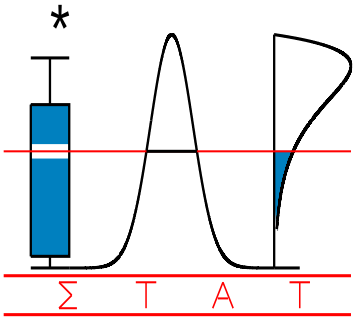
T E C H N I C A L

R E P O R T

0526

TWO-SAMPLE TESTS IN FUNCTIONAL DATA ANALYSIS STARTING FROM DISCRETE DATA

HALL, P. and I. VAN KEILEGOM



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

TWO-SAMPLE TESTS IN FUNCTIONAL DATA ANALYSIS STARTING FROM DISCRETE DATA

Peter Hall[1,2] and Ingrid Van Keilegom[2]

ABSTRACT. One of the ways in which functional data analysis differs from other areas of statistics is in the extent to which data are pre-processed prior to analysis. Functional data are invariably recorded discretely, although they are generally substantially smoothed as a prelude even to viewing by statisticians, let along further analysis. This has a potential to interfere with the performance of two-sample statistical tests, since the use of different tuning parameters for the smoothing step, or different observation times or subsample sizes (i.e. numbers of observations per curve), can mask the differences between distributions that a test is really trying to locate. In this paper, and in the context of two-sample tests, we take up this issue. Ways of pre-processing the data, so as to minimise the effects of smoothing, are suggested. We show theoretically and numerically that, by employing exactly the same tuning parameter (e.g. bandwidth) to produce each curve from its raw data, significant contributions to level inaccuracy and power loss can be avoided. Provided a common tuning parameter is used, it is often satisfactory to choose that parameter along conventional lines, as though the target was estimation of the continuous functions themselves, rather than testing hypotheses about them. Moreover, in this case, using a second-order smoother (such as local-linear methods), the subsample sizes can be almost as small as the square root of sample sizes before the effects of smoothing have any first-order impact on the results of a two-sample test.

KEYWORDS. Bandwidth, bootstrap, curve estimation, hypothesis testing, kernel, Cramér-von Mises test, local-linear methods, local-polynomial methods, nonparametric regression, smoothing.

SHORT TITLE. Tests for functional data.

[1] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia

[2] Institut de Statistique, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium

The research of Van Keilegom was supported by IAP research network grant nr. P5/24 of the Belgian government (Belgian Science Policy).

1. INTRODUCTION

Although, in functional data analysis, the data are treated as though they are in the form of curves, in practice they are invariably recorded discretely. They are subject to a pre-processing step, usually based on local-polynomial or spline methods, to transform them to the smooth curves to which the familiar algorithms of functional data analysis are applied. In many instances the pre-processing step is not of great importance. However, in the context of two-sample hypothesis testing it has the potential to significantly interfere with both power and level accuracy. Our aim in the present paper is to explore this issue, and suggest methods which allow the effects of smoothing to be minimised.

This problem has no real analogue in the context of two-sided tests applied to non-functional data. Although smoothing methods are sometimes used there, for example to produce alternatives to traditional two-sample hypothesis tests, they are not necessary for obtaining the data to which the tests are applied. Indeed, there is rightly a debate as to whether statistical smoothing should be used at all, in a conventional sense, when constructing two-sample hypothesis tests. Employing a small bandwidth (in theoretical terms, a bandwidth which converges to zero as sample size increases) can reduce statistical power unnecessarily, although from other viewpoints power can be increased. See, for example, the discussion by Ingster (1993), Fan (1994), Anderson et al. (1994), Fan (1998), Fan and Ullah (1999) and Li (1999).

By way of contrast, functional datasets are usually exchanged by researchers in post-processed form, after the application of a smoothing algorithm; that is seldom the case for data in related problems such as nonparametric regression. The widespread use of pre-process smoothing for functional data makes the effects of smoothing more insidious than usual, and strengthens motivation for understanding the impact that smoothing might have.

In the context of two-sample hypothesis testing, our main recommendation is appropriate in cases where the "subsample sizes" (that is, the numbers of points at which data are recorded for individual functions) do not vary widely. There we suggest that exactly the same tuning parameters be used to produce each curve from its raw data, for all subsamples in both datasets. For example, when using kernel-based methods this would mean using the same bandwidth in all cases; for splines it would mean using the same penalty, or the same knots. Such a choice ensures that, under the null hypothesis that the two samples of curves come from identical populations, the main effects of differing observation-time distributions and

differing subsample sizes (for the different curves) cancel. As a result, the effect that smoothing has on bias is an order of magnitude less than it would be if different bandwidths, tuned to the respective curves, were used. The latter approach can lead to both level inaccuracy and power loss for a two-sample test.

The constraint here that subsample sizes do not differ widely is actually quite mild. In asymptotic terms we ask only that the minimum size, divided by the maximum size, be bounded away from zero. If this condition is satisfied, and if the same bandwidth is used throughout; if the smoother is of second order, for example local-linear smoothing; and if the lesser of the two sample sizes is of smaller order than subsample size raised to the power $\frac{8}{5}$; then the effects of smoothing are negligible in a two-sample hypothesis test.

This condition is quite generous. It allows subsample size to be an order of magnitude smaller than sample size, without smoothing having a first-order effect on performance. By way of contrast, if different bandwidths are used for the different curves, tuned to the local regularities of those functions as well as to the respective subsample sizes, then problems can arise when the lesser of the two sample sizes is of smaller order than subsample size raised to the power $\frac{4}{5}$. This is a substantially more restrictive condition.

The fact that the common tuning parameter can be chosen by a conventional curve-estimation method, such as cross-validation or a plug-in rule, is an attractive feature of the proposal. New smoothing-parameter choice algorithms are not essential. Nevertheless, it can be shown that under more stringent assumptions a bandwidth of smaller size can be advantageous. See sections 3.3, 4 and 5.3 for discussion.

Recent work on two-sample hypothesis tests in nonparametric and semiparametric settings includes that of Louani (2000), Claeskens *et al.* (2003) and Cao and Van Keilegom (2005). Extensive discussion of methods and theory for functional data analysis is given by Ramsay and Silverman (1997, 2002). Recent contributions to hypothesis testing in this field include those of Fan and Lin (1998), Locantore *et al.* (1999), Spitzner *et al.* (2003), Cuevas *et al.* (2004) and Shen and Faraway (2004).

2. STATEMENT OF PROBLEM, AND METHODOLOGY

2.1. The data and the problem. We observe data

$$\begin{aligned} U_{ij} &= X_i(S_{ij}) + \delta_{ij}\,, \quad 1 \le i \le m\,, \quad 1 \le j \le m_i\,, \\ V_{ij} &= Y_i(T_{ij}) + \epsilon_{ij}\,, \quad 1 \le i \le n\,, \quad 1 \le j \le n_i\,, \end{aligned} \tag{2.1}$$

where $X_1, X_2, \ldots$ are identically distributed as X; $Y_1, Y_2, \ldots$ are identically distributed as Y; the δ_{ij}'s are identically distributed as δ; the ϵ_{ij}'s are identically dis-

tributed as ϵ; X and Y are both random functions, defined on the interval $\mathcal{I} = [0, 1]$; the observation errors, δ_{ij} and ϵ_{ij}, have zero means and uniformly bounded variances; the sequences of observation times, $S_{i1}, \ldots, S_{im_i}$ and $T_{i1}, \ldots, T_{in_i}$, are either regularly spaced on $\mathcal{I}$ or drawn randomly from a distribution (possibly different for each i, and also for S and T) having a density that is bounded away from zero on $\mathcal{I}$; and the quantities X_{i_1}, Y_{i_2}, $S_{i_1 j_1}$, $T_{i_2 j_2}$, δ_{i_1} and ϵ_{i_2}, for $1 \le i_1 \le m$, $1 \le j_1 \le m_{i_1}$, $1 \le i_2 \le n$ and $1 \le j_2 \le n_{i_2}$, are all totally independent.

Given the data at (2.1), we wish to test the null hypothesis, H_0, that the distributions of X and Y are identical. In many cases of practical interest, X and Y would be continuous with probability 1, and then H_0 would be characterised by the statement,

$$F_X(z) = F_Y(z) \quad \text{for all continuous functions } z\,,$$

where F_X and F_Y are the distributional functionals of X and Y, respectively:

$$F_X(z) = P\big\{X(t) \le z(t) \quad \text{for all} \quad t \in \mathcal{I}\big\}\,,$$
$$F_Y(z) = P\big\{Y(t) \le z(t) \quad \text{for all} \quad t \in \mathcal{I}\big\}\,.$$

In some problems, the time points S_{ij} and T_{ij} would be regularly spaced on grids of the same size. For example, they might represent the times of monthly observations of a process, such as the number of tons of a certain commodity exported by a particular country in month j of year i, in which case $m_i = n_i = 12$ for each i. (The differing lengths of different months could usually be ignored. In examples such as this we might wish first to correct the data for linear or periodic trends.) Here the assumption of independence might not be strictly appropriate, but the methods that we shall suggest will be approximately correct under conditions of weak dependence.

In other cases the values of m_i and n_i can vary from one index i to another, and in fact those quantities might be modelled as conditioned values of random variables. The observation times may also exhibit erratic variation. For example, in longitudinal data analysis, U_{ij} might represent a measurement of the condition of the ith type-X patient at the jth time in that patient's history. Since patients are seen only at times that suit them, then both the values of the observation times, and the number of those times, can vary significantly from patient to patient.

In general, unless we have additional knowledge about the distributions of X and Y (for example, both distributions are completely determined by finite parameter vectors), we cannot develop a theoretically consistent test unless the values of

m_i and n_i increase without bound as m and n increase. Therefore, the divergence of m_i and n_i will be a key assumption in our theoretical analysis.

