Limit theory and inference about conditional distributions *

Purevdorj Tuvaandorj and Victoria Zinde-Walsh
McGill University, CIREQ and CIRANO
purevdorj.tuvaandorj@mail.mcgill.ca
victoria.zinde-walsh@mcgill.ca

June 24, 2014

Abstract

We consider conditional distribution and conditional density functionals in the space of generalized functions. The approach follows Phillips (1985, 1991, 1995) who employed generalized functions to overcome non-differentiability in order to develop expansions. We obtain the limit of the kernel estimators for weakly dependent data, even under non-differentiability of the distribution function; the limit Gaussian process is characterized as a stochastic random functional (random generalized function) on the suitable function space. An alternative simple to compute estimator based on the empirical distribution function is proposed for the generalized random functional. For test statistics based on this estimator, limit properties are established. A Monte Carlo experiment demonstrates good finite sample performance of the statistics for testing logit and probit specification in binary choice models.

*The authors thank the participants of the 14th Advances in Econometrics Conference and especially Brendan Beare, and an anonymous referee, the editors Yoosoon Chang and Joon Park for helpful comments and suggestions. The authors gratefully acknowledge support of this research from the Fonds de recherche sur la société et la culture (Québec).

1 Introduction

This paper treats the conditional distribution (and conditional density) as elements in the spaces of generalized functions; this endows functions with well-defined generalized derivatives and permits expansions even when the functions are not differentiable in the ordinary sense. The generalized function approach was utilized in a number of papers by P.C.B. Phillips (Phillips (1985, 1991, 1995)), where the results were based on expansions in situations involving lack of differentiability. In the space of generalized functions the differentiation operator is always defined and is continuous; this is exploited in developing expansions. Here the methodology of generalized functions provides a means to derive limit processes for estimators of conditional distributions and densities.

The asymptotic properties of the kernel estimators of conditional distribution and conditional density are established in the literature only under restrictive smoothness and support assumptions. Existence and smoothness of the marginal density, $f_x(x)$, as well as differentiability (possibly twice - see, e.g. Li and Racine (2007) and Pagan and Ullah (1999)) of the joint distribution, $F_{x,y}(x,y)$, or of the conditional distribution, $F_{y|x}(x,y)$, is typically assumed. There are several problems with making such assumptions: first, both from a theoretical economic modeling perspective and from empirical evidence smoothness may be violated (kinks, spikes in hazard functions), second, smoothness assumptions are practically impossible to test, third, if smoothness is indeed violated there is no clear indication as to the pointwise behavior of the estimator under such a misspecification. The question then arises: what is the limit process of the kernel estimator of the conditional distribution and conditional density functions when the usual smoothness assumptions are violated?

The precise answer is provided when we consider these estimators in spaces of generalized functions; the limit process is a Gaussian random functional (generalized random function) on a given space of smooth functions; convergence is at a parametric rate. Although the limit process can no longer be interpreted pointwise (as in the case when e.g. the density does not exist), it provides the possibility for inference which utilizes convergence at a parametric rate to a Gaussian process. In the i.i.d. case the limit process was obtained in Zinde-Walsh (2013); here the result is extended to stationary mixing processes.

Our derivation suggests a simpler estimator that can be used in place of the kernel estimator; the estimator is based on the empirical distribution function that is easy to compute and has the same limit process as the kernel estimator in the space of generalized functions. It has the additional advantage of being unbiased.

We use the results about the limit process in generalized functions to propose test statistics for testing distributional hypotheses. We focus here on testing a parametric conditional distribution. The test differs from the conditional Kolmogorov (CK) test of Andrews (1997) and related tests that use the difference between the empirical joint distribution and the joint distribution constructed from the parametric conditional. We approach the conditional distribution via differentiation rather than integration (implicit in the CK test); this is made possible by generalized functions. Our statistic cannot exploit the sup norm since generalized functions are not defined pointwise, instead we construct a joint Gaussian process of values of the functional (indexed by several functions) and construct a χ^2 type statistic based on that (it would also be possible to use a supremum of the values of the functional). Our statistic similarly to CK suffers from the dependence arising from estimation of the parameters in the null distribution and requires performing bootstrap.

We perform a small numerical investigation of the approximation quality of the limit Gaussian process in finite samples and find that it provides a fairly good approximation to the distribution of the values of the functional (generalized function) for different choices of possible distributions even for very moderate sample sizes.

We also investigate the finite sample properties of our test for probit and logit specifications in a binary choice model. The Monte Carlo experiment provides good size properties (similar in behavior to those of CK in Andrews (1997)) in moderate samples. The test has good power in moderate sized samples; the alternatives we considered are similar to Stute and Zhu (2005).

The following notation is used throughout the paper: $\min\{a,b\} = a \wedge b$, $\xrightarrow{d}$ and $\xrightarrow{p}$ stand for convergence in distribution (weak convergence) and in probability, respectively, $\|\cdot\|$ denotes the Euclidean norm. For the space of d_v times differentiable functions on $\mathbb{R}^{d_v}$ or its open subspace W, define the differentiation operator $\partial^{d_v} \equiv \frac{\partial^{d_v}}{\partial v_1 ... \partial v_{d_v}}$ that provides for any function g the derivative $\frac{\partial^{d_v} g(v)}{\partial v_1 ... \partial v_{d_v}}$.

1.1 The space of generalized functions

The definitions of spaces of generalized functions that are used here are based on Gel'fand and Shilov (1964).

Consider a space of well-behaved "test" functions such as the space $D_\infty\left(\mathbb{R}^{d_x}\right)$ of infinitely differentiable functions with bounded support, or any of the spaces $D_m\left(\mathbb{R}^{d_x}\right)$ of m times continuously differentiable functions (with bounded support); sometimes the domain of definition can be an open subset W of $\mathbb{R}^{d_x}$, typically here W denotes $(0,1)^{d_x} \subset \mathbb{R}^{d_x}$. Denote the generic space by $D(W)$; convergence in $D(W)$ is defined as follows: a sequence $\psi_n \in D(W)$ converges to zero if all ψ_n are defined on a common bounded support in W and ψ_n as well as all the l-th order derivatives (with $l \leq m$ for D_m or all $l < \infty$ for D_∞) converge pointwise to zero. Another space of test functions is $S\left(\mathbb{R}^{d_x}\right)$ where the support of the infinitely differentiable functions is not necessarily bounded but the functions go to zero at infinity faster than any power (and all their derivatives have the same property); this was the space used in Phillips (1991, 1995).

The space of generalized functions is the dual space, $D^*(W)$, the space of linear continuous functionals on $D(W)$ with the value of the functional $f \in D^*(W)$ for a function $\psi \in D(W)$ denoted by (f,ψ). This is a linear space with the weak topology defined as follows: a sequence of elements of $D^*(W)$ converges if the sequence of values of the functionals converges for any function from $D(W)$: a sequence f_n converges to f if for any $\psi \in D(W)$ convergence $(f_n,\psi) \to (f,\psi)$ holds. Whenever some space $D_1 \subset D_2$, the inclusion $D_2^* \subset D_1^*$ holds for the dual spaces. The space of generalized functions in Phillips (1991, 1995), $S^*\left(\mathbb{R}^{d_x}\right)$, is included in $D^*\left(\mathbb{R}^{d_x}\right)$: $S^*\left(\mathbb{R}^{d_x}\right) \subset D^*\left(\mathbb{R}^{d_x}\right)$; the results here apply in that space as well.

Assume that functions in $D(W)$, $W \subseteq \mathbb{R}^{d_x}$, are suitably differentiable, e.g. at least d_x times continuously differentiable. Then for any $\psi \in D(W)$, $F \in D^*$ and the differentiation operator ∂^{d_x}, define a generalized derivative $f \in D^*$; $f = \partial^{d_x} F$ as the functional with values given by:

$$(f,\psi) = (-1)^{d_x}\left(F, \partial^{d_x}\psi\right). \tag{1.1}$$

The differentiation operator in the space of generalized functions is continuous.

Suppose that F is a probability distribution function on $\mathbb{R}^{d_x}$, then it is a regular

locally summable function with

$$(F,\psi)=\int\dots\int F(x_1,\dots,x_{d_x})\psi(x_1,\dots,x_{d_x})dx_1\dots dx_{d_x},$$

where the integral exists by the properties of F. Then the value on the right-hand side of (1.1) is provided by integration:

$$\left(F,\partial^{d_x}\psi\right)=\int\dots\int F(x_1,\dots,x_{d_x})\partial^{d_x}\psi(x_1,\dots,x_{d_x})dx_1\dots dx_{d_x}.$$

Thus the functional that gives the generalized derivative: $f=\partial^{d_x}F$, the generalized density function, is fully defined. If f exists as a regular function, then it is defined as a generalized function by

$$(f,\psi)=\int\dots\int f(x_1,\dots,x_{d_x})\psi(x_1,\dots,x_{d_x})dx_1\dots dx_{d_x};$$

it is then easy to check integrating by parts that (1.1) holds.

