Semiparametric Reduced-rank Regression

Ming-Chung Li and K. S. Chan *

October, 2001

Abstract

We have proposed a new semiparametric approach for inferring the structure of the regression function of a multivariate response Y on a multivariate covariate X. Our approach is partly motivated by the scientific needs for analyzing the common dynamic structure of nonlinear multivariate time series. The fundamental problem of developing efficient methods for exploring the functional form of the regression function of Y on X is very challenging owing to the curse of dimensionality which is further exacerbated by nonlinearity.

Here, we proposed a new model, the SemiPArametric Reduced-rank Regression (SPARR) model, for mitigating the curse of dimensionality by adapting the reduced-rank linear regression technique. The basic idea is to assume that the conditional mean function of Y given X depends on a small number of indices, each of which is a linear combination of X. Moreover, the link function linking the conditional mean of Y to the indices consists of linear combinations of a few generally nonlinear functions of the indices to be estimated nonparametrically. An alternative characterization of the SPARR model is that Y "loads" linearly on some nonlinear principal components which depend on some indices of X; the loading pattern may then facilitate an approach to classify the common regression structure among the components of Y. Moreover, the SPARR models provides a framework to study nonlinear co-integration relationships for multivariate time series.

We have proposed an estimation scheme for the SPARR model and derived, under suitable regularity conditions, the large-sample properties of the estimator for both the parametric and the nonparametric part of the model. We illustrated the new approach with the U.S. hog data and a modern panel of the Canada lynx data.

*Ming-Chung Li is Statistician, the EMMES Corporation, Rockville, MD 20850 (Email: mli@emmes.com). K. S. Chan is Professor, the University of Iowa, Department of Statistics & Actuarial Science, Iowa City, IA 52242, USA. (E-mail: kchan@stat.uiowa.edu). We thank Howell Tong and Nils Chr. Stenseth for their valuable comments.

Key Words: common structure, cross validation, local polynomial regression, nonlinear co-integration, panel of time series, strongly mixing.

1 Introduction

If the underlying data generating mechanism is well understood, one may carry out a substantive modeling of the data via some parametric model. However, more often than not, the underlying data mechanism is poorly understood. In the later case, it is pertinent to proceed with non-parametric modeling of the data. For high-dimensional data, directly adopting a nonparametric approach may suffer from the "*curse of dimensionality*". Thus many techniques have been proposed in the literature to overcome this problem, e.g., generalized additive model (Hastie and Tibshirani, 1990), projection pursuit regression (Friedman and Stuetzle, 1981), sliced inverse regression (Li, 1992), and partially linear single-index regression (Carroll et al., 1997; Xia et al., 1999).

Li and Chan (2001) adapt the reduced-rank linear regression technique (Reinsel and Velu, 1998) to develop parsimonious parametric nonlinear models for multivariate time series data. The reduced-rank method (Izenman, 1980; Reinsel and Velu, 1998) is a dimension reduction technique which "replaces" the covariate, say X, by some low-dimensional "indices" BX. Classical reduced-rank linear regression assumes a linear link function between the multivariate response and the indices. Here we generalize this basic idea by allowing for a general function relating the "indices" to the multivariate response. Specifically, we assume that the multivariate response depends linearly on a set of possibly nonlinear functionals of some indices; these functionals are modeled nonparametrically. This new approach generalizes the approach taken by Li and Chan (2001) where the functionals are specified as piecewise-linear. The multivariate response may consist of annual measurements taken over a number of sites; e.g., Stenseth et al. (1999) studied the spatial variations in the nonlinear dynamics of two panels of lynx data collected over Canada. By studying the pattern of how the multivariate response loads on the nonlinear functionals of the indices, the new model proposed here facilitates a semiparametric approach to study the spatial variations in the dynamics of a panel of time series. However, the new model may also be useful for modeling independent data.

The new model differs from the generalized partially linear single-index model (GPLSIM)

(Carroll et al., 1997) in that their model assumes single-index and univariate response. One advantage for the single-index model is that the predictor is one-dimensional so that it does not suffer from the curse of dimensionality. Here we consider vector indices (the dimension of which will be called the rank of the model). In other words, the new model is appropriate if it is thought that the multivariate responses depend on the covariates through some linear combinations of the covariates.

In §2, we describe the new semiparametric model and an estimation procedure of the model; Appendix A contains some useful formulas for implementing the estimation procedure. The limiting distributions of the parametric and the nonparametric parts of the model are derived in §3. Some simulation studies on the empirical performance of the proposed estimation method are reported in §4. Illustrations of the new model with the Canada lynx data and the U.S. hog data are given §5. All proofs are deferred to Appendices B and C.

2 The Model and an Estimation Procedure

Let Y_t and X_t be m and n-dimensional random vectors, both of which are assumed to have been standardized. The SemiPArametric Reduced-rank Regression (SPARR) model is defined as follows:

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

where C is an unknown $m \times r_1$ matrix and B an unknown $r_2 \times n$ matrix; we refer to r_1 and r_2 as the ranks of the model. The unknown (link) function f maps from R^{r_2} to R^{r_1}. Note that the covariate X_t may contain lagged values of Y_t.

The SPARR model is a rather general but fairly interpretable model. Let $Z_t = f(BX_t)$ which can be interpreted as some nonlinear factor process. Then (1) is equivalent to $Y_t = CZ_t + \epsilon_t$, which bears resemblance to a principal-component regression model. The matrix C is not unique because the model is unaltered if we post-multiply C by some non-singular matrix and pre-multiply Z_t by the inverse of the same matrix. A simple pattern may, however, emerge upon suitably rotating C, which may facilitate the interpretation of the (nonlinear) principal-component process Z_t; see §5. The nonlinear principal-component process depends on the covariates X through the lower-dimensional indices BX, with $f()$ providing a flexible nonparametric link between the indices and the principal-component process.

In general, model (1) is not identifiable: Let P and Q be two non-singular matrices of appropriate dimensions. Consider $Y_t = CPP^{-1}f(QQ^{-1}(BX_t))$. By setting $\tilde{C} = CP$, $\tilde{f}(\cdot) = P^{-1}f(Q(\cdot))$ and $\tilde{B} = Q^{-1}B$, we obtain another set of parameters indexing the same model. To ensure that model (1) is identifiable, we need to impose r_1^2 constraints on C or f, and r_2^2 constraints on f or B. One natural set of constraints is to require that after suitable permutation of the components in Y and X, the leading sub-square matrices of C and B are identity matrices, that is, $C = (I, C^{*T})^T$, $B = (I, B^*)$, where C^{*T} and B^* have $(m - r_1) \times r_1$ and $r_2 \times (n - r_2)$ parameters, respectively; the superscript ^{T} denotes the transpose. We shall henceforth adopt this parameterization of the SPARR model, unless stated otherwise. In practice, we can first fit a model with the above convenient parameterization, and then rotate the estimates of B and C for facilitating the interpretation of the fitted model. Alternatively, the responses may be classified in terms of their dynamics by first rotating C so that the principal-component process Z_t has uncorrelated components of unit variance, followed by applying a cluster analysis to the rotated C with the rows as the cases and the columns as variables. The idea is that each component series of the response variable is a linear combination of the principal-component process, and hence the Euclidean distance between any two rows of C measures the divergence of the dynamics of the two corresponding components of Y. Thus, the similarity in the underlying dynamics of the components of Y can be empirically explored by, e.g., a cluster analysis of the rotated C.

The SPARR model generalizes a number of existing parametric, nonparametric and semi-parametric 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).

We now consider the estimation of a SPARR model. Were the parameters (C, B) 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) for details. The last advantage is especially important for high dimensional time series data.

First, we assume that 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 cross-validatory weighted least square criterion function (the notation $||\cdot||$ denotes the L^2-norm):

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

where A_{0t} are $r_1 \times 1$ vectors and A_{1t} are $r_1 \times r_2$ matrices. The arguments minimizing $L(C, B, A_{0t}, A_{1t}, t = 1, \cdots, T, h)$ yield the estimators $\hat{C}, \hat{B}, \hat{A}_{0t}, \hat{A}_{1t}, t = 1, \cdots, T$, where $\hat{A}_{0t}$ estimates $f(BX_t)$ and $\hat{A}_{1t}$ estimates the first derivative matrix of $f(\cdot)$ evaluated at BX_t; $\hat{h}$ is the bandwidth estimate.

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; the bandwidth is determined by cross-validation. Specifically, for a given $x = X_t$, we model the data around x by

$$
Y_i = C[A_{0t} + A_{1t} B(X_i - X_t)] + \text{error} \tag{3}
$$

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

$$
\sum_{i \neq t} ||Y_i - C[A_{0t} + A_{1t} B(X_i - X_t)]||^2 \times K_h[B(X_i - X_t)]. \tag{4}
$$

It is desirable to estimate B, C and the As by simultaneously minimizing the preceding local least squares for all observations. A more tractable requirement is to minimize the sum of (4) over all t, which leads to the overall objective function defined by (2). Strictly speaking, the minimization of the objective function (2) only yields estimates of $f(\cdot)$ at BX's. However, given $\hat{C}, \hat{B}$ and $\hat{h}$, for any u, $f(u)$ can be estimated by $\hat{A}_0$ which minimizes

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

For the case of unknown ranks with a known bound, we can estimate the ranks by minimizing the arguments of the criterion function

$$
L(r_1, r_2) = \sum_t ||Y_t - \hat{C}\hat{f}(\hat{B}X_t)||^2 / \sum_t ||Y_t||^2 \tag{5}
$$

over a finite grid of r_1 and r_2, where $\hat{f}, \hat{B}$ and $\hat{C}$ are estimated by minimizing the objective function (2) with the ranks of C and B set as r_1 and r_2 respectively. $L(\cdot, \cdot)$ equals the fraction of unexplained total variances. (Recall that the Y's are standardized.) 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 (2), with further elaboration including useful formulas given in Appendix A.

