Published in final edited form as:
Stoch Process Their Appl. 2016 December ; 126(12): 3733–3759. doi:10.1016/j.spa.2016.04.005.

Asymptotic Normality of Quadratic Estimators

James Robins, **Lingling Li**, **Eric Tchetgen**, and **Aad van der Vaart**
Departments of Biostatistics and Epidemiology, School of Public Health, Harvard University, Mathematical Institute, Leiden University

Abstract

We prove conditional asymptotic normality of a class of quadratic U-statistics that are dominated by their degenerate second order part and have kernels that change with the number of observations. These statistics arise in the construction of estimators in high-dimensional semi- and non-parametric models, and in the construction of nonparametric confidence sets. This is illustrated by estimation of the integral of a square of a density or regression function, and estimation of the mean response with missing data. We show that estimators are asymptotically normal even in the case that the rate is slower than the square root of the observations.

Keywords

Quadratic functional; Projection estimator; Rate of convergence; U-statistic

1. Introduction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. random vectors, taking values in sets $\mathscr{X} \times \mathbb{R}$, for an arbitrary measurable space $(\mathscr{X}, \mathscr{A})$ and $\mathbb{R}$ equipped with the Borel sets. For given symmetric, measurable functions $K_n: \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ consider the U-statistics

$$U_n = \frac{1}{n(n-1)} \sum\sum_{1 \le r \ne s \le n} K_n(X_r, X_s) Y_r Y_s. \qquad (1)$$

Would the kernel $(x_1, y_1, x_2, y_2) \mapsto K_n(x_1, x_2)y_1y_2$ of the U-statistic be independent of n and have a finite second moment, then either the sequence $\sqrt{n}(U_n - \mathrm{E}U_n)$ would be asymptotically normal or the sequence $n(U_n - \mathrm{E}U_n)$ would converge in distribution to Gaussian chaos. The two cases can be described in terms of the Hoeffding decomposition $U_n = \mathrm{E}U_n + U_n^{(1)} + U_n^{(2)}$ of U_n, where $U_n^{(1)}$ is the best approximation of $U_n - \mathrm{E}U_n$ by a sum of the type $\sum_{i=1}^n h(X_r, Y_r)$ and $U_n^{(2)}$ is the remainder, a degenerate U-statistic (compare (28) in Section 5). For a fixed kernel K_n the linear term $U_n^{(1)}$ dominates as soon as it is nonzero, in

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

which case asymptotic normality pertains; in the other case $U_n^{(1)}=0$ and the U-statistic possesses a nonnormal limit distribution.

If the kernel depends on n, then the separation between the linear and quadratic cases blurs. In this paper we are interested in this situation and specifically in kernels K_n that concentrate as $n \to \infty$ more and more near the diagonal of $\mathscr{X} \times \mathscr{X}$. In our situation the variance of the U-statistics is dominated by the quadratic term $U_n^{(2)}$. However, we show that the sequence $(U_n - \mathrm{E}\, U_n)/\sigma(U_n)$ is typically still asymptotically normal. The intuitive explanation is that the U-statistics behave asymptotically as "sums across the diagonal $r = s$" and thus behave as sums of independent variables. Our formal proof is based on establishing conditional asymptotic normality given a binning of the variables X_r in a partition of the set $\mathscr{X}$.

Statistics of the type (1) arise in many problems of estimating a functional on a semiparametric model, with K_n the kernel of a projection operator (see [1]). As illustrations we consider in this paper the problems of estimating $\int g^2(x)\, dx$ or $\int f^2(x)\, dG(x)$, where g is a density and f a regression function, and of estimating the mean treatment effect in missing data models. Rate-optimal estimators in the first of these three problems were considered by [2, 3, 4, 5, 6], among others. In Section 3 we prove asymptotic normality of the estimators in [4, 5], also in the case that the rate of convergence is slower than $\sqrt{n}$, usually considered to be the "nonnormal domain". For the second and third problems estimators of the form (1) were derived in [1, 7, 8, 9] using the theory of second-order estimating equations. Again we show that these are asymptotically normal, also in the case that the rate is slower than $\sqrt{n}$.

Statistics of the type (1) also arise in the construction of adaptive confidence sets, as in [10], where the asymptotic normality can be used to set precise confidence limits.

Previous work on U-statistics with kernels that depend on n includes [14, 15, 16, 17, 18]. These authors prove unconditional asymptotic normality using the martingale central limit theorem, under somewhat different conditions. Our proof uses a Lyapounov central limit theorem (with moment $2 + \varepsilon$) combined with a conditioning argument, and an inequality for moments of U-statistics due to E. Giné. Our conditions relate directly to the contraction of the kernel, and can be verified for a variety of kernels. The conditional form of our limit result should be useful to separate different roles for the observations, such as for constructing preliminary estimators and for constructing estimators of functionals. Another line of research (as in [11]) is concerned with U-statistics that are well approximated by their projection on the initial part of the eigenfunction expansion. This has no relation to the present work, as here the kernels explode and the U-statistic is asymptotically determined by the (eigen) directions "added" to the kernel as the number of observations increases. By making special choices of kernel and variables Y_i, the statistics (1) can reduce to certain chisquare statistics, studied in [12, 13].

The paper is organized as follows. In Section 2 we state the main result of the paper, the asymptotic normality of U-statistics of the type (1) under general conditions on the kernels K_n. Statistical applications are given in Section 3. In Section 4 the conditions of the main theorem shown to be satisfied by a variety of popular kernels, including wavelet, spline, convolution, and Fourier kernels. The proof of the main result is given in Section 5, while proofs for Section 4 are given in an appendix.

The notation $a \lesssim b$ means $a \leq Cb$ for a constant C that is fixed in the context. The notations $a_n \sim b_n$ and $a_n \ll b_n$ mean that $a_n/b_n \to 1$ and $a_n/b_n \to 0$, as $n \to \infty$. The space $L_2(G)$ is the set of measurable functions $f: \mathscr{X} \to \mathbb{R}$ that are square-integrable relative to the measure G and $\|f\|_G$ is the corresponding norm. The product $f \times g$ of two functions is to be understood as the function $(x_1, x_2) \mapsto f(x_1)g(x_2)$, whereas the product $F \times G$ of two measures is the product measure.

2. Main result

In this section we state the main result of the paper, the asymptotic normality of the U-statistics (1), under general conditions on the kernels K_n and distributions of the vectors (X_r, Y_r). For $q > 0$ let

$$\mu(x)=\mathrm{E}(Y_1|X_1=x),$$

$$\mu_q(x)=\mathrm{E}(|Y_1|^q|X_1=x)$$

be versions of the conditional (absolute) moments of Y_1 given X_1. For simplicity we assume that μ_1 and and μ_2 are uniformly bounded. The marginal distribution of X_1 is denoted by G.

The kernels are assumed to be measurable maps $K_n: \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ that are symmetric in their two arguments and satisfy $\int\int K_n^2 d(G \times G)<\infty$ for every n. Thus the corresponding *kernel operators* (with abuse of notation denoted by the same symbol)

$$K_nf(x)=\int f(v)K_n(x,v)dG(v), \qquad (2)$$

are continuous, linear operators $K_n: L_2(G) \to L_2(G)$. We assume that their operator norms $\|K_n\| = \sup\{\|K_nf\|_G: \|f\|_G = 1\}$ are uniformly bounded:

$$\sup_n \|K_n\|<\infty. \qquad (3)$$

By the Banach-Steinhaus theorem this is certainly the case if $K_nf \to f$ in $L_2(G)$ as $n \to \infty$ for every $f \in L_2(G)$. The operator norms $\|K_n\|$ are typically much smaller than the $L_2(G\times G)$-norms of the kernels. The squares of the latter are typically of the same order of magnitude as the square $L_2(G \times G)$-norms weighted by $\mu_2 \times \mu_2$, which we denote by

$$k_n := \int\int K_n^2(x,y)(\mu_2 \times \mu_2)(x,y)d(G \times G)(x,y). \qquad (4)$$

We consider the situation that these square weighted norms are strictly larger than *n*:

$$\frac{k_n}{n} \to \infty. \qquad (5)$$

Under condition (5) the variance of the *U*-statistic (1) is dominated by the variance of the quadratic part of its Hoeffding decomposition. In contrast, if $k_n = n$, the linear and quadratic parts contribute variances of equal order. This case can be handled by the methods of this paper, but requires a special discussion on the joint limits of the linear and quadratic terms, which we omit. The remaining case $k_n \ll n$ leads to asymptotically linear *U*-statistics, and is well understood.

The remaining conditions concern the concentration of the kernels K_n to the diagonal of $\mathscr{X} \times \mathscr{X}$. We assume that there exists a sequence of finite partitions $\mathscr{X} = \cup_m \mathscr{X}_{n,m}$ in measurable sets such that

$$\frac{1}{k_n}\sum_m \int_{\mathscr{X}_{n,m}} \int_{\mathscr{X}_{n,m}} K_n^2(\mu_2 \times \mu_2)d(G \times G) \to 1, \qquad (6)$$

$$\frac{1}{k_n}\max_m \int_{\mathscr{X}_{n,m}} \int_{\mathscr{X}_{n,m}} K_n^2(\mu_2 \times \mu_2)d(G \times G) \to 0, \qquad (7)$$

$$\max_m G(\mathscr{X}_{n,m}) \to 0, \qquad (8)$$

$$\liminf_{n\to\infty} n \min_m G(\mathscr{X}_{n,m})>0. \qquad (9)$$

The sum in the first condition (6) is the integral of the square kernel (weighted by the function $\mu_2 \times \mu_2$) over the set $\cup_m(\mathscr{X}_{n,m} \times \mathscr{X}_{n,m})$ (shown in Figure 1). The condition requires this to be asymptotically equivalent to the integral k_n of this same function over the whole product space $\mathscr{X} \times \mathscr{X}$. The other conditions implicitly require that the partitioning sets are not too different and not too numerous.

A final condition requires implicitly that the partitioning is fine enough. For some $q > 2$, the partitions should satisfy

$$\frac{1}{k_n^{q/2}}\max_m \left(\frac{G(\mathscr{X}_{n,m})}{n}\right)^{q/2-1} \sum_m \int_{\mathscr{X}_{n,m}} \int_{\mathscr{X}_{n,m}} |K_n|^q(\mu_q \times \mu_q)d(G \times G) \to 0. \qquad (10)$$

This condition will typically force the number of partitioning sets to infinity at a rate depending on n and k_n (see Section 4). In the proof it serves as a Lyapounov condition to enforce normality.

The existence of partitions satisfying the preceding conditions depends mostly on the kernels K_n, and is established for various kernels in Section 4. The following theorem is the main result of the paper. Its proof is deferred to Section 5.

Let I_n be the vector with as coordinates $I_{n,1}, \ldots, I_{n,n}$ the indices of the partitioning sets containing $X_1, \ldots, X_n$, i.e. $I_{n,r} = m$ if $X_r \in \mathscr{X}_{n,m}$. Recall that the bounded Lipschitz distance generates the weak topology on probability measures.

Theorem 2.1

Assume that the function μ_2 *is uniformly bounded. If* (2) *and* (5) *hold and there exist finite partitions* $\mathscr{X} = \cup_m \mathscr{X}_{n,m}$ *such that* (6)–(10) *hold, then the bounded Lipschitz distance between the conditional law of* $(U_n - \mathrm{E}\,U_n)/\sigma(U_n)$ *given* I_n *and the standard normal distribution tends to zero in probability. Furthermore* $\operatorname{var} U_n \sim 2k_n/n^2$ *for* k_n *given in* (4).

The conditional convergence in distribution implies the unconditional convergence. It expresses that the randomness in U_n is asymptotically determined by the fine positions of the X_i within the partitioning sets, the numbers of observations falling in the sets being fixed by I_n.

In most of our examples the kernels are pointwise bounded above by a multiple of k_n, and (4) arises, because the area where K_n is significantly different from zero is of the order k_n^{-1}. Condition (10) can then be simplified to

$$\max_m G(\mathscr{X}_{n,m})\frac{k_n}{n} \to 0. \qquad (11)$$

Lemma 2.1

Assume that the functions μ_2 *and* μ_q *are bounded away from zero and infinity, respectively. If* $\|K_n\|_\infty \lesssim k_n$*, then* (10) *is implied by* (11).

Proof—The sum in (10) is bounded up to a constant by $\int |K_n|^q \, d(G \times G)$, which is bounded above by a constant times $k_n^{q-2}\int\int K_n^2 d(G \times G) \lesssim k_n^{q-1}$, by the definition of k_n.

3. Statistical applications

In this section we give examples of statistical problems in which statistics of the type (1) arise as estimators.

3.1. Estimating the integral of the square of a density

Let $X_1, \ldots, X_n$ be i.i.d. random variables with a density g relative to a given measure ν on a measurable space $(\mathscr{X}, \mathscr{A})$. The problem of estimating the functional $\int g^2 \, d\nu$ has been addressed by many authors, including [2], [6] and [19]. The estimators proposed by [4, 5], which are particularly elegant, are based on an expansion of g on an orthonormal basis e_1, e_2, … of the space $L_2(\mathscr{X}, \mathscr{A}, \nu)$, so that $\int g^2 d\nu = \sum_{i=1}^{\infty} \theta_i^2$, for $\theta_i = \int g e_i \, d\nu$ the Fourier coefficients of g. Because $\mathrm{E} e_i(X_1) e_i(X_2) = \theta_i^2$, the square Fourier coefficient θ_i^2 can be estimated unbiasedly by the U-statistic with kernel $(x_1, x_2) \mapsto e_i(x_1) e_i(x_2)$. Hence the truncated sum of squares $\sum_{i=1}^{k} \theta_i^2$ can be estimated unbiasedly by

$$U_n = \sum_{i=1}^{k} \frac{1}{n(n-1)} \sum_{r \neq s} \sum e_i(X_r) e_i(X_s).$$

This statistic is of the type (1) with kernel $K_n(x_1, x_2) = \sum_{i=1}^{k} e_i(x_1) e_i(x_2)$ and the variables $Y_1, \ldots, Y_n$ taken equal to unity.