2 Main Result

2.1 Conditional distribution and density as generalized functions

Here conditioning is limited to conditioning on a variable or vector in a joint distribution, that is given a joint distribution function $F_{x,y}(\cdot,\cdot)$ on $\mathbb{R}^{d_x}\times\mathbb{R}^{d_y}$ where d_x and d_y are the dimensions of x and y, define a (generalized) function $F_{y|x}(\cdot,\cdot)$ that represents the conditional distribution of y given x. The conditional distribution function may not exist for every point x. Denote by F_x and F_y the marginal distribution functions of x and y, respectively. Note that although support of the random y belongs to $\mathbb{R}^{d_y}$ it could be a discrete set of points, thus we do not restrict y to be continuously distributed.

The generalized function representation is derived by considering copula functions (see Zinde-Walsh (2013)). Represent $F_{x,y}(x,y)=C_{F_x,F_y}\left(F_x(x),F_y(y)\right)$, where $C_{F_x,F_y}(\cdot,\cdot)$ is the copula function (Sklar (1973)). If the marginal $F_x(x)$ is a continuous function, the copula function is uniquely defined in its first argument. If the marginal is not

continuous, the copula function is no longer unique; following Darsow, Nguyen, and Olsen (1992) in the case when the distribution is discrete or discrete/continuous, select the copula function obtained by linear interpolation. Then our interest is in

$$\lim_{\tilde{\Delta}\to 0} \frac{C_{F_x,F_y}(a+\tilde{\Delta},b) - C_{F_x,F_y}(a,b)}{\tilde{\Delta}}.$$

Since the copula function is differentiable with respect to each argument (and provides the appropriate value for the conditional distribution even at a mass point as shown in Darsow, Nguyen, and Olsen (1992)) the derivative $\frac{\partial C_{F_x,F_y}}{\partial F_x}$ is a regular function. However, the copula function is not easily obtainable and estimable and also often cannot be further differentiated to provide a conditional density. This is why we consider generalized derivatives that significantly simplify both derivation and estimation. Consider the copula function as a generalized function on $D((0,1)^{d_x}) \times D(\mathbb{R}^{d_y})$; as a functional applied to $\psi_{x,y} = \psi_x(x_1,\ldots,x_{d_x})\psi_y(y_1,\ldots,y_{d_y}) \in D((0,1)^{d_x}) \times D(\mathbb{R}^{d_y})$ it defines the value

$$\begin{aligned}
&\left(C_{F_x,F_y},\psi\right) \\
&= \int \ldots \int C_{F_x,F_y}\left(F_x, F_y(y)\right) \psi_x(F_{x_1},\ldots,F_{x_{d_x}})\psi_y(y_1,\ldots,y_{d_y}) dF_{x_1}\ldots dF_{x_{d_x}} dy_1 \ldots dy_{d_y} \\
&= \int \ldots \int F_{x,y}(x,y)\psi_x(F_{x_1},\ldots,F_{x_{d_x}})\psi_y(y_1,\ldots,y_{d_y}) dF_{x_1}\ldots dF_{x_{d_x}} dy_1 \ldots dy_{d_y}.
\end{aligned}$$

Define the conditional distribution $F_{y|x}$ as a functional on $D((0,1)^{d_x}) \times D(\mathbb{R}^{d_y})$ through the generalized derivative $\frac{\partial C_{F_x,F_y}}{\partial F_x}$. For any $\psi_{x,y} \in D((0,1)^{d_x}) \times D(\mathbb{R}^{d_y})$ the value of the functional is

$$\begin{aligned}
&\left(F_{y|x},\psi_{x,y}\right) \\
&= (-1)^{d_x} \int \ldots \int C_{F_x,F_y}\left(F_x, F_y(y)\right) \partial^{d_x}\psi_x(F_{x_1},\ldots,F_{x_{d_x}})\psi_y(y_1,\ldots,y_{d_y}) dF_{x_1}\ldots dF_{x_{d_x}} dy_1 \ldots dy_{d_y} \\
&= (-1)^{d_x} \int \ldots \int F_{x,y}\left(x,y\right) \partial^{d_x}\psi_x(F_{x_1},\ldots,F_{x_{d_x}})\psi_y(y_1,\ldots,y_{d_y}) dF_{x_1}\ldots dF_{x_{d_x}} dy_1 \ldots dy_{d_y}.
\end{aligned} \tag{2.1}$$

To define conditional density $f_{y|x}$ as a generalized function, one would have

$$\left(f_{y|x}, \psi_{x,y}\right) =$$
$$(-1)^{d_x+d_y} \int \dots \int F_{x,y}(x,y) \partial^{d_x} \psi_x(F_{x_1}, \dots, F_{x_{d_x}}) \partial^{d_y} \psi_y(y_1, \dots, y_{d_y}) dF_{x_1} \dots dF_{x_{d_x}} dy_1 \dots dy_{d_y}. \tag{2.2}$$

When $\left(F_{y|x}, \psi_x\right)$ exists as a pointwise function of y, one can consider the conditional distribution $F_{y|x}$ as a functional on $D\left(0,1\right)^{d_x}$ only. The representation is convenient in that it involves only distribution functions (and not the copula function). Using more concise notation for the multivariate functions and integrals, we express (2.1) as

$$\left(F_{y|x}, \psi_{x,y}\right) = (-1)^{d_x} \int \int F_{x,y}\left(x,y\right) \partial^{d_x} \psi_x(F_x) \psi_y(y) dF_x dy;$$

a similar formula holds for (2.2).

Note that, as in Zinde-Walsh (2013) it is possible to consider F_x to be either a vector function $F_x = (F_{x_1}, \dots, F_{x_{d_x}})'$ or a scalar function F_x of the vector $x = (x_1, \dots, x_{d_x})'$; in the latter case, without loss of generality we can express $F_{y|x}$ as a generalized function on $D((0,1)) \times D(\mathbb{R}^{d_y})$ as

$$\left(F_{y|x}, \psi_{x,y}\right) = -\int \int F_{x,y}\left(x,y\right) \partial \psi_x(F_x) \psi_y(y) dF_x dy. \tag{2.3}$$

2.2 The limit process of the kernel estimator of conditional distribution

Recall the usual kernel estimator of conditional distribution:

$$\begin{aligned} \tilde{F}_{y|x}\left(x,y\right) &= \frac{\sum_{t=1}^n G\left(\frac{y-y_t}{h_y}\right) k\left(\frac{x_t-x}{h}\right)}{\sum_{t=1}^n k\left(\frac{x_t-x}{h}\right)} \\ &= \frac{\frac{1}{n}\sum_{t=1}^n G\left(\frac{y-y_t}{h_y}\right) \frac{1}{h^{d_x}} k\left(\frac{x_t-x}{h}\right)}{\tilde{f}_x(x)}, \end{aligned} \tag{2.4}$$

where G is the integral of a kernel function g that satisfies assumptions similar to those on k otherwise it is assumed to be the indicator function $1(\cdot)$. The kernel functions satisfy the usual assumptions:

Assumption 1. *The kernel functions k and g are bounded, smooth, symmetric density functions on bounded support on $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$, respectively.*

To simplify exposition we assume that each component of vector x is associated with the same (scalar) bandwidth parameter h; it is not difficult to generalize to the case of distinct bandwidths.

In Zinde-Walsh (2013) the case of independent observations was examined. Here we consider assumptions that relax the independence condition allowing for some weak form of dependence to which Donsker's theorem (see e.g., van der Vaart and Wellner (1996)) applies. The following assumption summarizes the regularity condition.

Assumption 2. *The process $\left\{(x_t', y_t')'\right\}_{t=1}^{\infty} \in \mathbb{R}^{d_x+d_y}$ is strictly stationary and α-mixing with distribution function $F_{x,y}(x,y)$ and mixing coefficient $\alpha(\tau)$ that satisfies $\alpha(\tau) = O(\tau^{-1-\varepsilon})$ for some $\varepsilon > 0$.*

Denote the empirical distribution function by $\hat{F}_{\cdot}(\cdot)$; our proof employs limit results for empirical distributions. A remarkable result of Rio (2000) says that the mixing rate stated in Assumption 2, while being independent of the dimension of the vector $(x_t', y_t')'$, d_x+d_y, is sufficient for weak convergence of the empirical process $\sqrt{n}(\hat{F}_{x,y}(\cdot,\cdot)-F_{x,y}(\cdot,\cdot))$ to a tight Gaussian process. It is clear that the same property holds for $\sqrt{n}(\hat{F}_x(\cdot) - F_x(\cdot))$. Let U_x and $U_{x,y}$ be zero mean Gaussian processes with covariance functions

$$\Gamma_x(\xi_1,\xi_2) = \sum_{t=1}^{\infty} \mathrm{Cov}\left[1(x_1 \le \xi_1)1(x_t \le \xi_2)\right], \quad \xi_1,\xi_2 \in \mathbb{R}^{d_x} \tag{2.5}$$

and

$$\Gamma_{x,y}(\xi_1,\xi_2) = \sum_{t=1}^{\infty} \mathrm{Cov}[1((x_1', y_1')' \le \xi_1)1((x_t', y_t')' \le \xi_2)], \quad \xi_1,\xi_2 \in \mathbb{R}^{d_x+d_y} \tag{2.6}$$

respectively. Define for any v by d^v the d_x dimensional vector operator of differentiation

$$d^v = \left(\frac{\partial}{\partial v_1}, ..., \frac{\partial}{\partial v_{d_x}}\right).$$

Denote for a vector A and a vector B of dimension d_x by $J(A,B)$ the scalar product $\Sigma_{i=1}^{d_x} A_i B_i$. The multivariate integrals are over the appropriate spaces $D((0,1)^{d_x}) \times \mathbb{R}^{d_y}$.