Step 0: Fit two reduced-rank regressions $Y_t = C_0 B X_t + \epsilon_t$ and $Y_t = C B_0 X_t + \epsilon_t$ respectively of rank r_1 and r_2 to obtain initial estimates $\hat{C}^{(0)} = \hat{C}_0$ and $\hat{B}^{(0)} = \hat{B}_0$. The dimensions of C_0 and B_0 are $m \times r_1$ and $r_2 \times n$ respectively.

Step 1: Find $\hat{A}_{0t}$ and $\hat{A}_{1t}$ by minimizing the inner sum of the objective function in (2) 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

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

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. Note that B appearing in the kernel function is fixed at the value from the preceding iterate, lest the minimization problem becomes intractable. It can shown by adapting the proof of (B.17) in Appendix B 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. The normalization of $\hat{B}$ is implemented via the pivoting technique used in Gauss-Jordan elimination method (Press et al., 1992).

Step 3: Update C by minimizing the criterion

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

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: Repeat Steps 1 to 3 until the objective function converges.

3 Asymptotic Properties of the Estimator

We first derive the asymptotic distribution for $\hat{f}$ (and $\hat{f}'$) with the parameters B and C assumed known. Indeed, the proof shows that the same result applies if B and C are known up to an error of order $o_p\{(Th^{r_2})^{-1/2} + h^2\}$. The latter convergence rate holds if $\hat{B}$ and $\hat{C}$ differ from the true values by $O_P(1/\sqrt{T})$ and $Th^{r_2+4} = O(1)$. It is shown that $\hat{f}$ is asymptotically normal with a bias of order h^2 and the rate of convergence being $O_P\{(Th^{r_2})^{-1/2}\}$. Based on this result, 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 $\hat{B}$ and $\hat{C}$ have convergence rate of $O_P(T^{-1/2})$, then under suitable conditions, $\hat{B}$ and $\hat{C}$ are asymptotically normal. In summary, under some suitable conditions, the bandwidth can be chosen to ensure both the asymptotic normality of $\hat{B}$ and $\hat{C}$ and 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 B makes 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, ν_2, scalar ν_0, $r_1 \times 1$ vector $k_{2,f_0,h}$ and $m \times m$ matrix $\Sigma(u)$ by the following formulas:

$$
\begin{aligned}
k_2 &= \int ww^T K(w)dw; && (6)\\
\nu_0 &= \int K^2(w)dw; && (7)\\
\nu_2 &= \int ww^T K^2(w)dw; && (8)\\
k_{2,f_0,h}(u) &= h^2 \int (I_{r_1} \otimes w^T) f_0''(u) w K(w)dw; && (9)\\
\Sigma(u) &= Cov(Y|U_0 = u). && (10)
\end{aligned}
$$

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 (B.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

$$C = \begin{pmatrix} I \\ C^* \end{pmatrix}, \text{ and } B = (I, B^*). \tag{11}$$

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 C 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 ith 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 (11) 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 conditions that as $T \to \infty, h \to 0, Th^{r_2} \to \infty, Th^{r_2+4} = O(1)$. Under Condition 1, as $T \to \infty$,*

$$(Th^{r_2})^{1/2} \left(\begin{bmatrix} \hat{f}(u) - f_0(u) \\ h\{vec[\hat{f}'(u) - f_0'(u)]\} \end{bmatrix} - \frac{1}{2} \begin{bmatrix} k_{2,f_0,h}(u) \\ 0 \end{bmatrix} \right) \tag{12}$$

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

$$\Sigma_{g_0}(u) \equiv \begin{pmatrix} \Sigma^{11} & 0 \\ 0 & \Sigma^{22} \end{pmatrix} \tag{13}$$

where

$$\begin{aligned}
\Sigma^{11} &= \nu_0 (C_0^T C_0)^{-1} C_0^T \Sigma(u) C_0 (C_0^T C_0)^{-1} / g_0(u) \\
\Sigma^{22} &= (k_2^{-1} \nu_2 k_2^{-1}) \otimes ((C_0^T C_0)^{-1} (C_0^T \Sigma(u) C_0)(C_0^T C_0)^{-1}) / g_0(u).
\end{aligned}$$

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

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

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

$$\begin{aligned}
\text{squared bias} &\approx k^2_{2,f_0,h,j}(u)/4, && (14)\\
\text{covariance matrix} &\approx \frac{1}{Th^{r_2}} \Sigma_{f_0,j,j}(u) && (15)
\end{aligned}$$

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,

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

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

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

3.2 Asymptotic Distribution of the Parametric Part

We will assume that $\text{vec}(\hat{B}^*)$ and $\text{vec}(\hat{C}^*)$ are within some $T^{-1/2}$-neighborhood of respectively $\text{vec}(B_0^*)$ and $\text{vec}(C_0^*)$, i.e., $\text{vec}(\hat{B}^* - B_0^*) = O_p(T^{-1/2})$ and $\text{vec}(\hat{C}^* - C_0^*) = O_p(T^{-1/2})$. Let

$\epsilon_t = Y_t - C_0 f_0(B_0 X_t)$ and $U_0 = B_0 X$. 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 $U_t = B_0 X_t$ given Y_t exists and is uniformly bounded.

(iv) All moments of the error ϵ_t exists, 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 (11). 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 $T \to \infty$,*

$$T^{1/2} \begin{pmatrix} vec(\hat{B}^* - B_0^*) \\ vec(\hat{C}^* - C_0^*) \end{pmatrix} \xrightarrow{D} N(0, Q^{-1}P(Q^{-1})^T) \tag{16}$$

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 correspond to the components of X whose coefficients in the indices are fixed according to the constraint (11); X_t is similarly partitioned as $\begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix}$;

$$\begin{aligned}
P &= Var\{[g_0(U) - K_h(0)][\Lambda - E(\Lambda|U)C_0(C_0^T C_0)^{-1}C_0^T]\epsilon\}; \\
\Lambda &= \begin{pmatrix} (X_2 \otimes I_{r_2}) f_0'^T(B_0 X) C_0^T \\ 0_{r_1(m-r_1)\times r_1}, f_0(B_0 X) \otimes I_{m-r_1} \end{pmatrix}; \\
Q_1 &= E[g(U)\Lambda C_0 f_0'(U)(X_2^T \otimes I_{r_2})] - E[g(U)\Lambda C_0 f_0'(U) E(X_2^T \otimes I_{r_2}|U)] \\
&\quad - K_h(0)E[\Lambda C_0 f_0'(U)(X_2^T \otimes I_{r_2})] + K_h(0)E[\Lambda C_0 f_0'(U)E(X_2^T \otimes I_{r_2}|U)]; \\
Q_2 &= E\left[g(U)\Lambda \begin{pmatrix} 0_{r_1\times r_1(m-r_1)} \\ f_0^T(U) \otimes I_{m-r_1} \end{pmatrix}\right] - E\left[g(U)\Lambda C_0(C_0^T C_0)^{-1}C_0^T \begin{pmatrix} 0_{r_1\times r_1(m-r_1)} \\ f_0^T(U) \otimes I_{m-r_1} \end{pmatrix}\right]
\end{aligned}$$

$$
\begin{aligned}
& +K_h(0)E\left[\Lambda C_0(C_0^TC_0)^{-1}C_0^T\begin{pmatrix} 0_{r_1\times r_1(m-r_1)} \\ f_0^T(U)\otimes I_{m-r_1}\end{pmatrix}\right] \\
& -K_h(0)E\left[\Lambda\begin{pmatrix} 0_{r_1\times r_1(m-r_1)} \\ f_0^T(U)\otimes I_{m-r_1}\end{pmatrix}\right]; \\
Q \;\; = \;\; & (Q_1, Q_2).
\end{aligned}
$$

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 in Appendix C.

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 σ-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

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

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

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

so that $W_T - E_T$ is the sum of a martingale difference sequence, and that (B.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 Studies

We have experimented with three different models to check the empirical performance of the estimation method introduced in §2. To save space, we report below one typical case; see Li (2000) for further results. Consider the following model:

$$\begin{pmatrix} Y_{1t} \\ Y_{2t} \\ Y_{3t} \end{pmatrix} = C \begin{pmatrix} \sin\left(\frac{\pi(b_1 X_t - a)}{(b-a)}\right) - .6\sin(\pi\sqrt{3}b_2 X_t) \\ 2b_2 X_t \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \end{pmatrix} \tag{17}$$

where

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix},$$

$$B = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

The error vectors are iid $\mathcal{N}(0, .01I_m)$. The explanatory variables X_t are trivariate with independent uniform $(0,1/\sqrt{3})$ components. The constants $a = \sqrt{3}/2 - 1.645/\sqrt{12}$ 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). Upon normalization, the true C, B and f become

$$C^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad B^* = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \quad \text{and}$$

$$f^*(z^*) = f((z_1^*, z_2^*)') = \begin{pmatrix} \sin\left(\frac{\pi(z_1^*+z_2^*-a)}{(b-a)}\right) - .6\sin(\pi\sqrt{3}z_1^*) + 2z_1^* \\ \sin\left(\frac{\pi(z_1^*+z_2^*-a)}{(b-a)}\right) - .6\sin(\pi\sqrt{3}z_1^*) \end{pmatrix}$$

Data of sample size 200 were simulated from this model, and the number of replications is 100. We use Gaussian kernel in this and the next section. The sample mean and standard deviation of $(\hat{c}_{31}, \hat{c}_{32}, \hat{b}_{13}, \hat{b}_{23})$ are (0.007,0.987,0.002,0.996) and (0.013,0.025,0.023,0.025), respectively. Boxplots of these estimated values are shown in Figure 1(a). See Figure 1(b) for a comparison of the true function and a function estimate from a single replicate, as viewed from three angles. Based on Theorem 3.2, We have also constructed nominal 95% (individual) confidence intervals of c_{31}, c_{32}, b_{13} and b_{23}. The empirical coverage rates of these confidence intervals are respectively $94, 94, 91$ and 90 out of 100 replications. This confirms the validity of the results in Theorem 3.2.

