B(X_i - X_t)), \\ X_{it}^* &= X_i * K_h^{1/2}(B(X_i - X_t)). \end{aligned}$$

# C Proofs of Theorems 3.1 and 3.2

To save space, routine calculations are omitted from the proofs; see Li (2000) for details.

**Proof of Theorem 3.1:**

We will prove something stronger than Theorem 3.1, with $B$, $C$ and $D$ deviating from their true values by an error of order $o\{(Th^{r_2})^{-1/2} + h^2\}$. Let $c_T = (Th^{r_2})^{-1/2}$, $u = Bx$ and $U_i = BX_i$ and

$$X_i^* = \begin{pmatrix} I_{r_1} \\ \left(\frac{U_i - u}{h}\right) \otimes I_{r_1} \end{pmatrix}, A^* = \begin{pmatrix} c_T^{-1}\{A_0 - f_0(u)\} \\ c_T^{-1} h\{\mathrm{vec}(A_1 - f_0'(u))\} \end{pmatrix}.$$

Recall $\bar{f}_i = \bar{f}_i(u) = f_0(u) + f_0'(u)(U_i - u)$.

Since

$$\begin{aligned} & A_0 + A_1(U - u) \\ = \; & A_0 + \mathrm{vec}(A_1(U - u)) \\ = \; & (I_{r_1}, (U - u)^T \otimes I_{r_1}) \begin{pmatrix} A_0 \\ \mathrm{vec}(A_1) \end{pmatrix} \\ = \; & c_T (I_{r_1}, (\frac{U - u}{h})^T \otimes I_{r_1}) \begin{pmatrix} c_T^{-1}(A_0 - f_0(u)) \\ c_T^{-1} h(\mathrm{vec}(A_1 - f_0'(u))) \end{pmatrix} + f_0(u) + f_0'(u)(U - u) \\ = \; & c_T X_i^{*T} A^* + \bar{f}_i, \end{aligned}$$

the objective function for estimating $(f_0(u), f_0'(u))$ can be written as

$$-\sum_i tr[\{Y_i - C(DX_i + c_T X_i^{*T} A^* + \bar{f}_i)\}\{Y_i - C(DX_i + c_T X_i^{*T} A^* + \bar{f}_i)\}^T] K_h(B(X_i - x)) \quad \text{(C.1)}$$

Consider the normalized function

$$\begin{aligned} l_T(A^*) &= -h^{r_2}\sum_i tr[\{Y_i - C(DX_i + c_T X_i^{*T}A^* + \bar{f}_i)\}\{Y_i - C(DX_i + c_T X_i^{*T}A^* + \bar{f}_i)\}^T \\ &\quad -\{Y_i - C(DX_i + \bar{f}_i)\}\{Y_i - C(DX_i + \bar{f}_i)\}^T]K_h(B(X_i - x)) \end{aligned}$$

which is maximized by $\hat{A}^*$. By Taylor expansion and after some algebra, we have

$$\begin{aligned} l_T(A^*) &= \sum_{i=1}^{T} h^{r_2}[(c_T X_i^{*T}A^*)^T q_1(\bar{f}_i, Y_i - CDX_i) + \\ &\quad \frac{1}{2}(c_T X_i^{*T}A^*)^T q_2(\bar{f}_i, Y_i - CDX_i)(c_T X_i^{*T}A^*)]K_h(U_i - u) \\ &= A^{*T}W_T + \frac{1}{2}A^{*T}F_T A^* \end{aligned}$$

where

$$\begin{aligned} q_1(x, y) &= -\frac{\partial}{\partial x}tr[(y - Cx)(y - Cx)^T] = 2C^T(y - Cx) \\ q_2(x, y) &= \frac{\partial}{\partial x^T}q_1(x, y) = -2C^T C < 0 \end{aligned}$$

and

$$W_T = h^{r_2}c_T\sum_{i=1}^{T} X_i^* q_1(\bar{f}_i, Y_i - CDX_i)K_h(U_i - u), \tag{C.2}$$

$$F_T = h^{r_2}c_T^2\sum_{i=1}^{T} X_i^* q_2(\bar{f}_i, Y_i - CDX_i)X_i^{*T}K_h(U_i - u). \tag{C.3}$$

It can be shown (Li, 2000) that

$$F_T = -F + o_P(1), \tag{C.4}$$

where

$$F = F(u) = 2g_0(u)\begin{pmatrix} C_0^T C_0 & 0 \\ 0 & k_2 \otimes C_0^T C_0 \end{pmatrix}. \tag{C.5}$$

Therefore,

$$\hat{A}^* = F^{-1}W_T + o_P(1). \tag{C.6}$$

Hence the asymptotic normality of $\hat{A}^*$ will follow from that of $W_T$. Since $W_T$ is a sum of i.i.d. random vectors, we need to compute the first two moments and check conditions for the

Central Limit Theorem. First, we consider the Taylor expansion of $f_0 = (f_{j0})$. It follows from condition 1(iii) and the intermediate value theorem that

$$f_0(U) \;=\; f_0(u) + f_0'(u)(U-u) + \frac{1}{2}(I_{r_1} \otimes (U-u)^T) f_0''(\zeta)(U-u).$$

where $f_{j0}''(\zeta)$ is an $r_2 \times r_2$ matrix,

$$f_0''(\zeta) \equiv (f_{0,1}''^T(\zeta_1), \cdots, f_{0,r_1}''^T(\zeta_{r_2}))^T$$

is an $r_1 r_2 \times r_2$ matrix, and $\zeta$'s are some "intermediate" points between $u$ and $U$. Note that when $u$ coincides with $U$ so that $\zeta = u$, then

$$f_0''(u) \equiv (f_{0,1}''^T(u), \cdots, f_{0,r_1}''^T(u))^T. \tag{C.7}$$

This will be used in the derivation of $EW_T$ below. From the definition of $W_T$, we have,

$$\begin{aligned}
EW_T &= h^{r_2} c_T E(\sum_{i=1}^{T} X_i^* q_1(\bar{f}_i, Y_i - CDX_i) K_h(U_i - u)) \\
&= c_T^{-1} E\{X^* 2C^T [C_0 f_0(B_0 X) - C\bar{f}] K_h(U-u)\} \\
&= c_T^{-1} E\{X^* 2C^T C [f_0(B_0 X) - f_0(BX) + f_0(U) - f_0(u) \\
&\qquad\qquad -f_0'(u)(U-u)] K_h(U-u) + O(||C - C_0||) + O(||D - D_0||)\} \\
&\qquad \text{because } \bar{f} = f_0(u) + f_0'(u)(U-u) \text{ and condition 1(iii)} \\
&= c_T^{-1} g_0(u) \begin{pmatrix} C_0^T C_0 k_{2,f_0,h} \\ 0 \end{pmatrix} + O(c_T^{-1}||B - B_0||) + O(c_T^{-1}||C - C_0|| + T^{-1}||D - D_0||) \\
&\quad + o(h^2 c_T^{-1}).
\end{aligned} \tag{C.8}$$