The following theorem provides the limit process for the kernel estimator of the conditional distribution.

Theorem 1. *Suppose that Assumptions 1 and 2 hold, and the bandwidth parameter h satisfies $h = cn^{-\delta}$ for some $\delta > 1/4$. Then the estimator $\tilde{F}_{y|x}(x,y)$ as a generalized random function on $D((0,1)^{d_x}) \times D(\mathbb{R}^{d_y})$ converges at $n^{1/2}$ rate to the conditional distribution generalized function $F_{y|x}(x,y)$ defined by (2.1) as $n \to \infty$; the limit process for $\sqrt{n}(\tilde{F}_{y|x} - F_{y|x})$ is given by a $\psi_{x,y} \in D((0,1)^{d_x}) \times D(\mathbb{R}^{d_y})$ indexed random functional, $Q_{y|x}$ with*

$$(Q_{y|x}, \psi_{x,y}) = (-1)^{d_x} \left[\int \int F_{x,y} J\left(d^{F_x} \partial^{d_x} \psi_x(F_x), U_x\right) \psi_y(y) dF_x dy \right.$$
$$+ \int \int F_{x,y} \partial^{d_x} \psi_x(F_x) \psi_y(y) dU_x dy$$
$$\left. + \int \int \partial^{d_x} \psi_x(F_x) \psi_y(y) U_{x,y} dF_x dy \right] \tag{2.7}$$

where U_x and $U_{x,y}$ are Gaussian processes of dimensions d_x and $d_x + d_y$, and covariance functions given in (2.5) and (2.6) correspondingly to F_x and $F_{x,y}$; as a generalized random process the limit process $Q_{y|x}$ of $\sqrt{n}(\tilde{F}_{y|x} - F_{y|x})$ is Gaussian with mean functional zero and covariance bilinear functional C, given, for any (suppressing the subscripts x, y) $\psi_1, \psi_2 \in D((0,1)^{d_x}) \times D(\mathbb{R}^{d_y})$, by

$$(C, (\psi_1, \psi_2)) = \mathrm{Cov}[(Q_{y|x}, \psi_1), (Q_{y|x}, \psi_2)].$$

The proof of Theorem 1 generalizes the proof of Theorem 3 in Zinde-Walsh (2013) to weakly dependent data, and relies on expansions in the space of generalized functions to control the bias functional.

We obtain the asymptotic distribution of the kernel estimators for weakly dependent data, even under non-differentiability of the distribution function; the limit Gaussian process is characterized as a stochastic random functional (random generalized function) on the suitable function space; the usual kernel estimator (with suitable undersmoothing) converges at a parametric rate in the space of generalized functions.

Thus the question of what happens when the kernel estimator is used in the case when the assumptions that justify the asymptotic distribution pointwise are violated is

answered here: the estimator consistently estimates the generalized conditional distribution function. In consequence, the choice of bandwidth and kernel functions does not play a particular role for the asymptotic result to hold; the same limit process would be obtained if the kernel estimators were replaced by the empirical distribution functions. The latter is unbiased and more attractive from a computational viewpoint because the estimator of the generalized conditional distribution function becomes simply an average of functions evaluated at different observations, thus numerical integration is not required.

2.3 Estimators of the conditional distribution and conditional density that exploit the space of generalized functions

The proof of Theorem 1 demonstrates that the limit process of the usual kernel estimator in the space of generalized functions provides values for the functional that are identical to those obtained from an estimator based on empirical distribution function. Specifically, for $\hat{F}_{y|x}$ that is defined as a functional in the space of generalized functions by its values

$$\left(\hat{F}_{y|x}, \psi_{x,y}\right) =$$
$$(-1)^{d_x} \int \dots \int \hat{F}_{x,y}\left(x,y\right) \partial^{d_x} \psi_x(\hat{F}_{x_1}, \dots, \hat{F}_{x_{d_x}}) \psi_y(y_1, \dots, y_{d_y}) d\hat{F}_{x_1} \dots d\hat{F}_{x_{d_x}} dy_1 \dots dy_{d_y},$$

the limit process is the same as for $\tilde{F}_{y|x}$ defined in (2.7). This provides an estimator for the values of the functional which does not rely on a denominator, and is more stable and asymptotically unbiased.

As pointed out in Zinde-Walsh (2013), if the interest is only in the conditional distribution, we may restrict consideration to the value for ψ_x function only and consider the values $(-1)^{d_x} E_{\hat{F}_x}\left(\hat{F}_{x,y}\left(x,y\right) \partial^{d_x} \psi_x(\hat{F}_{x_1}, \dots, \hat{F}_{x_{d_x}})\right)$ as estimators for the functional $\left(F_{y|x}, \psi_x\right)$ for every y. This is appropriate since the generalized functions approach is applied to compensate for possible non-differentiability and in the conditional distribution differentiation is applied with respect to x only. Of course, if the interest is in conditional density, then differentiation with respect to y is also required and then the values of the functional need to be constructed for $\psi_{x,y}$, a function that is smooth in y

as well as in F_x. This gives us the value for the estimator of the density functional

$$\left(\tilde{f}_{y|x}, \psi_{x,y}\right)$$
$$= (-1)^{d_x+d_y} \int \dots \int \tilde{F}_{x,y}(x,y)\, \partial^{d_x}\psi_x(\tilde{F}_{x_1}, \dots, \tilde{F}_{x_{d_x}}) \partial^{d_y}\psi_y(y_1, \dots, y_{d_y}) d\tilde{F}_{x_1} \dots d\tilde{F}_{x_{d_x}} dy_1 \dots dy_{d_y}. \tag{2.8}$$

2.4 The limit process for the kernel estimator of generalized conditional density function

Similarly to Theorem 1 the following theorem provides the limit process for the estimator of the conditional density function defined in (2.8).

Theorem 2. *Let Assumptions 1 and 2 hold, and the bandwidth parameter h satisfy $h = cn^{-\delta}$ for some $\delta > 1/4$. The estimator $\tilde{f}_{y|x}(x,y)$ as a generalized random function on $D((0,1)^{d_x}) \times D(\mathbb{R}^{d_y})$ converges at $n^{1/2}$ rate to the conditional density generalized function $f_{y|x}(x,y)$ defined by (2.2) as $n \to \infty$; the limit process for $\sqrt{n}(\tilde{f}_{y|x} - f_{y|x})$ on $D((0,1)^{d_x}) \times D(\mathbb{R}^{d_y})$ is given by a $\psi_{x,y}$ indexed random functional, $M_{y|x}$ with*

$$\begin{aligned}(M_{y|x}, \psi_{x,y}) = (-1)^{d_x+d_y} \Bigg[& \int\int F_{x,y} J\left(d^{F_x}\partial^{d_x}\psi_x(F_x), U_x\right) \partial^{d_y}\psi_y(y) dF_x dy \\ & + \int\int F_{x,y}\partial^{d_x}\psi_x(F_x)\partial^{d_y}\psi_y(y) dU_x dy \\ & + \int\int \partial^{d_x}\psi_x(F_x) U_{x,y} \partial^{d_y}\psi_y(y) dF_x dy \Bigg],\end{aligned}$$

where U_x and $U_{x,y}$ are zero mean Gaussian processes of dimensions d_x and $d_x + d_y$, and covariance functions given in (2.5) and (2.6) correspondingly to F_x and $F_{x,y}$; as a generalized random process the limit process $M_{y|x}$ of $\sqrt{n}(\tilde{F}_{y|x} - F_{y|x})$ is Gaussian with mean functional zero and covariance bilinear functional C, given, for any ψ_1, ψ_2, by

$$(C, (\psi_1, \psi_2)) = \mathrm{Cov}[(M_{y|x}, \psi_1), (M_{y|x}, \psi_2)].$$

3 Testing a parametric form of conditional distribution

This section examines the problem of testing parametric functional form. The null hypothesis maintains that

$$H_0 : F_{y|x}(x,y) = F_{y|x}(x,y;\theta_0) \quad \text{for some} \quad \theta_0 \in \Theta \subseteq \mathbb{R}^p,$$

where $F_{y|x}(x,y;\theta_0)$ is a parametric conditional distribution function, and Θ denotes the parameter space. Let $\hat{\theta}$ denote the asymptotically normal maximum likelihood estimator (MLE) of θ_0. We consider a test statistic based on the difference between nonparametric and parametric linear functional estimators:

$$S_n(\hat{\theta}) = \sqrt{n} \begin{pmatrix} \left(\hat{F}_{y|x} - F_{y|x}(x,y;\hat{\theta}), \psi_{1x}\right) \\ \vdots \\ \left(\hat{F}_{y|x} - F_{y|x}(x,y;\hat{\theta}), \psi_{Lx}\right) \end{pmatrix}. \tag{3.1}$$