Table 1 reports the frequencies of the ranks which minimize the loss function defined by (5). The frequency of correct selection generally increases with the sample size, suggesting that the method is consistent.

5 Two Applications

5.1 U.S. Hog Data

The U.S. hog data have been previously analyzed by several authors, including Quenouille (1968), Box and Tiao (1977) and Velu et al. (1986). The hog data consist of annual observations from 1867 to 1948 on five variables, hog supply (Y_{1t}), hog price (Y_{2t}), corn price (Y_{3t}), corn supply (Y_{4t}) and farm wages (Y_{5t}); see Figure 3. These measurements were logarithmically transformed and linearly coded by Quenouille and, following Box and Tiao, the wage rate and hog price time series were shifted backward by one period. We denote the standardized transformed 5-variate time series as Y_t.

We first briefly summarize some pertinent results regarding the reduced-rank vector autoregressive model of the hog data; see Velu et al. (1986). Likelihood ratio tests for the vector autoregressive (VAR) order suggest an VAR(2) model for the hog data: $Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \epsilon_t$.

The determinant of the residual covariance matrix equals 6.33×10^{-16} with 50 d.f., with many elements of the covariance matrices being insignificant. Likelihood ratio tests for the rank suggest a rank-3 VAR(2) model for the hog data.

Henceforth in this example, X_t consists of the first two lags of Y_t. Lagged regression plots (Figures 2) illustrate some nonlinear patterns in the hog-supply and hog-price series. Here, we illustrate the SPARR model with the hog data; see Li (2000) for a more comprehensive analysis. Using the criterion function defined by (5), the ranks of the SPARR model are found to be $r_1 = 2$ and $r_2 = 1$ with the bandwidth estimated to be 0.1; see Table 2. The SPARR model explains 85.6% of the total variance of Y, as compared to 86.3% and 84.9% for the VAR(2) model and reduced rank VAR(2) model respectively.

An important use of the SPARR model is to classify the response variables in terms of their nonlinear dynamics. From eyeball inspection of Figure 3, the five series appear to move together. Interestingly, this phenomenon can be explored as follows. Below is $\hat{C}$ from the fitted SPARR model with $r_1 = 2, r_2 = 1$ (after varimax rotation):

$$C = \begin{bmatrix} 0.739 & 0.366 \\ (0.064) & (0.096) \\ 0.314 & 0.921 \\ (0.055) & (0.052) \\ -0.032 & 1.000 \\ (NA) & (NA) \\ 1.000 & 0.032 \\ (NA) & (NA) \\ 0.442 & 0.835 \\ (0.045) & (0.054) \end{bmatrix}.$$

Standard errors, enclosed in parentheses, are computed according to the results stated in Theorem 3.2. However, the hog data violate the strong mixing condition required by Theorem 3.2; hence, the validity of these standard errors requires further investigation. The four missing standard errors, denoted as NA, arise because they correspond to the parameters subject to the normalization constraint. The residuals as shown in Figure 4 appear to be stationary. The time series plot of f_1 (Figure 5(a)) resembles that of corn supply while the time series plot of f_2 resembles that of corn price, as can also be inferred from $\hat{C}$. Figure 5(b) suggests that f_i

are approximately piecewise linear functions of bx with the threshold at zero. (Note that B is written as b as it is a row vector.) Because $\hat{b}X_t$ is generally negative before 1907 and positive afterward, this suggests that the data may have different dynamics before and after 1907. In addition,

$$\hat{b} = \begin{bmatrix} 1.00 & 0.104 & -0.0979 & -0.174 & 1.97 & 0.0975 & 0.440 & 0.116 & 0.0132 & -1.840) \\ (NA) & (0.157) & (0.0795) & (0.0816) & (0.520) & (0.110) & (0.139) & (0.101) & (0.155) & (0.434) \end{bmatrix}$$

so that the index variable is essentially determined by the first lag of hog supply ($Y_{1,t-1}$), the second lag of hog price ($Y_{1,t-2}$) and the annual change of the (log) farm wages ($Y_{5,t-1} - Y_{5,t-2}$).

The preceding analysis of the SPARR model can be re-cast in terms of nonlinear co-integration. Let $I(0)$ denote a stationary process. Treating the stochastic errors of the SPARR model as exactly stationary, we can, from the estimated C matrix, write

$$\begin{aligned} Y_{4t} &= f_1(bX_t) + I(0) \\ Y_{3t} &= f_2(bX_t) + I(0). \end{aligned}$$

Let $Z_t = (Y_{1t}, Y_{2t}, Y_{5t})^T$ and $W_t = (Y_{4t}, Y_{3t})^T$. Then

$$\begin{aligned} Z_t &= \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{51} & c_{52} \end{pmatrix} \begin{pmatrix} f_1(bX_t) \\ f_2(bX_t) \end{pmatrix} + I(0) \\ &= \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{51} & c_{52} \end{pmatrix} \begin{pmatrix} Y_{4t} + I(0) \\ Y_{3t} + I(0) \end{pmatrix} + I(0) \\ &= \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{51} & c_{52} \end{pmatrix} \begin{pmatrix} Y_{4t} \\ Y_{3t} \end{pmatrix} + I(0), \\ &\equiv C^* \begin{pmatrix} Y_{4t} \\ Y_{3t} \end{pmatrix} + I(0), \end{aligned}$$

or $Z_t = C^*W_t + I(0)$ with $W_t = (f_1(bX_t), f_2(bX_t))^T + I(0)$. This implies a nonlinear co-integration structure on the hog data with corn price (Y_{3t}) and corn supply (Y_{4t}) driven by a nonlinear process; for further discussions on nonlinear co-integration see Granger and Hallman (1991) and Granger et al. (1997).

5.2 Lynx Data

We consider as a second example a modern panel of eight annual lynx pelt series, labeled as L15 to L22, which were collected over eight providences and regions in Canada from 1920 to 1994; see Stenseth et al. (1999) for a recent summary account of this data set, who reported that the dynamics of the eight series can be classified into three groups: Pacific-maritime, Continental and Atlantic-maritime. (See Tong (1990) for a review of the related lynx series collected in the Mackenzie river region.) Recently, Li and Chan (2001) fitted the REduced-rank Threshold Autoregressive (RETAR) model to the panel of lynx data and found broadly similar classification. However, the RETAR model assumes the underlying dynamics is piecewise linear, while Stenseth et al. (1999) applied a panel of independent Threshold Autoregressive models to analyze the common structure of the lynx data. Here, we consider the use of the SPARR model for classifying the series without imposing strong prior assumption on the functional form.

Let Y consist of the (standardized) log lynx counts L15 to L22, in this order. As in Stenseth et al. (1999) and Li and Chan (2001), we use the first two lags of Y as the explanatory variables, i.e., $X_t = (Y_{t-1}, Y_{t-2})'$. Table 3 shows that the ranks $r_1 = 2$ and $r_2 = 2$ minimize the cross-validatory percent of unexplained total variance, although $r_1 = 2$ and $r_2 = 3$ are a close second. Hence, we tentatively identify the model as a SPARR model with $r_1 = r_2 = 2$, i.e.,

$$\begin{aligned} Y_t &= Cf(BX_t) + \epsilon_t, \\ &= (c_1, c_2)\begin{pmatrix} f_1(b_1X_t, b_2X_t) \\ f_2(b_1X_t, b_2X_t) \end{pmatrix} + \epsilon_t \end{aligned} \tag{18}$$

where $C = (c_1, c_2), B = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$, $f = (f_1, f_2)'$, r_1=rank(C)=2, r_2=rank(B)=2. With the ranks set to be $r_1 = 2$ and $r_2 = 2$, the cross validation criterion defined by (2) chooses $\hat{h} = 0.4$. Below are $\hat{C}$ before and after uncorrelated-component rotation, i.e., the rotated nonlinear principal-component process $\{(f_1(b_1X_t, b_2X_t), f_2(b_1X_t, b_2X_t))^T\}$ has uncorrelated components that are of unit variance:

$$\begin{bmatrix} 0.792 & 0.305 \\ (0.126) & (0.109) \\ 0.000 & 1.000 \\ (NA) & (NA) \\ 0.590 & -0.021 \\ (0.140) & (0.153) \\ 1.004 & -0.232 \\ (0.176) & (0.158) \\ 0.939 & -0.645 \\ (0.182) & (0.193) \\ 0.990 & -0.518 \\ (0.200) & (0.184) \\ 1.000 & 0.000 \\ (NA) & (NA) \\ 0.796 & 0.344 \\ (0.123) & (0.112) \end{bmatrix}, \quad \begin{bmatrix} 0.0796 & 0.0357 \\ (0.0139) & (0.0127) \\ 0.0208 & 0.0959 \\ (NA) & (NA) \\ 0.0722 & -0.00751 \\ (0.0213) & (0.0179) \\ 0.0734 & -0.0037 \\ (0.0179) & (0.0127) \\ 0.0618 & -0.0398 \\ (0.0235) & (0.0211) \\ 0.0712 & -0.0233 \\ (0.0222) & (0.0126) \\ 0.100 & 0.0208 \\ (NA) & (NA) \\ 0.0980 & 0.0178 \\ (0.0219) & (0.0133) \end{bmatrix}.$$

The Euclidean distance between any two rows of the rotated C measures the dissimilarity of the two corresponding lynx series in terms of their dynamics. Applying the hierarchical clustering (with the options of complete linkage and the Euclidean distance) to the rotated $\hat{C}$ yields (Figure 6(a)) three clusters, namely, L16 (Yukon) in the first cluster, L15 (British Columbia), L21 (Ontario), and L22 (Quebec) in the second cluster of maritime provinces that may be further broken down into the Pacific-maritime and the Atlantic-maritime clusters, and L17 (North West Territories), L18 (Alberta), L19 (Saskatchewan) and L20 (Manitoba) in the third cluster of continental provinces. This classification of the eight lynx series is broadly similar to that reported by Stenseth et al. (1999) and Li and Chan (2001).

