

A semiparametric multivariate partially linear model: a difference approach

Lawrence D. Brown, Michael Levine, and Lie Wang

University of Pennsylvania, Purdue University, and MIT

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

Key words and phrases: Multivariate semiparametric model, difference-based method, asymptotic efficiency, partial linear model, random field.

1. Introduction

Semiparametric models have a long history in statistics and have received considerable attention in the last 30 – 40 years. They have also been a subject of continuing investigation in subject areas such as econometrics. The main reason they are considered is that sometimes the relationships between the response and predictors are very heterogeneous in the same model. Some of the relationships are clearly linear whereas other ones are much harder to categorize. In many situations, a small subset of variables is presumed to have an unknown relationship with the response that is modeled nonparametrically while the rest are assumed to have a linear relationship with it. As an example, Engle, Granger, Rice and Weiss (1986) studied the nonlinear relationship between temperature

and electricity usage where other related factors, such as income and price, are parameterized linearly. Shiller (1984) considered an earlier cost curve study in the utility industry using a partial linear model.

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

$$Y_i = a + X_i' \beta + f(U_i) + \varepsilon_i \quad (1.1)$$

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

The bibliography concerning the case of $q = 1$ is very extensive and we refer readers to Wang, Brown and Cai (2011) for details. The case where $q > 1$ has received much less attention in the past. Some of the papers that discussed that model are He and Shi (1996), Samarov, Spokoiny and Vial (2006), Schick (1996) and Müller, Schick and Wefelmeyer (2012). All of them considered random design case only.

In this paper, we consider the estimation of both parametric and nonparametric components. The difference sequence approach utilized in Wang, Brown and Cai (2011) is generalized so that it can be used when $q > 1$. In the fixed design case, the model is best enumerated using multivariate indices. Such a model is, effectively, a semiparametric random field model. Let n be the sample size; then, using differences of observations, a \sqrt{n} -consistent estimator of the parametric component and a \sqrt{n} -consistent estimator of the intercept are constructed; to obtain \sqrt{n} rate of convergence for the intercept a , the smoothness of a nonparametric component must exceed $q/2$. As is the case in Wang, Brown and Cai (2011), the correlation between differences has to be ignored and the

ordinary least squares approach must be used instead of the generalized least squares to obtain an optimal estimator. The reason for that is that the use of weighted least squares to estimate β in case of the difference model is conceptually analogous to use of ordinary least squares for the original model (1.1). In either case, the estimate $\hat{\beta}$ will be seriously biased due to the presence of the nonparametric component f . A similar remark is also made in Wang, Brown and Cai (2011) on p. 5. These estimators can be made asymptotically efficient if the order of the difference sequence is allowed to go to infinity. The estimator of the nonparametric component is defined by using a kernel regression on the residuals and is found to be $n^{-\alpha/(2\alpha+q)}$ consistent. The hypotheses testing problem for the linear coefficients is also considered and an F-statistic is constructed. The asymptotic power of the F-test is found to be the same as if the nonparametric component is known.

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

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

The paper is organized as follows. Section 2 discusses the fixed design case

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

2. Deterministic design

We consider the following semiparametric model

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

where $X_{\mathbf{i}} \in \mathbb{R}^p$, $U_{\mathbf{i}} \in S = [0, 1]^q \subset \mathbb{R}^q$, $\varepsilon_{\mathbf{i}}$ are iid zero mean random variables with variance σ^2 and finite absolute moments of the order $\delta + 2$ for some small $\delta > 0$: $E|\varepsilon_{\mathbf{i}}|^{\delta+2} < \infty$. The assumption of random $X_{\mathbf{i}}$'s is also very common when a multivariate nonparametric model is considered; generally speaking, the use of random $X_{\mathbf{i}}$'s corresponds to the assumption that $X_{\mathbf{i}}$'s are generated by observational (and not experimental) data, which is much more common in practice. In the model (2.1), $\mathbf{i} = (i_1, \dots, i_q)'$ is a multidimensional index; throughout this article, we will use bold font for all multivariate indices and a regular font for scalar ones. Each $i_k = 0, 1, \dots, m$ for $k = 1, \dots, q$; thus, the total sample size is $n = (m + 1)^q$. This assumption ensures that $m = o(n)$ as $n \rightarrow \infty$. We will say that two indices $\mathbf{i}^1 = (i_1^1, \dots, i_q^1) \leq \mathbf{i}^2 = (i_1^2, \dots, i_q^2)$ if $i_k^1 \leq i_k^2$ for any $k = 1, \dots, q$; the relationship between \mathbf{i}^1 and \mathbf{i}^2 is that of partial ordering. Also, for a multivariate index \mathbf{i} $|\mathbf{i}| = |i_1| + \dots + |i_q|$. Here we assume that $U_{\mathbf{i}}$ follows a fixed equispaced design: $U_{\mathbf{i}} = (u_{i_1}, \dots, u_{i_q})' \in \mathbb{R}^q$ where each coordinate is $u_{i_k} = \frac{i_k}{m}$. Also, β is an unknown p -dimensional vector of parameters and a is an unknown intercept term. We assume that $X_{\mathbf{i}}$'s are independent random vectors and that $X_{\mathbf{i}}$ is also independent of $\varepsilon_{\mathbf{i}}$; moreover, we denote the non-singular covariance matrix of X as Σ_X . For convenience, we also denote $N = \{0, \dots, m\}^q$. Note that in this model the intercept a cannot be absorbed in the design matrix X due to identifiability issues; in order to ensure that the model is identifiable, we have to require that an identifiability condition $\int_{[0,1]^q} f(u) du = 0$ is satisfied. Otherwise, one can add and subtract $\int_{[0,1]^q} f(u) du$ to the right hand side of the model with the new constant becoming $a' = a + \int_{[0,1]^q} f(u) du$. Finally, the version of (2.1) with $q = 1$ has been considered earlier in Wang, Brown and Cai (2011).

We will follow the same approach as Wang, Brown and Cai (2011), estimat-

ing first the vector coefficient β using the difference approach and then using residuals from that fit to estimate both the intercept a and the unknown function f . To obtain uniform convergence rates for the function f , some smoothness assumptions need to be imposed first. For this purpose, we consider functions f that belong to the Lipschitz ball class $\Lambda^\alpha(M)$ for some positive constant M that is defined as follows. For a q -dimensional index $\mathbf{j} = (j_1, \dots, j_q)$, we define $\mathbf{j}(l) = \{\mathbf{j} : |\mathbf{j}| = j_1 + \dots + j_q = l\}$. Then, for any function $f : \mathbb{R}^q \rightarrow \mathbb{R}$, $\frac{D^{\mathbf{j}(l)}f}{\partial u_1^{j_1} \dots \partial u_q^{j_q}}$ is defined for all \mathbf{j} such that $|\mathbf{j}| = l$. Then, the Lipschitz ball $\Lambda^\alpha(M)$ consists of all functions $f(u) : [0, 1]^q \rightarrow \mathbb{R}$ such that $|D^{\mathbf{j}(l)}f(u)| \leq M$ for $l = 0, 1, \dots, \lfloor \alpha \rfloor$ and $|D^{\mathbf{j}(\lfloor \alpha \rfloor)}f(v) - D^{\mathbf{j}(\lfloor \alpha \rfloor)}f(w)| \leq M\|v - w\|^{\alpha'}$ with $\alpha' = \alpha - \lfloor \alpha \rfloor$. Here and in the future, $\|\cdot\|$ stands for the regular l_2 norm in \mathbb{R}^q .

As in Cai, Levine and Wang (2009), our approach will be based on differences of observations $Y_{\mathbf{i}}$. The differences of an arbitrary order must be carefully defined when indices are multivariate. Let A be an arbitrary set in \mathbb{R}^q . It is clear that we need to specify a particular choice of observations that form a difference since there are many possibilities for a difference of any order “centered” around an observation $Y_{\mathbf{i}}$. As in Cai, Levine and Wang (2009) and Munk, Bissantz, Wagner and Freitag (2005), we choose a positive integer $\gamma \geq 2$ and select a set of q -dimensional indices $J = \{(0, \dots, 0), (1, \dots, 1), \dots, (\gamma, \dots, \gamma)\}$. For any vector $u \in \mathbb{R}^q$, a real number v and a set A , we define the set $B = u + vA = \{y \in \mathbb{R}^q : y = u + va, a \in A \subset \mathbb{R}^q\}$; then, we introduce a set R that consists of all indices $\mathbf{i} = (i_1, \dots, i_q)$ such that $R + J \equiv \{(\mathbf{i} + \mathbf{j}) | \mathbf{i} \in R, \mathbf{j} \in J\} \subset N$. In order to define a difference of observations of order γ , we define first a sequence of real numbers $\{d_j\}$ such that $\sum_{j=0}^{\gamma} d_j = 0$ and $\sum_{j=0}^{\gamma} d_j^2 = 1$. The latter assumption makes the sequence $\{d_j\}$ normalized. Moreover, denote $c_k = \sum_{i=0}^{\gamma-k} d_i d_{i+k}$. Note that the so-called polynomial sequence used in Wang, Brown, Cai and Levine (2008) with $d_j = \binom{\gamma}{j} (-1)^j / (\binom{2\gamma}{\gamma})^{1/2}$ satisfies this asymptotic requirement; moreover, it also satisfies an important property that $\sum_{j=0}^{\gamma} d_j j^k = 0$ for any power $k = 1, \dots, \gamma$. For the asymptotic optimality results that will be described later, the order of the difference sequence γ must go to infinity as $n \rightarrow \infty$. Then the difference of order γ “centered” around the point $Y_{\mathbf{i}}$, $\mathbf{i} \in R$ is defined as