We estimate $\left(F_{y|x}, \psi_x\right)$ employing (2.3) with empirical distribution functions, $\hat{F}_{x,y}$ and $\hat{F}_x$ to estimate the joint and the marginal distributions, thus

$$\int \hat{F}_{x,y}(x,y)\partial^{d_x}\psi_x(\hat{F}_x)d\hat{F}_x = \frac{1}{n}\sum_{t=1}^{n} \hat{F}_{x,y}(x_t,y)\partial^{d_x}\psi_x(\hat{F}_x(x_t)) \tag{3.2}$$

and

$$\int F_{y|x}(x,y;\hat{\theta})\psi_x(\hat{F}_x)d\hat{F}_x = \frac{1}{n}\sum_{t=1}^{n} F_{y|x}(x_t,y;\hat{\theta})\psi_x(\hat{F}_x(x_t)), \tag{3.3}$$

hence no numerical integration is required for the estimation of the linear functionals. Without loss of generality, we show the asymptotic normality for $L = 1$. From (3.2) and (3.3) we have for $y = \bar{y}$ fixed

$$\begin{aligned} S_n(\hat{\theta}) &= \sqrt{n}(\hat{F}_{y|x}(x,\bar{y}) - F_{y|x}(x,\bar{y};\hat{\theta}), \psi_x) \\ &= \sqrt{n}\left(\int \hat{F}_{x,y}(x,\bar{y})\partial^{d_x}\psi_x(\hat{F}_x)d\hat{F}_x - \int F_{y|x}(x,\bar{y};\hat{\theta})\psi_x(\hat{F}_x)d\hat{F}_x\right) \\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left(\hat{F}_{x,y}(x_t,\bar{y})\partial^{d_x}\psi_x(\hat{F}_x(x_t)) - F_{y|x}(x_t,\bar{y};\hat{\theta})\psi_x(\hat{F}_x(x_t))\right). \end{aligned}$$

To determine the asymptotic distribution of $S_n(\hat{\theta})$, write

$$\sqrt{n}(\hat{F}_{y|x}(x,\bar{y}) - F_{y|x}(x,\bar{y};\hat{\theta}),\psi_x) = \sqrt{n}(\hat{F}_{y|x}(x,\bar{y}) - F_{y|x}(x,\bar{y};\theta_0),\psi_x) \\ - \sqrt{n}(F_{y|x}(x,\bar{y};\hat{\theta}) - F_{y|x}(x,\bar{y};\theta_0),\psi_x). \qquad (3.4)$$

The limiting distribution of the first term on the right-hand side of (3.4) is given by $(Q_{y|x},\psi_x)$ in (2.7) . Upon linearization, the second term is shown to follow normal distribution asymptotically under some regularity conditions. It will then follow that $S_n(\hat{\theta})$ has a limit Gaussian distribution with a variance that can be consistently estimated by some estimator $\hat{\Sigma}$, and the statistic defined by

$$T_n(\hat{\theta}) = S_n(\hat{\theta})'\hat{\Sigma}^{-1}S_n(\hat{\theta}) \qquad (3.5)$$

is asymptotically χ^2_1 distributed under the null hypothesis. Since the statistic is asymptotically pivotal, bootstrap is a natural alternative to tests based on the limit χ^2 distribution and could provide refinement. Define

$$H_n(\theta) = \frac{1}{n}\sum_{t=1}^{n}\frac{\partial F_{y|x}}{\partial\theta'}(x_t,\bar{y};\theta)\psi_x(F(x_t)).$$

The formal regularity conditions under which we establish the limiting distribution of the parametric component of $S_n(\hat{\theta})$ are stated in the following assumption:

Assumption 3. *(a)* $\hat{\theta} \xrightarrow{p} \theta_0$ *with* $\theta_0 \in \text{interior}(\Theta)$, $\Theta \subseteq \mathbb{R}^p$. *(b)* $F_{y|x}(x,y;\theta)$ *is continuously differentiable with respect to* θ *in some neighborhood* $\mathcal{N}$ *of* θ_0 *and admits a density function* $f_{y|x}(x,y;\theta)$. *(c) The log-likelihood function* $\ell_n(\theta) = \sum_{t=1}^{n}\log f_{y|x}(x_t,y_t;\theta)$ *is twice continuously differentiable in* $\mathcal{N}$ *with first and second order derivatives* $\partial\ell_n(\theta)/\partial\theta$ *and* $\partial^2\ell_n(\theta)/\partial\theta\partial\theta'$ *respectively. (d)* $(1/\sqrt{n})\partial\ell_n(\theta_0)/\partial\theta \xrightarrow{d} N[0,I(\theta_0)]$ *where* $I(\theta_0)$ *is nonsingular and finite, and* $I(\theta)$ *is continuous at* θ_0. *(e)* $\sup_{\|h\|\le\delta}|-(1/n)\partial^2\ell_n(\theta_0 + h/\sqrt{n})/\partial\theta\partial\theta' - I(\theta_0)| = o_p(1)$ *where* $\delta > 0$ *is finite and* $h \in \mathbb{R}^p$. *(f)* $\sup_{\theta\in\mathcal{N}}\|H_n(\theta) - H(\theta)\| \xrightarrow{p} 0$ *with* $H(\theta)$ *continuous at* θ_0.

The regularity conditions concerning the likelihood function and the MLE are standard (see e.g., Theorem 3.1 of Newey and McFadden (1994)). Assumption 3 (f) is a local uniform convergence condition similar to (e) and is implied by the conditions under

which the uniform law of large numbers for $H_n(\theta)$ holds. For i.i.d. data, this condition could be replaced by local uniform integrability of $(\partial F_{y|x}/\partial\theta')(x_t, \bar{y}; \theta)$ on $\mathcal{N}$ as in Lemma 4.3 of Newey and McFadden (1994). The next result characterizes the asymptotic distribution of the linear functional estimator when the parameter θ_0 is estimated by the MLE $\hat{\theta}$.

Proposition 3. *Let Assumptions 2 and 3 hold. Under the null hypothesis that* H_0 : $F_{y|x}(x, y) = F_{y|x}(x, y; \theta_0)$ *for some* $\theta_0 \in \Theta$*, we have, as* $n \to \infty$

$$S_n(\hat{\theta}) \xrightarrow{d} (Q_{y|x}, \psi_x) - \left[Z + \int F_{y|x}(x, \bar{y}; \theta_0) J\left(d^{F_x}\psi_x(F_x), U_x\right) dF_x + \int F_{y|x}(x, \bar{y}; \theta_0)\psi_x(F_x) dU_x \right] \tag{3.6}$$

where $Z \sim N[0, H(\theta_0)I(\theta_0)^{-1}H(\theta_0)']$.

4 Simulations

This section presents simulation evidence on the performance of estimators and test statistics. We use (3.2) for the estimation of $\left(F_{y|x}, \psi_x\right)$:

$$(\hat{Q}_{y|x}, \psi_x) = -\int \hat{F}_{x,y}\partial\psi_x(\hat{F}_x)d\hat{F}_x = -\frac{1}{n}\sum_{t=1}^{n} \hat{F}_{x,y}(x_t, y)\partial\psi_x(\hat{F}_x(x_t))$$

following (2.3). All calculations were carried out in R Version 3.0.2 (R Development Core Team (2013)).