The (unreported) time series plot of f_1 resembles that of L21 (Ontario) while the time series plot of f_2 is similar to that of L16 (Yukon), as can also be inferred from $\hat{C}$. The two series, Ontario and Yukon, form an approximate bases for the other series, in the sense that the other series are approximated by a weighted averages of these two series. To further understand the shape of estimated function $f = (f_1, f_2)'$, we plot in Figure 6(b) two 3-D plots of f_i versus $U_1 = b_1' X_t$ and $U_2 = b_2' X_t$. The smoothed plots of the f's are from the loess function fits with the options of local polynomial of degree 1 and span= 0.75. The smooth plots suggest that f_1 is approximately a linear function and f_2 resembles a piecewise linear function, "confirming" the nonlinearity of the lynx dynamics.

6 Conclusion

We have demonstrated that the SPARR model provides a flexible approach for studying the relationship of a high-dimensional regression model with multivariate response and explanatory variables. In particular, in the case of panel time series data, the SPARR model may shed insights on grouping the data into groups of common dynamical structure. Future research problems include investigating the limiting properties of the estimators for nonstationary processes, and establishing the consistency of the rank selection procedure in §2.

A Some Formulas Useful for Implementing Steps 1 to 3.

For simplicity, the iteration index k will be suppressed throughout this appendix.

Step 1: The objective function can be rewritten as (with $X_t = x$):

$$\sum_{i \neq t} ||Y_i^* - \hat{C}(A_{0t}\ A_{1t}) \begin{pmatrix} w_{it}^{1/2} \\ \hat{B}(X_i - x) w_{it}^{1/2} \end{pmatrix} ||^2$$

where $w_{it} = K_h(\hat{B}(X_i - x))$. Let

$$\begin{aligned} Y^* &= (Y_1^*, \cdots, Y_{-t}^*, \cdots, Y_T^*) = (Y_1 w_{1t}^{1/2}, \cdots, Y_t w_{-tt}^{1/2}, \cdots, Y_T w_{Tt}^{1/2}) \\ X^* &= \begin{pmatrix} w_{1t}^{1/2} & \dots & w_{-tt}^{1/2} & \dots & w_{Tt}^{1/2} \\ \hat{B}(X_1 - x) w_{1t}^{1/2} & \dots & \hat{B}(X_t - x) w_{-tt}^{1/2} & \dots & \hat{B}(X_T - x) w_{Tt}^{1/2} \end{pmatrix} \end{aligned}$$

where the minus index signifies the omission of the corresponding term. After some calculus (see Reinsel and Velu, 1998, p. 157), it can be shown that

$$(\hat{A}_{0t}, \hat{A}_{1t}) = (\hat{C}^T \hat{C})^{-1} \hat{C}^T \hat{\Sigma}_{Y^*X^*} \hat{\Sigma}_{X^*X^*}^{-1} \tag{A.1}$$

where $\hat{\Sigma}_{Y^*X^*} = (1/T)(Y^* X^{*'}), \hat{\Sigma}_{X^*X^*} = (1/T)(X^* X^{*'})$.

Step 2: The objective function can be written as $\sum_{t, i \neq t} ||Y_{it}^{2*} - \hat{C}_{it}^{2*} B X_{it}^{2*}||^2$, where $Y_{it}^{2*} = Y_i w_{it}^{1/2} - \hat{C} \hat{A}_{0t} w_{it}^{1/2}, \hat{C}_{it}^{2*} = \hat{C} \hat{A}_{1t}, X_{it}^{2*} = (X_i - X_t) w_{it}^{1/2}$, and $w_{it} = K_h(\hat{B}^{(k-1)}(X_i - X_t))$. After some matrix calculus (Schott, 1997), it can be shown that $\hat{B}$ satisfies the normal equation:

$$\sum_{it} [-C_{it}^{2*T} Y_{it}^{2*} X_{it}^{2*T} + C_{it}^{2*T} C_{it}^{2*} B X_{it}^{2*} X_{it}^{2*T}] = 0, \tag{A.2}$$

which has the solution

$$\mathrm{vec}(B) = \left[\sum_i X_{it}^{2*} X_{it}^{2*T} \otimes C_{it}^{2*T} C_{it}^{2*}\right]^{-1} \cdot \left[\sum_i [\mathrm{vec}(C_{it}^{2*T} Y_{it}^{2*} X_{it}^{2*T})]\right]. \tag{A.3}$$

Step 3: The objective function can be rewritten as, $\sum_{t,i\neq t} ||Y_{it}^{3*} - C\ X_{it}^{3*}||^2$ where $Y_{it}^{3*} = Y_i w_{it}^{1/2}, X_{it}^{3*} = (\hat{A}_{0t} + \hat{A}_{1t}\hat{B}(X_i - X_t))w_{it}^{1/2}$. Hence the solution $\hat{C} = Y^{3*}X^{3*T}(X^{3*}X^{3*T})^{-1}$.

B 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 and C satisfying $||B - B_0|| = o\{(Th^{r_2})^{-1/2} + h^2\}$ and $||C - C_0|| = o\{(Th^{r_2})^{-1/2} + h^2\}$, where $||\cdot||$ denotes the Euclidean norm of a matrix. 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

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

Note that the case $i = t$ has negligible contribution to the above sum and hence included in the criterion function. Consider the normalized function

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

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

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

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

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

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

It can be shown (Li, 2000) that

$$D_T = -D + o_P(1), \tag{B.4}$$

where

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

Therefore,

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

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 ζ's are some "intermediate" points between u and U. Note that when u coincides with U so that $\zeta = u$, then

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

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

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

The variance of W_T equals

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

where

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

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

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

Therefore,

$$D^{-1}W_T - D^{-1}EW_T \xrightarrow{D} N(0, D^{-1}WD^{-1}), \tag{B.11}$$

or,

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

or,

$$c_T^{-1}\begin{pmatrix} [A_0 - f_0(u)] \\ h\{\text{vec}[A_1 - f_0'(u)]\} \end{pmatrix} - \frac{1}{2}c_T^{-1}\begin{pmatrix} (C_0^T C_0)^{-1} & 0 \\ 0 & (k_2 \otimes C_0^T C_0)^{-1} \end{pmatrix}\begin{pmatrix} C_0^T C_0 k_{2,f_0,h} \\ 0 \end{pmatrix}$$
$$+ o_P(c_T^{-1}h^2) + o_P(1) \xrightarrow{D} N(0, D^{-1}WD^{-1}).$$

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 and C_0 are known, we have

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

(b) For general B and C, we have

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

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

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

uniformly in $u \in \mathcal{D}$, where $D_T(u)$ and $D(u)$ are defined in (B.4) and (B.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

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

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

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

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

Consequently,

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

Claim 2:

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

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

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

Via Taylor expansion, we obtain

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

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

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 in Appendix C 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.)

$$
\begin{aligned}
& T^{-1}\sum_t (Y_t - C_0 a) K_h'(U_t - u_0)[(X_{2t}-x_0)^T \otimes I_{r_2}]\mathrm{vec}(\hat{B}^* - B_0^*) \\
= \quad & O_P(h/\sqrt{T} + \sqrt{\ln T}/(T\sqrt{h^{r_2}})) = o_P(1/\sqrt{T}). \qquad \text{(B.17)}
\end{aligned}
$$

Similarly, we have

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

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

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

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

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

This completes the proof of Claim 2. $\square$

Claim 3:

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

Proof: After some algebra, it can be shown that

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

$\square$

.

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

$$-\sum_t \sum_{i\neq t} ||Y_i - C[\hat{f}(BX_t; B^*, C^*) + \hat{f}'(BX_t; B^*, C^*)B(X_i - X_t)]||^2 K_h[B(X_i - X_t)].$$

Let

$$
\begin{aligned}
\hat{\Lambda}_{i,t} &= \begin{pmatrix} (X_{2t}\otimes I_{r_2})\hat{f}'^T(\hat{B}X_t;\hat{B}^*,\hat{C}^*)\hat{C}^T + [(X_{2i}-X_{2t})\otimes \hat{f}'^T(\hat{B}X_t;\hat{B}^*,\hat{C}^*)]\hat{C}^T \\ 0_{r_1(m-r_1)\times r_1}, -[\hat{f}(\hat{B}X_t;\hat{B}^*,\hat{C}^*) + \hat{f}'(\hat{B}X_t;\hat{B}^*,\hat{C}^*)\hat{B}(X_i-X_t)]\otimes I_{m-r_1} \end{pmatrix}, \\
\Lambda_{i,t} &= \begin{pmatrix} (X_{2t}\otimes I_{r_2})f_0'^T(B_0X_t)C_0^T + [(X_{2i}-X_{2t})\otimes f_0'^T(B_0X_t)]C_0^T \\ 0_{r_1(m-r_1)\times r_1}, [f_0(B_0X_t) + f_0'(B_0X_t)B_0(X_i-X_t)]\otimes I_{m-r_1} \end{pmatrix}, \\
\Lambda_t &= \begin{pmatrix} (X_{2t}\otimes I_{r_2})f_0'^T(B_0X_t)C_0^T \\ 0_{r_1(m-r_1)\times r_1}, f_0(B_0X_t)\otimes I_{m-r_1} \end{pmatrix}.
\end{aligned}
$$