$$D_{\mathbf{i}} = \sum_{\mathbf{j} \in J} d_{\mathbf{j}} Y_{\mathbf{i} + \mathbf{j}} \quad (2.2)$$

where $\mathbf{j} = (j, \dots, j)'$, $j = 0, \dots, \gamma$ is now a q -dimensional vector of identical coordinates and $d_{\mathbf{j}}$ is the same as d_j . Note that this particular choice of the set J makes numbering of difference coefficients $d_{\mathbf{j}}$ very convenient; since each q -dimensional index \mathbf{j} consists of only identical scalars, that particular scalar can be thought of as a scalar index of d ; thus, $\sum_{\mathbf{j} \in J} d_{\mathbf{j}}$ is the same as $\sum_{j=0}^{\gamma} d_j$. With this clarification in mind, we will use regular, and not the bold, font notation for indices of coefficients $d_{\mathbf{j}}$ from now on whenever the needed sum only includes the coefficients and, possibly, indices themselves; this guarantees that no confusion can occur.

Now, let $Z_{\mathbf{i}} = \sum_{\mathbf{j} \in J} d_{\mathbf{j}} X_{\mathbf{i}+\mathbf{j}}$, $\delta_{\mathbf{i}} = \sum_{\mathbf{j} \in J} d_{\mathbf{j}} f(U_{\mathbf{i}+\mathbf{j}})$, and $\omega_{\mathbf{i}} = \sum_{\mathbf{j} \in J} d_{\mathbf{j}} \varepsilon_{\mathbf{i}+\mathbf{j}}$, for any $\mathbf{i} \in R$. Then, by differencing the original model (2.1), one obtains

$$D_{\mathbf{i}} = Z'_{\mathbf{i}} \beta + \delta_{\mathbf{i}} + \omega_{\mathbf{i}} \quad (2.3)$$

for all $\mathbf{i} \in R$. The ordinary least squares solution for β can be written as

$$\hat{\beta} = \operatorname{argmin}_{\beta} \sum_{\mathbf{i} \in R} (D_{\mathbf{i}} - Z'_{\mathbf{i}} \beta)^2$$

Our interest lies in establishing consistency and asymptotic distribution for the least squares $\hat{\beta}$ as $n = (m+1)^q \rightarrow \infty$. We are going to prove the following result.

Theorem 2.1. *Let the distribution of $\varepsilon_{\mathbf{i}}$ have an absolute finite moment of order $2+\delta$ for some small $\delta > 0$. Also, let us assume that the marginal density function of $\varepsilon_{\mathbf{i}}$ $h(x)$ has a bounded variation over the real line. We also assume that the Lipschitz index α satisfies $\alpha > \frac{q}{2}$. Then,*

1. *if a difference sequence d_j of order $\gamma \geq \lfloor \alpha \rfloor$ such that $\sum_{j=0}^{\gamma} d_j = 0$, $\sum_{j=0}^{\gamma} d_j^2 = 1$, $\sum_{j=0}^{\gamma} d_j j^k = 0$ for $k = 1, \dots, \gamma$ is chosen, the resulting least squares solution is asymptotically normal in the sense that*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{L} N \left(0, \sigma^2 \Sigma_X^{-1} \left(1 + O \left(\frac{1}{\gamma} \right) \right) \right).$$

In the above, the term $O \left(\frac{1}{\gamma} \right)$ can be written explicitly as

$$2 \sum_{k=1}^{\gamma} \sum_{i=0}^{\gamma-k} d_i d_{i+k}.$$

2. The resulting least squares estimator $\hat{\beta}$ is not asymptotically efficient if the difference sequence order γ is finite. However, if we let $\gamma \rightarrow \infty$ while $\gamma = o(m)$ and $\sum_{j=0}^{\gamma} |d_j| j^l < \infty$ for some $l > q/2$, the asymptotic efficiency is achieved.

Our next step is to obtain properties of the estimated intercept \hat{a} . For simplicity of notation, let $\mathbf{n} = (n, \dots, n)$ be a q -dimensional index consisting of n 's. The natural estimator $\hat{a} = \frac{1}{n} \sum_{\mathbf{i} \leq \mathbf{n}} (Y_{\mathbf{i}} - X_{\mathbf{i}}' \hat{\beta})$ can be used. Its properties can be described in the following lemma.

Lemma 2.2. *Under the assumption of the uniform design on $S = [0, 1]^q$ and $\alpha > q/2$, we have*

$$\sqrt{n}(\hat{a} - a) \xrightarrow{L} N(0, \sigma^2)$$

Next, the estimation of the function f is an important task. One of the ways to do this is to apply a smoother to the residuals $r_{\mathbf{i}} = Y_{\mathbf{i}} - \hat{a} - X_{\mathbf{i}}' \hat{\beta}$; out of the many possible smoothers, we choose a multivariate kernel smoother defined as a product of the univariate kernels. More specifically, let $K(U^l)$ be a univariate kernel function for a specific coordinate U^l , $l = 1, \dots, q$ satisfying $\int K(U^l) dU^l = 1$ and having $[\alpha]$ vanishing moments. We choose the asymptotically optimal bandwidth $h = n^{-1/(2\alpha+q)}$ (see, for example, J. Fan and I. Gijbels (1995)). We define its rescaled version as $K_h(U^l) = h^{-1}K(h^{-1}U^l)$ so that the q -dimensional rescaled kernel is $K_h(U) = h^{-q} \prod_{l=1}^q K(h^{-1}U^l)$. Wang, Brown and Cai (2011) used Gasser-Müller kernel weights to smooth the residuals $r_{\mathbf{i}}$ in the one-dimensional case. In the multivariate case, it is clearly preferable to use some other approach to define weights that add up to 1; the classical Nadaraya-Watson approach is the one we choose. The Nadaraya-Watson kernel weights are defined as

$$W_{\mathbf{i},h}(U - U_{\mathbf{i}}) = \frac{K_h(U - U_{\mathbf{i}})}{\sum_{\mathbf{i} \leq \mathbf{n}} K_h(U - U_{\mathbf{i}})}.$$

Finally, the resulting kernel estimator of the function $f(U)$ can then be defined as

$$\hat{f}(U) = \sum_{\mathbf{i} \leq \mathbf{n}} W_{\mathbf{i},h}(U - U_{\mathbf{i}}) r_{\mathbf{i}}$$

Theorem 2.3. For any Lipschitz indicator $\alpha > 0$ and any $U_0 \in [0, 1]^q$, the estimator \hat{f} satisfies

$$\sup_{f \in \Lambda^\alpha(M)} \mathbb{E}[(\hat{f}(U_0) - f(U_0))^2] \leq Cn^{-2\alpha/(2\alpha+q)}$$

for a constant $C > 0$. Also, for any $\alpha > 0$,

$$\sup_{f \in \Lambda^\alpha(M)} \mathbb{E} \left[\int_{[0,1]^q} (\hat{f}(U) - f(U))^2 dU \right] \leq Cn^{-2\alpha/(2\alpha+q)}$$

Remark 2.4. He and Shi (1996) considered the model (1.1) for the random design case and provided an estimation approach for both parametric and non-parametric parts that uses a bivariate tensor-product B-splines based method; the resulting method was illustrated in detail for the case of $q = 2$. They also noted that the optimal result for the mean squared error of the nonparametric component requires that the degree of smoothness of that component r increases with the dimension q as $r > q/2$, similarly to our result obtained using the difference sequence approach.

3. Random Design Case

So far, we have only considered the deterministic setting whereby the function $f(U)$ is defined on $S = [0, 1]^q \in \mathbb{R}^q$. In the multivariate setting, this means using a grid with each observation $U_i = (u_{i_1}, \dots, u_{i_q})' \in \mathbb{R}^q$ and defining each coordinate as $u_{i_k} = \frac{i_k}{m}$. It is also interesting to consider the random design case where the argument $U \in \mathbb{R}^q$ is random and not necessarily independent of X . We note that in this case the use of multivariate indices does not result in any added convenience so we use the standard univariate ones.

Now, our model is again

$$Y_i = a + X_i' \beta + f(U_i) + \varepsilon_i \tag{3.1}$$

for $i = 1, \dots, n$; we also assume that $(X_i', U_i) \in \mathbb{R}^p \times \mathbb{R}^q$ are independent with an unknown joint density $g(x, u)$. Moreover, we assume that the conditional covariance matrix $\Sigma_* = \mathbb{E}[(X_1 - \mathbb{E}(X_1|U_1))(X_1 - \mathbb{E}(X_1|U_1))']$ is non-singular. Next, $\beta \in \mathbb{R}^p$ is the vector of coefficients, and ε_i are independent identically distributed random variables with mean zero and variance σ^2 that are independent of (X_i', U_i) . To make the model identifiable, we also need to assume that

$\mathbb{E}(f(U_i)) = 0$. Finally, an individual coordinate of the vector X_i will be denoted X_i^l , for $l = 1, \dots, p$.