The estimator U_n is unbiased for the truncated series $\sum_{i=1}^{k} \theta_i^2$, but biased for the functional of interest $\int g^2 d\nu = \sum_{i=1}^{\infty} \theta_i^2$. The variance of the estimator can be computed to be of the order $k/n^2 \vee 1/n$ (cf. (29) below). If the Fourier coefficients are known to satisfy $\sum_{i=1}^{\infty} \theta_i^2 i^{2\beta} \leq 1$, then the bias can be bounded by $\sum_{i=k+1}^{\infty} \theta_i^2 \leq k^{-2\beta}$, and trading square bias versus the variance leads to the choice $k = n^{1/(2\beta+1/2)}$.

In the case that $\beta > 1/4$, the mean square error of the estimator is $1/n$ and the sequence $\sqrt{n}(U_n - \int g^2 d\nu)$ can be shown to be asymptotically linear in the efficient influence function $2(g - \int g^2 \, d\nu)$ (see (28) with $\mu(x) = \mathrm{E}(Y_1 | X_1 = x) \equiv 1$ and [4], [5]). More interesting from our present perspective is the case that $0 < \beta < 1/4$, when the mean square error is of order $n^{-4\beta/(2\beta+1/2)} \gg 1/n$, and the variance of U_n is dominated by its second-order term. By Theorem 2.1 the estimator, centered at its expectation, and with the orthonormal basis (e_i) one of the bases discussed in Section 4, is still asymptotically normally distributed.

The estimator depends on the parameter β through the choice of k. If β is not known, then it would typically estimated from the data. Our present result does not apply to this case, but extension are thinkable.

3.2. Estimating the integral of the square of a regression function

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. random vectors following the regression model $Y_i = b(X_i) + \varepsilon_i$ for unobservable errors ε_i that satisfy $\mathrm{E}(\varepsilon_i | X_i) = 0$. It is desired to estimate $\int b^2 \, dG$ for G the marginal distribution of $X_1, \ldots, X_n$.

If the distribution G is known, then an appropriate estimator can take exactly the form (1), for K_n the kernel of an orthonormal projection on a suitable k_n-dimensional space in $L_2(G)$. Its asymptotics are as in Section 3.1.

Because an orthogonal projection in $L_2(G)$ can only be constructed if G is known, the preceding estimator is not available if G is unknown. If the regression function b is regular of order $\beta \geq 1/4$, then the parameter can be estimated at $\sqrt{n}$-rate (see [1]). In this section we consider an estimator that is appropriate if b is regular of order $\beta < 1/4$ and the design distribution G permits a Lebesgue density g that is bounded away from zero and sufficiently smooth.

Given initial estimators $\hat{b}_n$ and $\hat{g}_n$ for the regression function b and design density g, we consider the estimator

$$T_n=\frac{1}{n}\sum_{r=1}^{n}\left(\hat{b}_n(X_r)^2+2\hat{b}_n(X_r)(Y_r-\hat{b}_n(X_r))\right)+\frac{1}{n(n-1)}\sum\sum_{1\leq r\neq s\leq n}(Y_r-\hat{b}_n(X_r))K_{k_n,\hat{g}_n}(X_r,X_s)(Y_s-\hat{b}_n(X_s)).$$

(12)

Here $(x_1, x_2) \mapsto K_{k,g}(x_1, x_2)$ is a projection kernel in the space $L_2(G)$. For definiteness we construct this in the form (14), where the basis $e_1, \ldots, e_k$ may be the Haar basis, or a general wavelet basis, as discussed in Section 4. Alternatively, we could use projections on the Fourier or spline basis, or convolution kernels, but the latter two require twicing (see (16)) to control bias, and the arguments given below must be adapted.

The initial estimators $\hat{b}_n$ and $\hat{g}_n$ may be fairly arbitrary rate-optimal estimators if constructed from an independent sample of observations. (e.g. after splitting the original sample in parts used to construct the initial estimators and the estimator (12)). We assume this in the following theorem, and also assume that the norm of $\hat{b}_n$ in $C^{\beta}[0, 1]$ is bounded in probability, or alternatively, if the projection is on the Haar basis, that this estimator is in the linear span of $e_1, \ldots, e_{k_n}$. This is typically not a loss of generality.

Let $\hat{E}$ and $\widehat{\text{var}}$ denote expectation and variance given the additional observations. Set $\mu_q(x) = \mathrm{E}(|\varepsilon_1|^q|X_1 = x)$ and let $\|\cdot\|_3$ denote the L_3-norm relative to Lebesgue measure.

Corollary 3.1—*Let $\hat{b}_n$ and $\hat{g}_n$ be estimators based on independent observations that converge to b and g in probability relative to the uniform norm and satisfy* $\|\hat{b}_n - b\|_3 = O_P(n^{-\beta/(2\beta+1)})$ *and* $\|\hat{g}_n - g\|_3 = O_P(n^{-\gamma/(2\gamma+1)})$. *Let* μ_q *be finite and uniformly bounded for some* $q > 2$. *Then for* $b \in C^{\beta}[0, 1]$ *and strictly positive* $g \in C^{\gamma}[0, 1]$, *with* $\gamma \geq \beta$, *and for* k_n *satisfying* (5),

$$\left|\hat{\mathrm{E}}_{b,g}T_n-\int b^2dG\right|=O_P\left(\frac{1}{k_n}\right)^{2\beta}+O_P\left(\frac{1}{n}\right)^{2\beta/(2\beta+1)+\gamma/(2\gamma+1)},$$

$$\widehat{\mathrm{var}}_{b,g}T_n=\frac{2}{n^2}\int\int(\mu_2\times\mu_2)K^2_{k_n,g}d(G\times G)(1+o_P(1))=O_P\left(\frac{k_n}{n^2}\right).$$

Furthermore, the sequence $(T_n-\hat{\mathrm{E}}_{b,g}T_n)/\widehat{\mathrm{sd}}_{b,g}(T_n)$ *tends in distribution to the standard normal distribution.*

For $k_n = n^{1/(2\beta+1/2)}$ the estimator T_n of $\int b^2\, dG$ attains a rate of convergence of the order $n^{-2\beta/(2\beta+1/2)} + n^{-2\beta/(2\beta+1)-\gamma/(2\gamma+1)}$. If $\gamma > \beta/(4\beta^2 + \beta + 1/2)$, then this reduces to $n^{-4\beta/(1+4\beta)}$, which is known to be the minimax rate when g is known and b ranges over a ball in $C^\beta[0, 1]$, for $\beta \le 1/4$ (see [3] or [20]). For smaller values of γ the estimator can be improved by considering third or higher order U-statistics (see [9]).

3.3. Estimating the mean response with missing data

Suppose that a typical observation is distributed as $X = (YA, A, Z)$ for Y and A taking values in the two-point set $\{0, 1\}$ and conditionally independent given Z, with conditional mean functions $b(z) = P(Y = 1|Z = z)$ and $a(z)^{-1} = P(A = 1|Z = z)$, and Z possessing density g relative to some dominated measure ν.

In [7] we introduced a quadratic estimator for the *mean response* $EY = \int bg\, d\nu$, which attains a better rate of convergence than the conventional linear estimators. For initial estimators $\hat{a}_n$, $\hat{b}_n$ and $\hat{g}_n$, and $K_{k,\hat{a}_n,\hat{g}_n}$ a projection kernel in $L_2(g/a)$, this takes the form

$$\frac{1}{n}\sum_{r=1}^{n}\Big(A_r\hat{a}_n(Z_r)(Y_r-\hat{b}_n(Z_r))+\hat{b}_n(Z_r)\Big)-\frac{1}{n(n-1)}\sum\sum_{1\le r\ne s\le n}\Big(A_r(Y_r-\hat{b}_n(Z_r))K_{k_n,\hat{\alpha}_n,\hat{g}_n}(Z_r,Z_s)(A_s\hat{a}_n(Z_s)-1)\Big).$$

Apart from the (inessential) asymmetry of the kernel, the quadratic part has the form (1). Just as in the preceding section, the estimator can be shown to be asymptotically normal with the help of Theorem 2.1.

4. Kernels

In this section we discuss examples of kernels that satisfy the conditions of our main result. Detailed proofs are given in an appendix.

Most of the examples are kernels of *projections* K, which are characterised by the identity $Kf = f$, for every f in their range space. For a projection given by a kernel, the latter is equivalent to $f(x) = \int f(\upsilon)K(x, \upsilon)\, dG(\upsilon)$ for (almost) every x, which suggests that the measure $\upsilon \mapsto K(x, \upsilon)\, dG(\upsilon)$ acts on f as a Dirac kernel located at x. Intuitively, if the projection spaces increase to the full space, so that the identity is true for more and more f, then the kernels $(x, \upsilon) \mapsto K(x, \upsilon)$ must be increasingly dominated by their values near the diagonal, thus meeting the main condition of Theorem 2.1.

For a given orthonormal basis $e_1, e_2, \ldots$ of $L_2(G)$, the orthogonal projection onto lin $(e_1, \ldots, e_k)$ is the kernel operator $K_k: L_2(G) \to L_2(G)$ with kernel

$$K_k(x_1,x_2)=\sum_{i=1}^{k}e_i(x_1)e_i(x_2). \qquad (13)$$

It can be checked that it has operator norm 1, while the square L_2-norm $\int\int K_k^2 d(G\times G)=k$ of the kernel is k.

A given orthonormal basis e_1, e_2, … relative to a given dominating measure, can be turned into an orthonormal basis $e_1/\sqrt{g}, e_2/\sqrt{g}, \ldots$ of $L_2(G)$, for g a density of G. The kernel of the orthogonal projection in $L_2(G)$ onto lin $(e_1/\sqrt{g}, \ldots, e_k/\sqrt{g})$ is

$$K_{k,g}(x_1, x_2) = \frac{\sum_{i=1}^{k} e_i(x_1) e_i(x_2)}{\sqrt{g(x_1)}\sqrt{g(x_2)}}. \qquad (14)$$

If g is bounded away from zero and infinity, the conditions of Theorem 2.1 will hold for this kernel as soon as they hold for the kernel (13) relative to the dominating measure.

The orthogonal projection in $L_2(G)$ onto the linear span lin $(f_1, \ldots, f_k)$ of an arbitrary set of functions f_i possesses the kernel

$$K_k(x_1, x_2) = \sum_{i=1}^{k}\sum_{j=1}^{k} A_{i,j} f_i(x_1) f_j(x_2), \qquad (15)$$

for A the inverse of the $(k \times k)$-matrix with (i, j)-element $\langle f_i, f_j \rangle_G$. In statistical applications this projection has the advantage that it projects onto a space that does not depend on the (unknown) measure G. For the verification of the conditions of Theorem 2.1 it is useful to note that the matrix A is well-behaved if $f_1, \ldots, f_k$ are orthonormal relative to a measure G_0 that is not too different from G: from the identity $\alpha^T (\langle f_i, f_j \rangle_G) \alpha = \int (\sum_{i=1}^{k} \alpha_i f_i)^2 dG$, one can verify that the eigenvalues of A are bounded away from zero and infinity if G and G_0 are absolutely continuous with a density that is bounded away from zero and infinity.

Orthogonal projections K have the important property of making the inner product $\langle (I-K)f, f \rangle_G = \|(I-K)f\|_G^2$ quadratic in the approximation error. Nonorthogonal projections, such as the convolution kernels or spline kernels discussed below, lack this property, and may result in a large bias of an estimator. *Twicing kernels*, discussed in [21] as a means to control the bias of plug-in estimators, remedy this problem. The idea is to use the operator $K + K^* - KK^*$, where K^* is the adjoint of K: $L_2(G) \to L_2(G)$, instead of the original operator K. Because $I - K - K^* + KK^* = (I - K)(I - K^*)$, it follows that

$$\langle (I - K - K^* + KK^*) f, f \rangle_G = \langle (I-K)f, (I-K)f \rangle_G = \|(I-K)f\|_G^2.$$

If K is an orthogonal projection, then $K = K^*$ and the twicing kernel is $K + K^* - KK^* = K$, and nothing changes, but in general using a twicing kernel can cut a bias significantly.

If K is a kernel operator with kernel $(x_1, x_2) \mapsto K(x_1, x_2)$, then the adjoint operator is a kernel operator with kernel $(x_1, x_2) \mapsto K(x_2, x_1)$, and the twicing operator $K + K^* - KK^*$ is a kernel operator with kernel (which depends on G)

$$(x_1,x_2) \mapsto K(x_1,x_2)+K(x_2,x_1) - \int K(x_1,z)K(x_2,z)dG(z). \qquad (16)$$

4.1. Wavelets

Consider expansions of functions $f \in L_2(\mathbb{R}^d)$ on an orthonormal basis of compactly supported, bounded wavelets of the form

$$f(x)=\sum_{j\in\mathbb{Z}^d}\sum_{\upsilon\in\{0,1\}^d}\langle f,\psi_{0,j}^{\upsilon}\rangle\psi_{0,j}^{\upsilon}(x)+\sum_{i=0}^{\infty}\sum_{j\in\mathbb{Z}^d}\sum_{\upsilon\in\{0,1\}^d-\{0\}}\langle f,\psi_{i,j}^{\upsilon}\rangle\psi_{i,j}^{\upsilon}(x), \qquad (17)$$

where the base functions $\psi_{i,j}^{\upsilon}$ are orthogonal for different indices (i, j, υ) and are scaled and translated versions of the 2^d base functions $\psi_{0,0}^{\upsilon}$:

$$\psi_{i,j}^{\upsilon}(x)=2^{id/2}\psi_{0,0}^{\upsilon}(2^i x - j).$$

Such a higher-dimensional wavelet basis can be obtained as tensor products $\psi_{0,0}^{\upsilon}=\phi^{\upsilon_1}\times\cdots\times\phi^{\upsilon_d}$ of a given father wavelet ϕ^0 and and mother wavelet ϕ^1 in one dimension. See for instance Chapter 8 of [22].