Taking the first derivative of the objective function with respect to $(\text{vec}(B^*), \text{vec}(C^*))$, and via Taylor expansion we have

$$
\begin{aligned}
0 &= \frac{1}{\sqrt{T^3}}\sum_{t,i\neq t}\hat{\Lambda}_{i,t}\{Y_i - \hat{C}[\hat{f}(\hat{B}X_t;\hat{B}^*,\hat{C}^*) + \hat{f}'(\hat{B}X_t;\hat{B}^*,\hat{C}^*)\hat{B}(X_i-X_t)]\}K_h[\hat{B}(X_i-X_t)] \\
&\qquad + o_P(1/\sqrt{T}) \\
&= \frac{1}{\sqrt{T^3}}\sum_{t,i\neq t}\Lambda_{i,t}\{Y_i - C_0[f_0(B_0X_t) + f_0'(B_0X_t)B_0(X_i-X_t)]\}K_h[B_0(X_i-X_t)] \\
&\quad - \frac{1}{\sqrt{T^3}}\sum_{t,i\neq t}\Lambda_{i,t}C_0[\hat{f}(\hat{B}X_t;\hat{B}^*,\hat{C}^*) - f_0(B_0X_t)]K_h[B_0(X_i-X_t)] \\
&\quad - \frac{1}{\sqrt{T^3}}\sum_{t,i\neq t}\Lambda_{i,t}C_0[X_{2t}^T\otimes f_0'(B_0X_t) + (X_{2i}-X_{2t})^T\otimes f_0'(B_0X_t)]\text{vec}(\hat{B}^*-B_0^*)K_h[B_0(X_i-X_t)] \\
&\quad - \frac{1}{\sqrt{T^3}}\sum_{t,i\neq t}\Lambda_{i,t}\begin{pmatrix} 0_{r_1\times r_1(m-r_1)} \\ [f_0(B_0X_t) + f_0'(B_0X_t)B_0(X_i-X_t)]^T\otimes I_{m-r_1}\end{pmatrix}\text{vec}(\hat{C}^*-C^*) \\
&\qquad \times K_h[B_0(X_i-X_t)] \\
&\quad + o_P(1). \qquad\qquad \text{(B.19)}
\end{aligned}
$$

It follows from Claims 1-3 that

$$
\begin{aligned}
0 &= \frac{1}{\sqrt{T^3}}\sum_{t,i\neq t}\Lambda_{i,t}\{Y_i - C_0[f_0(U_t) + f_0'(U_t)B_0(X_i-X_t)]\}K_h[B_0(X_i-X_t)] \\
&\quad - \frac{1}{\sqrt{T^3}}\sum_{t,i\neq t}\Lambda_{i,t}C_0f_0'(U_t)(X_{2t}^T\otimes I_{r_2})\text{vec}(\hat{B}^*-B_0^*)K_h[B_0(X_i-X_t)] \\
&\quad - \frac{1}{\sqrt{T^3}}\sum_{t,i\neq t}\Lambda_{i,t}C_0((C_0^TC_0)^{-1}C_0^T[g(U_t)T]^{-1}\sum_j\{Y_j - C_0[f_0(U_j) \\
&\qquad + f_0'(U_j)B_0(X_j-X_t)]\}K_h[B_0(X_j-X_i)])K_h[B_0(X_i-X_t)] \\
&\quad + \frac{1}{\sqrt{T^3}}\sum_{t,i\neq t}\Lambda_{i,t}C_0f_0'(U_t)E(X_2^T\otimes I_{r_2}|U_t)\text{vec}(\hat{B}^*-B_0^*)K_h[B_0(X_i-X_t)]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{\sqrt{T^3}}\sum_{t,i\neq t}\Lambda_{i,t}C_0(C_0^TC_0)^{-1}C_0^T\begin{pmatrix}0_{r_1\times r_1(m-r_1)}\\ f_0^T(U_t)\otimes I_{m-r_1}\end{pmatrix}\mathrm{vec}(\hat{C}^*-C_0^*)K_h[B_0(X_i-X_t)]\\
&-\frac{1}{\sqrt{T^3}}\sum_{t,i\neq t}\Lambda_{i,t}C_0[X_{2i}^T\otimes f_0'(U_t)]\mathrm{vec}(\hat{B}^*-B_0^*)K_h[B_0(X_i-X_t)]\\
&-\frac{1}{\sqrt{T^3}}\sum_{t,i\neq t}\Lambda_{i,t}\begin{pmatrix}0_{r_1\times r_1(m-r_1)}\\ [f_0(U_t)+f_0'(U_t)B_0(X_i-X_t)]^T\otimes I_{m-r_1}\end{pmatrix}\mathrm{vec}(\hat{C}^*-C_0^*)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\times K_h[B_0(X_i-X_t)]\\
&+o_P(1). \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(B.20)}
\end{aligned}
$$

We split the sum in(B.20) into two sums:

$$\sum_t\sum_{i\neq t}=\sum_t\sum_i-\sum_t\sum_{i=t}=J_1-J_2.$$

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

$$
\begin{aligned}
J_1 \;=\;& \frac{1}{\sqrt{T}}\sum_t\Lambda_t\epsilon_tg(U_t)\\
&-\frac{1}{\sqrt{T}}\sum_t E(\Lambda|U_t)C_0(C_0^TC_0)^{-1}C_0^T\epsilon_tg(U_t)\\
&-\frac{1}{\sqrt{T}}\sum_t\Lambda_tC_0f_0'(U_t)(X_{2t}^T\otimes I_{r_2})g(U_t)\mathrm{vec}(\hat{B}^*-B_0^*)\\
&+\frac{1}{\sqrt{T}}\sum_t\Lambda_tC_0f_0'(U_t)E(X_2^T\otimes I_{r_2}|U_t)g(U_t)\mathrm{vec}(\hat{B}^*-B_0^*)\\
&+\frac{1}{\sqrt{T}}\sum_t\Lambda_tC_0(C_0^TC_0)^{-1}C_0^T\begin{pmatrix}0_{r_1\times r_1(m-r_1)}\\ f_0^T(U_t)\otimes I_{m-r_1}\end{pmatrix}g(U_t)\mathrm{vec}(\hat{C}^*-C_0^*)\\
&-\frac{1}{\sqrt{T}}\sum_t\Lambda_t\begin{pmatrix}0_{r_1\times r_1(m-r_1)}\\ f_0^T(U_t)\otimes I_{m-r_1}\end{pmatrix}g(U_t)\mathrm{vec}(\hat{C}^*-C_0^*)+o_P(1).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
J_2 \;=\;& \frac{1}{\sqrt{T}}\sum_t\Lambda_t\epsilon_tK_h(0)\\
&-\frac{1}{\sqrt{T}}\sum_t E(\Lambda|U_t)C_0(C_0^TC_0)^{-1}C_0^T\epsilon_tK_h(0)\\
&-\sqrt{T}E[\Lambda_tC_0f_0'(U)(X_2^T\otimes I_{r_2})]\mathrm{vec}(\hat{B}^*-B_0^*)K_h(0)\\
&+\sqrt{T}E[\Lambda_tC_0f_0'(U)E(X_2^T\otimes I_{r_2}|U)]\mathrm{vec}(\hat{B}^*-B_0^*)K_h(0)
\end{aligned}
$$

$$+\sqrt{T}E\left[\Lambda_t C_0(C_0^TC_0)^{-1}C_0^T\begin{pmatrix}0_{r_1\times r_1(m-r_1)}\\ f_0^T(B_0X)\otimes I_{m-r_1}\end{pmatrix}\right]\mathrm{vec}(\hat{C}^*-C_0^*)K_h(0)$$

$$-\sqrt{T}E\left[\Lambda_t\begin{pmatrix}0_{r_1\times r_1(m-r_1)}\\ f_0^T(B_0X)\otimes I_{m-r_1}\end{pmatrix}\right]\mathrm{vec}(\hat{C}^*-C_0^*)K_h(0)+o_P(1).$$

Hence (B.20) becomes

$$\begin{aligned}
0 \;=\;& \frac{1}{\sqrt{T}}\sum_t g(U_t)[\Lambda_t - E(\Lambda|U_t)C_0(C_0^TC_0)^{-1}C_0^T]\epsilon_t \\
&-\frac{1}{\sqrt{T}}\sum_t[\Lambda_t - E(\Lambda|U_t)C_0(C_0^TC_0)^{-1}C_0^T]\epsilon_t K_h(0) \\
&-\sqrt{T}\{E[g(U)\Lambda C_0 f_0'(U)(X_2^T\otimes I_{r_2})] - E[g(U)\Lambda C_0 f_0'(U)E(X_2^T\otimes I_{r_2}|U)] \\
&\quad -K_h(0)E[\Lambda C_0 f_0'(U)(X_2^T\otimes I_{r_2})] + K_h(0)E[\Lambda C_0 f_0'(U)E(X_2^T\otimes I_{r_2}|U)]\} \\
&\quad \times\mathrm{vec}(\hat{B}^*-B_0^*) \\
&-\sqrt{T}\left\{E\left[g(U)\Lambda\begin{pmatrix}0_{r_1\times r_1(m-r_1)}\\ f_0^T(U)\otimes I_{m-r_1}\end{pmatrix}\right] - E\left[g(U)\Lambda C_0(C_0^TC_0)^{-1}C_0^T\begin{pmatrix}0_{r_1\times r_1(m-r_1)}\\ f_0^T(U)\otimes I_{m-r_1}\end{pmatrix}\right]\right. \\
&\quad +K_h(0)E\left[\Lambda C_0(C_0^TC_0)^{-1}C_0^T\begin{pmatrix}0_{r_1\times r_1(m-r_1)}\\ f_0^T(U)\otimes I_{m-r_1}\end{pmatrix}\right] \\
&\left. -\;K_h(0)E\left[\Lambda\begin{pmatrix}0_{r_1\times r_1(m-r_1)}\\ f_0^T(U)\otimes I_{m-r_1}\end{pmatrix}\right]\right\}\times\mathrm{vec}(\hat{C}^*-C_0^*) \\
&+o_P(1).
\end{aligned}$$

Hence

$$\begin{aligned}
&\sqrt{T}Q\begin{pmatrix}\mathrm{vec}(\hat{B}^*-B_0^*)\\ \mathrm{vec}(\hat{C}^*-C_0^*)\end{pmatrix} \\
=\;& \frac{1}{\sqrt{T}}\sum_t[g(U_t)-K_h(0)][\Lambda_t - E(\Lambda|U_t)C_0(C_0^TC_0)^{-1}C_0^T]\epsilon_t \\
&+o_P(1),
\end{aligned}\tag{B.21}$$

from which (16) can be readily derived. □

C Lemma 1

Let $\{Y_t, \underline{X}_t\}$ be a bivariate time series, and ψ a measurable function. The following lemma essentially follows from Theorem 5 and Corollary 3 in Masry (1996); hence the proof is omitted.