The variance of $W_T$ equals

$$\begin{aligned}
&\mathrm{Var}(W_T) \\
=\;& h^{r_2} \mathrm{Var}[X^* q_1(\bar{f}, Y - CDX) K_h(U-u)] \\
=\;& 4 g_0(u) \begin{pmatrix} \nu_0 C_0^T \Sigma(u) C_0 & 0 \\ 0 & \nu_2 \otimes C_0^T \Sigma(u) C_0 \end{pmatrix} + O(h^2 + ||C - C_0|| + ||B - B_0|| + ||D - D_0||) \\
\equiv\;& W + o(1),
\end{aligned}$$

where

$$W = 4 g_0(u) \begin{pmatrix} \nu_0 C_0^T \Sigma(u) C_0 & 0 \\ 0 & \nu_2 \otimes C_0^T \Sigma(u) C_0 \end{pmatrix}. \tag{C.9}$$

Under Condition 1, it can be verified that the central limit theorem (Hamilton, 1994, p.194) holds for $\{W_T\}$, i.e.,

$$W_T - E(W_T) \xrightarrow{D} N(0, W) \tag{C.10}$$

Therefore,

$$F^{-1}W_T - F^{-1}EW_T \xrightarrow{D} N(0, F^{-1}WF^{-1}), \tag{C.11}$$

or,

$$\hat{A}^* - F^{-1}EW_T \xrightarrow{D} N(0, F^{-1}WF^{-1}), \tag{C.12}$$

or,

$$c_T^{-1}\begin{pmatrix} [A_0 - f_0(u)] \\ h\{\text{vec}[A_1 - f_0'(u)]\} \end{pmatrix} - \frac{1}{2}c_T^{-1}\begin{pmatrix} (C_0^TC_0)^{-1} & 0 \\ 0 & (k_2 \otimes C_0^TC_0)^{-1} \end{pmatrix}\begin{pmatrix} C_0^TC_0k_{2,f_0,h} \\ 0 \end{pmatrix}$$
$$+ o_P(c_T^{-1}h^2) + o_P(1) \xrightarrow{D} N(0, F^{-1}WF^{-1}).$$

This completes the proof of Theorem 3.1. □

**Proof of Theorem 3.2:**

We adopt the same notations as defined in the preceding proof.

Claim 1: (a) Assume $B_0$, $C_0$ and $D_0$ are known, we have

$$\begin{aligned} & \sup_{u\in\mathcal{D}} || \begin{pmatrix} \hat{f}(u) - f_0(u) \\ h\text{vec}[\hat{f}'(u) - f_0'(u)] \end{pmatrix} - c_TF^{-1}W_T|| \\ = \;& O_P(c_Th^2 + c_T\sqrt{\frac{\ln T}{Th^{r_2}}}). \end{aligned} \tag{C.13}$$

(b) For general $B$, $C$ and $D$, we have

$$\begin{aligned} & \sup_{u\in\mathcal{D}} || \begin{pmatrix} \hat{f}(u; B^*, C^*, D) - f_0(u) \\ h\text{vec}[\hat{f}'(u; B^*, C^*, D) - f_0'(u)] \end{pmatrix} || \\ = \;& O_P(h^2 + c_T||B^* - B_0^*|| + c_T||C^* - C_0^*|| + c_T||D - D_0|| + c_T\sqrt{\frac{\ln T}{Th^{r_2}}}). \end{aligned} \tag{C.14}$$

Proof: First of all, by using Theorem 2 of Masry (1996) and the fact that (Li, 2000) $EF_T = -F + o(1)$, we have

$$\begin{aligned} F_T(u) &= EF_T(u) + O_P(\sqrt{\frac{\ln T}{Th^{r_2}}}) \\ &= -F(u) + O_P(h^2 + ||B^* - B_0^*|| + ||C^* - C_0^*|| + ||D - D_0|| + \sqrt{\frac{\ln T}{Th^{r_2}}}) \end{aligned}$$

uniformly in $u \in \mathcal{D}$, where $F_T(u)$ and $F(u)$ are defined in (C.4) and (C.5) except that we here stress the dependence on $u$. There are two cases.

(a) For known $B_0$ and $C_0$, we have

$$\begin{aligned} 0 &= W_T + D_T A^* \\ &= W_T - D[1 + O_P(h^2 + \sqrt{\frac{\ln T}{Th^{r_2}}})]A^* \end{aligned}$$

implying that

$$A^* = D^{-1}W_T + O_P[h^2 + \sqrt{\frac{\ln T}{Th^{r_2}}})]. \tag{C.15}$$

Multiplying $c_T$ on both sides of (C.15), we obtain the result in (C.13).

(b) For unknown $B_0^*$ and $C_0^*$, via (C.8) and Theorem 2 in Masry (1996), we have

$$\begin{aligned} c_T W_T &= c_T[(W_T - EW_T) + EW_T] \\ &= h^2 + c_T\sqrt{\ln T/(Th^{r_2})}, \end{aligned}$$

hence (C.14).

Claim 2:

$$\begin{aligned} & \hat{f}(u_0; \hat{B}^*, \hat{C}^*, \hat{D}) - f_0(u_0) \\ = \; & (C_0^T C_0)^{-1} \frac{T^{-1}\sum_i C_0^T\{Y_i - C_0[D_0 X_i + f_0(u_0) + f_0'(u_0)(U_i - u_0)]\}K_h(U_i - u_0)}{g(u_0)} \\ & - f_0'(u_0) E(X_2^T \otimes I_{r_2} | U = u_0)\mathrm{vec}(\hat{B}^* - B_0^*) \\ & - (C_0^T C_0)^{-1} C_0^T \begin{pmatrix} 0_{r_1 \times r_1(m-r_1)} \\ \{E(R^T|U = u_0) + f_0^T(u_0)\} \otimes I_{m-r_1} \end{pmatrix} \mathrm{vec}(\hat{C}^* - C_0^*) \\ & - (C_0^T C_0)^{-1} C_0^T E(X^T \otimes C_0 | U = u_0)\mathrm{vec}(\hat{D} - D_0) \\ & + o_p(T^{-1/2}), \end{aligned} \tag{C.16}$$

where we recall that $R = D_0 X$.