2.2. Methodology for hypothesis testing. Our approach to testing H_0 is to compute estimators, $\widehat{X}_i$ and $\widehat{Y}_i$, of X_i and Y_i, by treating the problem as one of nonparametric regression and passing nonparametric smoothers through the datasets $(S_{i1}, U_{i1}), \ldots, (S_{im_i}, U_{im_i})$ and $(T_{i1}, V_{i1}), \ldots, (T_{in_i}, V_{in_i})$, respectively. Then, treating the functions $\widehat{X}_1, \ldots, \widehat{X}_m$ and $\widehat{Y}_1, \ldots, \widehat{Y}_n$ as independent and identically distributed observations of X and Y, respectively (under our assumptions they are at least independent), we construct a test of H_0.

For example, we might compute estimators $\widehat{F}_X$ and $\widehat{F}_Y$ of F_X and F_Y, respectively:

$$\widehat{F}_X(z) = \frac{1}{m} \sum_{i=1}^{m} I\big(\widehat{X}_i \le z\big)\,, \quad \widehat{F}_Y(z) = \frac{1}{n} \sum_{i=1}^{n} I\big(\widehat{Y}_i \le z\big)\,, \tag{2.2}$$

where the indicator $I(\widehat{X}_i \le z)$ is interpreted as $I\{\widehat{X}_i(t) \le z(t)$ for all $t \in \mathcal{I}\}$, and $I(\widehat{Y}_i \le z)$ is interpreted similarly. These quantities might be combined into a test statistic of Cramér-von Mises type, say

$$\widehat{T} = \int \big\{\widehat{F}_X(z) - \widehat{F}_Y(z)\big\}^2 \, \mu(dz)\,, \tag{2.3}$$

where μ denotes a probability measure on the space of continuous functions.

The integral here can be calculated by Monte Carlo simulation, for example as

$$\widehat{T}_N = \frac{1}{N} \sum_{i=1}^{N} \big\{\widehat{F}_X(M_i) - \widehat{F}_Y(M_i)\big\}^2\,, \tag{2.4}$$

where $M_1, \ldots, M_N$ are independent random functions with the distribution of M, say, for which $\mu(A) = P(M \in A)$ for each Borel set A in the space of continuous functions on $\mathcal{I}$. Of course, the M_i's are independent of $\widehat{F}_X$ and $\widehat{F}_Y$, and $\widehat{T}_N \to \widehat{T}$, with probability one conditional on the data, as $N \to \infty$.

2.3. Methodology for estimating X_i and Y_i. For brevity we shall confine attention to just one technique, local-polynomial methods, for computing $\widehat{X}_i$ and $\widehat{Y}_i$. (Results can be expected to be similar if one uses other conventional smoothers, for example splines.) Taking the degree of the polynomial to be odd, and estimating X_i, we compute the value $(\hat{a}_0, \ldots, \hat{a}_{2r+1})$ of the vector $(a_0, \ldots, a_{2r+1})$ that minimises

$$\sum_{j=1}^{m_i} \left\{U_{ij} - \sum_{k=0}^{2r-1} a_k \, (S_{ij} - t)^k \right\}^2 K\left(\frac{t - S_{ij}}{h_{Xi}}\right),$$

where $r \ge 1$ is an integer, h_{Xi} is a bandwidth, and K, the kernel function, is a bounded, symmetric, compactly supported probability density. Then, $\hat{a}_0 = \widehat{X}_i(t)$.

In the particular case $r = 1$ we obtain a local-linear estimator of $X_i(t)$,

$$\widehat{X}_i(t) = \frac{A_{i2}(t)\, B_{i0}(t) - A_{i1}(t)\, B_{i1}(t)}{A_{i0}(t)\, A_{i2}(t) - A_{i1}(t)^2}\,,$$

where

$$A_{ir}(t) = \frac{1}{m_i h_{Xi}} \sum_{j=1}^{m_i} \left(\frac{t - S_{ij}}{h_{Xi}}\right)^r K\left(\frac{t - S_{ij}}{h_{Xi}}\right),$$

$$B_{ir}(t) = \frac{1}{m_i h_{Xi}} \sum_{j=1}^{m_i} U_{ij} \left(\frac{t - S_{ij}}{h_{Xi}}\right)^r K\left(\frac{t - S_{ij}}{h_{Xi}}\right),$$

h_{Xi} denotes a bandwidth and K is a kernel function. The estimator $\widehat{Y}_i$ is constructed similarly. Local-linear methods have an advantage over higher-degree local-polynomial approaches in that they suffer significantly less from difficulties arising from singularity, or near-singularity, of estimators.

Treating X_i as a fixed function (that is, fixing i and conditioning on the stochastic process X_i); and assuming that X_i has $2(r + 1)$ bounded derivatives, and h_{Xi} is chosen of size $m_i^{-1/(2r+1)}$; the estimator $\widehat{X}_i$ converges to X_i at the mean-square optimal rate $m_i^{-2r/(2r+3)}$, as m_i increases. See, for example, Fan (1993), Fan and Gijbels (1996) and Ruppert and Wand (1994) for discussion of both practical implementation and theoretical issues.

2.4. Bandwidth choice. A number of potential bandwidth selectors are appropriate when all the subsample sizes m_i and n_j are similar and the bandwidths $h = h_{Xi} = h_{Yj}$ are identical. Theoretical justification for using a common bandwidth, when the goal is hypothesis testing rather than function estimation, will be given in section 3.

One approach to common bandwidth choice is to use a "favourite" method to compute an empirical bandwidth for each curve X_i and Y_j, and then take the average value to be the common bandwidth. Another technique, appropriate in the case of plug-in rules, is to use an average value of each of the components of a plug-in bandwidth selector, and assemble the average values, using the plug-in formula, to form the common bandwidth. A third approach, valid in the context of cross-validation, is to use a criterion which is the average of the cross-validatory criteria corresponding to the different curves. For each of these methods, "average" might be defined in a weighted sense, where the weights represent the respective subsample sizes.

2.5. Bootstrap calibration. Bootstrap calibration is along conventional lines, as follows. Having constructed smoothed estimators $\widehat{X}_i$ and $\widehat{Y}_i$ of the functions X_i and Y_i, respectively, pool them into the class

$$\mathcal{Z} = \{Z_1, \ldots, Z_{m+n}\} = \left\{\widehat{X}_1, \ldots, \widehat{X}_m\right\} \cup \left\{\widehat{Y}_1, \ldots, \widehat{Y}_n\right\}. \tag{2.5}$$

By sampling randomly, with replacement, from $\mathcal{Z}$, derive two independent resamples $\{X_1^*, \ldots, X_m^*\}$ and $\{Y_1^*, \ldots, Y_n^*\}$; compute

$$\bar{F}_X^*(z) = \frac{1}{m} \sum_{i=1}^{m} I\big(X_i^* \le z\big)\,, \quad \bar{F}_Y^*(z) = \frac{1}{n} \sum_{i=1}^{n} I\big(Y_i^* \le z\big)\,;$$

and finally, calculate

$$\overline{T}^* = \int \big\{\bar{F}_X^*(z) - \bar{F}_Y^*(z)\big\}^2 \, \mu(dz)\,. \tag{2.6}$$

Of course, $\bar{F}_X^*$ and $\bar{F}_Y^*$ are bootstrap versions of the actual empirical distribution functionals,

$$\bar{F}_X(z) = \frac{1}{m} \sum_{i=1}^{m} I\big(X_i \le z\big)\,, \quad \bar{F}_Y(z) = \frac{1}{n} \sum_{i=1}^{n} I\big(Y_i \le z\big)\,, \tag{2.7}$$

which we would have computed if we had access to the full-function data X_i and Y_i. Likewise, $\overline{T}$ is the ideal, but impractical, test statistic that we would have used if we had those data:

$$\overline{T} = \int \big\{\bar{F}_X(z) - \bar{F}_Y(z)\big\}^2 \, \mu(dz)\,. \tag{2.8}$$

Suppose the desired critical level for the test is $1 - \alpha$. By repeated resampling from $\mathcal{Z}$ we can compute, by Monte Carlo means, the critical point, $\hat{t}_\alpha$ say, given by

$$P\big(\overline{T}^* \ge \hat{t}_\alpha \;\big|\; \mathcal{Z}\big) = \alpha\,. \tag{2.9}$$

We reject the null hypothesis H_0, that there is no difference between the distributions of X and Y, if $\widehat{T} > \hat{t}_\alpha$.

At this point, some of the potential difficulties of two-sample hypothesis testing become clear. Regardless of how we smooth the data, the conditional expected value of $\bar{F}_X^* - \bar{F}_Y^*$, given $\mathcal{Z}$, is exactly zero. However, even if H_0 is correct, $E(\widehat{F}_X - \widehat{F}_Y)$ will generally not vanish, owing to the different natures of the datasets providing information about the functions X_i and Y_i (for example, different distributions of the sampling times), and the different ways we constructed $\widehat{X}_i$ and $\widehat{Y}_i$ from those data. Therefore, the test statistic $\widehat{T}$ suffers biases which are not reflected in its bootstrap form, $\overline{T}^*$, and which can lead to level inaccuracy and power loss for the test. Of course, this problem would vanish if we could use $\bar{F}_X - \bar{F}_Y$ in place of $\widehat{F}_X - \widehat{F}_Y$ for the test; that is, if we could employ $\overline{T}$ instead of $\widehat{T}$. But in practice, that is seldom possible.

2.6. Choice of the measure μ. Recall from (2.4) that our application of the measure μ, to calculate the statistic $\widehat{T}$, proceeds by simulating the stochastic process M of

which μ defines the distribution. Therefore it is necessary only to construct M. It is satisfactory to define M in terms of its Karhunen-Loève expansion,

$$M(t) = \mu(t) + \sum_{i=1}^{\infty} \zeta_i \, \phi_i(t) \, ,$$

where $\mu(t) = E(M(t))$, the functions ϕ_i form a complete orthonormal sequence on $\mathcal{I}$ and the random variables ζ_i are uncorrelated and have zero mean.