See Masry (1996) for undefined notations below; in particular, d and n defined in Masry (1996) correspond to r_2 and T defined in this paper.

Lemma 1: Under the conditions as stated in Theorem 5 and Corollary 3 in Masry (1996), which follows from Conditions 1-3, it holds that

$$\begin{aligned} & \frac{1}{n-d+1}\sum_{i=0}^{n-d}\psi(Y_{d+i})K_h(\underline{X}_i-\underline{x}) \\ = & \sum_{0\le|\underline{j}|\le p+1}\frac{1}{\underline{j}!}(D^{\underline{j}}m)(\underline{x})E\left[\left(\frac{\underline{X}_i-\underline{x}}{h}\right)^{\underline{j}}K_h(\underline{X}_i-\underline{x})\right]+o(h^{p+1})+O(\sqrt{\frac{\ln n}{nh^d}}). \end{aligned}$$

uniformly in $\underline{x}\in D$.

References

Box, G. E. P. and Tiao, G. C. (1977), "A canonical analysis of multiple time series," *Biometrika*, 64, 355–365.

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.

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.

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

Granger, C. W. J. and Hallman, J. (1991), "Nonlinear transformations of integrated time series," *Journal of Time Series Analysis*, 12, 207–224.

Granger, C. W. J., Inoue, T., and Morin, N. (1997), "Nonlinear stochastic trends," *Journal of Econometrics*, 81, 65–92.

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.

Izenman, A. J. (1980), "Assessing dimensionality in multivariate regression," in *Handbook of Statistics*, ed. Krishnaiah, P., New York: North Holland, pp. 571–592.

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), "Multivariate reduced-rank nonlinear time series modeling," *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.

Quenouille, M. H. (1968), *The Analysis of Multiple Time Series*, London: Griffin.

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.

Schott, J. R. (1997), *Matrix Analysis for Statistics*, New York: John Wiley.

Stenseth, N. C., Chan, K. S., Tong, H., Boonstra, R., Boutin, S., Krebs, C. J., Post, E., O'Donoghue, M., Yoccoz, N. G., Forchhammer, M. C., and Hurrell, J. W. (1999), "Common dynamic structure of Canadian lynx populations within three geo-climatic zones," *Science*, 285, 1071–1073.

Tong, H. (1990), *Non-linear Time Series: A Dynamical System Approach*, Oxford: Oxford University Press.

Velu, R., Reinsel, G. C., and Wichern, D. W. (1986), "Reduced rank models for multiple time series," *Biometrika*, 73, 105–18.

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

Table 1: Frequency (out of 100 replicates) for selecting ranks r_1=rank(C) and r_2=rank(B) of the SPARR model from the simulated data. T is the sample size.

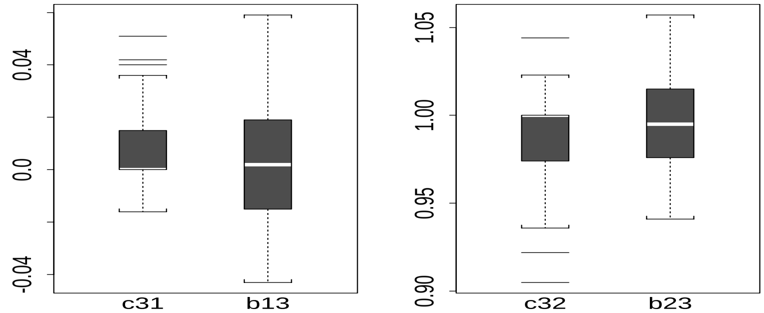
	T=100			T=200			T=400			T=800		
r_1/r_2	1	2	3	1	2	3	1	2	3	1	2	3
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	83	2	0	85	1	0	81	2	0	92	0
3	0	31	0	0	12	2	0	17	0	0	8	0

Table 2: Objective function for selecting r_1=rank(C) and r_2=rank(B) of the SPARR model for the U.S. hog data. The objective function is defined by (5). The entry **ER2**(**ER3**) denotes that the estimated B(C) matrix can not be normalized.

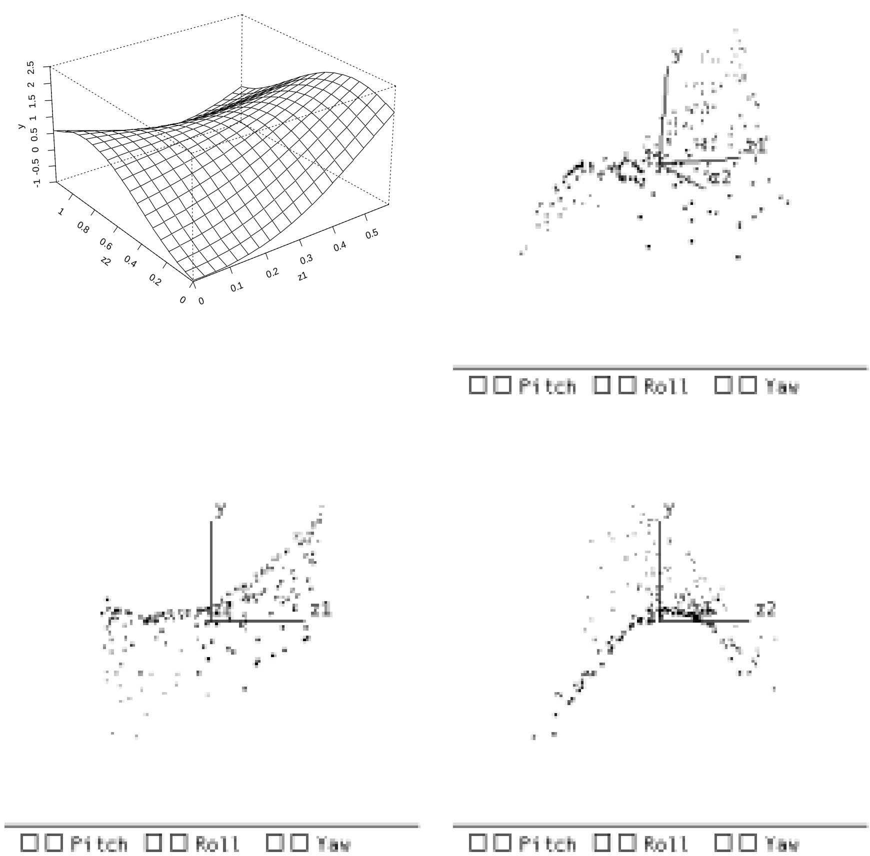
r_1/r_2	1	2	3	4
1	.955	.658	ER2	.348
2	.257	.463	.560	.463
3	ER3	.449	ER2	ER2
4	.322	.681	ER2	ER3
5	.583	.705	.625	ER2

Table 3: Objective function for selecting r_1=rank(C) and r_2=rank(B) of the SPARR model for the lynx data. The objective function is defined by (5).

r_1/r_2	1	2	3	4	5
1	.477	.490	.520	.652	.520
2	.461	.375	.376	.641	.654
3	.736	.525	.545	.429	.435
4	.671	.616	.419	.568	.472



(a) Boxplots of the coefficient estimates.



(b) The graph of the true function $f()$ and three views of an estimated function from a simulated realization.

Figure 1: Simulation studies of the model defined by (17).