Proof: Let $a = f_0(u_0)$ and $b = h\mathrm{vec}[f_0'(u_0)]$. The local linear estimates $\hat{a} = \hat{f}_0(u_0; \hat{B}^*, \hat{C}^*, \hat{D})$ and $\hat{b} = h\mathrm{vec}[\hat{f}_0'(u_0; \hat{B}^*, \hat{C}^*, \hat{D})]$ solve the following equation

$$0 = \frac{1}{T}\sum_i \begin{pmatrix} I_{r_1} \\ (\frac{\hat{U}_i - u_0}{h}) \otimes I_{r_1} \end{pmatrix} \hat{C}^T(Y_i - \hat{C}\{\hat{D}X_i + I_{r_1}\hat{a} + [(\frac{\hat{U}_i - u_0}{h})^T \otimes I_{r_1}]\hat{b}\})K_h(\hat{U}_i - u_0).$$

Via Taylor expansion, we obtain

$$0 = \frac{1}{T}\sum_i \begin{pmatrix} I_{r_1} \\ (\frac{U_i - u_0}{h}) \otimes I_{r_1} \end{pmatrix} C_0^T(Y_i - C_0\{R_i + I_{r_1}a + [(\frac{U_i - u_0}{h})^T \otimes I_{r_1}]b\})K_h(U_i - u_0)$$

$$
\begin{aligned}
&-\frac{1}{T}\sum_i \begin{pmatrix} I_{r_1} \\ (\frac{U_i-u_0}{h})\otimes I_{r_1} \end{pmatrix} C_0^T C_0 [I_{r_1}, (\frac{U_i-u_0}{h})^T \otimes I_{r_1}] \begin{pmatrix} \hat{a}-a \\ \hat{b}-b \end{pmatrix} K_h(U_i-u_0) \\
&-\frac{1}{T}\sum_i \begin{pmatrix} I_{r_1} \\ (\frac{U_i-u_0}{h})\otimes I_{r_1} \end{pmatrix} C_0^T C_0 f_0'(u_0)(X_{2i}^T \otimes I_{r_2})\mathrm{vec}(\hat{B}^*-B_0^*)K_h(U_i-u_0) \\
&+\frac{1}{T}\sum_i \begin{pmatrix} I_{r_1} \\ (\frac{U_i-u_0}{h})\otimes I_{r_1} \end{pmatrix} C_0^T (Y_i - C_0\{R_i + I_{r_1}a + [(\frac{U_i-u_0}{h})^T \otimes I_{r_1}]b)\})K_h'^T(U_i-u_0) \\
&\qquad \times [(X_{2i}-x_0)^T \otimes I_{r_2}]\mathrm{vec}(\hat{B}^*-B_0^*) \\
&-\frac{1}{T}\sum_i \begin{pmatrix} I_{r_1} \\ (\frac{U_i-u_0}{h})\otimes I_{r_1} \end{pmatrix} C_0^T \begin{pmatrix} 0_{r_1\times r_1(m-r_1)} \\ \{f_0^T(u_0)+R_i^T\}\otimes I_{m-r_1} \end{pmatrix} \mathrm{vec}(\hat{C}^*-C_0^*)K_h(U_i-u_0) \\
&+\frac{1}{T}\sum_i \begin{pmatrix} I_{r_1} \\ (\frac{U_i-u_0}{h})\otimes I_{r_1} \end{pmatrix} [(Y_i - C_0\{R_i + I_{r_1}a + [(\frac{U_i-u_0}{h})^T \otimes I_{r_1}]b\})^T \otimes I_{r_1}] \\
&\qquad \times K_h(U_i-u_0)\mathrm{vec}(\hat{C^*}^T - C_0^{*T}) \\
&-\frac{1}{T}\sum_i \begin{pmatrix} I_{r_1} \\ (\frac{U_i-u_0}{h})\otimes I_{r_1} \end{pmatrix} C_0^T (X_i^T \otimes C_0)\mathrm{vec}(\hat{D}-D_0)K_h(U_i-u_0) \\
&+O_p(\frac{1}{T}) + O_P(\frac{h^2}{\sqrt{T}} + \frac{c_T||\hat{B}^*-B_0^*||}{\sqrt{T}} + \frac{c_T||\hat{C}^*-C_0^*||}{\sqrt{T}} + \frac{c_T\sqrt{\ln T/(Th^{r_2})}}{\sqrt{T}}),
\end{aligned}
$$

where the first remainder term comes from the second order expansion of the parametric part in the Taylor expansion, while the second remainder term comes from the cross product of the parametric part and nonparametric part of the second order expansion in the Taylor expansion. The sum of remainder terms is $o_P(1/\sqrt{T})$ under the conditions $h \to 0, Th^{r_2} \to \infty$ and $\ln T/(Th^{r_2}) \to 0$. Moreover, it follows from Lemma 1 of Li and Chan (2001) with $p = 0$ and $K_h'$ replacing $K_h$ that (recall $X_t$ is partitioned as $\begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix}$ with $X_{1t}$ being $r_2$ dimensional.)

$$
\begin{aligned}
& T^{-1}\sum_t (Y_t - C_0a - C_0DX_t)K_h'(U_t-u_0)[(X_{2t}-x_0)^T \otimes I_{r_2}]\mathrm{vec}(\hat{B}^*-B_0^*) \\
= \;& O_P(h/\sqrt{T} + \sqrt{\ln T}/(T\sqrt{h^{r_2}})) = o_P(1/\sqrt{T}). \qquad \text{(C.17)}
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
& T^{-1}\sum_t [(Y_t - C_0a - C_0DX_t)^T \otimes I_{r_1}]K_h(U_i-u_0)\mathrm{vec}(\hat{C^*}^T - C_0^{*T}) \\
= \;& O_P(h/\sqrt{T} + \sqrt{\ln T}/(T\sqrt{h^{r_2}})) = o_P(1/\sqrt{T})
\end{aligned}
$$

under the conditions that $h \to 0$ and $\ln T/(Th^{r_2}) \to 0$. Hence,

$$\begin{aligned}
0 &= T^{-1}\sum_i C_0^T\{Y_i - C_0[R_i + f_0(u_0) + f_0'(u_0)(U_i - u_0)]\}K_h(U_i - u_0) - g(u_0)C_0^TC_0(\hat{a} - a_0) \\
&- g(u_0)C_0^TC_0f_0'(u_0)E[X_2^T \otimes I_{r_2}|U = u_0]\text{vec}(\hat{B}^* - B^*) \\
&- g(u_0)C_0^T\begin{pmatrix} 0_{r_1 \times r_1(m-r_1)} \\ \{f_0^T(u_0) + E(R^T|U = u_0)\} \otimes I_{m-r_1} \end{pmatrix}\text{vec}(\hat{C}^* - C_0^*) \\
&- g(u_0)C_0^TE(X^T \otimes C_0|U = u_0)\text{vec}(\hat{D} - D_0) + o_p(T^{-1/2})
\end{aligned}$$