One's first inclination is to try to order multivariate observations U_i in some way in order to form a difference sequence. This would be a direct analogy to what was done in Wang, Brown and Cai (2011). While there is a number of ways to do so (e.g. by using the lexicographical ordering that results in the complete, and not just partial, order), the resulting sequence is of little use in estimation of the function f at any particular point U . Speaking heuristically, the reason for that is that it is impossible to keep such an ordering and ensure that, at the same time, the points remain in a neighborhood of the point U . Due to this, such a direct generalization is impossible.

The above discussion suggests a different way out.

For a given point U_i , we choose all the points U such that the squared Euclidean norm $\|U_i - U\|^2 \leq \nu$ for some small $\nu > 0$. The number of such points will, of course, depend on ν and, moreover, it will be different for different points U_i as long as the marginal distribution of U_i is non-trivial. Let all of the points that are within the squared distance ν of U_i be denoted $U_{i,t}$ where $t = 0, \dots, \gamma_i(\nu)$. Then, a difference ‘‘centered’’ on the point U_i will be $\delta_i = \sum_{t=0}^{\gamma_i(\nu)} d_t f(U_{i,t})$. Applying this difference to both sides of (3.1), one obtains

$$D_i = Z_i' \beta + \delta_i + \omega_i \quad (3.2)$$

where $D_i = \sum_{t=0}^{\gamma_i(\nu)} d_t Y_{i+t}$, $Z_i = \sum_{t=0}^{\gamma_i(\nu)} d_t X_{i+t}$, and $\omega_i = \sum_{t=0}^{\gamma_i(\nu)} d_t \varepsilon_{i+t}$, $i = 1, \dots, n$. Note that, as opposed to the fixed design case, the difference sequence considered here is of a *variable* order that depends on the value of the marginal density function $g(U_i)$ at which the function f is to be estimated as well as the ‘‘tuning’’ parameter ν . For simplicity, we will suppress the dependence of the difference order on ν and write simply γ_i , unless indicated otherwise.

As before, the sequence is defined in such a way that $\sum_{j=0}^{\gamma_i} d_j = 0$, $\sum_{j=0}^{\gamma_i} d_j^2 = 1$, $\sum_{j=0}^{\gamma_i} d_j j^k = 0$ for $k = 1, \dots, \gamma_i$. We will also denote

$$c_{ij} = \sum_{t=0}^{\min(\gamma_i, \gamma_j) - (i-j)} d_t d_{t+(i-j)}$$

and $c_n = \sum_{i,j=1}^n c_{ij}^2$.

In the matrix form the model (3.2) can be written as

$$D = Z\beta + \delta + \omega \quad (3.3)$$

where Z is the matrix whose i th row is Z'_i , $D = (D_1, \dots, D_n)'$, $\omega = (\omega_1, \dots, \omega_n)'$, and $\delta = (\delta_1, \dots, \delta_n)'$. The least squares solution is, then,

$$\hat{\beta} = (Z'Z)^{-1}Z'D \quad (3.4)$$

Now, the following result can be established.

Theorem 3.5. *Let the marginal density function of U_i $g(u)$ be bounded everywhere away from zero uniformly on \mathbb{R}^q . Also, let the function $f(U) \in \Lambda^\alpha(M_f)$ and $h(U) \equiv \mathbb{E}(X|U) \in \Lambda^\rho(M_h)$. Define the difference based estimator of β as above in (3.4) with $\nu_n \rightarrow 0$ as $n \rightarrow \infty$. Then, as long as $\sup_{1 \leq i \leq n} \gamma_i(\nu) \nu^{2(\rho+\alpha)} \rightarrow 0$ while $n \rightarrow \infty$ and $\lim_{n \rightarrow \infty} c_n = 0$, the estimator $\hat{\beta}$ is asymptotically normal and efficient. More precisely,*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{L} N(0, \sigma^2 \Sigma_*^{-1})$$

where $\Sigma_* = \mathbb{E}[(X - \mathbb{E}(X|U))(X - \mathbb{E}(X|U))']$.

Remark 3.6. *Requiring that the marginal density function $g(u)$ be bounded uniformly away from zero is not the weakest possible assumption - moderate rates of growth to infinity can be permitted as well at the expense of making ν go to zero faster as $n \rightarrow \infty$. We do not pursue this question further here*

The asymptotic normality of the intercept, stated in the Lemma 2.2 in the fixed design case, is also valid in the random design case. More specifically, we can say that under the assumptions of Theorem 3.5, we have again $\sqrt{n}(\hat{a} - a) \xrightarrow{L} N(0, \sigma^2)$. One can also show that the analogue of Theorem 2.3 is also true in the random design case. For this to be valid, we need to require that the design density $g(u)$ be uniformly bounded everywhere away from zero over \mathbb{R}^q . Since the proofs of these results are very similar to proofs of Lemma 2.2 and Theorem 2.3, respectively, we do not give them here in full.

4. Linear component related tests

In this section we consider testing of linear hypotheses of the type $H_0 : C\beta = 0$ vs. $H_a : C\beta \neq 0$ for some full-rank $r \times p$ matrix C with $\text{rank}(C) = r$;

here r is the number of hypotheses tested. In this section, we only consider a fixed design case. It is assumed that the errors are independent and normally distributed, that is $\varepsilon_{\mathbf{i}} \sim N(0, \sigma^2)$ for some $\sigma^2 > 0$. A random design case is substantially more difficult and will be part of our future research. To estimate the error variance σ^2 , for any $\mathbf{i} \in R$ we define the estimated i th residual as $e_{\mathbf{i}} = D_{\mathbf{i}} - Z_{\mathbf{i}}' \hat{\beta} = D_{\mathbf{i}} - Z_{\mathbf{i}}' (\sum_{\mathbf{s} \in R} Z_{\mathbf{s}} Z_{\mathbf{s}}')^{-1} \sum_{\mathbf{s} \in R} Z_{\mathbf{s}} D_{\mathbf{s}}$ and, therefore, the estimated error variance as

$$\hat{\sigma}^2 = \frac{\sum_{\mathbf{i} \in R} e_{\mathbf{i}}^2}{n - \gamma - p} \quad (4.1)$$

Theorem 4.7. *Suppose $\alpha > q/2$ and $1 - d_0 = O(\gamma^{-1})$. In order to be able to test $H_0 : C\beta = 0$ vs. $H_1 : C\beta \neq 0$ where C is an $r \times p$ matrix with $\text{rank}(C) = r$, the test statistic*

$$F = \frac{\hat{\beta}' C' (C (\sum_{\mathbf{s} \in R} Z_{\mathbf{s}} Z_{\mathbf{s}}')^{-1} C')^{-1} C \hat{\beta} / r}{\hat{\sigma}^2}$$

is asymptotically distributed as $F(r, n - \gamma - p)$ distribution under the null hypothesis.

5. Simulation

We begin with the fixed design discussion. We select the sample size $n = 2500$, define $T_i \sim \text{Uniform}(0, 1)$ for $i = 1, \dots, n$ and consider two possible designs of the parametric component. In the first case, dimensionality of the linear component is $p = 1$ and the true coefficient is $\beta = 2$; the one-dimensional random variable $X_i \sim N(\mu_i, 1)$ for $i = 1, \dots, n$ with $\mu_i = T_i$. For the second case, we denote a 3×3 identity matrix I_3 . Then, we select $p = 3$, $\beta = (2, 2, 4)'$ and $X_i = (X_i^1, X_i^2, X_i^3)' \sim N((\mu_i, 2\mu_i, 4\mu_i^2)', I_3)$ where, again, $\mu_i = T_i$. In both cases, errors are generated from the standard normal distribution. We select the dimensionality of the functional argument to be $q = 2$ and use the sample size $n = 2500$. Four possible choices of functions are considered: $f_1(U) = U_1^2 + U_2^4$, $f_2 = 5 \sin(\pi(U_1 + U_2))$, $f_3 = \min(U_1, 1 - U_1) + \min(U_2, 1 - U_2)$ and $f_4(U) = f_4^1(U_1) * f_4^2(U_2)$ where $f_4^1(U_1) = |4 * U_1 - 2|$ and $f_4^2(U_2) = \frac{|4U_2 - 2| + 1}{2}$. The first two choices are taken from Yang and Tschernig (1999) where they were used to study bandwidth selection for the multivariate polynomial regression. The third function brings discontinuities in our experimental setting. The fourth is the so-called g-Sobol function, commonly used for sensitivity analysis (see, e.g.

Table 5.1: Fixed design case: the MSE's of estimate $\hat{\beta}$ over 200 replications with sample size $n = 2500$. The numbers inside parentheses are the standard deviations.