We shall be interested in functions f with support $\mathscr{X} = [0, 1]^d$. In view of the compact support of the wavelets, for each resolution level i and vector υ only to the order 2^{id} base elements $\psi_{i,j}^{\upsilon}$ are nonzero on $\mathscr{X}$; denote the corresponding set of indices j by J_i. Truncating the expansion at the level of resolution $i = I$ then gives an orthogonal projection on a subspace of dimension k of the order 2^{Id}. The corresponding kernel is

$$K_k(x_1,x_2)=\sum_{j\in J_0}\sum_{\upsilon\in\{0,1\}^d}\psi_{0,j}^{\upsilon}(x_1)\psi_{0,j}^{\upsilon}(x_2)+\sum_{i=0}^{I}\sum_{j\in J_i}\sum_{\upsilon\in\{0,1\}^d-\{0\}}\psi_{i,j}^{\upsilon}(x_1)\psi_{i,j}^{\upsilon}(x_1). \qquad (18)$$

Proposition 4.1—*For the wavelet kernel* (18) *with* $k = k_n = 2^{Id}$ *satisfying* $k_n/n \to \infty$ *and* $k_n/n^2 \to 0$ *conditions* (2), (6), (7), (8), (9) *and* (10) *are satisfied for any measure* G *on* $[0, 1]^d$ *with a Lebesgue density that is bounded and bounded away from zero and regression functions* μ_2 *and* μ_q *(for some* $q > 2$*) that are bounded and bounded away from zero.*

4.2. Fourier basis

Any function $f \in L_2[-\pi, \pi]$ can be represented through the Fourier series $f = \sum_{j\in\mathbb{Z}} f_j e_j$, for the functions $e_j(x)=e^{ijx}/\sqrt{2\pi}$ and the Fourier coefficients $f_j=\int_{-\pi}^{\pi} f e_j d\lambda$. The truncated series $f_k = \sum_{|j|\leq k} f_j e_j$ gives the orthogonal projection of f onto the linear span of the function $\{e_j : |j| \leq k\}$, and can be written as $K_k f$ for K_k the kernel operator with kernel (known as the Dirichlet kernel)

$$K_k(x_1,x_2)=\sum_{|j|\le k} e_j(x_1)e_j(x_2)=\frac{\sin((k+\frac{1}{2})(x_1-x_2))}{2\pi\ \sin(\frac{1}{2}(x_1-x_2))}. \qquad (19)$$

Proposition 4.2—*For the Fourier kernel* (19) *with* $k = k_n$ *satisfying* $n \ll k_n \ll n^2$ *conditions* (2), (6)–(10) *are satisfied for any measure G on* $\mathbb{R}$ *with a bounded Lebesgue density and regression functions* μ_2 *and* μ_q *(for some* $q > 2$*) that are bounded and bounded away from zero.*

4.3. Convolution

For a uniformly bounded function $\phi: \mathbb{R} \to \mathbb{R}$ with $\int |\phi|\, d\lambda < \infty$, and a positive number σ, set

$$K_\sigma(x_1,x_2)=\frac{1}{\sigma}\phi\left(\frac{x_1-x_2}{\sigma}\right) := \phi_\sigma(x_1-x_2). \qquad (20)$$

For $\sigma \downarrow 0$ these kernels tend to the diagonal, with square norm of the order σ^{-1}.

Proposition 4.3—*For the convolution kernel* (20) *with* $\sigma = \sigma_n$ *satisfying* $n^{-2} \ll \sigma_n \ll n^{-1}$ *conditions* (2), (6)–(10) *are satisfied for any measure G on* [0, 1] *with a Lebesgue density that is bounded and bounded away from zero and regression functions* μ_2 *and* μ_q *(for some* $q > 2$*) that are bounded and bounded away from zero.*

4.4. Splines

The *Schoenberg space* $S_r(T, d)$ of order r for a given knot sequence T: $t_0 = 0 < t_1 < t_2 < \cdots < t_l < 1 = t_{l+1}$ and vector of *defects* $d = (d_1, \ldots, d_l) \in \{0, \ldots, r - 1\}$ are the functions f: $[0, 1] \to \mathbb{R}$ whose restriction to each subinterval (t_i, t_{i+1}) is a polynomial of degree $r - 1$ and which are $r - 1 - d_i$ times continuously differentiable in a neighbourhood of each t_i. (Here “0 times continuously differentiable” means “continuous” and “−1 times continuously differentiable” means no restriction.) The Schoenberg space is a $k = r + \sum_i d_i$-dimensional vector space. Each “augmented knot sequence”

$$-t_{r+1} \le \cdots \le t_0=0<t_1<t_2<\cdots<t_l<1=t_{l+1} \le \cdots \le t_{l+r} \qquad (21)$$

defines a basis $N_1, \ldots, N_k$ of *B-splines*. These are nonnegative splines with $\sum_j N_j = 1$ such that N_j vanishes outside the interval (t'_j, t'_{j+r}). Here the “basic knots” (t'_j) are defined as the knot sequence (t_j), but with each $t_i \in (0, 1)$ repeated d_i times. See [23], pages 137, 140 and 145). We assume that $|t_{i-1} - t_i| \le |t_{-1} - t_0|$ if $i < 0$ and $|t_{i+1} - t_i| \le |t_{l+1} - t_l|$ if $i > l$.

The *quasi-interpolant operator* is a projection K_k: $L_1[0, 1] \to S_r(T, d)$ with the properties

$$\|f - K_k f\|_p \le C_r \|f - S_r(T,d)\|_p,$$

$$\|K_k f\|_p \leq C_r \|f\|_p.$$

for every $1 \leq p \leq \infty$ and a constant C_r depending on r only (see [23], pages 144–147). It follows that the projection K_k inherits the good approximation properties of spline functions, relative to any L_p-norm. In particular, it gives good approximation to smooth functions.

The quasi-interpolant operator K_k is a projection onto $S_l(T, d)$ (i.e. $K_k^2 = K_k$ and $K_k f = f$ for $f \in S_l(T, d)$), but not an orthogonal projection. Because the B-splines form a basis for $S_l(T, d)$, the operator can be written in the form $K_k f = \sum_j c_j(f) N_j$ for certain linear functionals $c_j : L_1[0, 1] \to \mathbb{R}$. It can be shown that, for any $1 \leq p \leq \infty$,

$$|c_j(f)| \leq C_r \frac{1}{(t'_{j+r} - t'_j)^{1/p}} \|f 1_{[t'_j, t'_{j+r}]}\|_p. \quad (22)$$

([23], page 145.) In particular, the functionals c_j belong to the dual space of $L_1[0, 1]$ and can be written as $c_j(f) = \int f c_j \, d\lambda$ for (with abuse of notation) certain functions $c_j \in L_\infty[0, 1]$. This yields the representation of K_k as a kernel operator with kernel

$$K_k(x_1, x_2) = \sum_{j=1}^{k} N_j(x_1) c_j(x_2). \quad (23)$$

Proposition 4.4—*Consider a sequence (indexed by l) of augmented knot sequences* (21) *with* $l^{-1} \lesssim t^l_{i+1} - t^l_i \lesssim l^{-1}$ *for every* $0 \leq i \leq l$ *and splines with fixed defects* $d_i = d$. *For the corresponding (symmetrized) spline kernel* (23) *with* $l = l_n$ *conditions* (2), (6), (7), (8), (9) *and* (10) *are satisfied if* $l_n/n \to \infty$ *and* $l_n/n^2 \to 0$ *for any measure G on* [0, 1] *with a Lebesgue density that is bounded and bounded away from zero and regression functions* μ_2 *and* μ_q *(for some* $q > 2$) that are bounded and bounded away from zero.

5. Proof of Theorem 2.1

For M_n the cardinality of the partition $\mathscr{X} = \cup_m \mathscr{X}_{n,m}$, let $N_{n,1}, \ldots, N_{n,M_n}$ be the numbers of X_r falling in the partitioning sets, i.e.

$$I_{n,r} = m \quad \text{if } X_r \in \mathscr{X}_{n,m},$$

$$N_{n,m} = \#(1 \leq r \leq n : I_{n,r} = m).$$

The vector $N_n = (N_{n,1}, \ldots, N_{n,M_n})$ is multinomially distributed with parameters n and vector of success probabilities $p_n = (p_{n,1}, \ldots, p_{n,M_n})$ given by

$$p_{n,m}=G(\mathscr{X}_{n,m}).$$

Given the vector $I_n = (I_{n,1}, \ldots, I_{n,n})$ the vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent with distributions determined by

$$X_r \text{ has distribution } G_{n,I_{n,r}} \text{ given by } dG_{n,I_{n,r}}=1_{\mathscr{X}_{n,I_{n,r}}} dG/p_{n,I_{n,r}} \tag{24}$$

$$Y_r \text{ has the same conditional distribution given } X_r \text{ as before.} \tag{25}$$

We define U-statistics V_n by restricting the kernel K_n to the set $\cup_m \mathscr{X}_{n,m} \times \mathscr{X}_{n,m}$, as follows:

$$V_n=\frac{1}{n(n-1)}\sum\sum_{1\le r\ne s\le n} K_n(x_r, X_s)Y_rY_s1_{(X_r,X_s)\in\cup_m\mathscr{X}_{n,m}\times\mathscr{X}_{n,m}}. \tag{26}$$

The proof of Theorem 2.1 consists of three elements. We show that the difference between U_n and V_n is asymptotically negligible due to the fact that the kernels shrink to the diagonal, we show that the statistics V_n are conditionally asymptotically normal given the vector of bin indicators I_n, and we show that the conditional and unconditional means and variances of V_n are asymptotically equivalent. These three elements are expressed in the following four lemmas, which should be understood all implicitly to assume the conditions of Theorem 2.1.

Lemma 5.1

$\mathrm{var}(U_n - V_n)/\ \mathrm{var}\ U_n \to 0.$

Lemma 5.2

$$\sup_x|\mathrm{P}((V_n-\mathrm{E}(V_n|I_n))/\mathrm{sd}(V_n|I_n)\le x|I_n)-\Phi(x)| \xrightarrow{P} 0.$$

Lemma 5.3

$$(\mathrm{E}V_n-\mathrm{E}(V_n|I_n))/\mathrm{sd}V_n \xrightarrow{P} 0.$$

Lemma 5.4

$$\mathrm{var}(V_n|I_n)/\ \mathrm{var}\ V_n \xrightarrow{P} 1.$$

5.1. Proof of Theorem 2.1

By Lemmas 5.1 and 5.3 the sequence $((U_n - \mathrm{E}U_n)-(V_n - \mathrm{E}(V_n| I_n)) / \mathrm{sd}\ V_n$ tends to zero in probability. Because conditional and unconditional convergence in probability to a constant is the same, we see that it suffices to show that $(V_n - \mathrm{E}(V_n| I_n))/ \mathrm{sd}\ V_n$ converges conditionally given I_n to the normal distribution, in probability. This follows from Lemmas 5.4 and 5.2.

The variance of U_n is computed in (29) in Section 5.2. By the Cauchy-Schwarz inequality (cf. (2)),

$$\langle K_n\mu, \mu\rangle_G^2 \le \|K_n\mu\|_G^2 \|\mu\|_G^2 \le \|K_n\|^2 \|\mu\|_G^4,$$

$$\|(K_n\mu)\sqrt{\mu_2}\|_G^2 \le \|\mu_2\|_\infty \|K_n\mu\|_G^2 \le \|\mu_2\|_\infty \|K_n\|^2 \|\mu\|_G^2.$$

Because μ_2 is bounded by assumption and the norms $\|K_n\|$ are bounded in n by assumption (2), the right sides are bounded in n. In view of (5) it follows that the first two terms in the final expression for the variance are of lower order than the third, whence

$$\text{var } U_n \sim \frac{2k_n}{n^2}. \quad (27)$$

5.2. Moments of U-statistics

To compute or estimate moments of U_n we employ the Hoeffding decomposition (e.g. [24], Sections 11.4 and 12.1) $U_n = \text{E}U_n + U_n^{(1)} + U_n^{(2)}$ of U_n given by

$$U_n^{(1)} = \frac{2}{n}\sum_{r=1}^{n}(K_n\mu(X_r)Y_r - \text{E}U_n), \quad (28)$$

$$U_n^{(2)} = \frac{1}{n(n-1)}\sum\sum_{1\le r\ne s\le n}\left[K_n(X_r, X_s)Y_rY_s - K_n\mu(X_r)Y_r - K_n\mu(X_s)Y_s + \text{E}U_n\right].$$

The variables $U_n^{(1)}$ and $U_n^{(2)}$ are uncorrelated, and so are all the variables in the single and double sums defining $U_n^{(1)}$ and $U_n^{(2)}$. It follows that

$$\begin{aligned}\text{var } U_n &= \frac{4}{n}\text{var}(K_n\mu(X_1)Y_1)\\ &+ \frac{2}{n(n-1)}\text{var}(K_n(X_1, X_2)Y_1Y_2 - K_n\mu(X_1)Y_1 - K_n\mu(X_2)Y_2)\\ &= \left[\frac{4}{n} - \frac{4}{n(n-1)}\right] \text{var}(K_n\mu(X_1)Y_1) + \frac{2}{n(n-1)}\text{var}(K_n(X_1, X_2)Y_1Y_2)\\ &= \frac{4(n-2)}{n(n-1)}\|(K_n\mu)\sqrt{\mu_2}\|_G^2 - \frac{4(n-2)+2}{n(n-1)}\langle K_n\mu, \mu\rangle_G^2 + \frac{2k_n}{n(n-1)}\end{aligned} \quad (29)$$

See equation (4) for the definition of k_n.

There is no similarly simple expression for higher moments of a *U*-statistic, but the following useful bound is (essentially) established in [25].

Lemma 5.5—(Giné, Latala, Zinn). *For any* $q \geq 2$ *there exists a constant* C_q *such that for any i.i.d. random variables* $X_1, \dots, X_n$ *and degenerate symmetric kernel K,*

$$\mathrm{E}\left|\frac{1}{n(n-1)}\sum\sum_{1\leq r\neq s\leq n} K(X_r,X_s)\right|^q \leq C_q n^{-q}\left(\mathrm{E}K^2(X_1,X_2)\right)^{q/2} \vee n^{-3q/2+1}\mathrm{E}|K(X_1,X_2)|^q \leq C_q n^{-q}\mathrm{E}|K(X_1,X_2)|^q.$$

Proof: The second inequality is immediate from the fact that the L_2-norm is bounded above by the L_q-norm, and $3q/2 - 1 \geq q$, for $q \geq 2$. For the first inequality we use (3.3) in [25] (and decoupling as explained in Section 2.5 of that paper) to see that the left side of the lemma is bounded above by a multiple of

$$n^{-q}\left(\mathrm{E}K^2(X_1,X_2)\right)^{q/2} \vee n^{-3q/2+1}\mathrm{E}(\mathrm{E}(K^2(X_1,X_2)|X_2))^{q/2} \vee n^{2-2q}\mathrm{E}|K(X_1,X_2)|^q.$$

Because L_q-norms are increasing in q, the second term on the right is bounded above by $n^{-3q/2+1}\mathrm{E}|K(X_1, X_2)|^q$, which is also a bound on the third term, as $n^{2-2q} \leq n^{-3q/2+1}$ for $q \geq 2$.