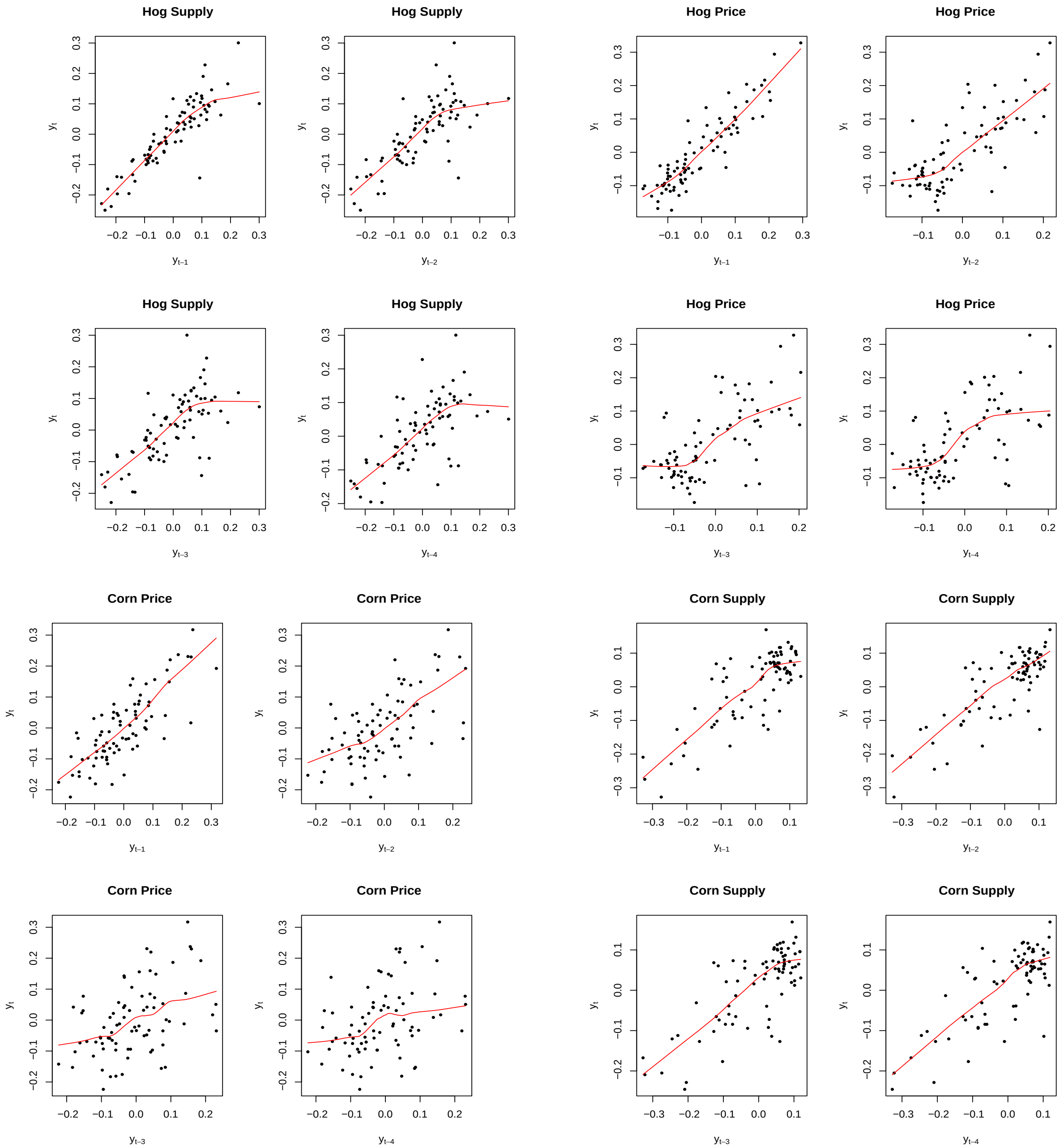


Figure 2: Lagged regression plots for the U.S. hog data

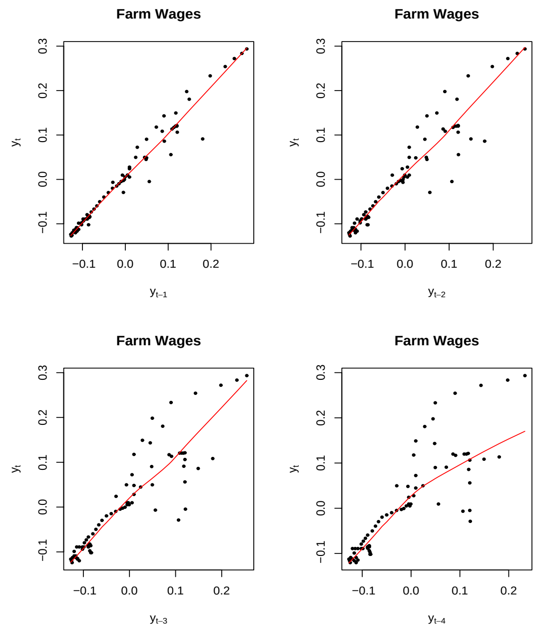


Figure 2: (continued)

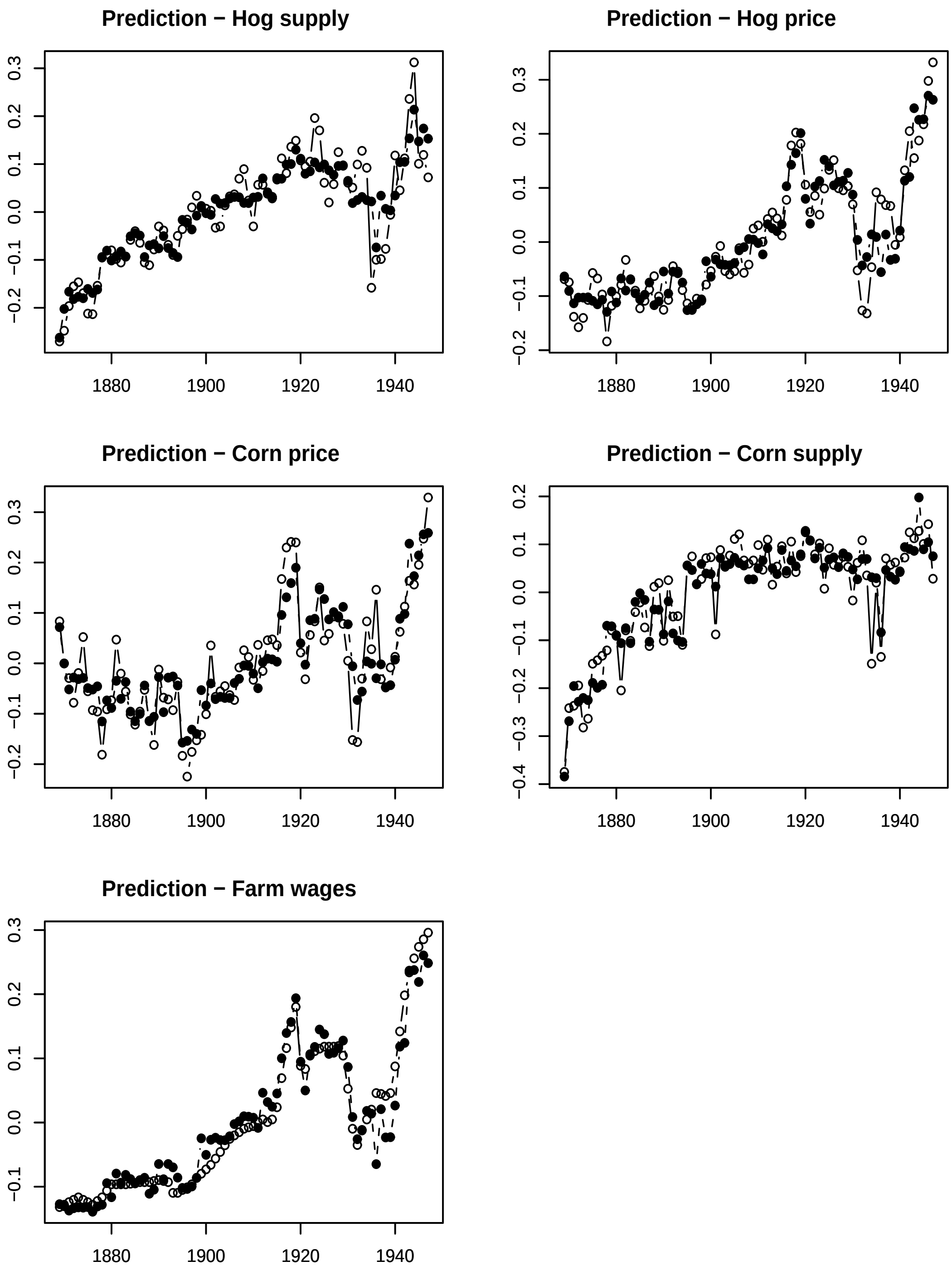


Figure 3: Within-sample 1-step ahead predictions from the SPARR model of the U.S. hog data, with $r_1 = 2$ and $r_2 = 1$. Observations are drawn as open circles, and predicted values as solid circles.