We shall take $\zeta_i = \theta_i \, \xi_i$, where $\theta_1 > \theta_2 > \ldots > 0$ are positive constants decreasing to zero, and the random variables ξ_i are independent and identically distributed with zero means and unit variances. The functions ϕ_i could be chosen empirically, to be the orthonormal functions corresponding to a principal component analysis of the dataset $\mathcal{Z}$ at (2.5). In this case they would form the sequence of eigenvectors of a linear operator, the kernel of which is the function

$$L(s,t) = \frac{1}{m+n} \sum_{i=1}^{m+n} \{Z_i(s) - \overline{Z}(s)\} \, \{Z_i(t) - \overline{Z}(t)\} \, ,$$

where $\overline{Z} = (m+n)^{-1} \sum_{i \le m+n} Z_i$. The constants θ_i would in that case be given by the square-root of the corresponding eigenvalues.

However, exposition is simpler if we take the ϕ_i's to be a familiar orthonormal sequence, such as the cosine sequence on $\mathcal{I}$:

$$\phi_1 \equiv 1 \, , \quad \phi_{i+1}(x) = 2^{1/2} \cos(i \pi x) \quad \text{for} \quad i \ge 1 \, .$$

(Recall that $\mathcal{I} = [0,1]$.) In particular, this makes it easier to describe the smoothness of the functions M. If $\theta_i = O(i^{-a})$ as $i \to \infty$, where $a > 3$; and if the common distribution of the ξ_i's is compactly supported; then there exist $C, c > 0$ such that, with probability 1, $|M''(s) - M''(t)| \le C \, |s-t|^c$. This is the level of regularity that our theoretical properties require of the distribution of M. Numerical work in section 4 argues that in practice it is adequate to take the process M to have a light-tailed distribution, such as the Gaussian; it is not necessary to assume the distribution is compactly supported.

3. THEORETICAL PROPERTIES

3.1. Overview. Section 3.2 gives a simple approximation, $\widetilde{T}$, to $\widehat{T}$; section 3.3 treats a centring term, D, which represents the main difference between $\widetilde{T}$ and $\overline{T}$; and section 3.4 describes asymmetric properties of $\overline{T}$. These steps in our argument culminate in section 3.5, which draws our main conclusions. In particular, section 3.5

combines results of sections 3.2–3.4 to give conditions under which the statistics $\widehat{T}$ and $\overline{T}$ have the same asymptotic properties. It also shows that the practical statistic $\widehat{T}$ leads to tests with the same asymptotic level as its idealised counterpart $\overline{T}$, and to the same asymptotic power against local alternatives.

3.2. Main approximation property. Let $\bar{F}_X$, $\bar{F}_Y$ and $\overline{T}$ be the quantities defined at (2.7) and (2.8). In section 2.5 we discussed the fact that the bootstrap form of $\overline{T}$ does not adequately reflect differences between the functionals $\widehat{F}_X$ and $\widehat{F}_Y$ on which the practicable test statistic $\widehat{T}$ is based. Our main result in the present section shows that the main aspects of these potential problems can be encapsulated quite simply in terms of a difference between expected values.

In particular, under very mild conditions, difficulties associated with stochastic variability of $\widehat{F}_X - \widehat{F}_Y$ are negligibly small; and the impact of the difference,

$$D = E(\bar{F}_X - \bar{F}_Y) - E(\widehat{F}_X - \widehat{F}_Y)\,,$$

between the mean of $\widehat{F}_X - \widehat{F}_Y$ and the mean of $\bar{F}_X - \bar{F}_Y$, can be summarised very simply. Theorem 1 below shows that $\widehat{T}$ is closely approximated by

$$\widetilde{T} = \int \bigl\{\bar{F}_X(z) - \bar{F}_Y(z) - D(z)\bigr\}^2\,\mu(dz)\,. \tag{3.1}$$

The sets of assumptions A.1 and A.2, used for Theorems 1 and 2 respectively, will be collected together in section 5.

Theorem 1. *If the measure μ has no atoms, and assumptions A.1 hold, then*

$$\bigl|\widehat{T}^{1/2} - \widetilde{T}^{1/2}\bigr| = o_p\bigl(m^{-1/2} + n^{-1/2}\bigr)\,. \tag{3.2}$$

To interpret this result, note that, under the null hypothesis that the distributions of X and Y are identical, $\overline{T}$ is of size $m^{-1} + n^{-1}$. (Further discussion of this point is given in section 3.4.) The quantity, $D(z)$, that distinguishes $\widetilde{T}$ from $\overline{T}$, can only increase this size. Result (3.2) asserts that the difference between $\widehat{T}^{1/2}$ and $\widetilde{T}^{1/2}$ is actually of smaller order than the asymptotic sizes of either $\widehat{T}^{1/2}$ or $\widetilde{T}^{1/2}$, and so $D(z)$ captures all the main issues that will affect the power, and level accuracy, of the statistic $\widehat{T}$, compared with $\overline{T}$.

3.3. Properties of $D(z)$. First we summarise properties of $D(z)$ when $(2r-1)$st degree local-polynomial estimators are employed. Using arguments similar to those we shall give in section 6.2, it may be shown that for functions z that are sufficiently smooth, and for each $\eta > 0$,

$$P\bigl(\widehat{X}_i \le z\bigr) = P(X_i \le z) + O\bigl\{h_{Xi}^{2r-\eta} + (m_i h_{Xi})^{\eta-1}\bigr\}\,. \tag{3.3}$$

This result, and its analogue for the processes $\widehat{Y}_i$ and Y_i, lead to the following result: Under the null hypothesis, and for each $\eta > 0$,

$$D(z) = O\biggl[\frac{1}{m^{1-\eta}} \sum_{i=1}^{m} \bigl\{h_{Xi}^{2r} + (m_i h_{Xi})^{-1}\bigr\} + \frac{1}{n^{1-\eta}} \sum_{i=1}^{n} \bigl\{h_{Yi}^{2r} + (n_i h_{Yi})^{-1}\bigr\}\biggr]. \quad (3.4)$$

Neglecting the effects of η; assuming that the subsample sizes m_i and n_i are close to a common value, ν say; and supposing that the bandwidths, h_{Xi} and h_{Yi}, are also taken of similar sizes; (3.4) suggests allowing those bandwidths to be of size $\nu^{-1/(2r+1)}$, in which case $D(z) = O(\nu^{-2r/(2r+1)})$.

While this approach is of interest, the extent of reduction in subsampling effects, under H_0, can often be bettered by taking the bandwidths $h = h_{Xi} = h_{Yj}$ to be identical, for each $1 \le i \le m$ and $1 \le j \le n$. That technique allows the quantities that contribute the dominant bias terms, involving h_{Xi}^{2r} and h_{Yi}^{2r}, in (3.3) and its analogue for the Y-sample, to cancel perfectly. That reduces the bias contribution, from the $2r$th to the $2(r+1)$st power of bandwidth.

Using identical bandwidths makes local-linear methods, which correspond to taking $r = 1$ in the formulae above, particularly attractive, for at least two reasons. Firstly, the contribution of bias is reduced to that which would arise through fitting third-degree, rather than first-degree, polynomials in the case of non-identical bandwidths, yet the greater robustness of first-degree fitting is retained. Secondly, the appropriate bandwidth is now close to $\nu^{-1/5}$, the conventional bandwidth size for estimating the functions X_i and Y_i as functions in their own right. This suggests that tried-and-tested bandwidth selectors, such as those discussed in section 2.4, could be used.

The mathematical property behind the common-bandwidth recommendation is the following more detailed version of (3.3), which for simplicity we give only in the local-linear case, i.e. $r = 1$. If $h = h_{Xi}$ for each i, and h is of size $\nu^{-1/5}$, or larger, then for each $\eta > 0$,

$$P\bigl(\widehat{X}_i \le z\bigr) = P\bigl(X_i + \tfrac{1}{2}\,\kappa_2\,h^2\,X_i'' \le z\bigr) + O\bigl\{h^{4-\eta} + (\nu h)^{\eta-1}\bigr\}\,, \quad (3.5)$$

uniformly in smooth functions z, where $\kappa_2 = \int u^2\,K(u)\,du$. If H_0 holds, and we use the same bandwidth, h, for the Y data as well as for the X data, then

$$P\bigl(X_i + \tfrac{1}{2}\,\kappa_2\,h^2\,X_i'' \le z\bigr) = P\bigl(Y_i + \tfrac{1}{2}\,\kappa_2\,h^2\,Y_i'' \le z\bigr)\,, \quad (3.6)$$

for each i and each function z. Therefore (3.5) implies that, under H_0, and assuming that the subsample sizes are all similar to ν,

$$P\bigl(\widehat{X}_i \le z\bigr) - P\bigl(\widehat{Y}_j \le z\bigr) = O\bigl\{h^{4-\eta} + (\nu h)^{\eta-1}\bigr\}\,,$$

for $1 \le i \le m$ and $1 \le j \le n$.

Optimising the right-hand side of (3.6) with respect to h suggests using a relatively conventional bandwidth selector of size $\nu^{-1/5}$. A small value of the right-hand side of (3.6) implies that D is close to zero, which in turn ensures that $\widehat{T}$ (which is close to $\widetilde{T}$, as shown in section 3.2) is well approximated by $\overline{T}$. That result implies that the bootstrap test is unlikely to reject H_0 simply because of poor choice of smoothing parameters; see section 5.2 for further discussion.

We conclude with a concise statement of (3.5). Given sequences a_m and b_m of positive numbers, write $a_m \asymp b_m$ to denote that a_m/b_m is bounded away from zero and infinity as $n \to \infty$. The reader is referred to section 5.2 for a statement of assumptions A.2 for Theorem 2. These include the condition that all bandwidths $h = h_{Xi}$ are identical, and $h \asymp \nu^{-1/q}$ where $3 < q < \infty$.

Theorem 2. *If $r = 1$ and assumptions A.2 hold then, for each $\eta > 0$,*

$$P\big(\widehat{X}_i \le z\big) - P\big(X_i + \tfrac{1}{2}\,\kappa_2\,h^2\,X_i'' \le z\big) = \begin{cases} O(\nu^{\eta-(q-1)^2/4q}) & \text{if } 3 < q \le 5 \\ O(\nu^{\eta-4/q}) & \text{if } q > 5\,, \end{cases} \tag{3.7}$$

where the "big oh" terms are of the stated orders uniformly in functions z that have two Hölder-continuous derivatives, and in $1 \le i \le m$.