Because $\hat{a} = \hat{f}(u_0; h, \hat{B}^*, \hat{C}^*)$ and $a_0 = f_0(u_0)$, the above equation implies that

$$\begin{aligned}
&\hat{f}(u_0; h, \hat{B}^*, \hat{C}^*) - f_0(u_0) \\
= \;& (C_0^TC_0)^{-1}\frac{T^{-1}\sum_i C_0^T\{Y_i - C_0[f_0(u_0) + f_0'(u_0)(U_i - u_0)]\}K_h(U_i - u_0)}{g(u_0)} \\
&-f_0'(u_0)E[X_2^T \otimes I_{r_2}|U = u_0]\text{vec}(\hat{B}^* - B_0^*) \\
&-(C_0^TC_0)^{-1}C_0^T\begin{pmatrix} 0_{r_1 \times r_1(m-r_1)} \\ \{f_0^T(u_0) + E(R^T|U = u_0)\} \otimes I_{m-r_1} \end{pmatrix}\text{vec}(\hat{C}^* - C_0^*) \\
&-(C_0^TC_0)^{-1}C_0^TE(X^T \otimes C_0|U = u_0)\text{vec}(\hat{D} - D_0) + o_p(T^{-1/2}).
\end{aligned}$$

This completes the proof of Claim 2. □

Claim 3:

$$\begin{aligned}
&\hat{f}(\hat{B}X_i; \hat{B}^*, \hat{C}^*, \hat{D}) - f_0(B_0X_i) \\
= \;& f_0'(B_0X_i)(X_{2i}^T \otimes I_{r_2})\text{vec}(\hat{B}^* - B_0^*) + \hat{f}(B_0X_i; \hat{B}^*, \hat{C}^*, \hat{D}) - f_0(B_0X_i) \\
&+o_P(T^{-1/2}).
\end{aligned} \tag{C.18}$$

Proof: After some algebra, it can be shown that

$$\begin{aligned}
&\hat{f}(\hat{B}X_i; \hat{B}^*, \hat{C}^*, \hat{D}) - f_0(B_0X_i) \\
= \;& \hat{f}(\hat{B}X_i; \hat{B}^*, \hat{C}^*, \hat{D}) - \hat{f}(B_0X_i; \hat{B}^*, \hat{C}^*, \hat{D}) + \hat{f}(B_0X_i; \hat{B}^*, \hat{C}^*, \hat{D}) - f_0(B_0X_i) \\
= \;& \hat{f}'(B_0X_i; \hat{B}^*, \hat{C}^*, \hat{D})(\hat{B}^* - B_0^*)X_{2i} + \hat{f}(B_0X_i; \hat{B}^*, \hat{C}^*, \hat{D}) - f_0(B_0X_i) + o_p(T^{-1/2}) \\
= \;& f_0'(B_0X_i)(X_{2i}^T \otimes I_{r_2})\text{vec}(\hat{B}^* - B_0^*) + \hat{f}(B_0X_i; \hat{B}^*, \hat{C}^*, \hat{D}) - f_0(B_0X_i) + o_P(T^{-1/2}).
\end{aligned}$$

□

To prove Theorem 3.2, recall that $(\hat{B}^*, \hat{C}^*, \hat{D})$ maximizes the objective function defined by

$$-\sum_t\sum_i ||Y_i - C[DX_t + \hat{f}(BX_t; B^*, C^*, D) + \hat{f}'(BX_t; B^*, C^*, D)B(X_i - X_t)]||^2K_h[B(X_i - X_t)],$$

where $D$ and $B$ are subject to the constraints that $DB^T = 0$. The constrained optimization problem can be turned into an unconstrained optimization via the Lagrangian approach, i.e., the objective function becomes

$$-\sum_t \sum_i ||Y_i - C[DX_t + \hat{f}(BX_t; B^*, C^*, D) + \hat{f}'(BX_t; B^*, C^*, D)B(X_i - X_t)]||^2 K_h[B(X_i - X_t)] + \lambda^T DB^T, \tag{C.19}$$

where $\lambda$ is the Lagrange multiplier; see Aitchison and Silvey (1958). For simplicity, write $\theta = (B^*, C^*, D)$ and $h(\theta) = DB^T$. Let

$$\hat{\Lambda}_{i,t} = \begin{pmatrix} (X_{2t} \otimes I_{r_2})\hat{f}'^T(\hat{B}X_t; \hat{B}^*, \hat{C}^*, \hat{D})\hat{C}^T + [(X_{2i} - X_{2t}) \otimes \hat{f}'^T(\hat{B}X_t; \hat{B}^*, \hat{C}^*, \hat{D})]\hat{C}^T \\ 0_{r_1(m-r_1)\times r_1}, -[\hat{D}X_t + \hat{f}(\hat{B}X_t; \hat{B}^*, \hat{C}^*, \hat{D}) + \hat{f}'(\hat{B}X_t; \hat{B}^*, \hat{C}^*, \hat{D})\hat{B}(X_i - X_t)] \otimes I_{m-r_1} \\ X_t \otimes \hat{C}^T \end{pmatrix},$$

$$\Lambda_{i,t} = \begin{pmatrix} (X_{2t} \otimes I_{r_2}) f_0'^T(B_0X_t)C_0^T + [(X_{2i} - X_{2t}) \otimes f_0'^T(B_0X_t)]C_0^T \\ 0_{r_1(m-r_1)\times r_1}, [R_t + f_0(B_0X_t) + f_0'(B_0X_t)B_0(X_i - X_t)] \otimes I_{m-r_1} \\ X_t \otimes C_0^T \end{pmatrix},$$

$$\Lambda_t = \begin{pmatrix} (X_{2t} \otimes I_{r_2}) f_0'^T(B_0X_t)C_0^T \\ 0_{r_1(m-r_1)\times r_1}, [R_t + f_0(B_0X_t)] \otimes I_{m-r_1} \\ X_t \otimes C_0^T \end{pmatrix}.$$

Taking the first derivative of the objective function with respect to $\theta$, and via Taylor expansion we have

$$\begin{aligned} 0 &= \frac{1}{\sqrt{T^3}} \sum_{t,i} \hat{\Lambda}_{i,t}\{Y_i - \hat{C}[\hat{D}X_t + \hat{f}(\hat{B}X_t; \hat{B}^*, \hat{C}^*) + \hat{f}'(\hat{B}X_t; \hat{B}^*, \hat{C}^*)\hat{B}(X_i - X_t)]\}K_h[\hat{B}(X_i - X_t)] \\ &\qquad + h'(\hat{\theta})\hat{\lambda} + o_P(1/\sqrt{T}) \\ &= \frac{1}{\sqrt{T^3}} \sum_{t,i} \Lambda_{i,t}\{Y_i - C_0[D_0X_t + f_0(B_0X_t) + f_0'(B_0X_t)B_0(X_i - X_t)]\}K_h[B_0(X_i - X_t)] \\ &\quad - \frac{1}{\sqrt{T^3}} \sum_{t,i} \Lambda_{i,t}C_0[\hat{f}(\hat{B}X_t; \hat{B}^*, \hat{C}^*, \hat{D}) - f_0(B_0X_t)]K_h[B_0(X_i - X_t)] \\ &\quad - \frac{1}{\sqrt{T^3}} \sum_{t,i} \Lambda_{i,t} \begin{pmatrix} 0_{r_1\times r_1(m-r_1)} \\ [D_0X_t + f_0(B_0X_t) + f_0'(B_0X_t)B_0(X_i - X_t)]^T \otimes I_{m-r_1} \end{pmatrix} \mathrm{vec}(\hat{C}^* - C_0^*) \\ &\qquad \times K_h[B_0(X_i - X_t)] \\ &\quad - \frac{1}{\sqrt{T^3}} \sum_{t,i} \Lambda_{i,t}(X_t^T \otimes C_0)\mathrm{vec}(\hat{D} - D) \times K_h[B_0(X_i - X_t)] \end{aligned}$$