	$f \equiv 0$	f_1	f_2	f_3	f_4
Case(1)	0.0004 (0.0006)	0.0005 (0.0007)	0.0006 (0.0008)	0.0006 (0.0008)	0.0005 (0.0006)
Case (2)	0.0028 (0.0022)	0.0028 (0.0024)	0.0038 (0.0034)	0.0030 (0.0025)	0.0026 (0.0024)

Saltelli (2000) and Touzani and Busby (2011)). It is strongly nonlinear and non-monotonic. Finally, we use a difference sequence with the parameter $\gamma = 4$.

First, we want to assess the influence of the unknown function f on the estimation of the linear component. We use 200 Monte-Carlo runs and the mean squared error is defined as $\|\hat{\beta} - \beta\|_2^2$ with $\|\cdot\|_2$ being the Euclidean norm. The results are presented in the Table (5.1). The first row corresponds to the first choice of one-dimensional parametric component and the second row corresponds to the choice of the three-dimensional parametric component. We look at the mean squared error of $\hat{\beta}$ when there is no functional component ($f \equiv 0$) as well as for all four choices of f described above.

The estimation procedure seems to function reasonably well even if the non-parametric component is highly nonlinear and nonmonotonic. The presence or absence of such a component makes almost no difference in the size of the mean squared error of the parametric component estimator $\hat{\beta}$.

As a next step, we estimate the nonparametric component for both choices of the parametric component design. For comparison purposes, we also illustrate it when the parametric component is equal to zero. We are using the multivariate Nadaraya-Watson estimator and select the optimal bandwidth using the cross-validation approach. Since the test functions used are not symmetric, different bandwidths are assumed for different coordinates. Note that the Priestley-Chao kernel used in Wang, Brown and Cai (2011) is not as convenient for multivariate settings and therefore we prefer not to use it in this case. The results are summarized in the Table (5.2). It is clear that the choice of the parametric design does not have any perceptible influence on estimation of nonparametric component. All of the function choices can be fairly precisely estimated, even those that are strongly nonlinear.

As a next step, we want to verify how well our estimation procedures perform

Table 5.2: Fixed design case: the MSE's of estimate \hat{f} over 200 replications with sample size $n = 2500$. The numbers inside parentheses are the standard deviations

	$f \equiv 0$	f_1	f_2	f_3	f_4
$\beta = 0$ known in advance	0.0009 (0.0018)	0.0081 (0.0024)	0.0347 (0.0047)	0.0073 (0.0020)	0.0162 (0.0032)
Case(1)	0.0013 (0.0021)	0.0096 (0.0036)	0.0372 (0.0060)	0.0093 (0.0039)	0.0185 (0.0047)
Case (2)	0.0013 (0.0019)	0.0094 (0.0031)	0.0371(0.0054)	0.0088 (0.0033)	0.0178(0.0044)

in the random design case. We will use the same selection of functions only now we assume that $V^j \sim Unif[0, 1]$, $j = 1, 2$ and each argument of the function f is a two-dimensional point $V_i = (V_i^1, V_i^2)$, $i = 1, \dots, n$. Thus, the functions considered are, again, $f_1(V) = (V^1)^2 + (V^2)^4$, $f_2 = 5 \sin(\pi(V^1 + V^2))$, $f_3 = \min(V^1, 1 - V^1) + \min(V^2, 1 - V^2)$ and $f_4(V) = f_4^1(V^1) * f_4^2(V^2)$ where $f_4^1(V^1) = |4 * V^1 - 2|$ and $f_4^2(V^2) = \frac{|4V^2 - 2| + 1}{2}$. In order to choose the order of the difference sequence, we use the nearest neighbor principle. Whenever $f(V)$ needs to be estimated, we specify first a small $\nu > 0$ and then select the difference based on the points V_i such that $\|V_i - V\|^2 \leq \nu$. In this particular case, we use $\nu = 0.05$. If there are no points in such a neighborhood of V , we take the smallest possible number of points which is 2 and select two of the nearest neighbors of the point V .

We begin, again, with estimation of the parametric component. There are 200 Monte-Carlo runs, the sample size is $n = 2500$, and the mean squared error is defined as $\|\hat{\beta} - \beta\|_2^2$ with $\|\cdot\|_2$ being the Euclidean norm. To illustrate the fact that accounting for the presence of a nonparametric component in the model is crucial, we also conduct estimation of the Euclidean component using simple linear least squares that disregards the presence of the function f . The results are summarized in the Table (5.3) whose last two rows illustrate what happens if the presence of the nonparametric component is ignored and the standard least squares estimation procedure is applied.

Note that the mere presence of a nonparametric component clearly does not have much influence on the estimation of the parametric part if our difference sequence method is applied. However, simply ignoring the presence of the nonparametric component and applying the standard least squares method produces bad results; indeed, the results in the last two rows of (5.3) are much worse than those in the first two rows with an exception of the first column. The rest of

Table 5.3: Random design case: the MSE's of estimate $\hat{\beta}$ over 200 replications with sample size $n = 2500$. The numbers inside parentheses are the standard deviations. The first two rows assume that the functional component has been taken into account

	$f \equiv 0$	f_1	f_2	f_3	f_4
Case(1)	0.0010 (0.0014)	0.0009 (0.0014)	0.0011 (0.0016)	0.0011 (0.0017)	0.0009 (0.0011)
Case (2)	0.0040 (0.0032)	0.0051 (0.0049)	0.0044 (0.0036)	0.0051 (0.0045)	0.0046 (0.0042)
Case(1)- functional component ignored	0.0001 (0.0001)	0.0306 (0.0030)	0.0021 (0.0018)	0.0251 (0.0025)	0.1084 (0.0066)
Case (2)- functional component ignored	0.0008 (0.0007)	0.0332 (0.0038)	0.0128 (0.0105)	0.0275 (0.0036)	0.1151 (0.0098)

Table 5.4: Random design case: the MSE's of estimate \hat{f} over 200 replications with sample size $n = 2500$. The numbers inside parentheses are the standard deviations

	$f \equiv 0$	f_1	f_2	f_3	f_4
$\beta = 0$ known in advance	0.0004 (0.0006)	0.0077 (0.0020)	0.0384 (0.0051)	0.0076 (0.0021)	0.0169 (0.0032)
Case(1)	0.0009 (0.0013)	0.0102 (0.0045)	0.0409 (0.0070)	0.0102 (0.0044)	0.0194 (0.0055)
Case (2)	0.0009 (0.0014)	0.0118 (0.0063)	0.04213(0.0091)	0.0121 (0.0078)	0.0208(0.0065)

mean squared errors in those two rows are several orders of magnitude larger than those in the first two rows of the Table (5.3). The difference is especially pronounced for g-Sobol function choice due to its obvious "roughness".

Our next check is the estimation of the nonparametric component in the random design case. Again, we are using the multivariate Nadaraya-Watson estimator and select the optimal bandwidth using the cross-validation approach. Since the test functions used are not symmetric, different bandwidths are assumed for different coordinates. For comparison, the nonparametric component has also been estimated in the case where $\beta = 0$. The sample size used is $n = 2500$ and there are 200 Monte-Carlo runs. We also use $\nu = 0.05$ to define the nearest neighborhood of any point U where the function f has to be estimated. The Table (5.4) summarizes mean squared errors (MSE's) of the estimated function f . As is true for the fixed design case, note that MSE's in each column are yet again quite close to each other and the performance of the estimator \hat{f} does not seem to depend a lot on the structure of X and β .

It is also a matter of substantial interest to check how the performance of the proposed method in the fixed design case depends on the length of the difference sequence used. More specifically, we focus on the Case (2) and the function $f = f_2$ and compute mean squared errors of the estimated function and the parametric component coefficients for several choices of the difference sequence

length. There are $n = 2500$ observations used and 200 Monte-Carlo replications have been used. The chosen lengths of the difference sequences are 2, 4, 8 and 16. The results are summarized in the Table (5.5). Note that the performance of our method seems to deteriorate as the length of the difference sequence grows. We believe that there are two reasons for that phenomenon. First, for sufficiently small difference sequence lengths, such as $\gamma = 2$ or $\gamma = 4$, the fact that we use a diagonal difference sequence to estimate the model parameters implies the use of only a small proportion of the total number of points in the q dimensional space. This effectively precludes us from observing the effect of bias reduction that is typically associated with the use of difference sequence method (see, for example, Wang, Brown, Cai and Levine (2008)). We conjecture that if a non-diagonal difference method were used, the effect would be clearer; however, the method would be much more computationally intensive, the notation much harder, and the eventual asymptotic efficiency of obtained estimators uncertain. Second, we need to keep in mind that, for the estimator of the parametric component β to be efficient, the parameter γ must go to infinity as $n \rightarrow \infty$. In addition, the rate of growth of γ as a function of n must be $o(n)$ as $n \rightarrow \infty$. Due to the above, it appears likely that for our sample size $n = 2500$, larger choices of γ , such as $\gamma = 8$ or $\gamma = 16$, may simply be too large in comparison with the sample size. This suggests that, for a fixed sample size, improved performance for increasing orders of a difference sequence appears highly unlikely.