4.1 Numerical evaluation of test function selection

We conduct a simple experiment to explore which test function provides better results for the distribution of estimated values of the functional. We generate several joint distributions similar to those considered by Donoho and Johnstone (1995): y_t is generated according to

$$y_t = \frac{g(x_t)}{s_g} + u_t, \quad x_i \sim \text{Uniform}[0, 1] \quad u_t \sim N[0, \sigma^2], \quad t = 1, \ldots, n,$$

where s_g is the standard deviation of $g(x_t)$. The functions used are

$$
\begin{aligned}
\text{HeaviSine}: \quad & g(x) = 4\sin 4\pi x - \operatorname{sign}(x-0.3) - \operatorname{sign}(0.72-x) \\
\text{Blocks}: \quad & g(x) = \sum_{j=1}^{11} h_j \varphi(x-\tau_j), \quad \varphi(x) = \frac{1}{2}(1+\operatorname{sign}(x)) \\
& \tau = (0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81) \\
& h = (4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 2.1, -4.2) \\
\text{Bumps}: \quad & g(x) = \sum_{j=1}^{11} h_j \varphi\left(\frac{x-\tau_j}{w_j}\right), \quad \varphi(x) = \frac{1}{(1+|x|)^4} \\
& \tau = (0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81) \\
& h = (4, 5, 3, 4, 5, 4.2, 2.1, 4.3, 3.1, 5.1, 4.2) \\
& w = (0.005, 0.005, 0.006, 0.01, 0.01, 0.03, 0.01, 0.01, 0.005, 0.008, 0.005) \\
\text{Doppler}: \quad & g(x) = \sqrt{x(1-x)}\sin\left(\frac{2\pi(1+\varepsilon)}{x+\varepsilon}\right), \quad \varepsilon = 0.05.
\end{aligned}
$$

In the simulation experiment, the following functions are used as test functions: quadratic $\psi_x(x) = x(1-x)1(0 \le x \le 1)$, quartic $\psi_x(x) = x^2(1-x)^2 1(0 \le x \le 1)$, and bump $\psi_x(x) = \exp(-1/x(1-x))1(0 \le x \le 1)$. The number of simulated samples is $N = 1000$, and $n = 50$. The values of y at which the linear functional $\left(F_{y|x}, \psi_x\right)$ is estimated are 21 equidistant grid points on $[-5, 5]$. We average the values of $\sqrt{n}$ times the absolute bias and $\sqrt{n}$ times the root mean squared error (RMSE) over the grid points.

The results are reported in Table 1. On comparing $\sigma = 2$ to $\sigma = 1$, we see that the absolute bias and the RMSE of the test functions are stable for Blocks and Bumps, greater for Doppler, and somewhat lower for HeaviSine when $\sigma = 2$. The bump function (see Figure 1) outperforms the quadratic and quartic functions in terms of bias and RMSE in all cases, so we shall proceed with this function for the example studied in Section 4.2.

Choosing a test function ψ_x with good properties is obviously an important problem that calls for a further investigation. In general, an optimal choice of test function is likely to be specific to the problem at hand, and a considerable effort would be required for determining such functions.

Table 1: Mean absolute bias and RMSE of estimators of $(F_{y|x}, \psi_x)$ over 1000 simulated samples of size $n = 50$

$\sigma = 1$	HeaviSine		Blocks		Bumps		Doppler	
Test Functions	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Quadratic	0.151	0.162	0.452	0.457	0.454	0.458	0.072	0.087
Quartic	0.032	0.035	0.096	0.097	0.097	0.098	0.009	0.014
Bump	0.007	0.008	0.020	0.021	0.020	0.021	0.002	0.003
$\sigma = 2$	HeaviSine		Blocks		Bumps		Doppler	
Test Functions	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Quadratic	0.122	0.146	0.473	0.481	0.465	0.472	0.248	0.262
Quartic	0.021	0.027	0.100	0.102	0.099	0.101	0.046	0.050
Bump	0.005	0.007	0.021	0.022	0.021	0.021	0.010	0.011

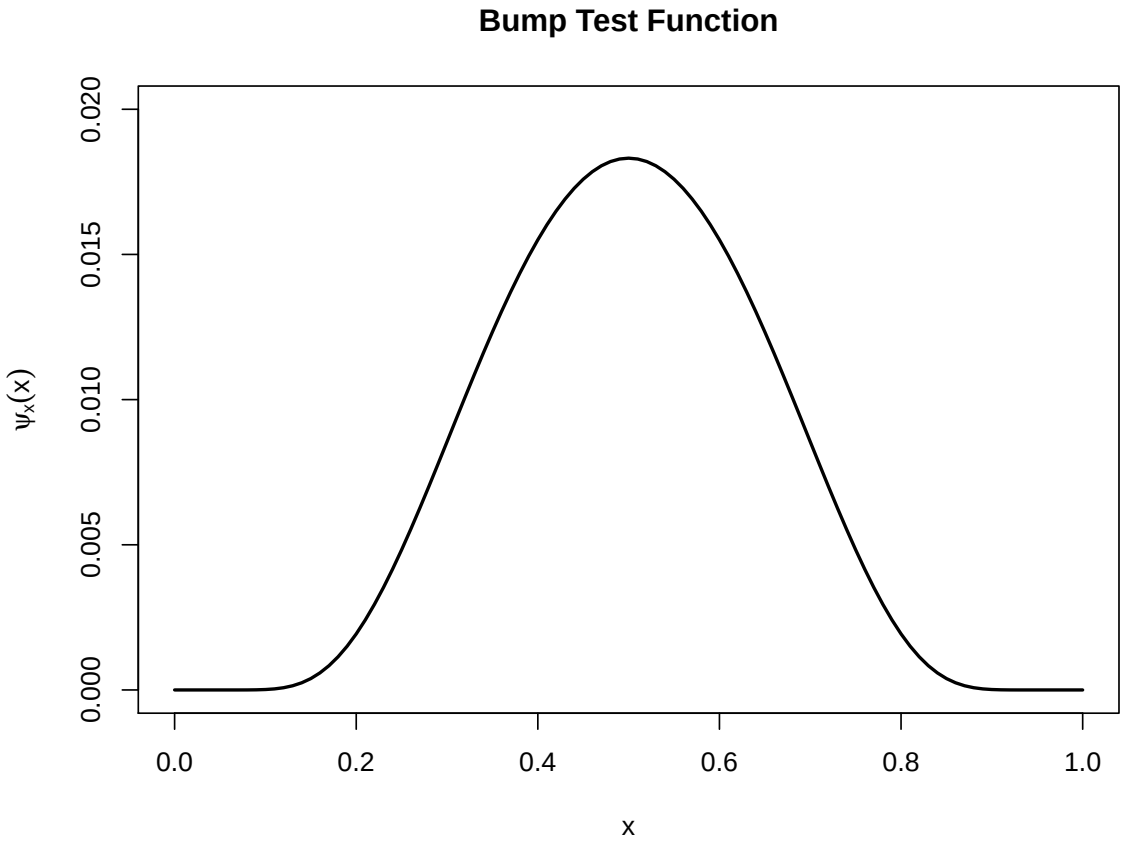

Figure 1: Bump function

4.2 A test of parametric specification in a binary choice model

The specification we use in the simulation is:

$$P(y_t = 1|x_t) = \Psi\left(x_t'\theta + c(\sum_{j=1}^{p}|x_{tj}|)\right) \quad t = 1, \ldots, n. \tag{4.1}$$

where $x_t = (x_{t1}, \ldots, x_{tp})' \sim N[0, I_p]$ and Ψ is either the logistic ($\Psi(u) = \exp(u)/(1 + \exp(u))$ or the standard normal (probit) CDF. The logit or probit model under the null hypothesis has $c = 0$. The alternative is generated with $c = 0.5, 1, 2, 3$. The parameter values are $\theta = (-1/\sqrt{2}, 1/\sqrt{2})'$ for $p = 2$, and $\theta = (-1/\sqrt{3}, 1/\sqrt{3}, -1/\sqrt{3})$ for $p = 3$. The logit design and the parameter configurations $c = 1, 2, 3$ follow Stute and Zhu (2005); we consider additionally the probit model and a closer alternative with $c = 0.5$.

Since, for $0 \leq \bar{y} < 1$, $F_{y|x}(x_t, \bar{y}; \theta) = P(y_t \leq \bar{y}|x_t, \theta) = P(y_t = 0|x_t, \theta)$, it suffices to determine the asymptotic distribution of the test statistic at a single value, say $\bar{y} = 0.5$, in which case the corresponding test statistics S_n and T_n are asymptotically normal and chi-square distributed, respectively. The test based on the statistic S_n is a percentile bootstrap type procedure; we generate $B = 199$ bootstrap samples $\{(x_t', y_t^*)'\}_{t=1}^n$ according to $P(y_t^* = 1) = \Psi(x_t'\hat{\theta})$ where $\hat{\theta}$ is the MLE of θ, and compute the bootstrap statistics $S_{n,j}^*, j = 1, \ldots, B$. We then reject the null hypothesis at level 5% if S_n is outside the $(0.025, 0.975)$ range of the bootstrap distribution. To construct the statistic T_n, the sample mean $\bar{S}^*$ of $S_{n,j}^*, j = 1, \ldots, B$ is subtracted off from S_n to correct the finite sample bias, and Σ is estimated by the sample covariance of the bootstrap samples, $\mathrm{Var}(S_n^*)$. The test rejects at 5% level if the statistic $T_n = (S_n - \bar{S}^*)^2/\mathrm{Var}(S_n^*)$ is greater than the 0.95 quantile of χ_1^2 distribution. The number of replications is $N = 1000$ as before.