$$+h'(\hat{\theta})\hat{\lambda} + o_P(1). \tag{C.20}$$

On the other hand, the first derivative of the objective function w.r.t. $\lambda$ yields the constraints $h(\hat{\theta}) = 0$. It follows from Claims 1-3 that

$$\begin{aligned}
0 \;=\; & \frac{1}{\sqrt{T^3}} \sum_{t,i} \Lambda_{i,t}\{Y_i - C_0[f_0(U_t) + f_0'(U_t)B_0(X_i - X_t)]\}K_h[B_0(X_i - X_t)] \\
& - \frac{1}{\sqrt{T^3}} \sum_{t,i} \Lambda_{i,t} C_0 f_0'(U_t)(X_{2t}^T \otimes I_{r_2})\mathrm{vec}(\hat{B}^* - B_0^*)K_h[B_0(X_i - X_t)] \\
& - \frac{1}{\sqrt{T^3}} \sum_{t,i} \Lambda_{i,t} C_0((C_0^T C_0)^{-1} C_0^T [g(U_t)T]^{-1} \sum_j \{Y_j - C_0[D_0 X_j + f_0(U_j) \\
& \qquad + f_0'(U_j)B_0(X_j - X_t)]\}K_h[B_0(X_j - X_i)])K_h[B_0(X_i - X_t)] \\
& + \frac{1}{\sqrt{T^3}} \sum_{t,i} \Lambda_{i,t} C_0 f_0'(U_t) E(X_2^T \otimes I_{r_2}|U_t)\mathrm{vec}(\hat{B}^* - B_0^*)K_h[B_0(X_i - X_t)] \\
& + \frac{1}{\sqrt{T^3}} \sum_{t,i} \Lambda_{i,t} C_0 (C_0^T C_0)^{-1} C_0^T \begin{pmatrix} 0_{r_1 \times r_1(m-r_1)} \\ \{f_0^T(U_t) + E(R^T|U_t)\} \otimes I_{m-r_1} \end{pmatrix} \mathrm{vec}(\hat{C}^* - C_0^*)K_h[B_0(X_i - X_t)] \\
& - \frac{1}{\sqrt{T^3}} \sum_{t,i} \Lambda_{i,t} \begin{pmatrix} 0_{r_1 \times r_1(m-r_1)} \\ [R_t + f_0(U_t) + f_0'(U_t)B_0(X_i - X_t)]^T \otimes I_{m-r_1} \end{pmatrix} \mathrm{vec}(\hat{C}^* - C_0^*) \\
& \qquad \times K_h[B_0(X_i - X_t)] \\
& + \frac{1}{\sqrt{T^3}} \sum_{t,i} \Lambda_{i,t} C_0 (C_0^T C_0)^{-1} C_0^T E(X^T \otimes C_0|U_t)\mathrm{vec}(\hat{D} - D_0) \\
& - \frac{1}{\sqrt{T^3}} \sum_{t,i} \Lambda_{i,t}\{X_i^T \otimes C_0\}\mathrm{vec}(\hat{D} - D_0) + h'(\hat{\theta})\hat{\lambda} + o_P(1).
\end{aligned} \tag{C.21}$$

Upon conditioning each summand given $U_t$ and using the technique of the proof of (48) in Carroll et al. (1995), it can be shown that (C.21) becomes

$$\begin{aligned}
0 \;=\; & \frac{1}{\sqrt{T}} \sum_t g(U_t)[\Lambda_t - E(\Lambda|U_t)C_0(C_0^T C_0)^{-1} C_0^T]\epsilon_t \\
& - \sqrt{T}\{E[g(U)\Lambda C_0 f_0'(U)(X_2^T \otimes I_{r_2})] - E[g(U)\Lambda C_0 f_0'(U) E(X_2^T \otimes I_{r_2}|U)]\} \\
& \qquad \times \mathrm{vec}(\hat{B}^* - B_0^*) \\
& - \sqrt{T}\left\{ E\left[ g(U)\Lambda \begin{pmatrix} 0_{r_1 \times r_1(m-r_1)} \\ \{R^T + f_0^T(U)\} \otimes I_{m-r_1} \end{pmatrix} \right] - E\left[ g(U)\Lambda C_0 (C_0^T C_0)^{-1} C_0^T \begin{pmatrix} 0_{r_1 \times r_1(m-r_1)} \\ (E(R^T|U) + f_0^T(U)) \otimes I_{m-r_1} \end{pmatrix} \right] \right\} \\
& \times \mathrm{vec}(\hat{C}^* - C_0^*) \\
& - \sqrt{T}\{E[g(U)\Lambda X^T \otimes C_0)] - E[g(U)\Lambda C_0 (C_0^T C_0)^{-1} C_0^T E(X^T \otimes C|U)]\} \\
& \qquad \times \mathrm{vec}(\hat{D} - D_0) + h'(\theta_0)\sqrt{T}\hat{\lambda} + o_P(1).
\end{aligned}$$

Also, the constraints $h(\hat{\theta}) = 0$ can be linearized as $(h'(\theta_0))^T\sqrt{T}(\hat{\theta} - \theta_0) + o_P(1) = 0$. Write $H = (h'(\theta_0))^T$, and it can be shown by routine calculus that $H = [(I_{r_2} \otimes D_{20})K_{r_2,n-r_2}, 0_{r_2 \times r_1,(m-r_1)\times r_1}, B_0 \otimes$

$I_{r_1}]$ where $D_0 = [D_{10}, D_{20}]$ is partitioned such that $D_{10}$ is of dimension $r_1 \times r_2$. Hence

$$\begin{pmatrix} Q & -H^T \\ -H & 0 \end{pmatrix} \begin{pmatrix} \hat{\theta} - \theta_0 \\ \hat{\lambda} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{T}} \sum_t g(U_t)[\Lambda_t - E(\Lambda|U_t)C_0(C_0^T C_0)^{-1}C_0^T]\epsilon_t \\ 0 \end{pmatrix} + o_P(1), \tag{C.22}$$

from which (17) can be readily derived. □

## References


Aitchison, J. and Silvey, S. D. (1958), "Maximum-likelihood estimation of parameters subject to restraints," *Annals of Statistics*, 29, 813–828.