To provide some empirical evidence for our conjectures, we also performed some extra simulations with a larger sample size $n = 10,000$ and $m = 500$, over 200 Monte-Carlo runs. The results show yet again that there is no improvement in the average mean squared error of estimated parametric component as the parameter γ is taking values 2, 4, 8, 16. Due to similarity of these results to the Table (5.5) in the manuscript, we did not include them in the final version.

Finally, we would also like to illustrate the performance of our testing procedure in the fixed design case. The hypothesis tested is $H_0 : \beta_0 = \beta_1 = 0$. Each cell of the Table (5.6) contains the number of times this null hypothesis has been rejected out of 200 Monte-Carlo runs and the average value of F statistic over these runs.

Table 5.5: The mean and standard deviation of the estimated coefficients and the average MSE of estimate \hat{f}_2 over 200 replications with sample size $n = 2500$ for different difference sequence lengths. The numbers inside parentheses are the standard deviations

	$\gamma = 2$	$\gamma = 4$	$\gamma = 8$	$\gamma = 16$
Mean(sd) of $\hat{\beta}_1$	0.0012 (0.0019)	0.0014 (0.0023)	0.0025 (0.0032)	0.0059 (0.0071)
Mean(sd) of $\hat{\beta}_2$	0.0008 (0.0016)	0.0013 (0.0019)	0.0024 (0.0035)	0.0056 (0.0078)
Mean(sd) of $\hat{\beta}_3$	0.0005 (0.0006)	0.0009 (0.0010)	0.0011 (0.0014)	0.0030 (0.0041)
MSE of \hat{f}_2	0.0360 (0.0054)	0.0366 (0.0059)	0.0376 (0.0062)	0.0419 (0.0104)

Table 5.6: The total number of rejects of F test over 200 replications at level 0.05. The numbers inside the parentheses are the mean value of F statistic.

	$f \equiv 0$	f_1	f_2	f_3	f_4
$\beta = (0, 0, 4)'$	14 (1.1442)	13 (0.9640)	6 (0.8737)	12 (1.1611)	13 (1.1899)
$\beta = (2, 2, 4)'$	200 (9916.93)	200 (9654.72)	200 (2765.968)	200 (9858.734)	200 (9537.464)

References

- Bolthausen, E. (1982). On the central limit theorem for stationary mixing random fields. *The Annals of Probability*, 1047-1050
- Cai, Tony T., Levine, M. and Wang, L. (2009) Variance function estimation in multivariate nonparametric regression with fixed design. *Journal of Multivariate Analysis* **100(1)**, 126-136.
- Doukhan, P. (1994). Mixing (pp. 15-23). Springer New York.
- Engle, R. F., Granger, C. W., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, **81(394)**, 310-320.
- Fan, J. (1996). Local polynomial modelling and its applications (Vol. 66). CRC Press.
- Gaetan, C., and Guyon, X. (2010). Spatial Statistics and Modeling, Series in Statistics. Springer.
- Guyon, X. (1995). Random fields on a network: modeling, statistics, and applications. Springer
- He, X., and Shi, P. (1996). Bivariate tensor-product B-splines in a partly linear model. *Journal of Multivariate Analysis*, **58(2)**, 162-181.
- Müller, U. U., Schick, A., and Wefelmeyer, W. (2012). Estimating the error distribution function in semiparametric additive regression models. *Journal of Statistical Planning and Inference*, **142(2)**, 552-566.
- Munk, A., Bissantz, N., Wagner, T., and Freitag, G. (2005). On difference based variance estimation in nonparametric regression when the covariate is high dimensional. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67(1)**, 19-41.
- Saltelli, A., Chan, K., and Scott, E. M. (Eds.). (2000). Sensitivity analysis (Vol. 134). New York: Wiley.

- Samarov, A., Spokoiny, V., and Vial, C. (2005). Component identification and estimation in nonlinear high-dimensional regression models by structural adaptation. *Journal of the American Statistical Association*, **100(470)**.
- Schick, A. (1996). Root-n consistent estimation in partly linear regression models. *Statistics and Probability Letters*, **28(4)**, 353-358.
- Shiller, R. J. (1984). Smoothness priors and nonlinear regression. *Journal of the American Statistical Association*, **79(387)**, 609-615.
- Touzani, S., and Busby, D. (2011). Multivariate wavelet kernel regression method.
- Tschernig, R., and Yang, L. (2000). Nonparametric lag selection for time series. *Journal of Time Series Analysis*, **21(4)**, 457-487.
- Wang, L., Brown, L.D. and Cai, Tony T. (2011). A difference based approach to the semiparametric partial linear model. *Electronic Journal of Statistics* **5**, 619-641.
- Wang, L., Brown, L. D., Cai, Tony T., and Levine, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics*, **36(2)**, 646-664.

6. Appendix

Proof of Theorem 2.1

Proof. As a first step, note that the solution has the usual form

$$\hat{\beta} = \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} D_{\mathbf{i}} \right)$$

and that

$$\begin{aligned} \hat{\beta} - \beta &= \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} [\omega_{\mathbf{i}} + \delta_{\mathbf{i}}] \right) \\ &= \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} \sum_{\mathbf{i} \in R} Z_{\mathbf{i}} \omega_{\mathbf{i}} + \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} \sum_{\mathbf{i} \in R} Z_{\mathbf{i}} \delta_{\mathbf{i}}. \end{aligned} \tag{5.1}$$

Note that the following notation is needed in order to characterize the covariance array of $\omega_{\mathbf{i}}$. For any two q -dimensional indices \mathbf{i}, \mathbf{j} we say that $|\mathbf{i} - \mathbf{j}| = l$ if for all $k = 1, \dots, q$ $|i_k - j_k| = l$. With that in mind, a set of pseudoresiduals $\omega_{\mathbf{i}}$, $\mathbf{i} \in R$ has a covariance array $\Psi = \{\Psi_{\mathbf{i}, \mathbf{j}}\}$ $\mathbf{i}, \mathbf{j} \in R$ with only the elements having the “index distance” $l \leq \gamma$ and $l \neq 1$ being non-zero. We denote those non-zero elements c_l for any $1 < l \leq \gamma$. Because $\omega_{\mathbf{i}}$'s for all $\mathbf{i} \in R$ are linear combinations of $\varepsilon_{\mathbf{i}}$, all of c_l 's will depend on the difference sequence $\{d_j\}$. More precisely, the covariance array Ψ has a typical element

$$\Psi_{\mathbf{i}, \mathbf{j}} = \begin{cases} \sigma^2, & \text{if } \mathbf{i} = \mathbf{j} \\ \sigma^2 c_l, & \text{if } |\mathbf{i} - \mathbf{j}| = l \leq \gamma \\ 0, & \text{otherwise} \end{cases}$$

We will examine the two terms in the above separately. First, it is clear that the expectation of the first term $\mathbb{E} \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} \sum_{\mathbf{i} \in R} Z_{\mathbf{i}} \omega_{\mathbf{i}} = 0$ and its conditional variance

$$\begin{aligned} & \text{Var} \left[\left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} \sum_{\mathbf{i} \in R} Z_{\mathbf{i}} \omega_{\mathbf{i}} \middle| Z_{\mathbf{i}}, \mathbf{i} \in R \right] \\ &= \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} \text{Var} \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} \omega_{\mathbf{i}} \right) \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1}. \end{aligned}$$

It is necessary to guarantee that the weak law of large numbers (WLLN) is valid for a sequence of random matrices $Z_{\mathbf{i}} Z'_{\mathbf{i}}$ where $\mathbf{i} \in R$. We can state such a law of large numbers in the following manner. Let K be an identical copy of the index set J ; then,

$$\frac{1}{n} \sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} = \frac{1}{n} \sum_{\mathbf{i} \in R} \left[\left(\sum_{\mathbf{j} \in J} d_{\mathbf{j}} X_{\mathbf{i}+\mathbf{j}} \right) \left(\sum_{\mathbf{k} \in K} d_{\mathbf{k}} X'_{\mathbf{i}+\mathbf{k}} \right) \right] \xrightarrow{p} \Sigma_X.$$

To establish such a law, we can rely on a straightforward generalization of the law of large numbers for random fields that has been proved in Guyon (1995) as Theorem 3.2.1. For example, since the cardinality of R is $|R| = (m+1-\gamma)^q$, we can quickly verify that conditions (3.1) and (3.2) on p. 108 of Guyon (1995) are satisfied. The validity of these two conditions guarantees that the geometry of the domain over which the law of large numbers is considered is not problematic.

All that remains now is a straightforward generalization of this result to a matrix valued random field.

To conclude that the term $(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} \omega_{\mathbf{i}}) \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1}$ is (conditionally on the set of $Z_{\mathbf{i}}$) asymptotically normal we need to use a central limit theorem for stationary random fields; for example, a version cited in Guyon (1995) that is originally due to Bolthausen (1982) seems suitable for our circumstances. In order to verify mixing conditions, it is useful to consider some characteristics of the random field $\omega_{\mathbf{i}}$, $\mathbf{i} \in R$ first. Note that a field $\omega_{\mathbf{i}} = \sum_{j \in J} d_j \varepsilon_{\mathbf{i}+\mathbf{j}}$ is a linear transformation of $\varepsilon_{\mathbf{i}}$; alternatively, it can also be viewed as an infinite moving average. This allows us to use some well-known results on mixing properties for linear fields that have been described in detail in Guyon (1995) and Doukhan (1994). Note that these results are much stronger than what is technically required here since our central limit theorem only describes the mean over a fairly simple set R .