3.4. Asymptotic distribution of $\overline{T}$, and power. If m and n vary in such a way that

$$m/n \to \rho \in (0, \infty) \quad \text{as} \quad m \to \infty\,, \tag{3.8}$$

and if, as prescribed by H_0, the distributions of X and Y are identical, then $\overline{T}$ satisfies

$$m\,\overline{T} \to \zeta \equiv \int \big\{\zeta_X(z) - \rho^{1/2}\,\zeta_Y(z)\big\}^2\,\mu(dz)\,, \tag{3.9}$$

where the convergence is in distribution and $\zeta_X(z)$ and $\zeta_Y(z)$ are independent Gaussian processes with zero means and the same covariance structures as the indicator processes $I(X \le z)$ and $I(Y \le z)$, respectively. In particular, the covariance of $\zeta_X(z_1)$ and $\zeta_X(z_2)$ equals $F_X(z_1 \wedge z_2) - F_X(z_1)\,F_X(z_2)$.

It follows directly from (3.9) that the asymptotic value of the critical point for an α-level test of H_0, based on $\overline{T}$, is the quantity u_α such that $P(\zeta > u_\alpha) = \alpha$. Analogously, the critical point $\hat{t}_\alpha$ for the bootstrap statistic $\overline{T}^*$ (see (2.6) and (2.9)) converges, after an obvious rescaling, to u_α as sample size increases: under H_0,

$$P(m\,\overline{T} > u_\alpha) \to \alpha\,, \quad m\,\hat{t}_\alpha \to u_\alpha\,, \tag{3.10}$$

where the second convergence is in probability. Of course, these are conventional properties of bootstrap approximations. In section 3.5 we shall discuss conditions

that are sufficient for the practical test statistic $\widehat{T}$, rather than its ideal form $\overline{T}$, to have asymptotically correct level; see (3.17).

Power properties under local alternatives are also readily derived. In particular, if F_Y is fixed and

$$F_X(z) = F_Y(z) + m^{-1/2}\, c\, \delta(z)\,, \tag{3.11}$$

where δ is a fixed function and c is a constant, then with convergence interpreted in distribution,

$$m\,\overline{T} \to \int \bigl\{\zeta_{Y1}(z) + c\,\delta(z) - \rho^{1/2}\,\zeta_{Y2}(z)\bigr\}^2\,\mu(dz)\,, \tag{3.12}$$

of which (3.9) is a special case. In (3.12), ζ_{Y1} and ζ_{Y2} are independent Gaussian processes each with zero mean and the covariance structure of ζ_Y.

From this result and the second part of (3.10) it is immediate that, provided δ is not almost surely zero with respect to μ measure, a test based on the ideal statistic $\overline{T}$, but using the bootstrap critical point $\hat{t}_\alpha$, is able to detect departures proportional to $m^{-1/2}\,\delta$:

$$\lim_{c\to\infty}\ \liminf_{m\to\infty}\ P_c\bigl(\overline{T} > \hat{t}_\alpha\bigr) = 1\,, \tag{3.13}$$

where P_c denotes probability under the model where F_Y is fixed and F_X is given by (3.11). In section 3.5 we shall note that if we use a common bandwidth, and if the subsample sizes are not too much smaller than the sample sizes m and n, then the same result holds true for the practicable statistic $\widehat{T}$.

Proofs of (3.9), (3.10) and (3.12) are straightforward. They do not require convergence of function-indexed empirical processes to Gaussian processes, and proceed instead via low-dimensional approximations to those empirical processes.

3.5. Sufficient conditions for $\widehat{T}$ and $\overline{T}$ to have identical asymptotic distributions under H_0. Assume (3.8) and the conditions of Theorem 2 for both the X and Y populations, and in particular that all the subsample sizes m_i and n_i are of the same order, in the sense that

$$\nu \asymp \min_{1\le i\le m} m_i \asymp \max_{1\le i\le m} m_i \asymp \min_{1\le i\le n} n_i \asymp \max_{1\le i\le n} n_i \tag{3.14}$$

as $m, n \to \infty$. Take h to be of conventional size, $h \asymp \nu^{-1/5}$. Then Theorem 2, and its analogue for the Y sample, imply that under H_0,

$$D(z) = O\bigl(\nu^{\eta-4/5}\bigr)\,, \tag{3.15}$$

uniformly in functions z with two Hölder-continuous derivatives, for each $\eta > 0$.

Theorem 1 implies that, in order for the practical statistic $\widehat{T}$, and its "ideal" version $\overline{T}$, to have identical asymptotic distributions, it is necessary only that $D(z)$ be of smaller order than the stochastic error of $\bar{F}_X - \bar{F}_Y$. Equivalently, if (3.8) holds, $D(z)$ should be of smaller order than $m^{-1/2}$, uniformly in functions z with two Hölder-continuous derivatives. For that to be true it is sufficient, in view of (3.15), that

$$m = O\big(\nu^{8/5-\eta}\big) \tag{3.16}$$

for some $\eta > 0$.

It follows from (3.10) that, provided (3.16) holds and the null hypothesis is correct,

$$P\big(m\,\widehat{T} > u_\alpha\big) \to \alpha\,. \tag{3.17}$$

This is the analogue of the first part of (3.10), for the practicable statistic $\widehat{T}$ rather than its ideal form $\overline{T}$. Similarly, result (3.13) holds, for $\widehat{T}$ rather than $\overline{T}$, if a common bandwidth is used and the subsample sizes satisfy (3.14). This confirms the ability of the practicable test, based on $\widehat{T}$, to detect semiparametric departures from the null hypothesis.

Condition (3.16) is surprisingly mild. It asserts that, in order for the effects of estimating X_i and Y_i to be negligible, it is sufficient that the subsample sizes m_i and n_i be of larger order than the $\frac{5}{8}$th root of the smaller of the two sample sizes, m and n.

4. NUMERICAL PROPERTIES

Suppose that S_{ij} $(1 \le i \le m; 1 \le j \le m_i)$ are independent and identically distributed (i.i.d.) and have a uniform distribution on $[0,1]$ and that T_{ij} $(1 \le i \le n; 1 \le j \le n_i)$ are i.i.d. with density given by $2 - b + 2(b-1)t$ for $0 \le t \le 1$ and $0 < b < 2$. Note that this density reduces to the uniform density when $b = 1$. We take $b = 1.2$. The errors δ_{ij} and ϵ_{ij} are independent and come from a normal distribution with mean zero and standard deviation $\sigma = 0.1$ and 0.3 respectively. Suppose that $X_1, \ldots, X_m$ are identically distributed as X, where $X(t) = \sum_{k \ge 1} c_k\, N_{kX}\, \psi_k(t)$, $c_k = e^{-k/2}$, N_{kX} $(k \ge 1)$ are i.i.d. standard normal random variables, and $\psi_k(t) = 2^{1/2} \sin\{(k-1)\pi t\}$ $(k > 1)$ and $\psi_1 \equiv 1$ are orthonormal basis functions. Similarly, let $Y_1, \ldots, Y_n$ be identically distributed as Y, where

$$Y(t) = \sum_{k=1}^{\infty} c_k\, N_{kY1}\, \psi_k(t) + a \sum_{k=1}^{\infty} a_k\, N_{kY2}\, \psi_k^*(t)\,,$$

N_{kY1} and N_{kY2} are i.i.d. standard normal variables, $a \ge 0$ controls the deviation

from the null model ($a = 0$ under H_0), $a_k = k^{-2}$, and

$$\psi_k^*(t) = \begin{cases} 1 & \text{if } k = 1 \\ 2^{1/2}\sin\{(k-1)\pi(2t-1)\} & \text{if } k \text{ is odd and } k > 1 \\ 2^{1/2}\cos\{(k-1)\pi(2t-1)\} & \text{if } k \text{ is even} \end{cases}$$

are orthonormal basis functions. For practical reasons, we truncate the infinite sum in the definition of $X(t)$ and $Y(t)$ at $k = 15$. Define $U_{ij} = X_i(S_{ij}) + \delta_{ij}$ $(1 \le i \le m; 1 \le j \le m_i)$ and $V_{ij} = Y_i(T_{ij}) + \epsilon_{ij}$ $(1 \le i \le n; 1 \le j \le n_i)$. Finally, $M_1, \ldots, M_N$ are independent and have the same distribution as M, where $M(t) = \sum_{k\ge1} c_k N_{kZ}\,\phi_k(t)$, N_{kZ} are i.i.d. standard normal variables and $\phi_k(t) = 2^{1/2}\cos\{(k-1)\pi t\}$ $(k > 1)$ and $\phi_1 \equiv 1$ are orthonormal functions. We take $N = 50$ and truncate the infinite sum after 15 terms. The simulation results are based on 500 samples, and the critical values of the test are obtained from 250 bootstrap samples. The functions $X(t)$ and $Y(t)$ are estimated by means of local-linear smoothing. The bandwidth is selected by minimising the following cross-validation type criterion:

$$h = \underset{h}{\text{argmin}} \left[(m_i m)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{m_i} \{U_{ij} - \widehat{X}_i(S_{ij})\}^2 + (n_i n)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \{V_{ij} - \widehat{Y}_i(T_{ij})\}^2 \right].$$

The function K is the biquadratic kernel, $K(u) = (15/16)(1-u^2)^2 I(|u| \le 1)$.

The results for $m = n = 15$, 25, 50 and $m_1 = n_1 = 20$ and 100 are summarised in Figure 1. The level of significance is $\alpha = 0.05$ and is indicated in the figure. The graphs show that under the null hypothesis the level is well respected and the power increases for larger values of m, n, m_1, n_1 and a. The value of the subsample sizes m_1 and n_1 has limited impact on the power, whereas this is clearly not the case for the sample sizes m and n. Other settings that are not reported here (equal variances in the two populations, bigger sample and subsample sizes, ...) show similar behavior for the power curves.