We can apply the preceding inequality to the degenerate part of the Hoeffding decomposition (28) of U_n and combine it with the Marcinkiewicz-Zygmund inequality to obtain a bound on the moments of U_n.

Corollary 5.1—*For any* $q \geq 2$ *there exists a constant* C_q *such that for the U-statistic given by* (1) *and* (28),

$$\mathrm{E}|U_n^{(1)}|^q \leq C_q n^{-q/2}\int |K_n\mu|^q \mu_q dG,$$

$$\mathrm{E}|U_n^{(2)}|^q \leq C_q n^{-q}\left(\int\int K_n^2\mu_2 \times \mu_2 dG \times G\right)^{q/2} \vee C_q n^{-3q/2+1}\int\int |K_n|^q \mu_q \times \mu_q dG \times G.$$

Proof: The first inequality follows from the Marcinkiewicz-Zygmund inequality and the fact that $\mathrm{E}|Z - \mathrm{E}Z|^q \leq 2^q\mathrm{E}|Z|^q$, for any random variable Z. To obtain the second we apply Lemma 5.5 to $U_n^{(2)}$, which is a degenerate *U*-statistic with kernel $K_n(X_1, X_2)Y_1Y_2 - \Pi_n(X_1, X_2, Y_1, Y_2)$, for Π_n the sum of the conditional expectations of $K_n(X_1, X_2)Y_1Y_2$ relative to (X_1, Y_1) and (X_2, Y_2) minus $\mathrm{E}U_n$. Because (conditional) expectation is a contraction for the L_q-norm ($\mathrm{E}|\mathrm{E}(Z|\mathscr{A})|^q \leq \mathrm{E}|Z|^q$ for any random variable Z and conditioning σ-field $\mathscr{A}$), we can bound the L_2- and L_q-norms of the degenerate kernel, appearing in the bound obtained from Lemma 5.5, by a constant (depending on q) times the L_2- of L_q-norm of the kernel $K_n(X_1, X_2)Y_1Y_2$.

5.3. Proof of Lemma 5.1

The statistic $U_n - V_n$ is a U-statistic of the same type as U_n, except that the kernel K_n is replaced by $K_n(1 - 1_{\mathscr{X}_n})$ for $\mathscr{X}_n = \cup_m (\mathscr{X}_{n,m} \times \mathscr{X}_{n,m})$. The variance of $U_n - V_n$ is given by formula (29), but with K_n replaced by the kernel operator with kernel $K_{n,n} = K_n(1 - 1_{\mathscr{X}_n})$. The corresponding kernel operator is $K_{n,n} f = K_n f - \sum_m K_n(f1_{\mathscr{X}_{n,m}})1_{\mathscr{X}_{n,m}}$, and hence

$$\begin{aligned}\frac{1}{2}\|K_{n,n}f\|_G^2 &\leq \|K_n f\|_G^2 + \left\|\sum_m K_n(f1_{\mathscr{X}_{n,m}})1_{\mathscr{X}_{n,m}}\right\|_G^2\\ &\leq \|K_n f\|_G^2 + \sum_m \|K_n(f1_{\mathscr{X}_{n,m}})\|_G^2\\ &\leq \|K_n\|^2\|f\|_G^2 + \sum_m \|K_n\|^2\|f1_{\mathscr{X}_{n,m}}\|_G^2 \leq 2\|K_n\|^2\|f\|_G^2.\end{aligned}$$

It follows that the operator norms $\|K_{n,n}\|_2$ of the operators $K_{n,n}$ are uniformly bounded in n (cf. equation (3) for the operators K_n). Applying decomposition (29) to the kernel $K_{n,n}$ we see that var($U_n - V_n$) = $O(n^{-1}) + 2k_{n,n}/n^2$, where $k_{n,n}$ is the $L_2(G \times G)$-norm $k_{n,n}$ of the kernel $K_{n,n}$ weighted by $\mu_2 \times \mu_2$, as in (4) but with K_n replaced by $K_{n,n}$. By assumption (6) the norm $k_{n,n}$ is negligible relative to the same norm (denoted k_n) of the original kernel. Because the variance of U_n is asymptotically equivalent to $2k_n/n^2$ and $k_n/n \to \infty$, this proves the claim.

5.4. Proof of Lemma 5.2

The variable V_n can be written as the sum $V_n = \sum_m V_{n,m}$, for

$$V_{n,m} = \frac{1}{n(n-1)} \sum\sum_{1 \leq r \neq s \leq n} K_n(X_r, X_s) Y_r Y_s 1_{(X_r, X_s) \in \mathscr{X}_{n,m} \times \mathscr{X}_{n,m}}. \tag{30}$$

Given the vector of bin-indicators I_n the observations (X_r, Y_r) are independently generated from the conditional distributions in which X_r is conditioned to fall in bin $\mathscr{X}_{n,I_{n,r}}$, as given in (24)–(25). Because each variable $V_{n,m}$ depends only on the observations (X_r, Y_r) for which X_r falls in bin $\mathscr{X}_{n,m}$, the variables $V_{n,1}, \ldots, V_{n,M_n}$ are conditionally independent. The conditional asymptotic normality of V_n given I_n can therefore be established by a central limit theorem for independent variables.

The variable $V_{n,m}$ is equal to $N_{n,m}(N_{n,m} - 1)/(n(n-1))$ times a U-statistic of the type (1), based on $N_{n,m}$ observations (X_r, Y_r) from the conditional distribution where X_r is conditioned to fall in $\mathscr{X}_{n,m}$. The corresponding kernel operator is given by

$$K_{n,m} f(x) = \int K_n(x, \upsilon) f(\upsilon) 1_{\mathscr{X}_{n,m} \times \mathscr{X}_{n,m}}(x, \upsilon) \frac{dG(\upsilon)}{p_{n,m}} = \frac{K(f1_{\mathscr{X}_{n,m}})(x) 1_{\mathscr{X}_{n,m}}(x)}{p_{n,m}}. \tag{31}$$

We can decompose each $V_{n,m}$ into its Hoeffding decomposition $V_{n,m}=\mathrm{E}(V_{n,m}|I_n)+V_{n,m}^{(1)}+V_{n,m}^{(2)}$ relative to the conditional distribution given I_n. We shall show that

$$\mathrm{E}\left(\frac{|\sum_m V_{n,m}^{(1)}|}{\mathrm{sd}(V_n|I_n)}|I_n\right)\xrightarrow{P}0. \qquad (32)$$

To prove Lemma 5.2 it then suffices to show that the sequence $\sum_m V_{n,m}^{(2)}/\mathrm{sd}(V_n|I_n)$ converges conditionally given I_n weakly to the standard normal distribution, in probability. By Lyapounov's theorem, this follows from, for some $q > 2$,

$$\frac{\sum_m \mathrm{E}(|V_{n,m}^{(2)}|^q|I_n)}{\mathrm{sd}(V_n|I_n)^q}\xrightarrow{P}0. \qquad (33)$$

By Lemma 5.4 the conditional standard deviation sd(V_n| I_n) is asymptotically equivalent in probability to the unconditional standard deviation, and by Lemma 5.1 this is equivalent to sd U_n, which is equivalent to $\sqrt{2k_n/n^2}$. Thus in both (32) and (33) the conditional standard deviation in the denominator may be replaced by $\sqrt{2k_n/n^2}$.

In view of the first assertion of Corollary 5.1,

$$\mathrm{var}(V_{n,m}^{(1)}|I_n)\le C_2\left(\frac{N_{n,m}(N_{n,m}-1)}{n(n-1)}\right)^2 N_{n,m}^{-1}\int\left|\frac{K_n(\mu 1_{\chi_{n,m}})}{p_{n,m}}\right|^2\mu_2 1_{\chi_{n,m}}\frac{dG}{p_{n,m}}.$$

By Lemma 5.6 (below, note that $(np_{n,m})^2 \lesssim (np_{n,m})^3$ in view of (9)) the expectation of the right side is bounded above by a constant times

$$\frac{(np_{n,m})^3}{n^2(n-1)^2p_{n,m}^3}\|\mu_2\|_\infty\|K_n(\mu 1_{\chi_{n,m}})\|_G^2\le\frac{1}{n}\|\mu_2\|_\infty\|K_n\|^2\|\mu 1_{\chi_{n,m}}\|_G^2.$$

In view of (2) the sum over m of this expression is bounded above by a multiple of $1/n$, which is $o(k_n/n^2)$ by assumption (5). Because $\mathrm{E}(V_{n,m}^{(1)}|I_n)=0$, this concludes the proof of (32).

In view of the second assertion of Corollary 5.1,

$$\mathrm{E}|V_{n,m}^{(2)}|^q \le C_q\left(\frac{N_{n,m}(N_{n,m}-1)}{n(n-1)}\right)^q \times \\ \left[N_{n,m}^{-q}\left(\int\int K_n^2\mu_2\times\mu_2 1_{\chi_{n,m}\times\chi_{n,m}}\frac{dG\times G}{p_{n,m}^2}\right)^{q/2} \\ \vee N_{n,m}^{-3q/2+1}\int\int |K_n|^q\mu_q\times\mu_q 1_{\chi_{n,m}\times\chi_{n,m}}\frac{dG\times G}{p_{n,m}^2}\right].$$

By Lemma 5.6 the expectation of the right side is bounded above by a constant times

$$\frac{(np_{n,m})^q}{n^q(n-1)^q p_{n,m}^q}\left(\int\int K_n^2\mu_2\times\mu_2 1_{\chi_{n,m}\times\chi_{n,m}}dG\times G\right)^{q/2}+\frac{(np_{n,m})^{q/2+1}}{n^q(n-1)^q p_{n,m}^2}\int\int |K_n|^q\mu_q\times\mu_q 1_{\chi_{n,m}\times\chi_{n,m}}dG\times G.$$

With $\alpha_{n,m}(q) = \int\int |K_n|^q \mu_q \times \mu_q 1_{\mathcal{X}_{n,m}\times\mathcal{X}_{n,m}} dG \times G$ it follows that

$$\sum_m \frac{\mathrm{E}|V_{n,m}^{(2)}|^q}{(k_n/n^2)^{q/2}} \lesssim \sum_m\left(\frac{\alpha_{n,m}(2)}{k_n}\right)^{q/2}+\sum_m\left(\frac{p_{n,m}}{n}\right)^{q/2-1}\frac{\alpha_{n,m}(q)}{k_n^{q/2}} \\ \le \max_m\left(\frac{\alpha_{n,m}(2)}{k_n}\right)^{q/2-1}\sum_m\frac{\alpha_{n,m}(2)}{k_n}+\max_m\left(\frac{p_{n,m}}{n}\right)^{q/2-1}\sum_m\frac{\alpha_{n,m}(q)}{k_n^{q/2}}.$$

The right side tends to zero by assumptions (6), (7) and (10). This concludes the proof of (33).

5.5. Proof of Lemma 5.3

Only pairs (X_r, X_s) that fall in one of the sets $\mathcal{X}_{n,m} \times \mathcal{X}_{n,m}$ contribute to the double sum (26) that defines V_n. Given I_n there are $N_{n,m}(N_{n,m} - 1)$ pairs that fall in $\mathcal{X}_{n,m}$ and the distribution of the corresponding vectors (X_r, Y_r), (X_s, Y_s) is determined as in (24)–(25). From this it follows that

$$\mathrm{E}(V_n|I_n)=\frac{1}{n(n-1)}\sum_m N_{n,m}(N_{n,m}-1)\int\int K_n\mu\times\mu 1_{\chi_{n,m}\times\chi_{n,m}}\frac{dG\times G}{p_{n,m}^2}.$$

Defining the numbers $\alpha_{n,m} = \int\int K_n \mu \times \mu 1_{\mathcal{X}_{n,m}\times\mathcal{X}_{n,m}} dG \times G$, we infer that

$$\mathrm{E}(V_n|I_n)-\mathrm{E}V_n=\sum_m\left(\frac{N_{n,m}(N_{n,m}-1)}{n(n-1)p_{n,m}^2}-1\right)\alpha_{n,m}.$$

By the Cauchy-Schwarz inequality, the numbers $\alpha_{n,m}$ satisfy

$$|\alpha_{n,m}| \le \|K_n(\mu 1_{\chi_{n,m}})\|_G\|\mu 1_{\chi_{n,m}}\|_G \le \|K_n\|\|\mu 1_{\chi_{n,m}}\|_G^2 \lesssim \|K_n\|\|\mu\|_\infty^2 p_{n,m}.$$

In particular $\sum_m |\alpha_{n,m}| \lesssim 1$. In view of (2) the numbers s_n^2 given in (34) (below) are of the order $M_n/n^2 + 1/n$. Lemma 5.7 (below) therefore implies that the right side of the second last display is of the order $O_P(\sqrt{M_n}/n+1/\sqrt{n})=O(1/\sqrt{n})$, because (9) implies that $M_n \lesssim n$. By assumption (5) this is smaller than $\sqrt{k_n/n^2}$, which is of the same order as sd V_n.

5.6. Proof of Lemma 5.4

By (29) applied to the variables $V_{n,m}$ defined in (30),

$$\begin{aligned}\mathrm{var}(V_n|I_n)=&\sum_m \mathrm{var}(V_{n,m}|I_n)\\ =&\sum_m\left(\frac{N_{n,m}(N_{n,m}-1)}{n(n-1)}\right)^2\left[\frac{4(N_{n,m}-2)}{N_{n,m}(N_{n,m}-1)}\|(K_{n,m}\mu)\sqrt{\mu_2}\|^2_{G_{n,m}}\right.\\ &\left.-\frac{4(N_{n,m}-2)+2}{N_{n,m}(N_{n,m}-1)}\langle K_{n,m}\mu,\mu\rangle^2_{G_{n,m}}+\frac{2k_{n,m}}{N_{n,m}(N_{n,m}-1)}\right],\end{aligned}$$

where the operator $K_{n,m}$ is given in (31), the distribution $G_{n,m}$ is defined in (24), and

$$k_{n,m}=\int\int K_n^2\mu_2\times\mu_2 1_{\mathscr{X}_{n,m}\times\mathscr{X}_{n,m}}\frac{dG\times G}{p_{n,m}^2}=:\frac{\alpha_{n,m}(2)}{p_{n,m}^2}.$$

We can split this into three terms. By Lemma 5.6 the expected value of the first term is bounded by a multiple of

$$\sum_m\frac{(np_{n,m})^3}{n^2(n-1)^2p_{n,m}^3}\|\mu_2\|_\infty\|K_n(\mu 1_{\mathscr{X}_{n,m}})\|_G^2\le\frac{1}{n}\|\mu_2\|_\infty\|K_n\|^2\|\mu\|_G^2.$$

Similarly the expected value of the absolute value of the second term is bounded by a multiple of

$$\sum_m\frac{(np_{n,m})^3}{n^2(n-1)^2p_{n,m}^4}\left\langle K_n(\mu 1_{\mathscr{X}_{n,m}}),\mu 1_{\mathscr{X}_{n,m}}\right\rangle_G^2\le\sum_m\frac{1}{np_{n,m}}\|K_n\|^2\|\mu 1_{\mathscr{X}_{n,m}}\|_G^4\le\frac{1}{n}\|K_n\|^2\|\mu\|_\infty^2\|\mu\|_G^2.$$

These two terms divided by k_n/n^2 tend to zero, by (5).