We report the estimated size and powers of the tests in Tables 2 and 3. We generally note that though both tests are undersized, the S_n test provides an improvement bringing the size closer to nominal. Also, the sizes of the tests improve as the sample size increases. Despite the low probability of rejection under the null, the power is substantial in both models. The percentile bootstrap test based on S_n dominates the asymptotic test T_n in terms of the power. Furthermore, in every case, the powers of the tests S_n and T_n are greater than those reported by Stute and Zhu (2005) for the logit model with $c = 1, 2, 3$.

Table 2: The null rejection frequency under (4.1) with $c = 0$ at 5% level

Logit	$p = 2$		$p = 3$	
Test statistics	S_n	T_n	S_n	T_n
$n = 50$	0.032	0.025	0.032	0.030
$n = 100$	0.030	0.039	0.041	0.030
Probit	$p = 2$		$p = 3$	
Test Statistics	S_n	T_n	S_n	T_n
$n = 50$	0.039	0.033	0.025	0.022
$n = 100$	0.051	0.041	0.034	0.031

Table 3: Powers of tests for H_0: (4.1) with $c = 0$ against H_1: (4.1) with $c = 0.5, 1, 2, 3$. at 5% level

Logit model								
Test statistics	S_n	T_n	S_n	T_n	S_n	T_n	S_n	T_n
$p = 2$	$c = 0.5$		$c = 1$		$c = 2$		$c = 3$	
$n = 50$	0.315	0.207	0.623	0.486	0.737	0.645	0.752	0.658
$n = 100$	0.521	0.423	0.790	0.735	0.827	0.805	0.830	0.803
$p = 3$	$c = 0.5$		$c = 1$		$c = 2$		$c = 3$	
$n = 50$	0.659	0.533	0.915	0.873	0.945	0.919	0.950	0.927
$n = 100$	0.924	0.883	0.992	0.989	0.996	0.992	0.996	0.990
Probit model								
Test statistics	S_n	T_n	S_n	T_n	S_n	T_n	S_n	T_n
$p = 2$	$c = 0.5$		$c = 1$		$c = 2$		$c = 3$	
$n = 50$	0.526	0.369	0.741	0.656	0.749	0.661	0.756	0.675
$n = 100$	0.746	0.677	0.844	0.815	0.837	0.806	0.849	0.811
$p = 3$	$c = 0.5$		$c = 1$		$c = 2$		$c = 3$	
$n = 50$	0.881	0.789	0.944	0.905	0.947	0.935	0.957	0.946
$n = 100$	0.979	0.972	0.997	0.989	0.991	0.988	0.995	0.992

5 Conclusion

This paper establishes the asymptotic properties of estimators of conditional distribution and density functions for stationary mixing data by treating them as elements of the space of generalized functions. Accordingly, no smoothness restriction is needed for the validity of asymptotic distribution of the proposed estimators; the possible irregularities associated with the distribution and density functions are smoothed out through the use of an appropriate test function. The computationally inexpensive linear functional estimator that is based on the empirical distribution performs satisfactorily for a moderate size sample. The percentile bootstrap procedure based on the proposed estimator is applied to testing functional form in the binary choice model and has good size and power properties in the simulations conducted.

References

Andrews, D. W. K. (1997): "A Conditional Kolmogorov Test," *Econometrica*, 65, 1097–1128.

——— (1999): "Estimation When a Parameter is on a Boundary," *Econometrica*, 67, 1341–1383.

Darsow, W. F., B. Nguyen, and E. T. Olsen (1992): "Copulas and Markov Processes," *Illinois Journal of Mathematics*, 36(4), 600–642.

Donoho, D. L., and I. M. Johnstone (1995): "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statisitcal Association*, 90(432), 1200–1224.

Gel'fand, I. M., and G. E. Shilov (1964): *Generalized Functions*, vol. 1-2: Properties and Operations. Academic Press, San Diego.

Li, Q., and J. S. Racine (2007): *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. IV, chap. 36, pp. 2111–2245. Elsevier Science.

Pagan, A., and A. Ullah (1999): *Nonparametric Econometrics*, Themes in Modern Econometrics. Cambridge, U.K.

Phillips, P. C. B. (1985): "A Theorem on the Tail Behaviour of Probability Distributions with an Application to the Stable Family," *Canadian Journal of Economics*, 18(1), 58–65.

——— (1991): "A Shortcut to LAD Estimator Asymptotics," *Econometric Theory*, 7, 450–463.

——— (1995): "Robust Nonstationary Regression," *Econometric Theory*, 11, 450–463.

R Development Core Team (2013): *R: A Language and Environment for Statistical Computing*R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rio, E. (2000): *Théorie asymptotique des processus aléatoires faiblement dépendants*, vol. 31 of *Mathématiques and Applications*. Springer.

Sklar, A. (1973): "Random Variables, Joint Distribution Functions, and Copulas," *Kybernetika*, 9(6), 449–460.

Stute, W., and L.-X. Zhu (2005): "Nonparametric Checks for Single-Index Models," *The Annals of Statistics*, 33(3), 1048–1083.

van der Vaart, A. W., and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer.

Zinde-Walsh, V. (2013): "Nonparametric Functionals as Generalized Functions," *arXiv:1303.1435*.

A Proofs

Theorem 1. For any $\psi_x \in D((0,1)^{d_x})$, by the lemma in Zinde-Walsh (2013) the value of the functional $(\tilde{F}_{y|x}, \psi_x)$ is well defined by

$$(\tilde{F}_{y|x}, \psi_x) = (-1)^{d_x} \int \frac{1}{n}\sum_{t=1}^{n} G\left(\frac{y-y_t}{h_y}\right) K\left(\frac{x_t - x}{h}\right) \partial^{d_x}\psi_x\left(\frac{1}{n}\sum_{t=1}^{n} K\left(\frac{x_t-x}{h}\right)\right) d\left(\frac{1}{n}\sum_{t=1}^{n} K\left(\frac{x_t-x}{h}\right)\right),$$

where G and K are integrals of the kernel functions g and k. Write

$$\int \tilde{F}_{x,y}\partial^{d_x}\psi_x(\tilde{F}_x)d\tilde{F}_x = \int \hat{F}_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)d\hat{F}_x + \left[\int \tilde{F}_{x,y}\partial^{d_x}\psi_x(\tilde{F}_x)d\tilde{F}_x - \int \hat{F}_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)d\hat{F}_x\right].$$

Calculating the bias by expansions, we find

$$\int \tilde{F}_{x,y}\partial^{d_x}\psi_x(\tilde{F}_x)d\tilde{F}_x - \int \hat{F}_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)d\hat{F}_x = O_p(h^2).$$

Clearly,

$$\begin{aligned}\int \hat{F}_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)d\hat{F}_x = &\left[\int F_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)dF_x + \int F_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)d(\hat{F}_x - F_x)\right] \\ &+ \left[\int (\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi_x(\hat{F}_x)dF_x + \int (\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi(\hat{F}_x)d(\hat{F}_x - F_x)\right].\end{aligned} \tag{A.1}$$

By Taylor expansion,

$$\partial^{d_x}\psi_x(\hat{F}_x(x)) = \partial^{d_x}\psi_x(F_x(x)) + \sum_{i=1}^{d_x}\frac{\partial}{\partial F_{x_i}}\partial^{d_x}\psi_x\left(F_x\right)\left(\hat{F}_{x_i}\left(x\right) - F_{x_i}(x)\right) + r(\hat{F}_x(x) - F_x(x)), \tag{A.2}$$

where $x^2 r(x) \to 0$ as $x \to 0$, and by the Mean Value theorem

$$\partial^{d_x}\psi_x(\hat{F}_x(x)) = \partial^{d_x}\psi_x(F_x(x)) + \sum_{i=1}^{d_x}\frac{\partial}{\partial F_{x_i}}\partial^{d_x}\psi_x\left(\bar{F}_x\right)\left(\hat{F}_{x_i}\left(x\right) - F_{x_i}(x)\right) \tag{A.3}$$

where $\bar{F}_x \in (0,1)^{d_x}$ is a mixture of $\hat{F}_x$ and F_x. Then the first term in (A.1) can be written as

$$\begin{aligned}\int F_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)dF_x = &\int F_{x,y}\partial^{d_x}\psi_x(F_x)dF_x + \int F_{x,y}J\left(d^{F_x}\partial^{d_x}\psi_x\left(F_x\right), \hat{F}_x - F_x\right)dF_x \\ &+ \int F_{x,y}\partial^{d_x}\psi_x(F_x)r(\hat{F}_x - F_x)dF_x.\end{aligned}$$