In section 3.3 we explained why it is recommended to take $h_{X_i} = h_{Y_j} = h$. We shall now verify in a small simulation study that identical bandwidths indeed lead to higher power. Consider the same model as above, except that now the standard deviation of the errors δ_{ij} and ϵ_{ij} is 0.2 and 0.5 respectively. Take $m = n = 15$, 25, and 50, $m_1 = 20$ and $n_1 = 100$. Figure 2 shows the power curves for this model. The rejection probabilities are obtained using either identical bandwidths (estimated by means of the above cross-validation procedure) or using different bandwidths for each sample (estimated by means of a cross-validation procedure for each sample). The graph suggests that under H_0 the empirical level is close to the nominal level in both cases. The power is however considerably lower when different bandwidths are used than when the same bandwidth is used for both samples.

Finally, we apply the proposed method to temperature data over the course of a year, taken from 35 weather stations across Canada. The raw data curves are shown in Figure 3. Each data point in the graph represents the mean temperature recorded by a weather station for the entire month, collected over 30 years. The data are taken from Ramsay and Silverman (2002). A detailed description and analysis of these data can be found there. Here, we are interested in whether weather patterns in three different regions of Canada (namely the Atlantic, Pacific and Continental region) are equal. The sample sizes for the three regions are 14, 5 and 13 respectively and the subsample size is 12 for all curves (one observation per month).

First we obtain the optimal bandwidth for each combination of two regions by means of the above cross-validation procedure. In each case we find $h = 3.0$. Next, we need to determine an appropriate measure μ or equivalently an appropriate process M. For this, we follow the procedure described in section 2.6 and let $\phi_i(t)$ and θ_i^2 be the eigenfunctions and eigenvalues corresponding to a principal component analysis (PCA) of the dataset, and truncate the infinite sum at four terms. The variables ξ_i $(1 \leq i \leq 4)$ are taken as independent standard normal variables, and the mean function $\mu(t)$ is estimated empirically. The estimation of these functions is carried out by using the PCA routines available on J. Ramsay's homepage (http://ego.psych.mcgill.ca/misc/fda/). Next, based on the so-obtained process M, we calculate the test statistics for each comparison and approximate the corresponding p-values from 1000 resamples. The p-values are 0.394 for the atlantic region compared with the pacific region, 0.003 for atlantic versus continental, and 0.070 for pacific versus continental.

5. ASSUMPTIONS FOR SECTION 3

5.1. Conditions for Theorem 1. The assumptions are the following, comprising A.1: for some $\eta > 0$,

$$\min\Big(\min_{1\leq i\leq m} m_i, \min_{1\leq i\leq n} n_i\Big) \to \infty\,, \tag{5.1}$$

$$\max_{1\leq i\leq m} h_{Xi} + \max_{1\leq i\leq n} h_{Yi} \to 0\,, \quad \min_{1\leq i\leq m} \big(m_i^{1-\eta}\, h_{Xi}\big) + \min_{1\leq i\leq n} \big(n_i^{1-\eta}\, h_{Yi}\big) \to \infty\,, \tag{5.2}$$

$$K \text{ is a bounded, symmetric, compactly-supported probability density,} \tag{5.3}$$

$$\text{the observation times } S_{ij} \text{ and } T_{ij} \text{ are independent random variables, identically distributed for each } i, \text{ and with densities that are bounded away from zero uniformly in } i \text{ and in population type.} \tag{5.4}$$

Assumption (5.1) asks that the subsample sizes m_i and n_i diverge to infinity in a uniform manner as m and n grow. This is a particularly mild condition; we

do not expect a subsample to provide asymptotically reliable information about the corresponding random function, X_i or Y_i, unless it is large. The first part of (5.2) asks that the bandwidths h_{Xi} and h_{Yi} are uniformly small. Again this is a minimal condition, since bandwidths that converge to zero are necessary for consistent estimation of X_i and Y_i. Likewise, the second part of (5.2) is only a little stronger than the assumption that the variances of the estimators of X_i and Y_i decrease uniformly to zero.

Assumptions (5.3) and (5.4) are conventional. The latter is tailored to the case of random design, as too is (5.9) below; in the event of regularly sapced design, both can be replaced by simpler conditions.

5.2. Conditions for Theorem 2. We shall need the following notation. Given $C > 0$ and $r \in (1, 2]$, write $\mathcal{C}_r(C)$ for the class of differentiable functions, z, on $\mathcal{I}$ for which: (a) $\|z'\|_\infty \le C$; (b) if $1 < r < 2$, $|z'(s) - z'(t)| \le C\,|s-t|^{r-1}$ for all $s, t \in \mathcal{I}$; and (c) if $r = 2$, z has two bounded derivatives and $\|z''\|_\infty \le C$. Given $d > 0$, put $W_d = X + d\,X''$, where X denotes a generic X_i, and let $f_{W_d(s)}$ denote the probability density of $W_d(s)$.

The assumptions leading to Theorem 2 are the following, comprising A.2:

$$\text{the kernel } K \text{ is a symmetric, compactly-supported probability density with two Hölder-continuous derivatives}\,; \tag{5.5}$$

$$\nu \asymp \min_{1 \le i \le m} m_i \asymp \max_{1 \le i \le m} m_i \quad \text{as} \quad m \to \infty\,; \tag{5.6}$$

$$\text{for some } 0 < \eta < 1 \text{ and all sufficiently large } m,\ m^\eta \le \nu \le m^{1/\eta}\,; \tag{5.7}$$

the common bandwidth, $h = h_{Xi}$, satisfies,

$$\text{for some } \eta > 0\,, \quad h \asymp \nu^{-1/q} \quad \text{where} \quad 3 < q < \infty\,; \tag{5.8}$$

the respective densities f_i of the observation times S_{ij} satisfy,

$$\sup_{1 \le i < \infty} \sup_{t \in \mathcal{I}} |f_i''(t)| < \infty\,, \quad \inf_{1 \le i < \infty} \inf_{t \in \mathcal{I}} f_i(t) > 0\,; \tag{5.9}$$

the random function X has four bounded derivatives, with

$$E\big(|\delta|^s\big) < \infty\,, \quad E\Big\{ \sup_{t \in \mathcal{I}} \max_{r=1,\ldots,4} \big|X^{(r)}(t)\big|^s \Big\} < \infty \quad \text{for each} \quad s > 0\,; \tag{5.10}$$

$$\sup_{|d| \le c} \sup_{s \in \mathcal{I}} \big\| f_{W_d(s)} \big\|_\infty < \infty \quad \text{for some} \quad c > 0\,; \tag{5.11}$$

for $1 \le p \le 2$ and $c > 0$, and for each $C > 0$,

$$P(W_d \le z + y) = P(W_d \le z) + \int_{\mathcal{I}} y(s)\, P\big\{W_d \le z \;\big|\; W_d(s) = z(s)\big\} \times f_{W_d(s)}\{z(s)\}\, ds + O\big(\|y\|_\infty^p\big)\,, \tag{5.12}$$

uniformly in $z \in \mathcal{C}_2(C)$, in $y \in \mathcal{C}_p(C)$, and in d satisfying $|d| \le c$.

Conditions (5.5)–(5.11) are conventional and self-explanatory. Condition (5.12) can be derived formally by taking the limit, as $r \to \infty$, of more familiar formulae for the probability $P\{W(s_i) \le z(s_i) \text{ for } 1 \le i \le r\}$, where $W = W_d$. To obtain (5.12) in detail, in the case $p = 2$ and under the assumption that W'' is well-defined and continuous, we can first calculate the probabilities conditional on W'', and argue as follows.

Note that $W(t) = V_0 + t\,V_1 + u(t)$, where $u(t) = t^2 \int_{0<s<1} W''(st)\,(1-s)\,ds$ and $V_j = W^{(j)}(0)$. Of course, u is held fixed if we condition on W''. Write Q for probability measure conditional on W'', define $v = z - u$, and let $\mathcal{A}(v, y)$ denote the set of pairs (v_0, v_1) such that $v_0 + t\,v_1 \le v(t) + y(t)$ for all $t \in \mathcal{I}$. Write $\mathcal{B}_1(v, y)$ [respectively, $\mathcal{B}_2$] for the set of (v_0, v_1) such that for some t, $y(t) > 0$ and $v(t) < v_0 + t\,v_1 \le v(t) + y(t)$ [$y(t) < 0$ and $v(t) + y(t) < v_0 + t\,v_1 \le v(t)$]. Define $\mathcal{D}_1(v, y) = \mathcal{A}(v, y) \cap \mathcal{B}_1(v, y)$ and $\mathcal{D}_2(v, y) = \mathcal{A}(v, 0) \cap \mathcal{B}_2(v, y)$. Then,

$$\begin{aligned} &Q(W \le z + y) - Q(W \le z) \\ &\qquad = Q\big\{(V_0, V_1) \in \mathcal{D}_1(v, y)\big\} - Q\big\{(V_0, V_1) \in \mathcal{D}_2(v, y)\big\}\,. \end{aligned} \tag{5.13}$$

The regions $\mathcal{D}_1(v, y)$ and $\mathcal{D}_2(v, y)$ shrink to sets of measure zero, their probabilities decreasing at rate $O(\|y\|_\infty)$, as $\|y\|_\infty \to 0$. In particular, both the probabilities on the right-hand side of (5.13) converge to zero. The fact that the probabilities are defined in terms of the vector (V_0, V_1) makes it possible to verify, working only in terms of the bivariate distribution of (V_0, V_1) conditional on W'', the formula that can be obtained formally by discrete approximation:

$$Q(W \le z + y) = Q(W \le z) + \int_{\mathcal{I}} y(s)\, Q\big\{W \le z \;\big|\; W(s) = z(s)\big\} \times f_{W(s)\,|\,W''}\{z(s)\}\, ds + O\big(\|y\|_\infty^2\big)\,.$$

Taking expectations in the distribution of W'', this can be used to derive (5.12) for classes of stochastic processes X. Examples include polynomial functions of multivariate Gaussian processes with sufficiently smooth sample paths.