Cai, Z., Fan, J., and Yao, Q. (2000), "Functional-coefficient regression models for nonlinear time series," *Journal of the American Statistical Association*, 95, 941–956.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1995), "Generalized partially linear single-index models," *Discussion Paper #9506*, Institute of Statistics, Catholic University of Louvain, Louvain-la-Neuve, Belgium.

— (1997), "Generalized partially linear single-index models," *Journal of the American Statistical Association*, 92, 477–489.

Chan, K. S. and Li, M. C. (2002), "Discussion of a paper by Xia, Y. et al." *Journal of the Royal Statistical Society, B*, 64, 395–396.

Eubank, R. L. (1988), *Spline smoothing and nonparametric regression*, New York: Marcek Dekker.

Fan, J. and Gijbels, I. (1996), *Local polynomial modeling and its applications*, London: Chapman and Hall.

Fan, J. and Zhang, W. Y. (1999), "Statistical estimation in varying coefficient models," *Annals of Statistics*, 27, 1491–1518.

Friedman, J. H. and Stuetzle, W. (1981), "Projection pursuit regression," *Journal of the American Statistical Association*, 76, 817–823.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton, New Jersey: Princeton University Press.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Li, K. C. (1992), "On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma," *Journal of the American Statistical Association*, 87, 1025–1039.

Li, M.-C. (2000), "Multivariate non-linear time series modeling," *Unpublished PhD Thesis*, The University of Iowa.

Li, M.-C. and Chan, K. S. (2001), "Semiparametric reduced-rank regression," *Technical Report*, The University of Iowa, Department of Statistics and Actuarial Science.

Masry, E. (1996), "Multivariate local polynomial regression for time series: uniform strong consistency and rates," *Journal of Time Series Analysis*, 17, 571–599.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992), *Numerical Recipes in C: The Art of Scientific Computing*, second ed. Cambridge University Press, Cambridge.

Reinsel, G. C. and Velu, R. (1998), *Multivariate-Reduced rank regression*, New York: Springer-Verlag.

Rosenblatt, M. (1956), "A central limit theorem and strong mixing conditions," *Proc. Nat. Acad. Sci.*, 4, 43–47.

Ruppert, D. and Wand, M. P. (1994), "Multivariate locally weighted least squares regression," *Annals of Statistics*, 22, 1346–1370.

Turkington, D. A. (2002), *Matrix calculus and zero-one matrices*, Cambridge: Cambridge University Press.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, New York: Springer.

Wand, M. P. (2002), "Vector differential calculus in statistics," *The American Statistician*, 56, 55–62.

Xia, Y., Tong, H., and Li, W. K. (1999), "On extended partially linear single-index models," *Biometrika*, 86, 831–842.

Xia, Y., Tong, H., Li, W. K., and Zhu, L. (2002), "An adaptive estimation of dimension reduction space (with Discussion)," *Journal of the Royal Statistical Society, B*, 64, 363 – 410.

Table 1: Frequency for estimating $r_1$=rank(C) and $r_2$=rank(B) for the simulated plsparr model.

| | T=50 | | | | T=100 | | | | T=200 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $r_1/r_2$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 15 | 70 | 0 | 0 | 0 | 95 | 0 | 0 | 0 | 99 | 0 | 0 |
| 3 | 5 | 10 | 0 | 0 | 0 | 5 | 0 | 0 | 0 | 1 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 2: Objective function for selecting $r_1$=rank(C) and $r_2$=rank(B) of the PLRR model for the Hong Kong pollution data. The objective function is defined by (6) divided by $\sum_t \|Y_t\|^2$.

| | 16-dimensional covariate | | | | 8-dimensional covariate | | | |
|---|---|---|---|---|---|---|---|---|
| $r_1/r_2$ | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | .614 (.1) | .610 (.747) | .603 (.583) | .616 (.492) | .616 (.05) | .618 (.747) | .615 (1.17) | .626 (.820) |
| 2 | .504 (.2) | .505 (.374) | .501 (.583) | .525 (.492) | .508 (.2) | .495 (.374) | .515 (.583) | .523 (.492) |
| 3 | .457 (.4) | .456 (.623) | .464 (.437) | .486 (.655) | .469 (.4) | .472 (.498) | .475 (.437) | .518 (.820) |
| 4 | .443 (.3) | .444 (.374) | .460 (.437) | .479 (.492) | .441 (.3) | .457 (.498) | NA | NA |