By Lemma 5.1 and (27) we have that var $V_n \sim 2k_n/n(n-1)$, which in term is asymptotically equivalent to $2\sum_m \alpha_{n,m}(2)/n(n-1)$, by (6). It follows that

$$\begin{aligned}\mathrm{var}(V_n|I_n)-\mathrm{var}\ V_n=&2\sum_m\frac{N_{n,m}(N_{n,m}-1)}{n^2(n-1)^2}k_{n,m}-2\frac{k_n}{n(n-1)}+o\left(\frac{k_n}{n^2}\right)\\ =&2\sum_m\left(\frac{N_{n,m}(N_{n,m}-1)}{n(n-1)p_{n,m}^2}-1\right)\frac{\alpha_{n,m}(2)}{n(n-1)}+o\left(\frac{k_n}{n^2}\right).\end{aligned}$$

Here the coefficients $\alpha_{n,m}(2)/k_n$ satisfy the conditions imposed on $\alpha_{n,m}$ in Corollary 5.2, in view of (6) and (7). Therefore this corollary shows that the expression on the right is $o_P(k_n/n^2)$.

5.7. Auxiliary lemmas on multinomial variables

Lemma 5.6—*Let N be binomially distributed with parameters* (n, p). *For any* $r \geq 2$ *there exists a constant* C_r *such that* $\mathrm{E}N^r 1_{N\geq 2} \leq C_r((np)^r \vee (np)^2)$.

Proof: For $r = \underline{r} + \delta$ with $\underline{r}$ an integer and $0 \leq \delta < 1$ there exists a constant C_r with $N^r 1_{N\geq 2} \leq C_r N^\delta N(N-1)\cdots(N-\underline{r}+1) + C_r N^\delta N(N-1)$ for every N. Hence

$$\mathrm{E}N^r 1_{N\geq 2} \leq C_r \sum_{k=2}^{n} k^\delta \left(k(k-1)\cdots(k-\underline{r}+1)+k(k-1)\right)\begin{pmatrix} n \\ k \end{pmatrix} p^k(1-p)^{n-k} = C_r\left((np)^{\underline{r}}\mathrm{E}N_1^\delta + (np)^2\mathrm{E}N_2^\delta\right),$$

for N_1 and N_2 binomially distributed with parameters $n - \underline{r}$ and p and $n - 2$ and p, respectively. By Jensen's inequality $\mathrm{E}N_j^\delta \leq (\mathrm{E}N_j)^\delta$, which is bounded above by $(np)^\delta$, yielding the upper bound $C_r((np)^r + (np)^{2+\delta})$. If $np \leq 1$, then this is bounded above by $2C_r(np)^2$ and otherwise by $2C_r(np)^r$.

The next result is a law of large numbers for a quadratic form in multinomial vectors of increasing dimension. The proof is based on a comparison of multinomial variables to Poisson variables along the lines of the proof of a central limit theorem in [12].

Lemma 5.7—*For each n let* N_n *be multinomially distributed with parameters* $(n, p_{n,1}, \ldots, p_{n,M_n})$ *with* $\max_m p_{n,m} \to 0$ *as* $n \to \infty$ *and* $\liminf_{n\to\infty} n \min_m p_{n,m} > 0$. *For given numbers* $\alpha_{n,m}$ *let*

$$s_n^2 = \frac{2}{n^2}\sum_m \frac{\alpha_{n,m}^2}{p_{n,m}^2} + \frac{4}{n}\sum_m p_{n,m}\left(\frac{\alpha_{n,m}}{p_{n,m}} - \sum_m \alpha_{n,m}\right)^2. \quad (34)$$

Then

$$\sum_m \alpha_{n,m}\left(\frac{N_{n,m}(N_{n,m}-1)}{n(n-1)p_{n,m}^2} - 1\right) = O_P\left(s_n + \frac{\sum_m |\alpha_{n,m}|}{\sqrt{n}}\right).$$

Proof: Because $\sum_m \alpha_{n,m}((n-1)/n - 1) = \sum_m \alpha_{n,m}(-1/n)$, it suffices to prove the statement of the lemma with $n(n-1)$ replaced by n^2. Using the fact that $\sum_m N_{n,m} = n$ we can rewrite the resulting quadratic form as, with $\lambda_{n,m} = np_{n,m}$,

$$\sum_m \alpha_{n,m}\left(\frac{N_{n,m}(N_{n,m}-1)}{n^2 p_{n,m}^2} - 1\right) = \sqrt{2}\sum_m \frac{\alpha_{n,m}}{\lambda_{n,m}} C_2(N_{n,m}, \lambda_{n,m}) + 2\sum_m \sqrt{\lambda_{n,m}}\left(\frac{\alpha_{n,m}}{\lambda_{n,m}} - \frac{\sum_m \alpha_{n,m}}{n}\right) C_1(N_{n,m}, \lambda_{n,m}),$$

for C_1 and C_2 the Poisson-Charlier polynomials of degrees 1 and 2, given by

$$C_1(x,\lambda)=\frac{x-\lambda}{\sqrt{\lambda}}, \quad C_2(x,\lambda)=\frac{x(x-1)-2\lambda x+\lambda^2}{\sqrt{2}\lambda}.$$

Together with $x \mapsto C_0(x) = 1$ the functions $x \mapsto C_1(x, \lambda)$ and $x \mapsto C_2(x, \lambda)$ are the polynomials 1, x, x^2 orthonormalized for the Poisson distribution with mean λ by the Gramm-Schmidt procedure. For $X = (X_1, \dots, X_{M_n})$ let

$$T_n(X)=\sum_m \frac{\alpha_{n,m}}{\lambda_{n,m}} C_2(X_m,\lambda_{n,m}) +\sum_m \sqrt{2\lambda_{n,m}} \left(\frac{\alpha_{n,m}}{\lambda_{n,m}} - \frac{\sum_m \alpha_{n,m}}{n}\right) C_1(X_m,\lambda_{n,m}).$$

Thus up to a factor $\sqrt{2}$ the statistic $T_n(N_n)$ is the quadratic form of interest.

If the variables $N_{n,1}, \dots, N_{n,M_n}$ were independent Poisson variables with mean values $\lambda_{n,m}$, then the mean of $T_n(N_n)$ would be zero and the variance would be given by $s_n^2/2$, and hence in that case $T_n(N_n) = O_P(s_n)$. We shall now show that the difference between multinomial and Poisson variables is of the order $\sum_m |\alpha_{n,m}|/\sqrt{n}$.

To make the link between multinomial and Poisson variables, let $\tilde{n}$ be a Poisson variable with mean n and given $\tilde{n} = k$ let $\tilde{N}_n = (\tilde{N}_{n,1}, \dots, \tilde{N}_{n,M_n})$ be multinomially distributed with parameters k and $p_n = (p_{n,1}, \dots, p_{n,M_n})$. The original multinomial vector N_n is then equal in distribution to $\tilde{N}_n$ given $\tilde{n} = n$. Furthermore, the vector $\tilde{N}_n$ is unconditionally Poisson distributed as in the preceding paragraph, whence, for any $M_n \to \infty$,

$$\mathrm{P}(|T_n(\tilde{N}_n)|>M_n s_n) \to 0.$$

The left side is bigger than

$$\sum_{k:|k-n|\leq \sqrt{n}} \mathrm{P}(|T_n(\tilde{N}_n)|>M_n s_n|\tilde{n}=k)\mathrm{P}(\tilde{n}=k) \;\geq \min_{k:|k-n|\leq \sqrt{n}} \mathrm{P}(|T_n(N_n(k))|>M_n s_n)\mathrm{P}(|\tilde{n}-n| \leq \sqrt{n}),$$

where the vector $N_n(k)$ is multinomial with parameters k and p_n. Because the sequence $(\tilde{n}-n)/\sqrt{n}$ tends to a standard normal distribution as $n \to \infty$, the probability $\mathrm{P}\left(|\tilde{n}-n| \leq \sqrt{n}\right)$ tends to the positive constant $\Phi(1) - \Phi(-1)$. We conclude that the sequence of minima on the right tends to zero. The probability of interest is the term with $k = n$ in the minimum. Therefore the proof is complete once we show that the minimum and maximum of the terms are comparable.

To compare the terms with different k we couple the multinomial vectors $N_n(k)$ on a single probability space. For given $k < k'$ we construct these vectors such that $N_n(k')=N_n(k)+N_n'(k'-k)$ for $N_n'(k'-k)$ a multinomial vector with parameters $k' - k$

and p_n independent of $N_n(k)$. For any numbers N and N' we have that $C_2(N+N',\lambda)-C_2(N,\lambda)=\left((N')^2+2NN'-N'(1+2\lambda)\right)/(\sqrt{2}\lambda)$. Therefore,

$$\begin{aligned}&\mathrm{E}\left|\sum_m\frac{\alpha_{n,m}}{\lambda_{n,m}}C_2(N_{n,m}(k'),\lambda_{m,n})-\sum_m\frac{\alpha_{n,m}}{\lambda_{n,m}}C_2(N_{n,m}(k),\lambda_{m,n})\right|\\&\quad\leq\sum_m\frac{|\alpha_{n,m}|}{\lambda_{n,m}}\frac{\mathrm{E}|N'_{n,m}(k'-k)^2+2N'_{n,m}(k'-k)N_{n,m}(k)-N'_{n,m}(k'-k)(1+2\lambda_{n,m})|}{\sqrt{2}\lambda_{n,m}}.\end{aligned}$$

For $|k'-n|\leq\sqrt{n}$ and $|k-n|\leq\sqrt{n}$ the binomial variable $N_{n,m}(k'-k)$ has first and second moment bounded by a multiple of $\sqrt{n}p_{n,m}$ and $np_{n,m}^2$. From this the right side of the display can be seen to be of the order $\sum_m|\alpha_{n,m}|O(n^{-1/2}):=\rho_n$. Similarly, we have $C_1(N+N',\lambda)-C_1(N,\lambda)=N'/\sqrt{\lambda}$ and

$$\mathrm{E}|\sum_m\sqrt{\lambda_{n,m}}\left(\frac{\alpha_{n,m}}{\lambda_{n,m}}-\frac{\sum_m\alpha_{n,m}}{n}\right)(C_1(N_{n,m}(k'),\lambda_{n,m})-C_1(N_{n,m}(k),\lambda_{n,m}))\Big|$$

can be seen to be of the order $\sum_m|\alpha_{n,m}/\lambda_{n,m}-\sum_m\alpha_{n,m}/n|\sqrt{n}p_{n,m}$, which is also of the order ρ_n.

We infer from this that $\mathrm{E}|T_n(N_n(k))-T_n(N_n(n))=O(\rho_n)$, uniformly in $|k-n|\leq\sqrt{n}$, and therefore

$$\begin{aligned}&\mathrm{P}(|T_n(N_n(n))|>M_n(s_n+\rho_n))\\&\quad\leq\mathrm{P}(|T_n(N_n(k))|>M_ns_n)+\mathrm{P}(|T_n(N_n(n))-T_n(N_n(k))|>M_n\rho_n)\\&\quad\leq\mathrm{P}(|T_n(N_n(k))|>M_ns_n)+o(1),\end{aligned}$$

uniformly in $|k-n|\leq\sqrt{n}$, for every $M_n\to\infty$, by Markov's inequality. In the preceding paragraph it was seen that the minimum of the right side over k with $|k-n|\leq\sqrt{n}$ tends to zero for any $M_n\to\infty$. Hence so does the left side.

Under the additional condition that

$$\frac{1}{s_n^2}\max_m\left[\frac{\alpha_{n,m}^2}{n^2p_{n,m}^2}+\frac{p_{n,m}}{n}\left(\frac{\alpha_{n,m}}{p_{n,m}}-\sum_m\alpha_{n,m}\right)^2\right]\to0,$$

it follows from Corollary 4.1 in [12] that the sequence s_n^{-1} times the quadratic form in the preceding lemma tends in distribution to the standard normal distribution. Thus in this case the order claimed by the lemma is sharp as soon as $n^{-1/2}\sum_m|\alpha_{n,m}|$ is not bigger than s_n.

Corollary 5.2—*For each n let N_n be multinomially distributed with parameters* $(n, p_{n,1}, \ldots, p_{n,M_n})$ *with* $\liminf_{n\to\infty} n \min_m p_{n,m} > 0$. *If* $\alpha_{n,m}$ *are numbers with* $\sum_m |\alpha_{n,m}| = O(1)$ *and* $\max_m |\alpha_{n,m}| \to 0$ *as* $n \to \infty$*, then*

$$\sum_m \alpha_{n,m}\left(\frac{N_{n,m}(N_{n,m}-1)}{n(n-1)p_{n,m}^2}-1\right)\xrightarrow{P} 0.$$

Proof: Since $np_{n,m} \gtrsim 1$ by assumption the numbers s_n defined in (34) satisfy

$$s_n^2 \le 2\sum_m \frac{\alpha_{n,m}^2}{n^2 p_{n,m}^2} + 4\sum_m \frac{\alpha_{n,m}^2}{np_{n,m}} \lesssim \sum_m \alpha_{n,m}^2.$$

The corollary is a consequence of Lemma 5.7.