5.3. Error reduction in more specialised cases. Writing $W = W_d$ and letting $f_{W(s_1),\ldots,W(s_k)}$ denote the joint density of $W(s_1), \ldots, W(s_k)$, the expansion at (5.12)

can sometimes (e.g. in the case of sufficiently smooth Gaussian processes) be taken to greater length:

$$\begin{aligned}
&P(W \le z + y) \\
&\quad = P(W \le z) + \int_{\mathcal{I}} y(s)\, P\{W \le z \,\big|\, W(s) = z(s)\}\, f_{W(s)}\{z(s)\}\, ds \\
&\qquad + \tfrac{1}{2} \iint_{\mathcal{I}^2} y(s_1)\, y(s_2)\, P\Big\{W \le z \,\Big|\, W(s_1) = z(s_1), W(s_2) = z(s_2)\Big\} \\
&\qquad \times f_{W(s_1),W(s_2)}\{z(s_1), z(s_2)\}\, ds_1\, ds_2 \\
&\qquad + \tfrac{1}{6} \iiint_{\mathcal{I}^3} y(s_1)\, y(s_2)\, y(s_3)\, P\Big\{W \le z \,\Big|\, W(s_1) = z(s_1), \\
&\qquad\qquad\qquad\qquad W(s_2) = z(s_2), W(s_3) = z(s_3)\Big\} \\
&\qquad \times f_{W(s_1),W(s_2),W(s_3)}\{z(s_1), z(s_2), z(s_3)\}\, ds_1\, ds_2\, ds_3 + \ldots\,, \qquad (5.14)
\end{aligned}$$

where, if the expansion were stopped after terms of kth power in y, the remainder would be of order $\|y\|_\infty^{k+1}$. As in the proof of Theorem 2 we shall give in section 6.2, let us replace y by a zero-mean stochastic process, Δ say, that is independent of W and for which (a) $\Delta(s_1)$, $\Delta(s_2)$ and $\Delta(s_3)$ are independent if each pair (s_i, s_j) with $i \ne j$ satisfies $|s_i - s_j| > C\,h$, and (b) $E\{\Delta(s)^2\} = O\{(\nu h)^{-1}\}$ uniformly in s. (Here, $C > 0$ depends only on the support of the kernel, K.) Condition on Δ, then develop the expansion at (5.14), then take expectations in the distribution of Δ.

After these steps, the absolute value of the term that is quadratic in y is seen to be bounded by a constant multiple of

$$(\nu h)^{-1} \int_{\mathcal{I}} ds_1 \int_{s_2\,:\,|s_2 - s_1| \le C\,h} f_{W(s_1),W(s_2)}\{z(s_1), z(s_2)\}\, ds_2\,, \qquad (5.15)$$

and higher-order terms may be dealt with similarly. Changing variable in (5.15) from s_2 to $u = s_2 - s_1$, and assuming the function z is sufficiently smooth and that W is a sufficiently smooth Gaussian process, it may be proved that the quantity in (5.15) is bounded above by a constant multiple of $\nu^{-1}h^{-1/2}$. This enables the upper bound at (3.7) to be improved to: $O(\nu^{\eta-(2q-1)/2q})$ if $q \le \frac{9}{2}$, and $O(\nu^{\eta-4/q})$ if $q > \frac{9}{2}$. This shows the potential for smaller bandwidths than $\nu^{-1/5}$ to produce lower coverage errors in the case of sufficiently regular processes, such as Gaussian processes, but it does not offer an improved bound when h is taken of conventional size, i.e. $\nu^{-1/5}$.

6. PROOFS OF MAIN PROPERTIES

6.1. Proof of Theorem 1. Define $\Delta_{Xi}(z) = I(\widehat{X}_i \le z) - I(X_i \le z)$, $\Delta_{Yi}(z) = I(\widehat{Y}_i \le z) - I(Y_i \le z)$, $\Delta_X = m^{-1} \sum_{i \le m} \Delta_{Xi}$ and $\Delta_Y = n^{-1} \sum_{i \le n} \Delta_{Yi}$, where indicator

functions are interpreted as in section 2. It suffices to show that

$$\int E\{\Delta_X(z) - E\Delta_X(z)\}^2\,\mu(dz) = o\big(m^{-1}\big)\,, \tag{6.1}$$
$$\int E\{\Delta_Y(z) - E\Delta_Y(z)\}^2\,\mu(dz) = o\big(n^{-1}\big)\,,$$

and it is adequate to prove (6.1). To this end, note that with $I_i = I(X_i \le z)$ and $\hat{I}_i = I(\widehat{X}_i \le z)$ we have,

$$\int E\{\Delta_X(z) - E\Delta_X(z)\}^2\,\mu(dz) = m^{-2}\sum_{i=1}^{m}\int \mathrm{var}\{\Delta_{Xi}(z)\}\,\mu(dz)$$
$$\le m^{-2}\sum_{i=1}^{m}\int E\big\{(\hat{I}_i - I_i)^2\big\}\,\mu(dz) = m^{-2}\sum_{i=1}^{m}\int \pi_i(z)\,\mu(dz)\,, \tag{6.2}$$

where $\pi_i(z) = P(\,\text{just one of "}X_i \le z\text{" and "}\widehat{X}_i \le z\text{" is true}\,)$. Now, for each $\eta > 0$,

$$\pi_i(z) \le P(\|\widehat{X}_i - X_i\|_\infty > \eta) + P(\|X - z\|_\infty \le \eta)\,. \tag{6.3}$$

Provided the bandwidths h_{Xi} satisfy (5.2) and (3.2), there exists a sequence $\eta = \eta(m)$ decreasing to zero, such that

$$\max_{1\le i\le m} P(\|\widehat{X}_i - X_i\|_\infty > \eta) \to 0 \tag{6.4}$$

uniformly in $1 \le i \le m$. Since $\eta \to 0$ and μ has no atoms, then

$$\int P(\|X - z\|_\infty \le \eta)\,\mu(dz) \to 0\,. \tag{6.5}$$

Combining (6.3)–(6.5) we deduce that

$$\sum_{i=1}^{m}\int \pi_i(z)\,\mu(dz) = o(m)$$

as $m \to \infty$. This result and (6.2) imply (6.1).

6.2. Proof of Theorem 2. Define $\kappa_r = \int u^r\,K(u)\,du$ and $\alpha_{ir}(t) = E\{A_{ir}(t)\}$. Under assumptions (5.5)–(5.9),

$$P\left\{\max_{1\le i\le m}\max_{1\le r\le r_0}\sup_{t\in\mathcal{I}}|A_{ir}(t) - \alpha_{ir}(t)| > (\nu h)^{\eta-1/2}\right\} = O\big(m^{-B}\big)\,,$$

for each $B, \eta > 0$ and $r_0 \ge 1$. More simply, $\alpha_{ir}(t) = \kappa_r\,f_i(t) + O(h^2)$, uniformly in i and t, if r is even, and $\alpha_{ir}(t) = O(h)$, uniformly in i and t, if r is odd. Hence,

$$\text{for even } r\,,\quad P\left\{\max_{1\le i\le m}\sup_{t\in\mathcal{I}}|A_{ir}(t) - \kappa_r\,f_i(t)| > (\nu h)^{\eta-1/2} + B_1\,h^2\right\} = O\big(m^{-B}\big)\,, \tag{6.6}$$

$$\text{for odd } r\,, \quad P\Big\{ \max_{1\le i\le m} \sup_{t\in\mathcal{I}} |A_{ir}(t)| > (\nu h)^{\eta-1/2} + B_1\, h \Big\} = O\big(m^{-B}\big)\,, \tag{6.7}$$

for each $B, \eta > 0$ and some $B_1 > 0$.

Define

$$\Delta_{ir}(t) = \frac{1}{m_i h_{Xi}} \sum_{j=1}^{m_i} \delta_{ij} \left(\frac{t - S_{ij}}{h_{Xi}}\right)^r K\left(\frac{t - S_{ij}}{h_{Xi}}\right),$$

$$\Delta_i(t) = \frac{A_{i2}(t)\,\Delta_{i0}(t) - A_{i1}(t)\,\Delta_{i1}(t)}{A_{i0}(t)\,A_{i2}(t) - A_{i1}(t)^2}\,.$$

By Taylor expansion of X_i,

$$\begin{aligned}\widehat{X}_i(t) = X_i(t) &+ \tfrac{1}{2}\,h^2\,X_i''(t)\,\frac{A_{i2}(t)^2 - A_{i1}(t)\,A_{i3}(t)}{A_{i0}(t)\,A_{i2}(t) - A_{i1}(t)^2} \\ &+ \tfrac{1}{6}\,h^3\,X_i'''(t)\,\frac{A_{i2}(t)\,A_{i3}(t) - A_{i1}(t)\,A_{i4}(t)}{A_{i0}(t)\,A_{i2}(t) - A_{i1}(t)^2} + \Delta_i(t) + R_{i1}(t)\,, \end{aligned} \tag{6.8}$$

where, for each $B, \eta > 0$,

$$P\Big\{ \max_{1\le i\le m} \sup_{t\in\mathcal{I}} |R_{i1}(t)| > h^{4-\eta} \Big\} = O\big(m^{-B}\big) \tag{6.9}$$

for each $B, \eta > 0$, and we have employed (5.10) and Markov's inequality to obtain (6.9). Using (6.6) and (6.7) to simplify the ratios on the right-hand side of (6.8), applying (6.9) to bound $R_{i1}(t)$, and employing (5.10) and Markov's inequality to bound $\sup_{t\in\mathcal{I}} |X_i^{(r)}|$ for $r = 2, 3$, we deduce that for each $\eta > 0$,

$$\widehat{X}_i(t) = X_i(t) + \tfrac{1}{2}\,\kappa_2\,h^2\,X_i''(t) + \Delta_i(t) + R_{i2}(t)\,, \tag{6.10}$$

where, for each $B, \eta > 0$,

$$P\Big\{ \max_{1\le i\le m} \sup_{t\in\mathcal{I}} |R_{i2}(t)| > h^{4-\eta} + h^2\,(\nu\,h)^{\eta-1/2} \Big\} = O\big(m^{-B}\big)\,. \tag{6.11}$$

Put $\xi = \xi(m) = h^{4-\eta} + h^2\,(\nu\,h)^{\eta-1/2}$ and $Q_i = X_i + \frac{1}{2}\,\kappa_2\,h^2\,X_i''$; that is, Q_i equals the version of W_d that arises when $d = \frac{1}{2}\,\kappa_2\,h^2$ and $X = X_i$. Then, by (6.10) and (6.11),

$$P\big(\widehat{X}_i \le z\big) \begin{cases} \le P(Q_i + \Delta_i \le z + \xi) + O(m^{-B}) \\ \ge P(Q_i + \Delta_i \le z - \xi) + O(m^{-B})\,, \end{cases} \tag{6.12}$$

for each $B, C, \eta > 0$, where the remainders $O(m^{-B})$ are of that size uniformly in functions $z \in \mathcal{C}_2(C)$ and in $1 \le i \le m$.