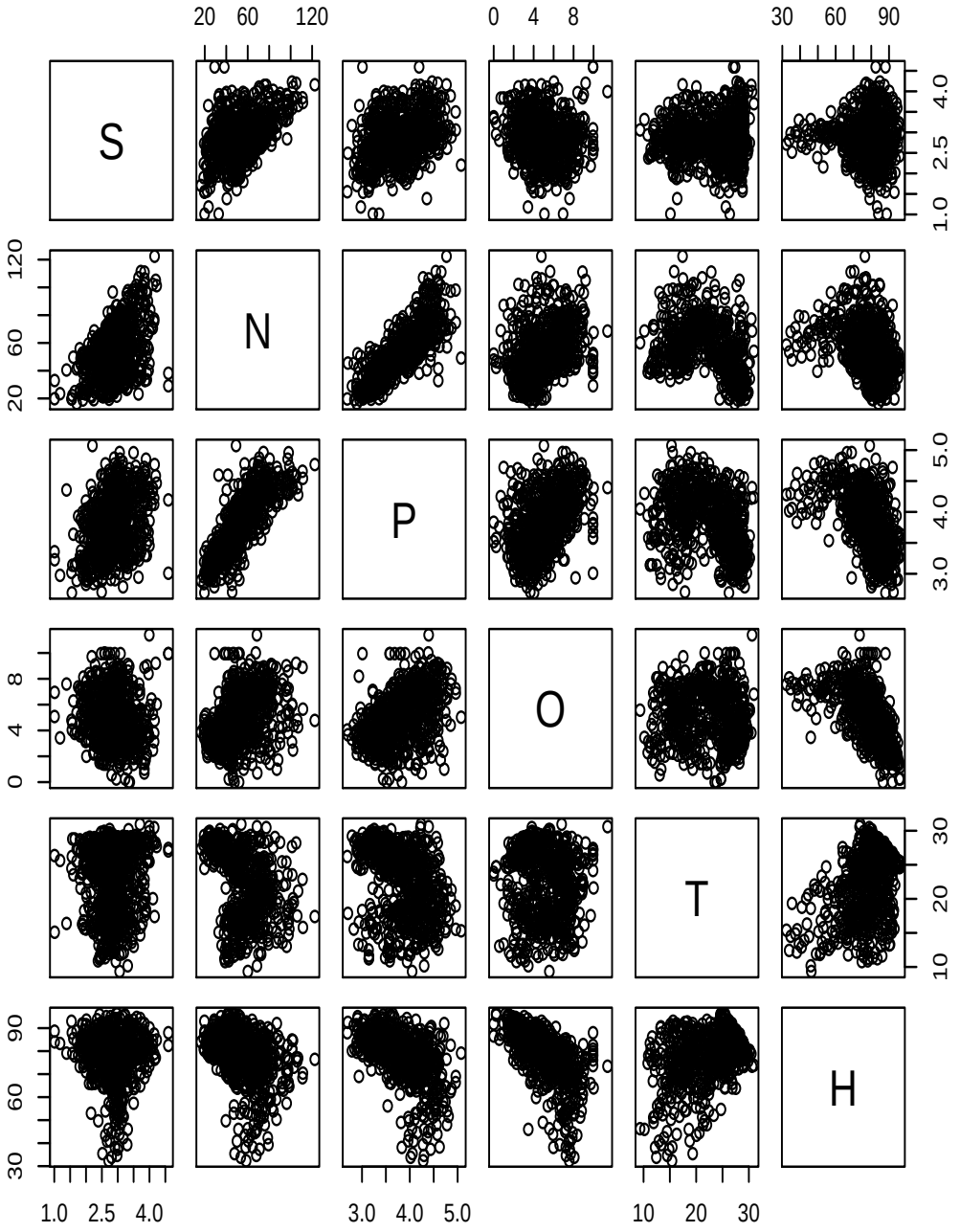


Figure 1: Scatter diagrams of the pollutant and weather variables.

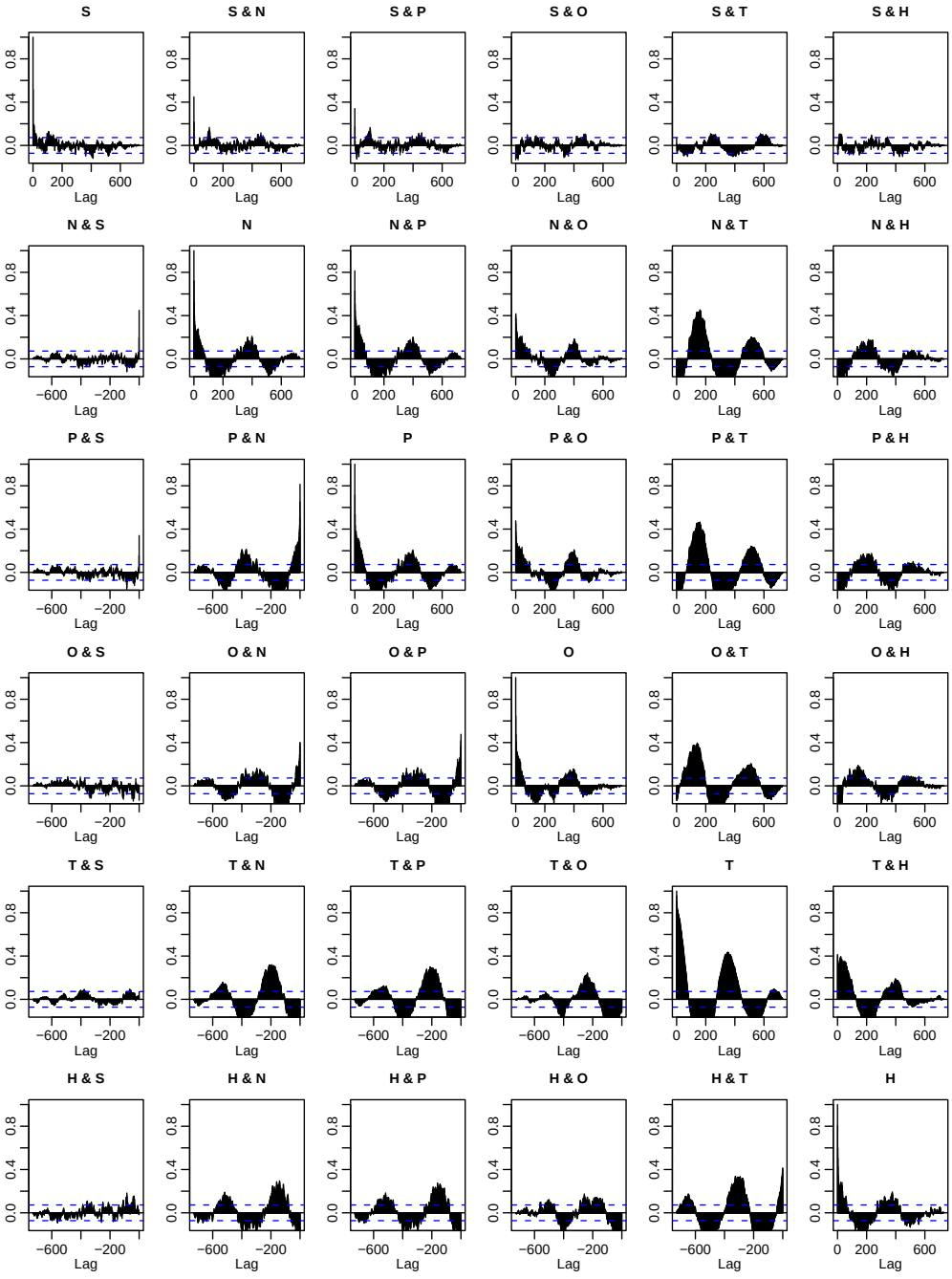


Figure 2: Cross- and Auto-correlation of the pollutant and weather variables.