Using (A.3), the remaining terms in (A.1) become

$$\begin{aligned}\int F_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)d(\hat{F}_x - F_x) = &\int F_{x,y}\partial^{d_x}\psi_x(F_x)d(\hat{F}_x - F_x) \\ &+ \int F_{x,y}J\left(d^{F_x}\partial^{d_x}\psi_x\left(\bar{F}_x\right), \hat{F}_x - F_x\right)d(\hat{F}_x - F_x),\end{aligned} \tag{A.4}$$

$$\int (\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi_x(\hat{F}_x)dF_x = \int (\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi_x(F_x)dF_x + \int (\hat{F}_{x,y} - F_{x,y})J\left(d^{F_x}\partial^{d_x}\psi_x\left(\bar{F}_x\right), \hat{F}_x - F_x\right)dF_x \tag{A.5}$$

and

$$\int (\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi_x(\hat{F}_x)d(\hat{F}_x - F_x) = \int (\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi_x(F_x)d(\hat{F}_x - F_x) + \int (\hat{F}_{x,y} - F_{x,y})J\left(d^{F_x}\partial^{d_x}\psi_x\left(\bar{F}_x\right), \hat{F}_x - F_x\right)d(\hat{F}_x - F_x). \tag{A.6}$$

By properties of $\psi_x \in D(W)$, the function $d^{F_x}\partial^{d_x}\psi_x\left(\bar{F}_x\right)$ is bounded. By weak convergence of the empirical distribution function for multivariate α-mixing processes (see Theorem 7.3 of Rio (2000)),

$$\sqrt{n}(\hat{F}_{x,y} - F_{x,y}) \xrightarrow{d} U_{x,y} \quad \sqrt{n}(\hat{F}_x - F_x) \xrightarrow{d} U_x,$$

where $U_{x,y}$ and U_x are Gaussian processes with dimension $d_x + d_y$ and d_x respectively. Then we can express $\sqrt{n}\left(\hat{F}_{y|x} - F_{y|x}, \psi_x\right)$ as

$$Q_\psi\left(\sqrt{n}(\hat{F}_x - F_x), \sqrt{n}(\hat{F}_{x,y} - F_{x,y})\right) + \sqrt{n}R\left(\sqrt{n}(\hat{F}_x - F_x), \sqrt{n}(\hat{F}_{x,y} - F_{x,y})\right),$$

where

$$\begin{aligned} &Q_\psi\left(\sqrt{n}(\hat{F}_x - F_x), \sqrt{n}(\hat{F}_{x,y} - F_{x,y})\right) \\ &= \int F_{x,y}\partial^{d_x}\psi_x\left(F_x\right)d\sqrt{n}(\hat{F}_x - F_x) + \int \sqrt{n}(\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi_x\left(F_x\right)dF_x \\ &\quad + \int F_{x,y}J\left(d^{F_x}\partial^{d_x}\psi_x\left(F_x\right), \sqrt{n}(\hat{F}_x - F_x)\right)dF_x \end{aligned}$$

and $R\left(\cdot,\cdot\right)$ is a bounded function. Since the function Q_ψ is continuous in its arguments, by substituting the limiting Gaussian processes for the arguments of $Q_\psi\left(\cdot,\cdot\right)$ we obtain $\sqrt{n}(\hat{F}_{y|x} - F_{y|x}, \psi_x) \xrightarrow{d} (Q_{y|x}, \psi_x) = Q_\psi\left(U_x, U_{x,y}\right)$. For any (omitting the subscript x)

$\psi_1, \ldots, \psi_l \in D(W)$, the joint limiting Gaussian process for

$$\sqrt{n}\left(\hat{F}_{y|x} - F_{y|x}, \psi_1\right), \ldots, \sqrt{n}\left(\hat{F}_{y|x} - F_{y|x}, \psi_l\right)$$

is similarly derived from the joint process of $Q_{\psi_1}(U_x, U_{x,y}), \ldots, Q_{\psi_l}(U_x, U_{x,y})$. The mean of the process is zero since Q_ψ is linear in its arguments and the covariance is given by $\mathrm{Cov}\left(Q_{\psi_1}(U_x, U_{x,y}), Q_{\psi_2}(U_x, U_{x,y})\right) = \mathrm{Cov}\left((Q_{y|x}, \psi_1), (Q_{y|x}, \psi_2)\right)$. Since by assumption $h^2 = o(n^{-\frac{1}{2}})$, the limiting process on $D(W)$ is fully described by $Q_{y|x}$. Since $\psi_{x,y}$ is a product function, this implies (2.7). □

Theorem 2. By proceeding similarly to Theorem 1, we have, for any $\psi \in D((0,1)^{d_x}) \times D(\mathbb{R}^{d_y})$,

$$(\tilde{f}_{y|x}, \psi_{x,y}) = (-1)^{d_x+d_y} \int\int \tilde{F}_{x,y}(x,y)\partial^{d_x}\psi_x(\tilde{F}_x)\partial^{d_y}\psi_y(y)d\tilde{F}_x dy.$$

Decompose

$$\begin{aligned}
&\int\int \tilde{F}_{x,y}(x,y)\partial^{d_x}\psi_x(\tilde{F}_x)\partial^{d_y}\psi_y(y)d\tilde{F}_x dy \\
&\quad = \int\int \hat{F}_{x,y}(x,y)\partial^{d_x}\psi_x(\hat{F}_x)\partial^{d_y}\psi_y(y)d\hat{F}_x dy \\
&\qquad + \left[\int\int \tilde{F}_{x,y}(x,y)\partial^{d_x}\psi_x(\tilde{F}_x)\partial^{d_y}\psi_y(y)d\tilde{F}_x dy - \int\int \hat{F}_{x,y}(x,y)\partial^{d_x}\psi_x(\hat{F}_x)\partial^{d_y}\psi_y(y)d\hat{F}_x dy\right].
\end{aligned}$$

The second term in the bracket is $O_p(h^2)$. Write the first term as

$$\begin{aligned}
\int\int \hat{F}_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)\partial^{d_y}\psi_y d\hat{F}_x dy &= \int\int F_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)\partial^{d_y}\psi_y dF_x dy && \text{(A.7)} \\
&\quad + \int\int F_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)\partial^{d_y}\psi_y d(\hat{F}_x - F_x)dy && \text{(A.8)} \\
&\quad + \int\int (\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi_x(\hat{F}_x)\partial^{d_y}\psi_y dF_x dy && \text{(A.9)} \\
&\quad + \int\int (\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi_x(\hat{F}_x)\partial^{d_y}\psi_y d(\hat{F}_x - F_x)dy. && \text{(A.10)}
\end{aligned}$$

For (A.7), using (A.2) we have

$$\begin{aligned}
&\int\int F_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)\partial^{d_y}\psi_y dF_x dy \\
&= \int\int F_{x,y}\partial^{d_x}\psi_x(F_x)\partial^{d_y}\psi_y dF_x dy + \int\int F_{x,y} J\left(d^{F_x}\partial^{d_x}\psi_x\left(F_x\right), \hat{F}_x - F_x\right)\partial^{d_y}\psi_y dF_x dy \\
&\quad + \int\int F_{x,y} r(\hat{F}_x - F_x)\partial^{d_y}\psi_y dF_x dy.
\end{aligned}$$

The first term in the right hand side of the above expression is the true value of the functional. When scaled by factor $\sqrt{n}$, the second term converges in distribution to $\int\int F_{x,y} J\left(d^{F_x}\partial^{d_x}\psi_x\left(F_x\right), U_x\right)\partial^{d_y}\psi_y dU_x dy$ and the third term is negligible. From (A.3), (A.8) and (A.9) become

$$\begin{aligned}
&\int\int F_{x,y}\partial^{d_x}\psi_x(\hat{F}_x)\partial^{d_y}\psi_y d(\hat{F}_x - F_x)dy \\
&= \int\int F_{x,y}\partial^{d_x}\psi_x(F_x)\partial^{d_y}\psi_y d(\hat{F}_x - F_x)dy \\
&\quad + \int\int F_{x,y} J\left(d^{F_x}\partial^{d_x}\psi_x\left(\bar{F}_x\right), \hat{F}_x - F_x\right)\partial^{d_y}\psi_y d(\hat{F}_x - F_x)dy
\end{aligned}$$

and

$$\begin{aligned}
&\int\int (\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi_x(\hat{F}_x)\partial^{d_y}\psi_y dF_x dy \\
&= \int\int (\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi_x(F_x)\partial^{d_y}\psi_y dF_x dy \\
&\quad + \int\int (\hat{F}_{x,y} - F_{x,y}) J\left(d^{F_x}\partial^{d_x}\psi_x\left(\bar{F}_x\right), \hat{F}_x - F_x\right)\partial^{d_y}\psi_y dF_x dy.
\end{aligned}$$