For the next step we need to "ridge" the quantity Δ_i, so as to avoid aberrations caused by its denominator, $D_i(t) = A_{i0}(t)\,A_{i2}(t) - A_{i1}(t)^2$, being too close to zero for some t. We may choose $\sigma > 0$ so small that the probability that the event $\mathcal{E}_i$,

that $|D_i(t)| > \sigma$ for all $1 \le i \le m$ and all $t \in \mathcal{I}$, equals $1 - O(m^{-B})$ for each $B > 0$; call this result (R). Let I_i denote the indicator function of the event that $\mathcal{E}_i$ holds. In view of (R), we may replace Δ_i by $\Delta_i\, I_i$, in (6.12), without affecting the veracity of that result:

$$P\big(\widehat{X}_i \le z\big) \begin{cases} \le P(Q_i + \Delta_i\, I_i \le z + \xi) + O(m^{-B}) \\ \ge P(Q_i + \Delta_i\, I_i \le z - \xi) + O(m^{-B})\,, \end{cases} \tag{6.13}$$

uniformly in $z \in \mathcal{C}_2(C)$ and in $1 \le i \le m$.

Assumption (5.8) implies that for each $p \in (1, 2)$ satisfying $p < (q-1)/2$, and each $B, C, \eta > 0$, $P\{\Delta_i \in \mathcal{C}_r(C)\} = 1 - O(m^{-B})$ for all $r > p$ sufficiently close to p. Noting this property; applying (5.12) and (5.11), with (W_d, y) there replaced by $(Q_i, \Delta_i\, I_i)$ and with $p < \min\{2, (q-1)/2\}$; observing that $\Delta_i\, I_i$ is independent of Q_i, so we may first condition on $\Delta_i\, I_i$ and then take expectation in the distribution of $\Delta_i\, I_i$; and noting that $E(\Delta_i\, I_i) = 0$, so that, after expectations are taken, the linear term in (5.12) vanishes; we deduce that, for either choice of the $\pm$ signs,

$$P(Q_i + \Delta_i\, I_i \le z \pm \xi) = P(Q_i \le z \pm \xi) + O\big(E\|\Delta_i\, I_i\|_\infty^p\big)\,, \tag{6.14}$$

uniformly in $z \in \mathcal{C}_2(C)$ and in $1 \le i \le m$. Now, $E\|\Delta_i\, I_i\|_\infty^p = O\{(\nu h)^{\eta - p/2}\}$, for each $\eta > 0$. Substituting these bounds into (6.14), and recalling that $\xi = \xi(m) = h^{4-\eta} + h^2\,(\nu\, h)^{\eta - 1/2}$, we deduce that for each $\eta > 0$,

$$P(Q_i + \Delta_i\, I_i \le z \pm \xi) = P(Q_i \le z \pm \xi) + O\big\{(\nu h)^{\eta - p/2} + h^{4-\eta}\big\}\,, \tag{6.15}$$

uniformly in $z \in \mathcal{C}_2(C)$ and in $1 \le i \le m$.

Now apply (5.12) once more, again with W_d replaced by Q_i but this time with $p \in (1, 2]$ and $y \equiv \pm\xi$, obtaining,

$$P(Q_i \le z \pm \xi) = P(Q_i \le z) + O(\xi)\,, \tag{6.16}$$

uniformly in $z \in \mathcal{C}_2(C)$ and in $1 \le i \le m$. Combining (6.13)–(6.16) we deduce that, uniformly in $z \in \mathcal{C}_2(C)$ and in $1 \le i \le m$,

$$P(\widehat{X}_i \le z) = P(Q_i \le z) + O\big\{(\nu h)^{\eta - p/2} + h^{4-\eta}\big\}. \tag{6.17}$$

If $q > 5$ in (5.8) then the arguments above apply with $p = 2$, and so (6.17) implies: $P(\widehat{X}_i \le z) = P(Q_i \le z) + O(\nu^{\eta - 4/q})$ for each $\eta > 0$. If $q \le 5$ then the arguments hold for each $1 < p < (q-1)/2$, and so (6.17) entails: $P(\widehat{X}_i \le z) = P(Q_i \le z) + O(\nu^{\eta - (q-1)^2/4q})$ for each $\eta > 0$.

REFERENCES

ANDERSON, N.H., HALL, P. AND TITTERINGTON, D.M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *J. Multivariate Anal.* **50**, 41–54.

CAO, R. AND VAN KEILEGOM, I. (2005). Empirical likelihood tests for two-sample problems via nonparametric density estimation. *Canad. J. Statist.*, to appear.

CLAESKENS, G., JING, B.-Y., PENG, L. AND ZHOU, W. (2003). Empirical likelihood confidence regions for comparison distributions and ROC curves. *Canad. J. Statist.* **31**, 173–190.

CUEVAS, A., FEBRERO, M. AND FRAIMAN, R. (2004). An anova test for functional data. *Comput. Statist. Data Anal.* **47**, 111–122.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196–216.

FAN, J. AND GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications.*. Chapman and Hall, London.

FAN, J. AND LIN, S.-K. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.* **93**, 1007–1021.

FAN, Y.Q. (1994). Testing the goodness-of-fit of a parametric density-function by kernel method. *Econometric Theory* **10**, 316–356.

FAN, Y.Q. (1998). Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameter. *Econometric Theory* **14**, 604–621.

FAN, Y.Q. AND ULLAH, A. (1999). On goodness-of-fit tests for weakly dependent processes using kernel method. *J. Nonparametr. Statist.* **11**, 337–360.

INGSTER, Y.I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. I, II. *Math. Methods Statist.* **2**, 85–114, 171–189.

LI, Q. (1999). Nonparametric testing the similarity of two unknown density functions: Local power and bootstrap analysis. *J. Nonparametr. Statist.* **11**, 189–213.

LOCANTORE, N., MARRON, J.S., SIMPSON, D.G., TRIPOLI, N., ZHANG, J.T. AND COHN, K.L. (1999). (With discussion.) Robust principal component analysis for functional data. *Test* **8**, 1–73.

LOUANI, D. (2000). Exact Bahadur efficiencies for two-sample statistics in functional density estimation. *Statist. Decisions* **18**, 389–412.

RAMSAY, J.O. AND SILVERMAN, B.W. (1997). *Functional Data Analysis.* Springer, New York.

RAMSAY, J.O. AND SILVERMAN, B.W. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* Springer, New York.

RUPPERT, D. AND WAND, M.P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346–1370.

SHEN, Q. AND FARAWAY, J. (2004). An F test for linear models with functional responses. *Statistica Sinica* **14**, 1239–1257.

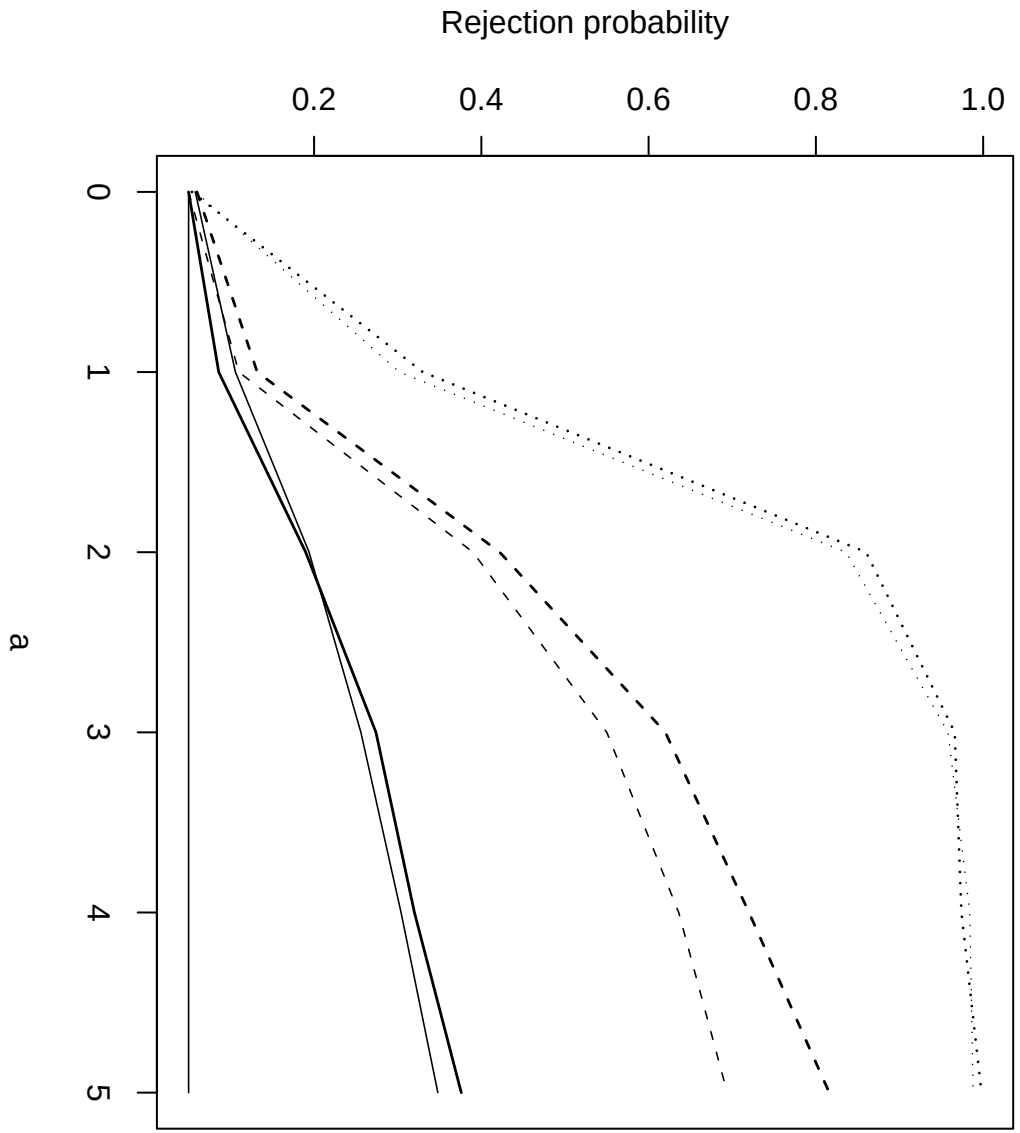
SPITZNER, D.J., MARRON, J.S. AND ESSICK, G.K. (2003). Mixed-model functional ANOVA for studying human tactile perception. *J. Amer. Statist. Assoc.* **98**, 263–272.