First, a brief introduction into strong mixing coefficients for a random field is needed. For the two sets $A, B \in \mathbb{R}$ we define the distance between them as $d(A, B) = \inf_{x \in A, y \in B} d(x, y)$ with d being a Euclidean metric in \mathbb{R} . Also, let $|A|$ and $|B|$ be the cardinality of sets A and B , respectively. For a random field X , a subset $X_C = \{X_t : t \in C\}$ for some subset C of q -dimensional indices is called a C -marginal of X . Let κ_A and κ_B be σ -algebras generated by X_A and X_B , respectively. For the two σ -algebras κ_A and κ_B , a strong mixing coefficient is defined as $\alpha(\kappa_A, \kappa_B) := \sup\{|P(U)P(V) - P(U \cap V)|; U \in \kappa_A, V \in \kappa_B\}$. Then, for two sets A and B a strong mixing coefficient $\alpha_X(A, B) := \alpha(\kappa_A, \kappa_B)$. Let u and v be two nonnegative integers; a special version of the general mixing coefficient $\alpha_X(A, B)$ for sufficiently ‘‘distant’’ sets A and B can be defined as $\alpha_X(k; u, v) = \sup\{\alpha_X(A, B) : d(A, B) \geq k, |A| \leq u, |B| \leq v\}$. Note that $\alpha_X(k; u, v)$ is an increasing function with respect to both u and v . Finally, we also denote $\alpha_X(k; u, \infty) = \sup_v \alpha_X(k; u, v)$.

To ensure that the central limit theorem is valid, we need to show that the strong mixing coefficient $\alpha_X(k; 2, \infty)$ of the field X decays sufficiently fast to satisfy the condition $\sum_{k \geq 1} k^{q-1} \alpha_X(k; 2, \infty)^{\zeta/2+\zeta}$ for some $\zeta > 0$. To do that, we will use Corollary 1 of the Theorem 1 of Doukhan (1994, pp 78-79) for the multivariate case (i.e., when $q > 1$). To ensure that all of the conditions mentioned in

the Theorem 1 are true, it is necessary to make certain assumptions on both the difference sequence $\{d_j\}$ and on the distribution function h of the independent random variables ε_i first. More specifically, we need to require that

- ε_i has a uniformly bounded absolute moment of order $2 + \delta$: $\sup_i E |\varepsilon_i|^{2+\delta} < \infty$ for some $\delta > 0$
- The density function h of ε_i possesses the following regularity property:

$$\int_{\mathbb{R}} |h(z+x) - h(z)| dz \leq C|x|$$

for some positive C that does not depend on x . This requirement is satisfied if the density function $h(x)$ has a bounded variation on a real line.

- The difference sequence d_j must satisfy the so-called invertibility condition (Guyon, 1995) that requires the existence of a sequence a_j such that the product of the two associated diagonal matrices $D = \text{diag}\{d_j\}$ and $A = \text{diag}\{a_j\}$ $DA = I$ with I being the unity matrix. To guarantee that this is true, it is necessary to require that for some $k > q/2$

$$\sum_i |i|^k |d_i| < \infty. \quad (5.2)$$

The reason we need to require this is because if we define $d(z) = \sum_{j=0}^{\gamma} d_j z^j$, then (5.2) guarantees the existence of an absolutely convergent Fourier series for a complex-valued function $a(z) = d^{-1}(z)$.

It is easy to see that, since $d_j = 0$ if $j > \gamma = o(n)$, the above mentioned Corollary 1 of Doukhan (1994) implies that the strong mixing coefficient $\alpha_X(2k) \equiv \sup_{u,v} \alpha_X(2k; u, v)$ decays even faster than exponential rate; therefore, according to the Remark 1 to the Central Limit Theorem (3.3.1) of Guyon (1995), this guarantees (conditional) asymptotic normality of the term $\left(\sum_{i \in R} Z_i Z_i'\right)^{-1} \sum_{i \in R} Z_i \omega_i$.

To establish the asymptotic variance of the first term, we find that the variance

$$\frac{1}{n} \text{Var} \left(\sum_{i \in R} Z_i \omega_i \right) = \frac{1}{n} \mathbb{E} \left[\sum_{i, j \in R} Z_i Z_j \omega_i \omega_j \right] = \Sigma_X \left(1 + 2 \sum_{k=1}^{\gamma} c_k^2 \right)$$

Finally, the conditions imposed on the difference coefficients above lead to $\sum_{k=1}^{\gamma} c_k^2 = O\left(\frac{1}{\gamma}\right)$ and we have for the conditional variance of the first term in (5.1) $\Sigma_X^{-1}(1 + 2\sum_{k=1}^{\gamma} c_k^2) = \Sigma_X^{-1}\left(1 + O\left(\frac{1}{\gamma}\right)\right)$.

Now we will treat the 2nd term $\left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}}\right)^{-1} \sum_{\mathbf{i} \in R} Z_{\mathbf{i}} \delta_{\mathbf{i}}$. As a first step, we note that the expected value of this term is $\mathbb{E} \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}}\right)^{-1} \sum_{\mathbf{i} \in R} Z_{\mathbf{i}} \delta_{\mathbf{i}} = 0$ due to the definition of the difference sequence d_j . Now we need to examine the variance term which is defined by $\mathbb{E} \left[\left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} \delta_{\mathbf{i}}\right) \left(\sum_{\mathbf{l} \in R} Z'_{\mathbf{l}} \delta_{\mathbf{l}}\right) \right]$.

Clearly,

$$\mathbb{E} \left[\left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} \delta_{\mathbf{i}}\right) \left(\sum_{\mathbf{l} \in R} Z'_{\mathbf{l}} \delta_{\mathbf{l}}\right) \right] = \left[\sum_{\mathbf{i} \in R} \delta_{\mathbf{i}}^2 - c_k \sum_{\mathbf{i} \in R} \delta_{\mathbf{i}} \sum_{\mathbf{j} \in J} \delta_{\mathbf{i}+\mathbf{j}} \right] \Sigma_X$$

Analyzing $\delta_{\mathbf{i}}$, $\mathbf{i} \in R$, it is convenient first to introduce the differential operator $D_{y,z}$ for any two arbitrary vectors $y, z \in \mathbb{R}^q$ as $D_{y,z} = \sum_{k=1}^q (y_k - z_k) \frac{\partial}{\partial x_k}$ with x_k being the generic k th argument of a q -dimensional function. Then, by using Taylor's formula to expand $f(U_{\mathbf{i}+\mathbf{j}})$ around $U_{\mathbf{i}}$, we find that, for any $\mathbf{i} \in R$,

$$\begin{aligned} \delta_{\mathbf{i}} &= \sum_{\mathbf{j} \in J} d_{\mathbf{j}} \left[\frac{\sum_{l=1}^{[\alpha]} D_{U_{\mathbf{i}+\mathbf{j}}, U_{\mathbf{i}}}^l f(U_{\mathbf{i}})}{l!} \right. \\ &\quad \left. + \int_0^1 \frac{(1-u)^{[\alpha]-1}}{([\alpha]-1)!} \left[D_{U_{\mathbf{i}+\mathbf{j}}, U_{\mathbf{i}}}^{[\alpha]} f(U_{\mathbf{i}} + u(U_{\mathbf{i}+\mathbf{j}} - U_{\mathbf{i}})) - D_{U_{\mathbf{i}+\mathbf{j}}, U_{\mathbf{i}}}^{[\alpha]} f(U_{\mathbf{i}}) \right] du \right] \end{aligned} \quad (5.3)$$

Following the same line of argument as in Cai, Levine and Wang (2009), we can conclude that, if the order of difference sequence $\gamma \geq [\alpha]$, the first additive term above is equal to zero due to properties of the polynomial difference sequence. Using the Lipschitz property of the function f , it can be shown that $\delta_{\mathbf{i}} \leq M \left(\frac{m}{n}\right)^{\alpha/q}$. Due to this, it is clear that

$$\sum_{\mathbf{i} \in R} \delta_{\mathbf{i}}^2 - c_k \sum_{\mathbf{i} \in R} \delta_{\mathbf{i}} \sum_{\mathbf{j} \in J} \delta_{\mathbf{i}+\mathbf{j}} = O(n^{1-2\alpha/q} m^{2\alpha/q})$$

and, therefore, as $n \rightarrow \infty$ we have $n \text{Var} \left(\left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}}\right)^{-1} Z' \delta \right) = O\left(\left(\frac{m}{n}\right)^{2\alpha/q}\right) \Sigma_X^{-1}$. The combination of the results for the two terms of (5.1) produces asymptotic normality of the least squares estimator. \square