6. Proofs for Section 3

Proof of Corollary 3.1

We consider the distribution of T_n conditionally given the observations used to construct the initial estimators $\hat b_n$ and $\hat g_n$. By passing to subsequences of n, we may assume that these sequences converge almost surely to b and g relative to the uniform norm. In the proof of distributional convergence the initial estimators $\hat b_n$ and $\hat g_n$ may therefore be understood to be deterministic sequences that converge to limits b and g.

The estimator (12) is a sum $T_n = T_n^{(1)} + T_n^{(2)}$ of a linear and quadratic part. The (conditional) variance of the linear term $T_n^{(1)}$ is of the order $1/n$, which is of smaller order than k_n/n^2. It follows that $(T_n^{(1)} - \mathrm{E}T_n^{(1)})/(\sqrt{k_n}/n)$ tends to zero in probability.

To study the quadratic part $T_n^{(2)}$ we apply Theorem 2.1 with the kernel K_n of the theorem taken equal to the present $K_{k_n,\hat g_n}$ and the Y_r of the theorem taken equal to the present $Y_r - \hat b_n(X_r)$. For given functions b_1 and g_1, set

$$\mu_q(b_1)(x) = \mathrm{E}(|Y_1 - b_1(X_1)|^q | X_1 = x) = \mathrm{E}(|\varepsilon_1 + (b - b_1)(x)|^q | X_1 = x),$$

$$k_n(b_1, g_1) = \int\int (\mu_2(b_1) \times \mu_2(b_1)) K_{k_n,g_1}^2 \, d(G \times G).$$

The function $\mu_q(\hat b_n)$ converges uniformly to the function $\mu_q(b)$, which is uniformly bounded by assumption, for $q = 1$, $q = 2$ and some $q > 2$. Furthermore $K_{k_n,\hat g_n} = K_{k_n,g}\sqrt{g \times g/\hat g_n \times \hat g_n}$, where the function $g \times g/\hat g_n \times \hat g_n$ converges uniformly to one. Therefore, the conditions of Theorem 2.1 (for the case that the observations are non-i.i.d.; cf. the remark following the theorem) are satisfied by Theorem 4.1 or 4.2. Hence the sequence $(T_n^{(2)} - \mathrm{E}T_n^{(2)})/\sqrt{\hat{k}_n/n^2}$ tends to a standard normal distribution, for $\hat{k}_n = k_n(\hat{b}_n, \hat{g}_n)$. From the conditions on the initial estimators it follows that $\hat{k}_n/k_n(b, g) \to 1$. Here $k_n(b, g)$ is of the order the dimension k_n of the kernel.

Let $T_n(b_1, g_1)$ be as T_n, but with the initial estimators $\hat{b}_n$ and $\hat{g}_n$ replaced by b_1 and g_1. Its expectation is given by

$$e(b_1, g_1) = \mathrm{E}_{b,g} T_n(b_1, g_1) = \int b_1^2 dG + \int 2b_1(b-b_1) dG + \int\int (b-b_1)\times(b-b_1) K_{k_n,g_1} dG\times G.$$

In particular $e(b, g) = \int b^2\, dG$. Using the fact that $K_{k_n,g}$ is an orthogonal projection in $L_2(G)$ we can write

$$\begin{aligned} e(b_1, g_1) - e(b, g) &= -\int (b_1 - b)^2 dG + \int\int (b-b_1)\times(b-b_1) K_{k_n,g_1} dG\times G \\ &= -\|(I - K_{k_n,g})(b_1 - b)\|_G^2 \\ &+ \int\int (b-b_1)\times(b-b_1)(K_{k_n,g_1} - K_{k_n,g}) dG\times G. \end{aligned} \quad (35)$$

By the definition of $K_{k_n,g}$ the absolute value of the first term on the right can be bounded as

$$\left\|(b-b_1) - \mathrm{lin}\left(\frac{e_1}{\sqrt{g}}, \dots, \frac{e_k}{\sqrt{g}}\right)\right\|_G^2 = \|(b-b_1)\sqrt{g} - \mathrm{lin}\,(e_1, \dots, e_k)\|_\lambda^2.$$

By assumption b is β-Hölder and g is γ-Hölder for some $\gamma \geq \beta$ and bounded away from zero. Then $b\sqrt{g}$ is β-Hölder and hence its uniform distance to lin $(e_1, \dots, e_k)$ is of the order $(1/k)^\beta$. If the norm of $\hat{b}_n$ in $C^\beta[0, 1]$ is bounded, then we can apply the same argument to the functions $\hat{b}_n\sqrt{g}$, uniformly in n, and conclude that the expression in the display with $\hat{b}_n$ instead of b_1 is bounded above by $O_P(1/k_n)^{2\beta}$. If the projection is on the Haar basis and $\hat{b}_n$ is contained in lin $(e_1, \dots, e_{k_n})$, then the approximation error can be seen to be of the same order, from the fact that the product of two projections on the Haar basis is itself a projection on this basis.

For $h = (\sqrt{g} - \sqrt{g_1})/\sqrt{gg_1}$ we can write

$$\frac{1}{\sqrt{g_1(x_1)}\sqrt{g_1(x_2)}} - \frac{1}{\sqrt{g(x_1)}\sqrt{g(x_2)}} = h(x_1)\left(\frac{1}{\sqrt{g_1(x_2)}}\right) + h(x_2)\left(\frac{1}{\sqrt{g(x_1)}}\right).$$

If multiplied by a symmetric function in (x_1, x_2) and integrated with respect to $G \times G$, the arguments x_1 and x_2 in the second term can be exchanged. The second term on the right in (35) can therefore be written

$$\begin{aligned}&\left\langle K_{k_n,\lambda}\left((b-b_1)h\right),(b-b_1)\left(\frac{1}{\sqrt{g_1}}+\frac{1}{\sqrt{g}}\right)\right\rangle_G\\&\lesssim \left\|K_{k_n,\lambda}\left((b-b_1)h\right)\right\|_{G,3/2}\left\|b-b_1\right\|_{G,3}\\&\lesssim \left\|(b-b_1)h\right\|_{G,3/2}\left\|b-b_1\right\|_{G,3}\lesssim \left\|b-b_1\right\|_{\lambda,3}\left\|h\right\|_{\lambda,3}\left\|b-b_1\right\|_{\lambda,3}.\end{aligned}$$

Here $\|\cdot\|_{G,3}$ is the $L_3(G)$-norm, we use the fact that L_2-projection on a wavelet basis decreases L_p-norms for $p = 3/2$ up to constants, and the multiplicative constants depend on uniform upper and lower bounds on the functions g_1 and g. We evaluate this expression for $b_1 = \hat{b}_n$ and $g_1 = \hat{g}_n$, and see that it is of the order $O(\|\hat{b}_n - b\|_3^2\|\hat{g}_n - g\|_3)$.

Finally we note that $\hat{\mathrm{E}}_{b,g}T_n = e(\hat{b}_n, \hat{g}_n)$ and combine the preceding bounds.

Acknowledgments

The research leading to these results has received funding from the European Research Council under ERC GrantAgreement320637.

7. Appendix: proofs for Section 4

Lemma 7.1

The kernel of an orthogonal projection on a k-dimensional space has operator norm $\|K_k\|_2 = 1$, *and square* $L_2(G\times G)$*-norm* $\int\int K_k^2\,d(G\times G)=k$.

Proof

The operator norm is one, because an orthogonal projection decreases norm and acts as the identity on its range. It can be verified that the kernel of a kernel operator is uniquely defined by the operator. Hence the kernel of a projection on a k-dimensional space can be written in the form (13), from which the L_2-norm can be computed.

Proof of Proposition 4.1—We can reexpress the wavelet expansion (17) to start from level I as

$$f(x)=\sum_{j\in\mathbb{Z}^d}\sum_{\upsilon\in\{0,1\}^d}\langle f,\psi_{I,j}^{\upsilon}\rangle\psi_{I,j}^{\upsilon}(x)+\sum_{i=I+1}^{\infty}\sum_{j\in\mathbb{Z}^d}\sum_{\upsilon\in\{0,1\}^d-\{0\}}\langle f,\psi_{i,j}^{\upsilon}\rangle\psi_{i,j}^{\upsilon}(x).$$

The projection kernel K_k sets the coefficients in the second sum equal to zero, and hence can also be expressed as

$$K_k(x_1,x_2)=\sum_{j\in J_I}\sum_{\upsilon\in\{0,1\}^d}\psi_{I,j}^{\upsilon}(x_1)\psi_{I,j}^{\upsilon}(x_2).$$

The double integral of the square of this function over $\mathbb{R}^{2d}$ is equal to the number of terms in the double sum (cf. (13) and the remarks following it), which is $O(2^{Id})$. The support of only a small fraction of functions in the double sum intersects the boundary of $\mathscr{X}$. Because also the density of G and the function μ_2 are bounded above and below, it follows that the weighted double integral k_n of K_k^2 relative to G as in (4) is also of the exact order $O(2^{Id})$.

Each function $(x_1, x_2) \mapsto \psi^{\upsilon}_{I,j}(x_1)\psi^{\upsilon}_{I,j}(x_2)$ has uniform norm bounded above by 2^{Id} times the uniform norm of the base wavelet of which it is a shift and dilation. A given point (x_1, x_2) belongs to the support of fewer than C_1^d of these functions, for a constant C_1 that depends on the shape of the support of the wavelets. Therefore, the uniform norm of the kernel K_k is of the order k_n.

By assumption each function $\psi^{\upsilon}_{I,j}$ is supported within a set of the form $2^{-I}(C + j)$ for a given cube C that depends on the type of wavelet, for any υ. It follows that the function $(x_1, x_2) \mapsto \psi^{\upsilon}_{I,j}(x_1)\psi^{\upsilon}_{I,j}(x_2)$ vanishes outside the cube $2^{-I}(C + j) \times 2^{-I}(C + j)$. There are $O(2^{Id})$ of these cubes that intersect $\mathscr{X} \times \mathscr{X}$; these intersect the diagonal of $\mathscr{X} \times \mathscr{X}$, but may be overlapping. We choose the sets $\mathscr{X}_{n,m}$ to be blocks (cubes) of l_n^d adjacent cubes $2^{-I}(C + j)$, giving $M_n = O(k_n/l_n^d)$ sets $\mathscr{X}_{n,m}$. [In the case $d = 1$, the "cubes" are intervals and they can be ordered linearly; the meaning of "adjacent" is then clear. For $d > 1$ cubes are "adjacent" in d directions. We stack l_n cubes $2^{-I}(C + j)$ in each direction, giving cubes $\mathscr{X}_{n,m}$ of sides with lengths l_n times the length of a cube $2^{-I}(C + j)$.]

Because the kernels are bounded by a multiple of k_n, condition (10) is implied by (11), in view of Lemma 2.1, The latter condition reduces to $M_n^{-1}k/n \to 0$, the probabilities $G(\mathscr{X}_{n,m})$ being of the order $1/M_n$.

The set of cubes $2^{-I}(C + j)$ that intersects more than one set $\mathscr{X}_{n,m}$ is of the order $M_n^{1/d}k_n^{1-1/d}$. To see this picture the set $\mathscr{X}$ as a supercube consisting of the M cubes $\mathscr{X}_{n,m}$, stacked together in a $M^{1/d} \times \cdots \times M^{1/d}$-pattern. For each coordinate $i = 1, \ldots, d$ the stack of cubes $\mathscr{X}$ can be sliced in $M^{1/d}$ layers each consisting of $(M^{1/d})^{d-1}$ cubes $\mathscr{X}_{m,n}$, which are $l_n(k_n^{1/d})^{d-1} = l_n^d(M_n^{1/d})^{d-1}$ cubes $2^{-I}(C + j)$. The union of the boundaries of all slices ($i = 1, \ldots, d$ and $M_n^{1/d}$ slices for each i) contains the union of the boundaries of the sets $\mathscr{X}_{n,m}$. The boundary between two particular slices is intersected by at most $C_2(k_n^{1/d})^{d-1}$ cubes $2^{-I}(C + j)$, for a constant C_2 depending on the amount of overlap between the cubes. Thus in total of the order $dM_n^{1/d}(k_n^{1/d})^{d-1}$ cubes intersect some boundary.

If $K_k(x_1, x_2) \neq 0$, then there exists j and υ with $\psi^{\upsilon}_{I,j}(x_1)\psi^{\upsilon}_{I,j}(x_2) \neq 0$, which implies that there exists j such that $x_1, x_2 \in 2^{-I}(C + j)$. If the cube $2^{-I}(C + j)$ is contained in some $\mathscr{X}_{n,m}$, then $(x_1, x_2) \in \mathscr{X}_{n,m} \times \mathscr{X}_{n,m}$. In the other case $2^{-I}(C + j)$ intersects the boundary of some $\mathscr{X}_{n,m}$. It follows that the set of (x_1, x_2) in the complement of $\cup_m \mathscr{X}_{n,m} \times \mathscr{X}_{n,m}$ where $K_k(x_1, x_1) \neq 0$ is contained in the union U of all cubes $2^{-I}(C + j)$ that intersect the boundary of some $\mathscr{X}_{n,m}$. The integral of K_k^2 over this set satisfies

$$\frac{1}{k_n}\int\int_U K_k^2 d(G\times G) \lesssim \tfrac{1}{k_n}k_n^2(G\times G)(U) \ \lesssim \tfrac{1}{k_n}k_n^2 M_n^{1/d}k_n^{1-1/d}\left(\tfrac{1}{k_n}\right)^2=\left(\tfrac{M_n}{k_n}\right)^{1/d}.$$

Here we use that $G(2^{-I}(C+j)) \lesssim 1/k_n$. This completes the verification of (6).

By the spatial homogeneity of the wavelet basis, the contributions of the sets $\mathscr{X}_{n,m} \times \mathscr{X}_{n,m}$ to the integral of K_k^2 are comparable in magnitude. Hence condition (7) is satisfied for any $M_n \to \infty$.

In order to satisfy conditions (8) and (9) we must choose $M_n \to \infty$ with $M_n \lesssim n$. This is compatible with choices such that $M_n/k_n \to 0$ and $M_n^{-1}k/n \to 0$.

Proof of Proposition 4.2—Because K_k is an orthogonal projection on a $(2k+1)$-dimensional space, Lemma 7.1 gives that the operator norm satisfies $\|K_k\| = 1$ and that the numbers k_n as in (4) but with $\mu_2 = 1$ are equal to $\int\int K_k^2 d\lambda d\lambda=2k+1$.