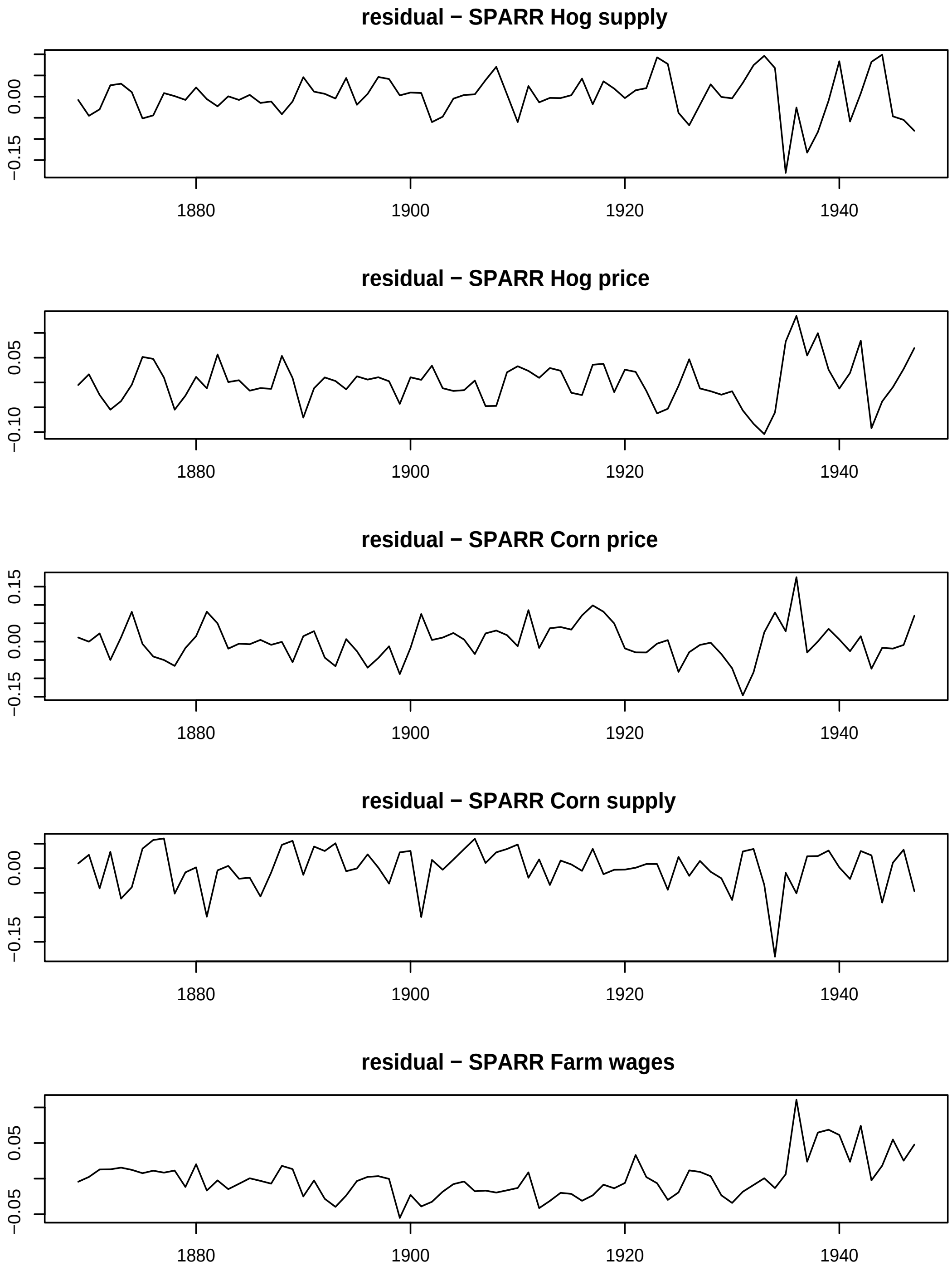
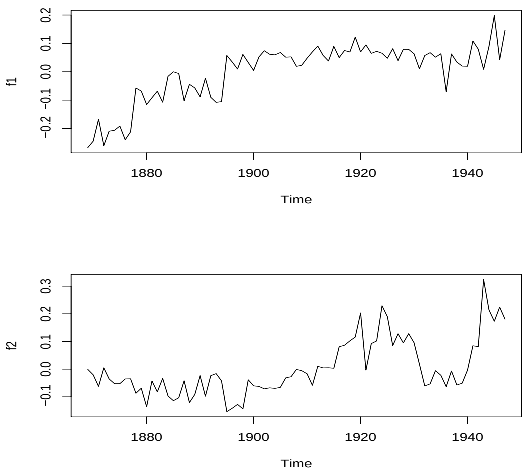
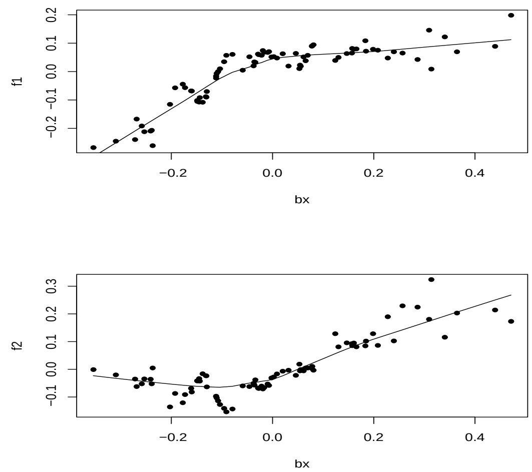


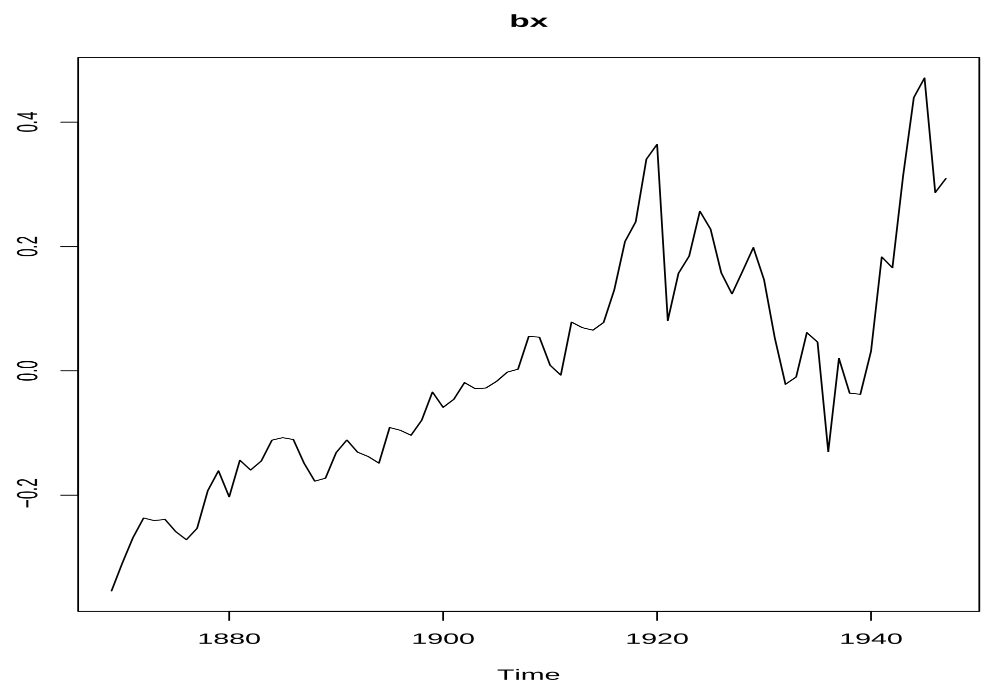
Figure 4: Residual plots for the U.S. hog data from the SPARR model with $r_1 = 2, r_2 = 1$.



(a) Time series plots $\hat{f}_1$ and $\hat{f}_2$

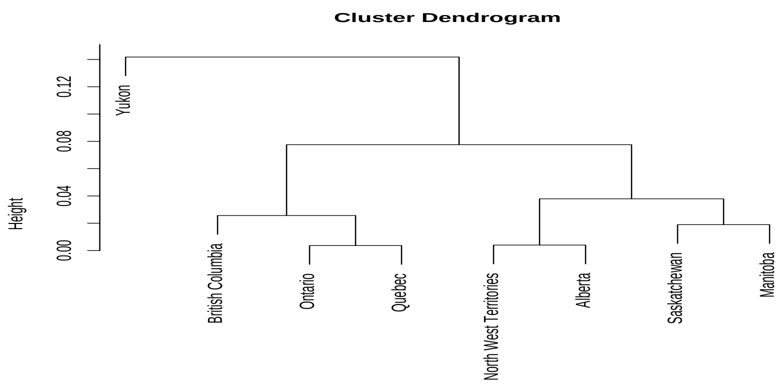


(b) Scatter plots of $\hat{f}_1$ and $\hat{f}_2$. Solid line is obtained from lowess.

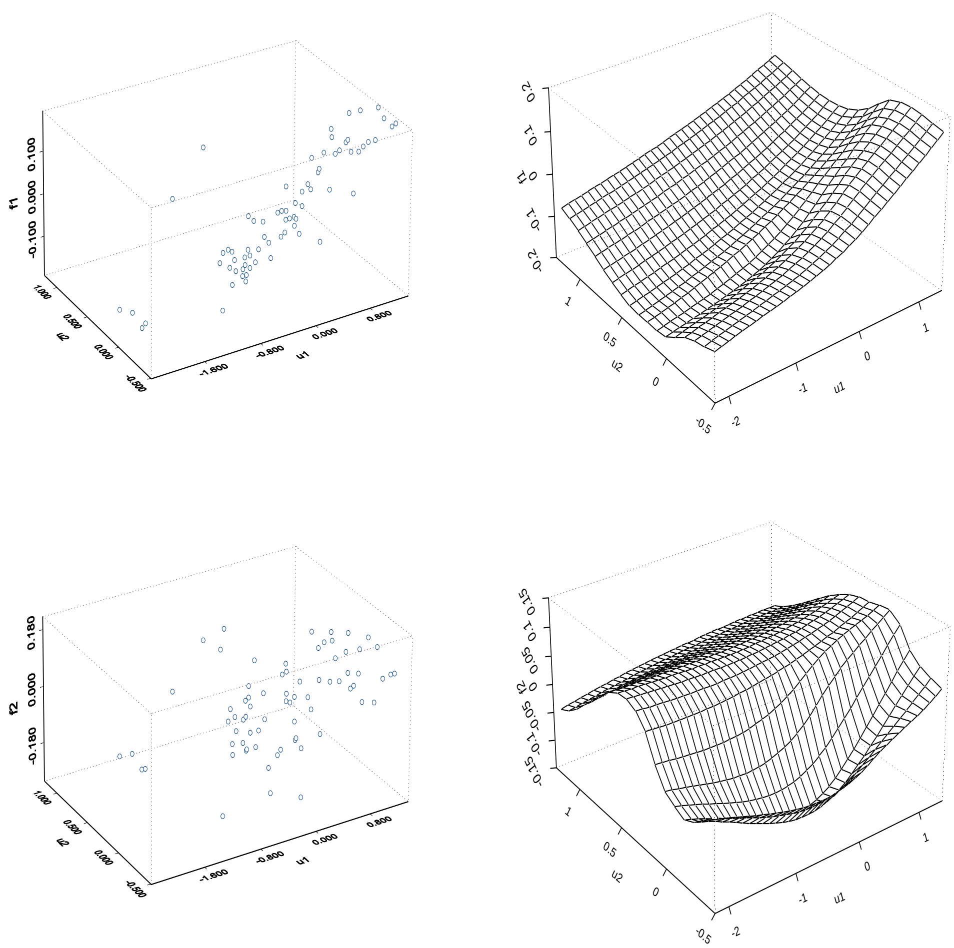


(c) Time series plot of $\hat{b}X_t$.

Figure 5: The fitted SPARR model for the US hog data, with $r_1 = 2$ and $r_2 = 1$.



(a) Cluster analysis of the lynx dynamics.



(b) Smoothed plots and 3-D scatter plots of $\hat{f}_1$ and $\hat{f}_2$.

Figure 6: The fitted SPARR model for the Lynx data, with $r_1 = 2$ and $r_2 = 2$.