Proof of Lemma 2.2

Proof. First, notice that, $a = \frac{1}{n} \sum_{\mathbf{i} \leq \mathbf{n}} (Y_{\mathbf{i}} - X_{\mathbf{i}}' \beta) - \frac{1}{n} \sum_{\mathbf{i} \leq \mathbf{n}} f(U_{\mathbf{i}}) + o_p(1)$; due to this, we have $\hat{a} - a = \frac{1}{n} \sum_{\mathbf{i} \leq \mathbf{n}} X_{\mathbf{i}}' (\hat{\beta} - \beta) + \frac{1}{n} \sum_{\mathbf{i} \leq \mathbf{n}} f(U_{\mathbf{i}}) + o_p(1)$. Recall that the function $f(\cdot) \in \Lambda^\alpha(M)$ and, therefore, $\frac{1}{n} \sum_{\mathbf{i} \leq \mathbf{n}} f(U_{\mathbf{i}}) = O(n^{-\alpha/q})$. This suggests that, if the ratio $\alpha/q > 1/2$, the asymptotic property of \hat{a} is driven by the $\frac{1}{n} \sum_{\mathbf{i} \leq \mathbf{n}} X_{\mathbf{i}}' (\hat{\beta} - \beta)$ only. This is also reasonable from the practical viewpoint - if the function $f(\cdot)$ is sufficiently smooth, its influence on the asymptotic behavior of \hat{a} is negligible; moreover, the degree of smoothness required depends on the dimensionality q . \square

Proof of Theorem 2.3

Proof. We will only prove the first statement since the derivation of the second statement is very similar. The proof follows closely that of Theorem 3 in Wang, Brown and Cai (2011) and so we only give its outlines. First, note that the residual $r_{\mathbf{i}} = f(U_{\mathbf{i}}) + \varepsilon_{\mathbf{i}} + a - \hat{a} + X_{\mathbf{i}}' (\beta - \hat{\beta})$ and, therefore, the estimate $\hat{f}(U) = \hat{f}_1(U) + \hat{f}_2(U)$ where $\hat{f}_1(U) = \sum_{\mathbf{i} \leq \mathbf{n}} W_{\mathbf{i},h}(U - U_{\mathbf{i}})[f(U_{\mathbf{i}}) + \varepsilon_{\mathbf{i}}]$ while $\hat{f}_2(U) = \sum_{\mathbf{i} \leq \mathbf{n}} W_{\mathbf{i},h}(U - U_{\mathbf{i}})[X_{\mathbf{i}}' (\beta - \hat{\beta})] + a - \hat{a}$. From the standard multivariate nonparametric regression results we know that for any $U_0 \in [0, 1]^q$ $\sup_{f \in \Lambda^\alpha(M)} \mathbb{E}[(\hat{f}_1(U_0) - f_1(U_0))^2] \leq C n^{-2\alpha/(2\alpha+q)}$ for some constant $C > 0$. On the other hand, clearly $\sum_{\mathbf{i} \leq \mathbf{n}} W_{\mathbf{i},h}^2(U - U_{\mathbf{i}}) = O\left(\frac{1}{nh^q}\right) = O(n^{-2\alpha/(2\alpha+q)})$. Therefore,

$$\begin{aligned} \mathbb{E}(\hat{f}_2(U_0))^2 &= \mathbb{E} \left[\left(\sum_{\mathbf{i} \leq \mathbf{n}} W_{\mathbf{i},h}(U - U_{\mathbf{i}}) X_{\mathbf{i}}' (\beta - \hat{\beta}) \right)^2 \right] \\ &\leq \sum_{\mathbf{i} \leq \mathbf{n}} W_{\mathbf{i},h}(U - U_{\mathbf{i}})^2 \mathbb{E}(X_{\mathbf{i}}' (\beta - \hat{\beta}))^2 = O\left(n^{-2\alpha/(2\alpha+q)}\right). \end{aligned} \quad (5.4)$$

Since \hat{a} converges to a at the usual parametric rate of $n^{1/2}$, the statement of the theorem is true. \square

Proof of Theorem 3.5

Proof. To analyze asymptotic behavior of this distribution it is useful, as before, to split the bias into two terms:

$$\hat{\beta} - \beta = (Z'Z)^{-1}Z'\delta + (Z'Z)^{-1}Z'\omega$$

and analyze these two terms separately. Starting with the second term, it is clear immediately that the conditional expectation $\mathbb{E}((Z'Z)^{-1}Z'\omega|Z) = 0$. Now, we need to look at the conditional variance of this term. Clearly, $\text{Var}((Z'Z)^{-1}Z'\omega|Z) = (Z'Z)^{-1}Z'\Psi Z(Z'Z)^{-1}$ where $\Psi = \text{Var}(\omega)$ is a matrix with a typical element $\Psi_{ij} = \sigma^2 \sum_{t=0}^{\min(\gamma_i, \gamma_j)} d_t d_{t+(i-j)}$. Note that the special case is $\Psi_{ii} = \sigma^2$ due to properties of the difference sequence we just specified. Therefore, the conditional distribution is

$$(Z'Z)^{-1}Z'\omega \sim N(0, (Z'Z)^{-1}Z'\Psi Z(Z'Z)^{-1})$$

Now, we need to analyze conditional variance. The first step is to investigate the behavior of expectations $\mathbb{E}Z'Z$ and $\mathbb{E}Z'\Psi Z$. First, we have $\mathbb{E}(Z_i Z_i') = \sum_{t=0}^{\gamma_i} d_t^2 \text{Var}(X_{i+t}|U) + [\sum_{t=0}^{\gamma_i} d_t h(U_{i,t})]' [\sum_{t=0}^{\gamma_i} d_t h(U_{i,t})]$. For non-equal indices, the analogous statement is

$$\begin{aligned} \mathbb{E}(Z_i Z_{i+j}') &= \sum_{l=0}^{\min(\gamma_i, \gamma_j)} d_{j+l} d_l \mathbb{E}(\text{Var} X_{i+j+l}|U) \\ &+ \left[\sum_{t=0}^{\gamma_i} d_t h(U_{i,t}) \right]' \left[\sum_{t=0}^{\gamma_j} d_t h(U_{i+j,t}) \right] \end{aligned} \quad (5.5)$$

Since the matrix $Z'Z = \sum_{i=1}^n Z_i Z_i'$, we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}(Z'Z) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E} Z_i Z_i' = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\text{Var}(X_i|U)] = \Sigma_*$$

because the second contributing term is bounded as

$$\left| \left[\sum_{t=0}^{\gamma_i} d_t h(U_{i,t}) \right]' \left[\sum_{t=0}^{\gamma_i} d_t h(U_{i,t}) \right] \right| \leq \gamma_i \nu^{2\rho}; \quad (5.6)$$

due to the boundedness of the marginal density $g(u)$, the length of the difference sequence is always $o(n)$. Therefore, as $n \rightarrow \infty$, the above upper bound will go to zero as $n \rightarrow \infty$ and $\nu \rightarrow 0$, respectively, no matter the point U it is centered around. Consequently, the second term disappears. In a similar way, for the expectation of the term $\mathbb{E}Z'\Psi Z$ we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}(Z'\Psi Z) = \sigma^2 \left(1 - \sum_{i,j=1}^n c_{ij}^2 \right) \Sigma_*$$

Let $U = (U_1, \dots, U_n)'$; to analyze the last term $(Z'Z)^{-1}Z'\delta$ we note first that its conditional variance is $Var((Z'Z)^{-1}Z'\delta|Z) = (Z'Z)^{-1}Z'\delta\delta'Z(Z'Z)^{-1}$. The behavior of the expectation $\mathbb{E}Z'Z$ has been already analyzed before; thus, it only remains to analyze the behavior of the expectation $\mathbb{E}Z'\delta\delta'Z$. We obtain

$$\begin{aligned}
& \frac{1}{n} \mathbb{E} \sum_{i,j,k,l} Z'_i \delta_j \delta'_k Z_l \\
&= \frac{1}{n} \mathbb{E} \left\{ \mathbb{E} \left\{ \sum_{i,j,k,l} Z'_i \delta_j \delta'_k Z_l | U \right\} \right\} = \frac{1}{n} \mathbb{E} \left\{ \sum_{i,j,k,l} \mathbb{E}(Z'_i | U) \delta_j \delta'_k \mathbb{E}(Z_l | U) \right\} \\
&= \frac{1}{n} \mathbb{E} \left\{ \sum_{i,j,k,l} \mathbb{E} \left(\sum_{t=0}^{\gamma_i} d_t X'_{i+t} | U \right) \delta_j \delta'_k \mathbb{E} \left(\sum_{t=0}^{\gamma_l} d_t X_{l+t} | U \right) \right\} \\
&= \frac{1}{n} \mathbb{E} \left\{ \sum_{i,j,k,l} \left(\sum_{t=0}^{\gamma_i} d_t h(U_{i,t}) \right) \delta_j \delta'_k \left(\sum_{t=0}^{\gamma_l} d_t h(U_{l,t}) \right) \right\}
\end{aligned}$$

By definition of differences that we use here, and since both $\gamma_i = o(n)$ and $\gamma_l = o(n)$, we obtain

$$\frac{1}{n} \mathbb{E} Z' \delta \delta' Z \leq \frac{1}{n} \nu^{2\rho+2\alpha} * o(n^2) \leq o(n) \nu^{2(\rho+\alpha)} \quad (5.7)$$

The (5.7) implies that, in order for the parametric part of the model (2.1) to be estimable, the expression above must go to zero as $n \rightarrow \infty$; for example, if $\nu = O(n^{-1})$, we obtain $\rho + \alpha > \frac{1}{2}$ which is the condition stated in Wang, Brown and Cai (2011).