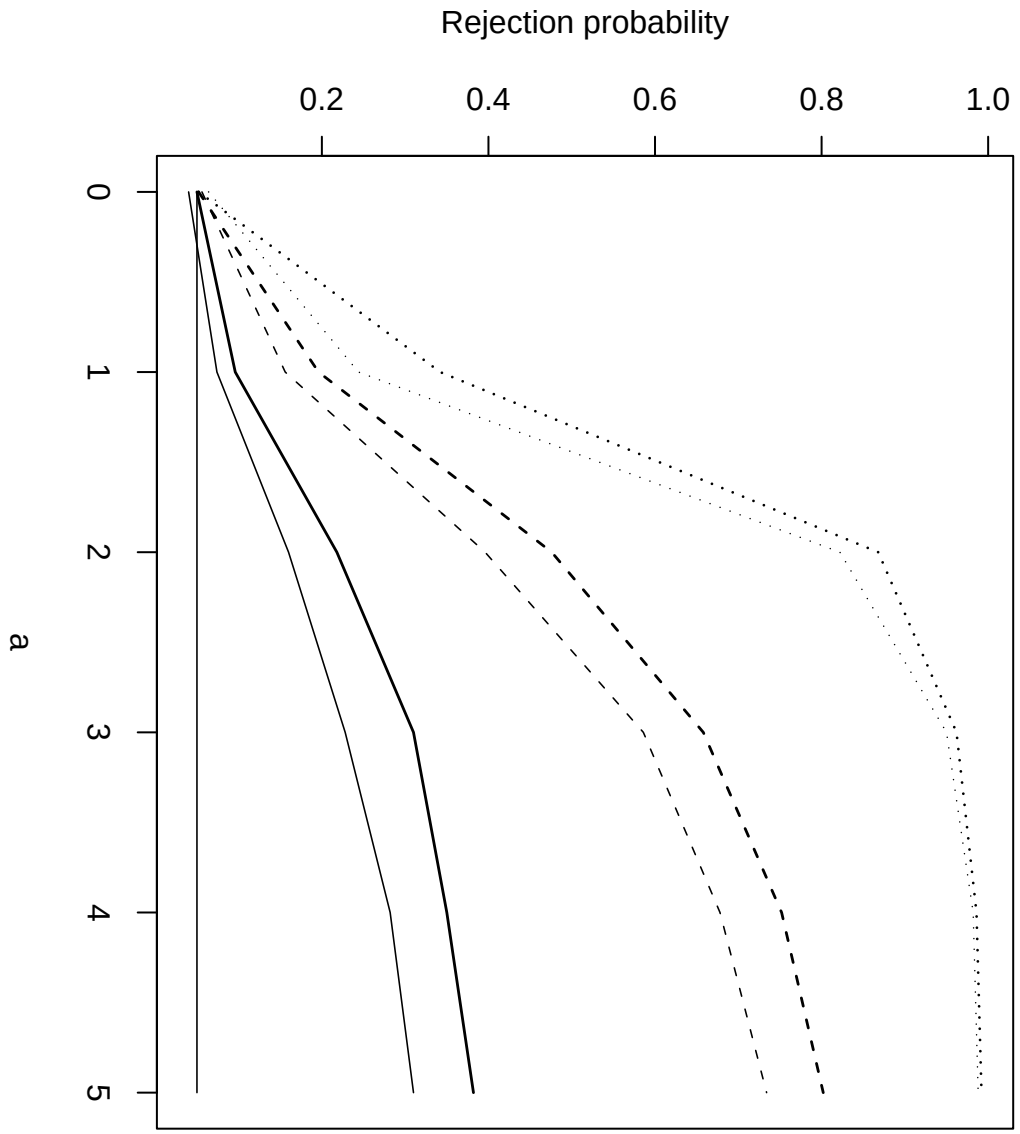
ZHANG, B. (2000). Estimating the treatment effect in the two-sample problem with auxiliary information. *J. Nonparametr. Statist.* **12**, 377–389.

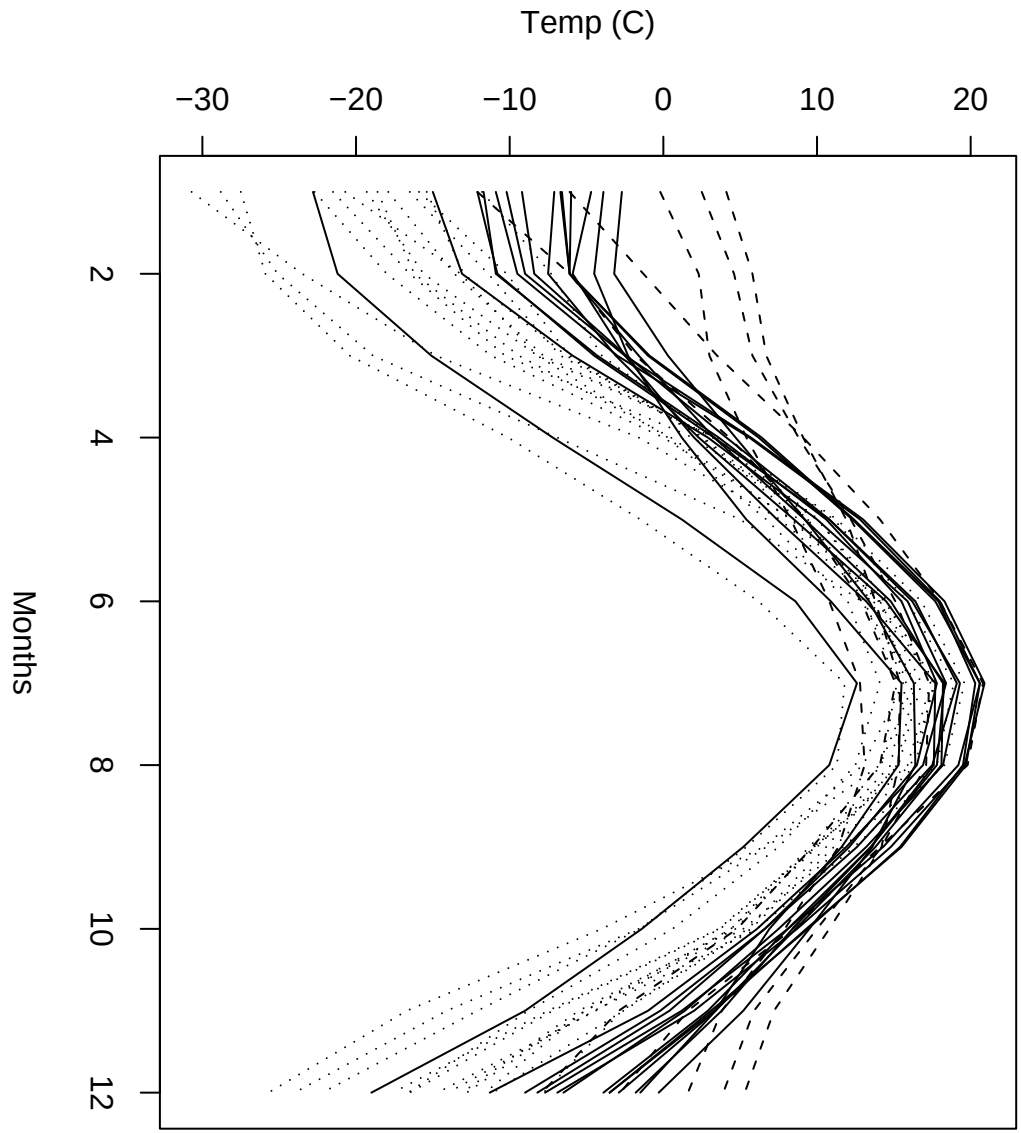
Caption for Figure 1: *Rejection probabilities for* $m = n = 15$ *(full curve),* $m = n = 25$ *(dashed curve) and* $m = n = 50$ *(dotted curve). The thin curves correspond to* $m_1 = n_1 = 20$*, the thick curves to* $m_1 = n_1 = 100$*. The null hypothesis holds for* $a = 0$*.*

Caption for Figure 2: *Rejection probabilities for* $m = n = 15$ *(full curve),* $m = n = 25$ *(dashed curve) and* $m = n = 50$ *(dotted curve). The thin curves are obtained by using different bandwidths for each sample, the thick curves use the same bandwidth. In all cases,* $m_1 = 20$ *and* $n_1 = 100$*. The null hypothesis holds for* $a = 0$*.*

Caption for Figure 3: *Raw data for mean monthly temperatures at 35 Canadian weather stations. The full curves correspond to data for Atlantic stations, dashed curves for Pacific stations and dotted curves for Continental stations.*


Rejection probability
0.2
0.4
0.6
0.8
1.0
a
0
1
2
3
4
5


Rejection probability
0.2
0.4
0.6
0.8
1.0
a
0
1
2
3
4
5


Temp (C)
−30
−20
−10
0
10
20
Months
2
4
6
8
10
12