Only the first terms on right hand side of the above expressions converge in distribution to nondegenerate random variables given by

$$\int\int F_{x,y}\partial^{d_x}\psi_x(F_x)\partial^{d_y}\psi_y(y)dU_x dy \quad \text{and} \quad \int\int U_{x,y}\partial^{d_x}\psi_x(F_x)\partial^{d_y}\psi_y(y)dF_x dy$$

respectively. Finally, we see that (A.10) is negligible by writing

$$\begin{aligned}
&\int\int (\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi_x(\hat{F}_x)\partial^{d_y}\psi_y d(\hat{F}_x - F_x)dy \\
&= \int\int (\hat{F}_{x,y} - F_{x,y})\partial^{d_x}\psi_x(F_x)\partial^{d_y}\psi_y d(\hat{F}_x - F_x)dy
\end{aligned}$$

$$+ \int \int (\hat{F}_{x,y} - F_{x,y}) J \left(d^{F_x} \partial^{d_x} \psi_x \left(\bar{F}_x \right), \hat{F}_x - F_x \right) \partial^{d_y} \psi_y d(\hat{F}_x - F_x) dy.$$

Taken together,

$$\begin{aligned}\sqrt{n}(\tilde{f}_{y|x} - f_{y|x}, \psi_{x,y}) \xrightarrow{d} (-1)^{d_x+d_y} \Bigg[& \int \int F_{x,y} J \left(d^{F_x} \partial^{d_x} \psi_x \left(F_x \right), U_x \right) \partial^{d_y} \psi_y dU_x dy \\ & + \int \int F_{x,y} \partial^{d_x} \psi_x(F_x) \partial^{d_y} \psi_y(y) dU_x dy \\ & + \int \int \partial^{d_x} \psi_x(F_x) U_{x,y} \partial^{d_y} \psi_y(y) dF_x dy \Bigg].\end{aligned}$$

□

Proposition 3. Recall the decomposition (3.4):

$$\begin{aligned}& \sqrt{n} \left[\int \hat{F}_{x,y}(x, \bar{y}) \partial^{d_x} \psi_x(\hat{F}_x) d\hat{F}_x - F_{y|x}(x, \bar{y}; \hat{\theta}) \psi_x(\hat{F}_x) d\hat{F}_x \right] \\ & = \sqrt{n} \left[\int \hat{F}_{x,y}(x, \bar{y}) \partial^{d_x} \psi_x(\hat{F}_x) d\hat{F}_x - \int F_{x,y}(x, \bar{y}) \partial^{d_x} \psi_x(F_x) dF_x \right] \\ & \quad - \sqrt{n} \left[\int F_{y|x}(x, \bar{y}; \hat{\theta}) \psi_x(\hat{F}_x) d\hat{F}_x - \int F_{y|x}(x, \bar{y}; \theta_0) \psi_x(F_x) dF_x \right].\end{aligned}$$

By Theorem 1, the first component converges in distribution to $(Q_{y|x}, \psi_x)$. Write the second term as

$$\begin{aligned}& \sqrt{n} \left[\int F_{y|x}(x, \bar{y}; \hat{\theta}) \psi_x(\hat{F}_x) d\hat{F}_x - \int F_{y|x}(x, \bar{y}; \theta_0) \psi_x(F_x) dF_x \right] \\ & = \sqrt{n} \int F_{y|x}(x, \bar{y}; \theta_0) \psi_x(F_x) d(\hat{F}_x - F_x) && \text{(A.11)} \\ & \quad + \sqrt{n} \int F_{y|x}(x, \bar{y}; \theta_0) \left(\psi_x(\hat{F}_x) - \psi_x(F_x) \right) d\hat{F}_x && \text{(A.12)} \\ & \quad + \sqrt{n} \int \left(F_{y|x}(x, \bar{y}; \hat{\theta}) - F_{y|x}(x, \bar{y}; \theta_0) \right) \psi_x(\hat{F}_x) d\hat{F}_x. && \text{(A.13)}\end{aligned}$$

We find the limiting distribution of each term in the above expression separately. For (A.11), we have

$$\sqrt{n} \int F_{y|x}(x, \bar{y}; \theta_0) \psi_x(F_x) d(\hat{F}_x - F_x) \xrightarrow{d} \int F_{y|x}(x, \bar{y}; \theta_0) \psi_x(F_x) dU_x \qquad \text{(A.14)}$$

as $n \to \infty$. On writing (A.12) as

$$\begin{aligned}
&\sqrt{n} \int F_{y|x}(x, \bar{y}; \theta_0) \left(\psi_x(\hat{F}_x) - \psi_x(F_x) \right) d\hat{F}_x \\
&= \sqrt{n} \int F_{y|x}(x, \bar{y}; \theta_0) \left(\psi_x(\hat{F}_x) - \psi_x(F_x) \right) dF_x \\
&\quad + \sqrt{n} \int F_{y|x}(x, \bar{y}; \theta_0) \left(\psi_x(\hat{F}_x) - \psi_x(F_x) \right) d(\hat{F}_x - F_x)
\end{aligned}$$

and applying the Mean Value theorem $\psi_x(\hat{F}_x) - \psi_x(F_x) = J\left(d^{F_x} \psi_x(\bar{F}_x), \hat{F}_x - F_x \right)$ and $\sqrt{n}(\hat{F}_x - F_x) \xrightarrow{d} U_x$ to each term in the last equation,

$$\begin{aligned}
&\sqrt{n} \int F_{y|x}(x, \bar{y}; \theta_0) \left(\psi_x(\hat{F}_x) - \psi_x(F_x) \right) d\hat{F}_x \\
&= \int F_{y|x}(x, \bar{y}; \theta_0) J\left(d^{F_x} \psi_x(F_x), \sqrt{n}(\hat{F}_x - F_x) \right) dF_x + o_P(1) \\
&\xrightarrow{d} \int F_{y|x}(x, \bar{y}; \theta_0) J\left(d^{F_x} \psi_x(F_x), U_x \right) dF_x. \qquad \text{(A.15)}
\end{aligned}$$

Assumption 3 (a)-(e) imply $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$ (see Andrews (1999) Theorem 1, van der Vaart and Wellner (1996) Chapter 3.4). Expanding the score function around the true parameter value and using Assumption 3 (d) and (e), we have $\sqrt{n}(\hat{\theta} - \theta_0) = I(\theta_0)^{-1}(1/\sqrt{n})\partial \ell_n(\theta_0)/\partial\theta + o_p(1) \xrightarrow{d} N[0, I(\theta_0)^{-1}]$. Then (A.13) becomes

$$\begin{aligned}
&\sqrt{n} \int \left(F_{y|x}(x, \bar{y}; \hat{\theta}) - F_{y|x}(x, \bar{y}; \theta_0) \right) \psi_x(\hat{F}_x) d\hat{F}_x \\
&= \frac{1}{n} \sum_{t=1}^{n} \sqrt{n} \left(F_{y|x}(x_t, \bar{y}; \hat{\theta}) - F_{y|x}(x_t, \bar{y}; \theta_0) \right) \psi_x(\hat{F}_x(x_t)) \\
&= \frac{1}{n} \sum_{t=1}^{n} \frac{\partial F_{y|x}}{\partial \theta'}(x_t, \bar{y}; \bar{\theta}) \psi_x(\hat{F}_x(x_t)) \sqrt{n}(\hat{\theta} - \theta_0) \\
&= \frac{1}{n} \sum_{t=1}^{n} \frac{\partial F_{y|x}}{\partial \theta'}(x_t, \bar{y}; \bar{\theta}) \psi_x(F_x(x_t)) \sqrt{n}(\hat{\theta} - \theta_0) + o_p(1) \\
&= H_n(\bar{\theta}) I(\theta_0)^{-1} \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\theta_0)}{\partial \theta} + o_p(1),
\end{aligned}$$

where the second equality is due to the Mean Value expansion around θ_0 with $\bar{\theta}$ being a convex combination of $\hat{\theta}$ and θ_0, and the third equality follows by expanding $\psi_x(\hat{F}_x(x_t))$ around $\psi_x(F_x(x_t))$ and using $\sqrt{n}(\hat{F}_x - F_x) = O_p(1)$, $\sqrt{n}(\hat{\theta} - \theta_0) = O_p(1)$ in conjunction

with Assumptions 3 (a) and (b). Since from Assumption 3 (a) and (g)

$$H_n(\bar{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial F_{y|x}}{\partial \theta'}(x_t, \bar{y}; \bar{\theta})\psi_x(F(x_t)) \xrightarrow{p} H(\theta_0),$$

using Slutsky's lemma we obtain

$$\sqrt{n}\int \left(F_{y|x}(x, \bar{y}; \hat{\theta}) - F_{y|x}(x, \bar{y}; \theta_0)\right) \psi_x(\hat{F}_x)d\hat{F}_x \xrightarrow{d} Z \sim N[0, H(\theta_0)I(\theta_0)^{-1}H(\theta_0)'] \tag{A.16}$$

as $n \to \infty$. Finally, it follows from (A.14), (A.15) and (A.16) that under the null hypothesis

$$S_n(\hat{\theta}) \xrightarrow{d} (Q_{y|x}, \psi_x) - \left(Z + \int F_{y|x}(x, \bar{y}; \theta_0) J\left(d^{F_x}\psi_x(F_x), U_x\right) dF_x + \int F_{y|x}(x, \bar{y}; \theta_0)\psi_x(F_x)dU_x\right)$$

as required.

□