Finally, we need to verify that all of the variances $\lim_{n \rightarrow \infty} Var \left(\frac{1}{n} Z' Z \right) = \lim_{n \rightarrow \infty} Var \left(\frac{1}{n} Z' \Psi Z \right) = \lim_{n \rightarrow \infty} Var \left(\frac{1}{n} Z' \delta \delta' Z \right) = 0$; all of the variances here are understood elementwise.

As an example, the first case gives the variance of the kl th element as $Var \left\{ \frac{1}{n} \sum_{i,j=1}^p \left\{ \sum_{t=0}^{\gamma_i} d_t X_{k+t}^i \sum_{t=0}^{\gamma_l} d_t X_{l+t}^j \right\} \right\}$; therefore, $\lim_{n \rightarrow \infty} Var \left(\frac{1}{n} Z' Z \right) = 0$ due to the existence of non-singular Σ_* as long as $\gamma_i = o(n)$ for any point U_i around which the respective difference is defined (due to the assumptions of the theorem). The same is also true for the second limit - one only needs to use the assumption on the elements of the covariance matrix Ψ as well. Finally, the third limit also goes to zero due to the Lipschitz property of the function $f(U)$. \square

Proof of Theorem 4.7

Proof. From our previous results, we know that the estimator $\hat{\beta}$ is asymptotically normal and efficient; in other words, it satisfies $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{L} N(0, \sigma^2 \Sigma_X^{-1})$. This immediately implies that $\sqrt{n}(C\hat{\beta} - C\beta) \xrightarrow{L} N(0, \sigma^2 C \Sigma_X^{-1} C')$. This, of course, suggests that, as in Wang, Brown and Cai (2011), we can define the test statistic based on $\frac{n}{\sigma^2} \hat{\beta}' C' (C \Sigma_X^{-1} C') C \hat{\beta}$; however, neither σ^2 nor Σ_X are known in real applications and, therefore, need to be estimated. To estimate Σ_X , we recall from the proof of Theorem (2.1) that $\frac{1}{n} \sum_{\mathbf{s} \in R} Z_{\mathbf{s}} Z'_{\mathbf{s}} \xrightarrow{P} \Sigma_X$ and, therefore, $\frac{1}{n} \sum_{\mathbf{s} \in R} Z_{\mathbf{s}} Z'_{\mathbf{s}}$ can be used as an estimate of Σ_X . The resulting test statistic would be $\frac{1}{\sigma^2} \hat{\beta}' C' \left(C \left(\sum_{\mathbf{s} \in R} Z_{\mathbf{s}} Z'_{\mathbf{s}} \right)^{-1} C' \right)^{-1} C \hat{\beta}$ that looks like a classical χ^2 type statistics asymptotically. However, σ^2 is also not known and needs to be estimated as well.

Let us start with the numerator. As in Wang, Brown and Cai (2011), introduce an array (essentially, a linear operator) $L : \mathbb{R}^n \rightarrow \mathbb{R}^{|R|}$ such that $L_{\mathbf{i}, \mathbf{j}} = d_{\mathbf{j}-\mathbf{i}}$ for any $0 \leq |\mathbf{i}-\mathbf{j}| \leq \gamma$ where $|\mathbf{i}-\mathbf{j}|$ is defined as on p. 7, $\mathbf{i} \in R$ and $\mathbf{j} \in \{0, \dots, m\}^q$; when $|\mathbf{i}-\mathbf{j}|$ is less than zero or greater than γ , we assume that $L_{\mathbf{i}, \mathbf{j}} = 0$. Another useful array that we use is a unity array (operator) $J : \mathbb{R}^n \rightarrow \mathbb{R}^{|R|}$ with $J_{\mathbf{i}, \mathbf{i}} = 1$ for any $\mathbf{i} \in R$ and 0 otherwise. Define an array $\omega = \{\omega_{\mathbf{i}}\}$ for all $\mathbf{i} \in R$. Using these definitions, we have $\omega = L\varepsilon = J\varepsilon + (L - J)\varepsilon = \omega_1 + \omega_2$ where $\omega_1 = J\varepsilon$ and $\omega_2 = (L - J)\varepsilon$. Clearly, ω_1 is a collection of uncorrelated normal random variables: $\omega_1 \sim N(0, \sigma^2 I_R)$ where I_R is a unity array with both indices varying over R . At the same time, $\omega_2 \sim N(0, \sigma^2 (L - J)(L - J)')$. Under the additional assumption of $1 - d_0 = O(\gamma^{-1})$, it is not hard to verify that each element of the covariance array of ω_2 is of the order $O(\gamma^{-1})$ and that, therefore, ω_2 tends to zero in probability as $n \rightarrow \infty$.

Note that

$$\begin{aligned} \hat{\beta} &= \beta + \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} \sum_{\mathbf{i} \in R} Z'_{\mathbf{i}} \delta_{\mathbf{i}} + \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} \sum_{\mathbf{i} \in R} Z'_{\mathbf{i}} \omega_{\mathbf{i}} \\ &= \beta + \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} \sum_{\mathbf{i} \in R} Z'_{\mathbf{i}} \delta_{\mathbf{i}} + \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} Z' \omega_1 + \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} Z' \omega_2. \end{aligned}$$

Therefore, under the null hypothesis we have $C\hat{\beta} = C\beta + C \left(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z'_{\mathbf{i}} \right)^{-1} \sum_{\mathbf{i} \in R} Z'_{\mathbf{i}} \delta_{\mathbf{i}} +$

$C(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z_{\mathbf{i}}')^{-1} Z' \omega_1 + C(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z_{\mathbf{i}}')^{-1} Z' \omega_2$. Following the proof of Theorem 1, we conclude that the term $(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z_{\mathbf{i}}')^{-1} \sum_{\mathbf{i} \in R} Z_{\mathbf{i}}' \delta_{\mathbf{i}}$ converges to zero in probability as $n \rightarrow \infty$; since under our assumptions each element of the covariance array of ω_2 is of the order $O(\gamma^{-1})$ we can consider just the term $C(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z_{\mathbf{i}}')^{-1} Z' \omega_1 \sim N(0, \sigma_2 C(\sum_{\mathbf{i} \in R} Z_{\mathbf{i}} Z_{\mathbf{i}}')^{-1} C')$.

To analyze the denominator, we substitute first $D_{\mathbf{i}} = Z_{\mathbf{i}}' \beta + \delta_{\mathbf{i}} + \omega_{\mathbf{i}}$ in the definition of a typical residual $e_{\mathbf{i}}$ and then, looking at (4.1), we realize that the δ related term $\sum_{\mathbf{i} \in R} |\delta_{\mathbf{i}} - Z_{\mathbf{i}}' (\sum_{\mathbf{s} \in R} Z_{\mathbf{s}} Z_{\mathbf{s}}')^{-1} \sum_{\mathbf{s} \in R} Z_{\mathbf{s}} \delta_{\mathbf{s}}|^2$ converges to zero in probability if $\alpha > \frac{q}{2}$. The ‘‘crossproduct’’ term that contains both $\delta_{\mathbf{i}}$ and $\omega_{\mathbf{i}}$ will also tend to zero in probability as $n \rightarrow \infty$ under the same circumstances. Therefore, we only need to analyze the behavior of the term

$$H\omega \equiv \sum_{\mathbf{i} \in R} \left| \omega_{\mathbf{i}} - Z_{\mathbf{i}}' \left(\sum_{\mathbf{s} \in R} Z_{\mathbf{s}} Z_{\mathbf{s}}' \right)^{-1} \sum_{\mathbf{s} \in R} Z_{\mathbf{s}} \omega_{\mathbf{s}} \right|^2 \quad (5.8)$$

To analyze the expression (5.8), one first needs to notice that the operator H is the projector of the rank $n - \gamma - p$ due to the regularity properties of the contrast process $\sum_{\mathbf{i} \in R} [D_{\mathbf{i}} - Z_{\mathbf{i}}' \beta]^2$; see, for example, Guyon (2009) pp. 271-274 for the details. Due to this, we conclude that the estimate $\hat{\sigma}^2$ has $\chi^2(n - \gamma - p)$ distribution and that it is independent from the numerator of the test statistic. \square

Lawrence D. Brown, Department of Statistics, University of Pennsylvania

E-mail: (lbrown@wharton.upenn.edu)

Michael Levine, Department of Statistics, Purdue University

E-mail: (mlevins@purdue.edu)

Lie Wang, Department of Mathematics, MIT

E-mail: (liewang@math.mit.edu)

Lawrence D. Brown’s work has been partially supported by NSF Grant DMS-1512084

Michael Levine's work has been partially supported by NSF Grant DMS-1208994

Lie Wang's work has been partially supported by NSF Grant DMS-1005539