By the change of variables $x_1 - x_2 = u$, $x_1 + x_2 = \upsilon$ we find, for any $\varepsilon \in (0, \pi]$, and $K_k(x_1, x_2) = D_k(x_1 - x_2)$,

$$\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} 1_{|x_1-x_2|>\varepsilon}K_k^2(x_1,x_2)dx_1dx_2=2\int_{\varepsilon}^{2\pi}\int_{u-2\pi}^{2\pi-u}D_k^2(u)\frac{1}{2}dvdu \ =2\int_{\varepsilon}^{2\pi}D_k^2(u)(2\pi-u)du.$$

By the symmetry of the Dirichlet kernel about π we can rewrite $\int_{\pi}^{2\pi}D_k^2(u)(2\pi-u)du$ as $\int_0^{\pi}D_k^2(u)udu$. Splitting the integral on the right side of the preceding display over the intervals $(\varepsilon, \pi]$ and $(\pi, 2\pi]$, and rewriting the second integral, we see that the preceding display is equal to

$$2\int_{\varepsilon}^{\pi}D_k^2(u)(2\pi-u)du+2\int_0^{\pi}D_k^2(u)udu=4\pi\int_{\varepsilon}^{\pi}D_k^2(u)du+2\int_0^{\varepsilon}D_k^2(u)udu.$$

For $\varepsilon = 0$ this expression is equal to the square L_2-norm of the kernel K_k, which shows that $4\pi\int_0^{\pi}D_k^2(u)du=2k+1$. On the interval (ε, π) the kernel D_k is bounded above by $\left(2\pi\ \sin(\frac{1}{2}\varepsilon)\right)^{-1}$. Therefore, the preceding display is bounded above by

$$\frac{4\pi}{\left(2\pi\ \sin(\frac{1}{2}\varepsilon)\right)^2}\int_{\varepsilon}^{\pi}du+2\varepsilon\int_0^{\varepsilon}D_k^2(u)du \lesssim \frac{1}{\sin^2(\frac{1}{2}\varepsilon)}+\varepsilon k.$$

We conclude that, for small $\varepsilon > 0$,

$$\frac{1}{2k+1}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}1_{|x_1-x_2|>\varepsilon}K_k^2(x_1,x_2)dx_1dx_2 \lesssim \varepsilon+\frac{1}{\varepsilon^2 k}.$$

This tends to zero as $k \to \infty$ whenever $\varepsilon = \varepsilon_k \downarrow 0$ such that $\varepsilon \gg 1/\sqrt{k}$.

We choose a partition $(-\pi, \pi] = \cup_m \mathscr{X}_{n,m}$ in $M_n = 2\pi/\delta$ intervals of length δ for $\delta \to 0$ with $\delta \gg \varepsilon$ and ε satisfying the conditions of the preceding paragraph. Then the complement of $\cup_m \mathscr{X}_{n,m} \times \mathscr{X}_{n,m}$ is contained in $\{(x_1, x_2): |x_1 - x_2| > \varepsilon\}$ except for a set of $2(M_n - 1)$ triangles, as indicated in Figure 3. In order to verify (6) it suffices to show that $(2k + 1)^{-1}$ times the integral of K_k^2 over the union of the triangles is negligible. Each triangle has sides of length of the order ε, whence, for a typical triangle Δ, by the change of variables $x_1 - x_2 = u$, $x_2 = \upsilon$, and an interval I of length of the order ε,

$$\int\int_{\Delta} K_k^2(x_1, x_2)dx_1dx_2 \lesssim \int_I \int_0^{\varepsilon} D_k^2(u)dud\upsilon \lesssim \varepsilon(2k+1).$$

Hence (6) is satisfied if $2(M_n - 1)\varepsilon \to 0$, i.e. $\varepsilon \ll \delta$.

Because $\int_{\mathscr{X}_{n,m}}\int_{X_{n,m}} K_k^2 d(\lambda \times \lambda)$ is independent of m, (7) is satisfied as soon as the number of sets in the partitions tends to infinity.

Because $0 \le K_k \le 2k + 1$, condition (10) is implied by (11), which is satisfied if $\delta \ll n/k$.

The desired choices $1/\sqrt{k} \ll \varepsilon \ll \delta \ll n/k$ are compatible, as by assumption $k/n^2 \to 0$.

Proof of Proposition 4.3—Without loss of generality we can assume that $\int |\phi|\, d\lambda = 1$. By a change of variables

$$\int K_\sigma^2 d(G \times G) = \frac{1}{\sigma}\int \phi^2(\upsilon) \int g(x - \sigma\upsilon)g(x)dxd\upsilon.$$

Here $|\int g(x - \sigma\upsilon)g(x)\, dx| \le \|g\|_\infty$ and, as $\sigma \downarrow 0$,

$$\left|\int g(x - \sigma\upsilon)g(x)dx - \int g^2(x)dx\right| \le \|g\|_\infty \int |g(x - \sigma\upsilon) - g(x)|dx \to 0,$$

for every fixed υ, by the L_1-continuity theorem. We conclude by the dominated convergence theorem that $\sigma\int K_\sigma^2 d(G \times G) \to \int g^2 d\lambda \int \phi^2 d\lambda$. Because μ_2 is bounded away from 0 and ∞, the numbers k_n defined in (4) are of the exact order σ^{-1}.

By another change of variables, followed by an application of the Cauchy-Schwarz inequality, for any $f \in L_2(G)$,

$$\int (K_\sigma f)^2 dG = \int \left(\int \phi(\upsilon)(fg)(x - \sigma\upsilon)d\upsilon\right)^2 dG(x) \le \|g\|_\infty^2 \int\int |\phi|(\upsilon)(f^2g)(x-\sigma\upsilon)d\upsilon dx = \|g\|_\infty^2 \int f^2 g d\lambda.$$

Therefore, the operator norms of the operators K_σ are uniformly bounded in $\sigma > 0$.

We choose a partition $\mathbb{R} = \cup_m \mathscr{X}_{n,m}$ consisting of two infinite intervals $(-\infty, -a]$ and (a, ∞) and a regular partition of the interval $(-a, a]$ in such a way that every partitioning set satisfies $G(\mathscr{X}_{n,m}) \leq \delta$. We can achieve this with a partition in $M_n = O(1/\delta)$ sets.

Because $|K_\sigma|$ is bounded by a multiple of σ^{-1}, condition (10) is implied by (11), which takes the form $\delta/(\sigma n) \rightarrow 0$, in view of Lemma 2.1.

For an arbitrary partitioning set $\mathscr{X}_{n,m}$,

$$\sigma\int_{\mathscr{X}_{n,m}}\int_{\mathscr{X}_{n,m}} K_\sigma^2 d(G\times G) \leq \int_{\mathscr{X}_{n,m}}\int \phi^2(\upsilon)g(x-\sigma\upsilon)g(x)d\upsilon dx \ \leq \|g\|_\infty \int \phi^2(\upsilon)d\upsilon G(\mathscr{X}_{n,m}).$$

It follows that (7) is satisfied as soon as $\delta \rightarrow 0$.

Finally, we verify condition (6) in two steps. First, for any $\varepsilon \downarrow 0$, by the change of variables $x_1 - x_2 = \upsilon$, $x_2 = x$,

$$\sigma \underset{|x_1-x_2|>\varepsilon}{\int\int} K_\sigma^2 d(G\times G) = \underset{|\upsilon|>\varepsilon/\sigma}{\int\int} \phi^2(\upsilon)g(x-\sigma\upsilon)g(x)dxd\upsilon \ \leq \|g\|_\infty \int_{|\upsilon|>\varepsilon/\sigma} \phi^2(\upsilon)d\upsilon.$$

This converges to zero as $\sigma \rightarrow 0$ for any $\varepsilon = \varepsilon_\sigma > 0$ with $\varepsilon \gg \sigma$. Second, for $\varepsilon \ll \delta$ the complement of the set $\cup_m \mathscr{X}_{n,m} \times \mathscr{X}_{n,m}$ is contained in $\{(x_1, x_2): |x_1 - x_2| > \varepsilon\}$ except for a set of $2(M_n - 1)$ triangles, as indicated in Figure 3. In order to verify (6) it suffices to show that σ times the integral of K_σ^2 over the union of the triangles is negligible. Each triangle has sides of length of the order ε, whence, for a typical triangle Δ, with projection I on the x_1-axis,

$$\sigma \underset{\Delta}{\int\int} K_\sigma^2 d(G\times G) \lesssim \int_I \int_{|\upsilon|<\varepsilon/\sigma} \phi^2(\upsilon)g(x-\sigma\upsilon)g(x)d\upsilon dx \leq \varepsilon\|g\|_\infty \int \phi^2(\upsilon)d\upsilon.$$

The total contribution of all triangles is $2(M_n - 1)$ times this expression. Hence (6) is satisfied if $2(M_n - 1)\varepsilon \rightarrow 0$, i.e. $\varepsilon \ll \delta$.

The preceding requirements can be summarized as $\sigma \ll \varepsilon \ll \delta \ll \sigma n$, and are compatible.

Proof of Proposition 4.4—Inequality (22) implies that $c_j(f) = 0$ for every f that vanishes outside the interval (t_j', t_{j+r}'), whence the representing function g_j is supported on this interval. It follows that the function $(x_1, x_2) \mapsto N_j(x_1)c_j(x_2)$ vanishes outside the square $[t_j', t_{j+r}'] \times [t_j', t_{j+r}']$, which has area of the order l^{-2}. We form a partition $(0, 1] = \cup_m \mathscr{X}_{n,m}$ by selecting subsets $0 = s_0^l < s_1^l < \cdots < s_{M_n}^l = 1$ of the basic knot sequences such that $M_n^{-1} \lesssim s_{i+1}^l - s_i^l \lesssim M_n^{-1}$ for every i and define $\mathscr{X}_{n,m} = (s_{m-1}^l, s_m^l]$. The numbers M_n are chosen integers much smaller than l_n, and we may set $s_i^l = t_{ip}^l$ for $p = \lfloor l_n/M_n \rfloor$.

Because K_k is a projection on $S_r(T, d)$ and the function $x_1 \mapsto K_k(x_1, x_2)$ is contained in $S_r(T, d)$ for every x_2, it follows that $\int K_k(x_1, x_2)K_k(x_1, x_2)\, dx_1 = K_k(x_2, x_2)$ for every x_2, and hence

$$\int\int K_k(x_1,x_2)^2 dx_1 dx_2 = \int K_k(x_1,x_1)d\lambda(x_1) = \int \sum_j N_j(x_1)c_j(x_1)dx_1 = \sum_j c_j(N_j) = \sum_j 1 = k,$$

because the identities $N_i = K_k N_i = \sum_j c_j(N_i)N_j$ imply that $c_j(N_i) = \delta_{ij}$ by the linear independence of the B-splines. Because the density of G and the function μ_2 are bounded above and below the $L_2(G \times G)$-norm k_n as in (4) is of the same order as the dimension $k_n = r + l_n d$ of the spline space.

Inequality (22) implies that the norm of the linear map c_j, which is the infinity norm $\|c_j\|_\infty$ of the representing function, is bounded above by a constant times $(t'_{j+r} - t'_j)^{-1}$, which is of the order k_n. Therefore,

$$\frac{1}{k_n}\int\int_{(\cup_m \mathscr{X}_{n,m}\times\mathscr{X}_{n,m})^c} K_n^2(\mu_2\times\mu_2)d(G\times G) \\ \lesssim \frac{1}{k_n}k_n^2\|\mu_2\|_\infty^2 \lambda\left(\bigcup_j (t'_j, t'_{j+r}] \times (t'_j, t'_{j+r}] - \bigcup_m (s_{m-1}, s_m] \times (s_{m-1}, s_m]\right).$$

The set in the right side is the union of M_n cubes of areas not bigger than the area of the sets $(t'_j, t'_{j+r}] \times (t'_j, t'_{j+r}]$, which is bounded above by a constant times k_n^{-2}. (See Figure 7.) The preceding display is therefore bounded above by

$$\frac{1}{k_n}k_n^2\|\mu_2\|_\infty^2 M_n \frac{1}{k_n^2}.$$

For $M_n/k_n \rightarrow 0$ this tends to zero. This completes the verification of (6).

The verification of the other conditions follows the same lines as in the case of the wavelet basis.

References

1. Robins, J.; Li, L.; Tchetgen, E.; van der Vaart, A. Probability and statistics: essays in honor of David A. Freedman, Vol. 2 of Inst. Math. Stat. Collect., Inst. Math. Statist. Beachwood, OH: 2008. Higher order influence functions and minimax estimation of nonlinear functionals; p. 335-421.URL http://dx.doi.org/10.1214/193940307000000527
2. Bickel PJ, Ritov Y. Estimating integrated squared density derivatives: sharp best order of convergence estimates. Sankhyā Ser. A. 1988; 50(3):381–393.
3. Birgé L, Massart P. Estimation of integral functionals of a density. Ann. Statist. 1995; 23(1):11–29.
4. Laurent B. Efficient estimation of integral functionals of a density. Ann. Statist. 1996; 24(2):659–681.
5. Laurent B. Estimation of integral functionals of a density and its derivatives. Bernoulli. 1997; 3(2):181–211.

6. Laurent B, Massart P. Adaptive estimation of a quadratic functional by model selection. Ann. Statist. 2000; 28(5):1302–1338.

7. Robins J, Li L, Tchetgen E, van der Vaart AW. Quadratic semiparametric von Mises calculus. Metrika. 2009; 69(2–3):227–247. URL http://dx.doi.org/10.1007/s00184-008-0214-3. [PubMed: 23087487]

8. van der Vaart A. Higher order tangent spaces and influence functions. Statist. Sci. 2014; 29(4):679–686.

9. Tchetgen E, Li L, Robins J, van der Vaart A. Higher order estimating equations for high-dimensional semiparametric models. preprint.

10. Robins J, van der Vaart A. Adaptive nonparametric confidence sets. Ann. Statist. 2006; 34(1):229–253.

11. Mikosch T. A weak invariance principle for weighted U-statistics with varying kernels. J. Multivariate Anal. 1993; 47(1):82–102.