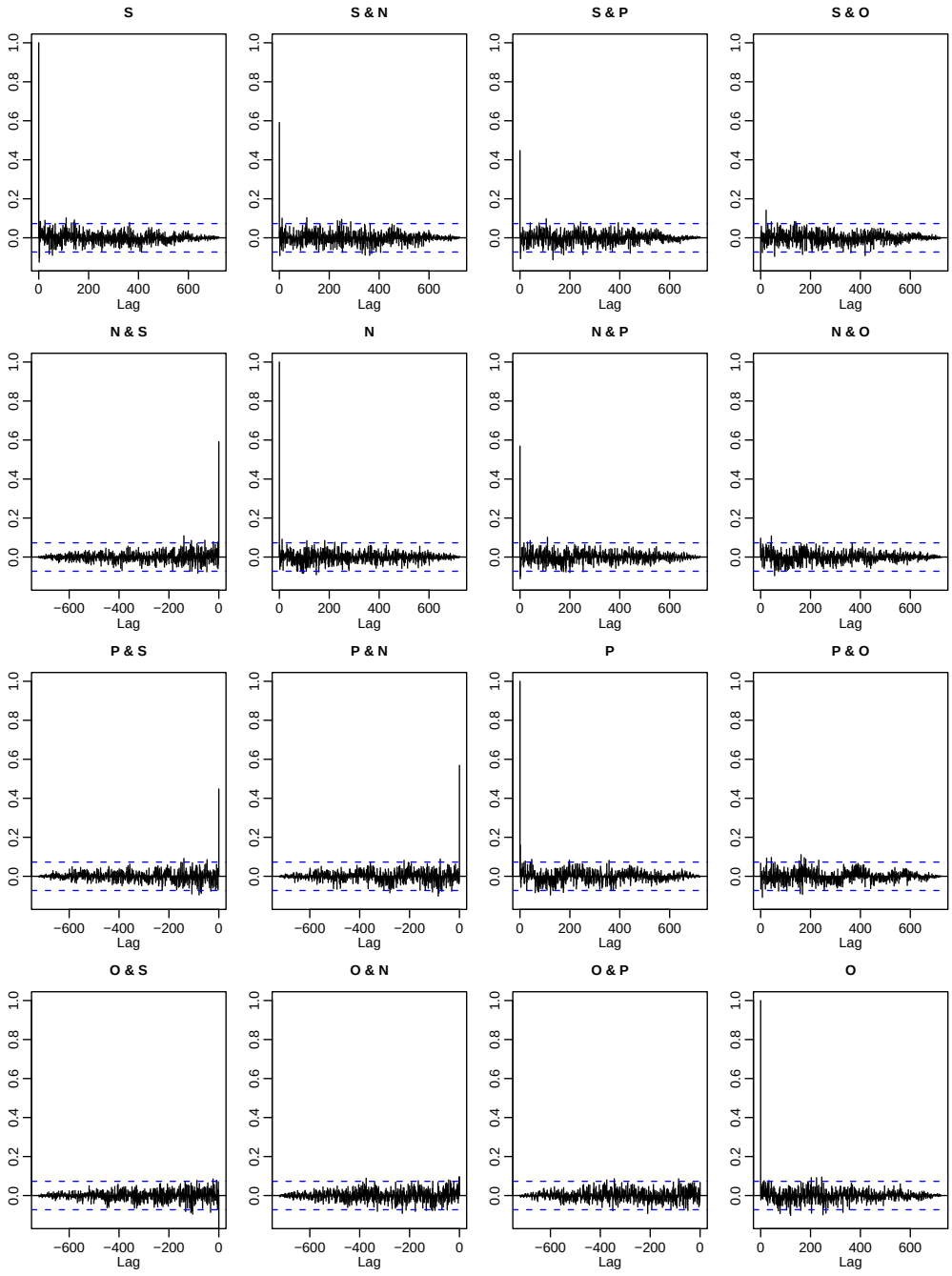


Figure 3: Cross- and Auto-correlation of the residuals from the PLRR model with $r_1 = 3$ and $r_2 = 1$ fitted to the Hong Kong pollution data.

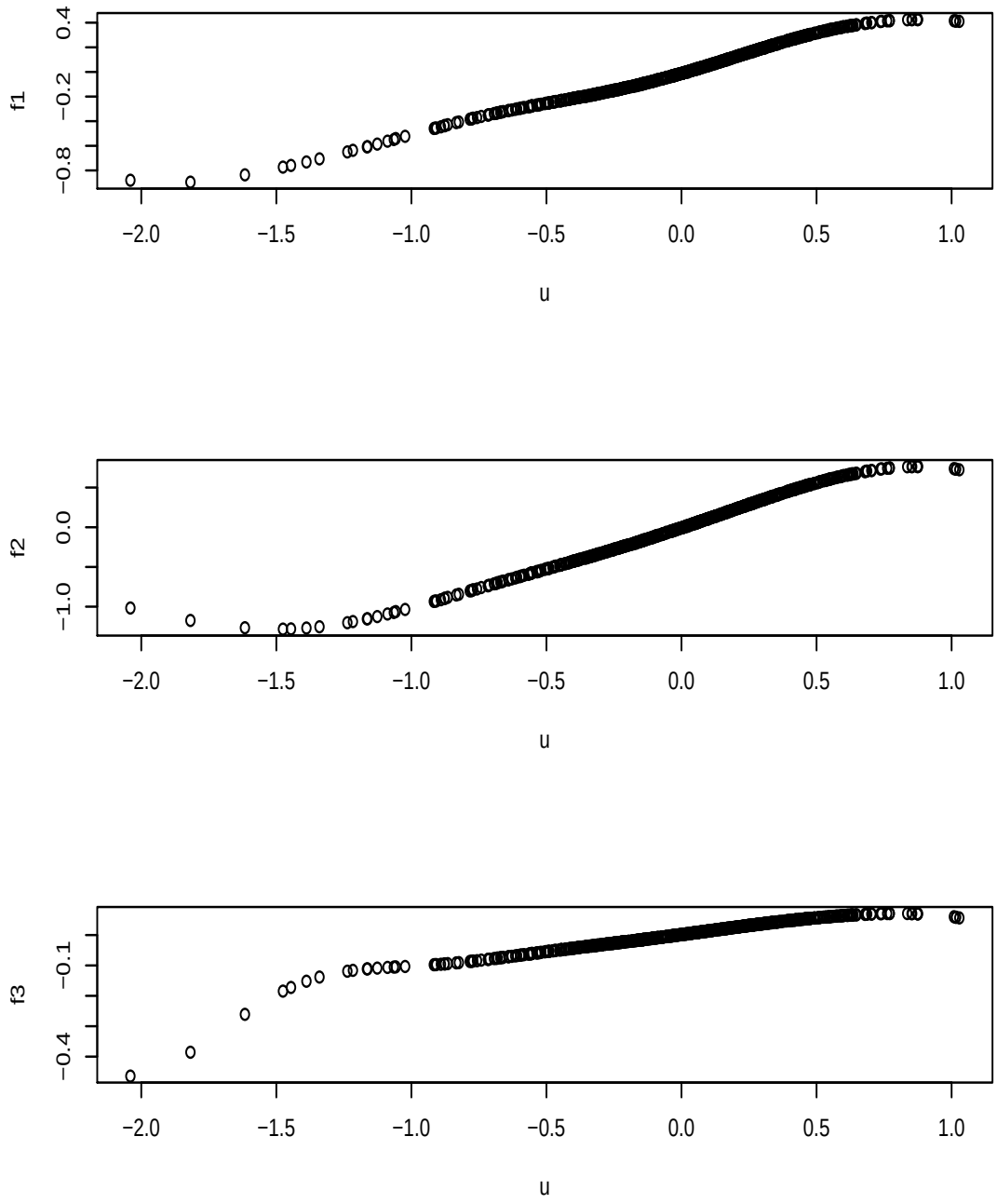


Figure 4: Scatter diagrams of $f(U)$ versus $U = BX_t$, for the PLRR model with $r_1 = 3$ and $r_2 = 1$ fitted to the Hong Kong pollution data.