12. Morris C. Central limit theorems for multinomial sums. Ann. Statist. 1975; 3:165–188.

13. Ermakov MS. Asymptotic minimaxity of chi-squared tests. Teor. Veroyatnost. i Primenen. 1997; 42(4):668–695.

14. Weber NC. Central limit theorems for a class of symmetric statistics. Math. Proc. Cambridge Philos. Soc. 1983; 94(2):307–313. URL http://dx.doi.org/10.1017/S0305004100061168.

15. Bhattacharya RN, Ghosh JK. A class of *U*-statistics and asymptotic normality of the number of *k*-clusters. J. Multivariate Anal. 1992; 43(2):300–330. URL http://dx.doi.org/10.1016/0047-259X(92)90038-H.

16. Jammalamadaka SR, Janson S. Limit theorems for a triangular scheme of U-statistics with applications to inter-point distances. Ann. Probab. 1986; 14(4):1347–1358.

17. de Jong P. A central limit theorem for generalized quadratic forms. Probab. Theory Related Fields. 1987; 75(2):261–277. URL http://dx.doi.org/10.1007/BF00354037.

18. de Jong P. A central limit theorem for generalized multilinear forms. J. Multivariate Anal. 1990; 34(2):275–289. URL http://dx.doi.org/10.1016/0047-259X(90)90040-O.

19. Kerkyacharian G, Picard D. Estimating nonquadratic functionals of a density using Haar wavelets. Ann. Statist. 1996; 24(2):485–507.

20. Robins J, Tchetgen Tchetgen E, Li L, van der Vaart A. Semiparametric minimax rates. Electron. J. Stat. 2009; 3:1305–1321. URL http://dx.doi.org/10.1214/09-EJS479.

21. Newey WK, Hsieh F, Robins JM. Twicing kernels and a small bias property of semiparametric estimators. Econometrica. 2004; 72(3):947–962.

22. Daubechies, I. Ten lectures on wavelets, Vol. 61 of CBMS-NSF Regional Conference Series in Applied Mathematics; Society for Industrial and Applied Mathematics (SIAM); Philadelphia, PA. 1992.

23. DeVore, RA.; Lorentz, GG. Constructive approximation, Vol. 303 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Berlin: Springer-Verlag; 1993.

24. van der Vaart, AW. Asymptotic statistics, Vol. 3 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press; 1998.

25. Giné, E.; Latala, R.; Zinn, J. High dimensional probability, II (Seattle, WA 1999), Vol. 47 of Progr. Probab. Boston, MA: Birkhäuser Boston; 2000. Exponential and moment inequalities for *U*-statistics; p. 13-38.

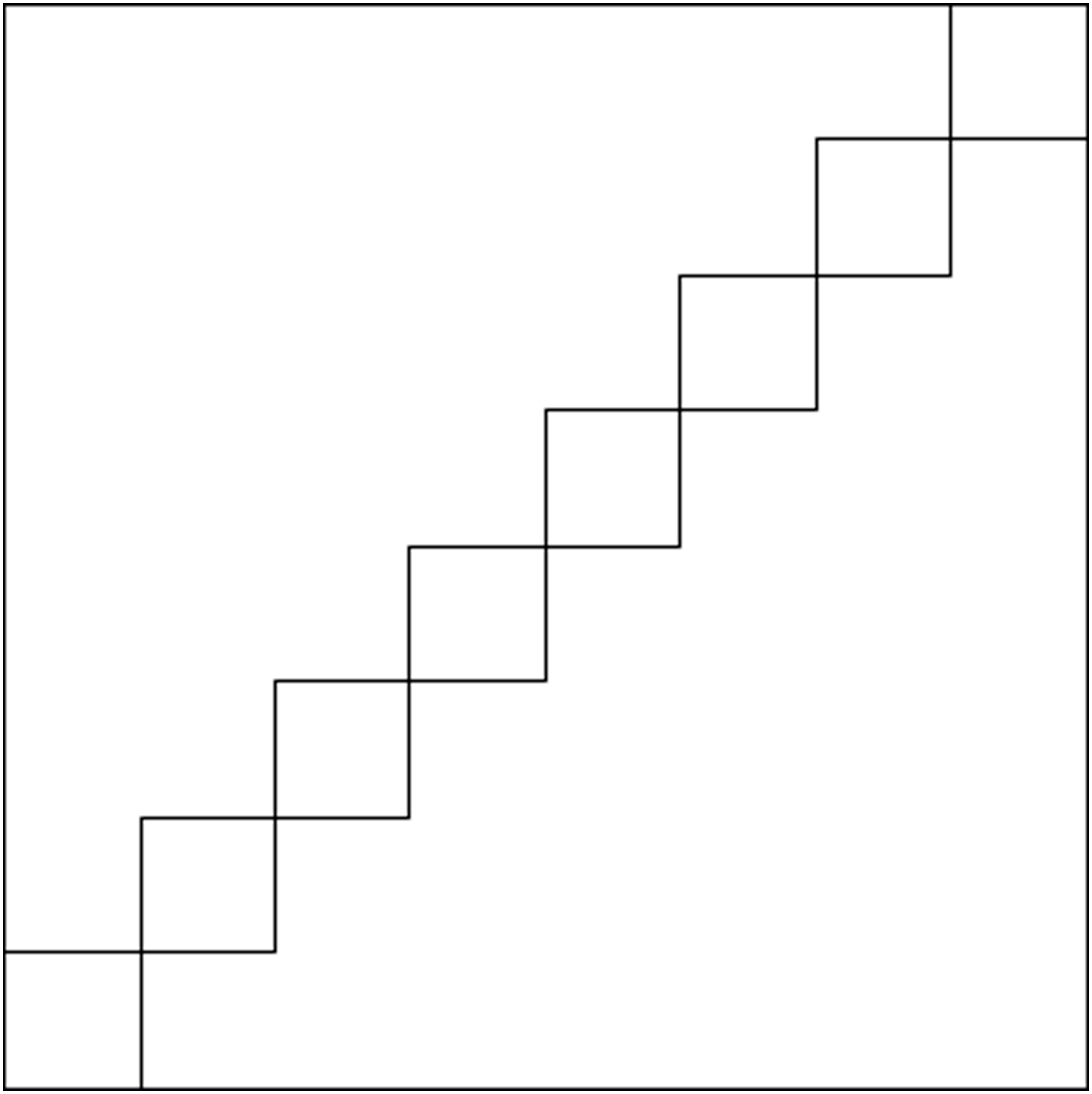

Figure 1.

The diagonal of $\mathscr{X} \times \mathscr{X}$ covered by the set $\cup_m(\mathscr{X}_{n,m} \times \mathscr{X}_{n,m})$.

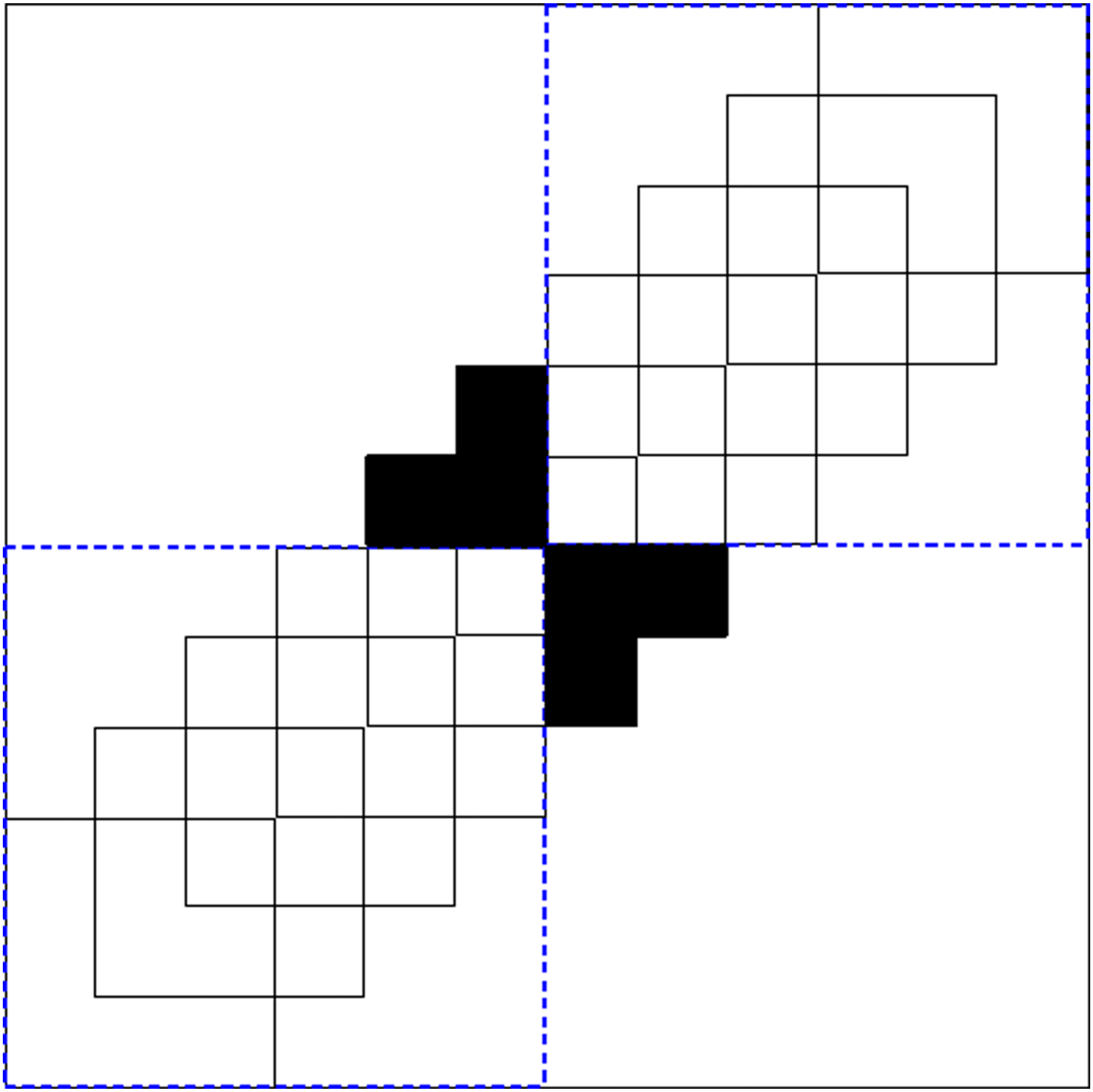

Figure 2.

The support cubes of the wavelets and the bigger cubes $\mathscr{X}_{n,m} \times \mathscr{X}_{n,m}$.

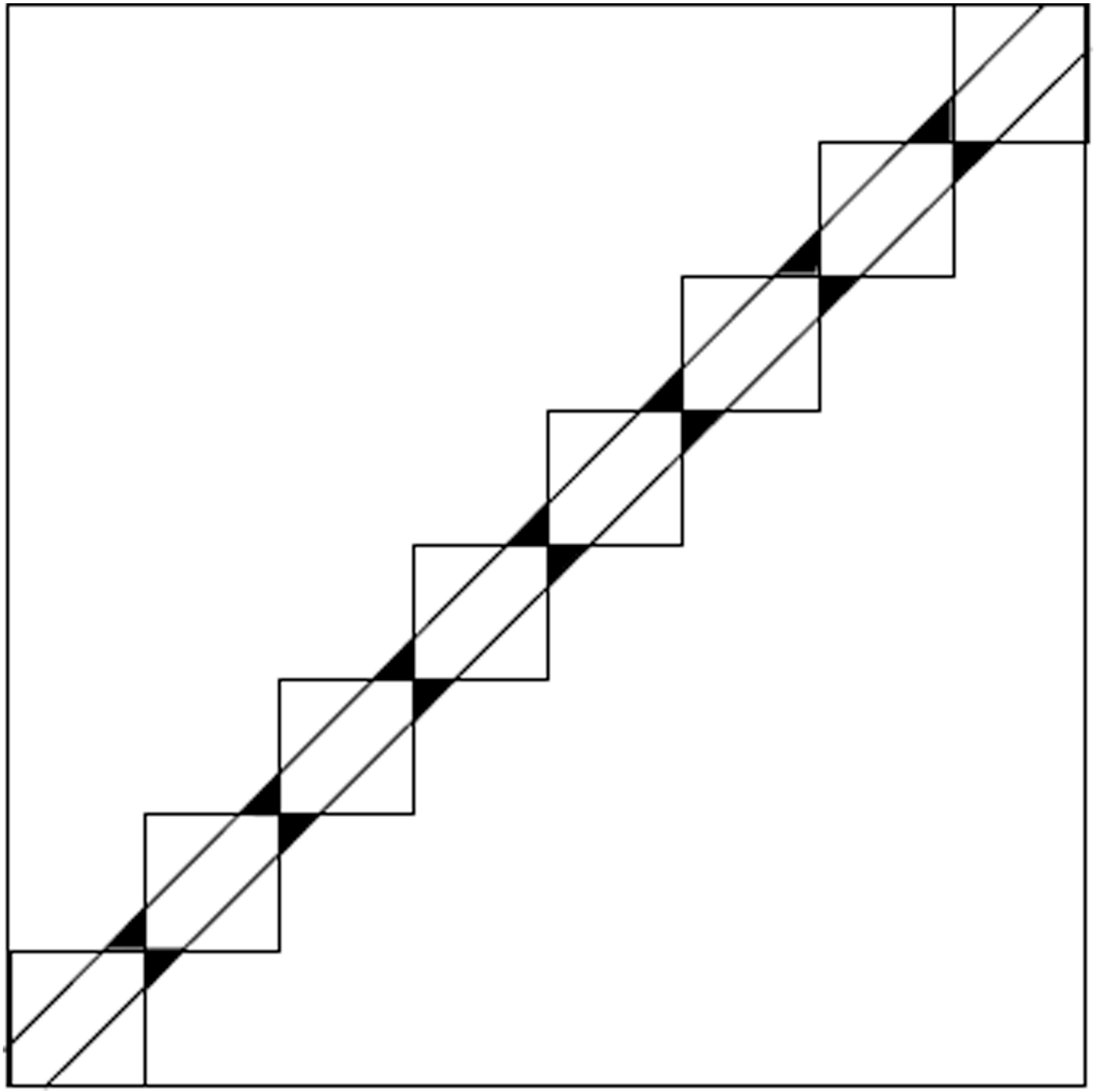

Figure 3.
The triangles used in the proofs of Theorems 4.2 and 4.3, and the sets $\mathscr{X}_{n,m} \times \mathscr{X}_{n,m}$.