Journal of Nonparametric Statistics, 2013
Vol. 25, No. 3, 635–645, http://dx.doi.org/10.1080/10485252.2013.810217

Local significant differences from nonparametric two-sample tests

Tarn Duong[a,b,c]*

[a]*Theoretical and Applied Statistics Laboratory (LSTA), University Pierre and Marie Curie – Paris 6, F-75005 Paris, France;* [b]*Institute of Translational Neurosciences, Pitié-Salpêtrière Hospital, F-75005 Paris, France;* [c]*Molecular Mechanisms of Intracellular Transport Laboratory, Institut Curie, CNRS, UMR144, F-75248 Paris, France*

(*Received 6 September 2012; final version received 23 May 2013*)

We establish a framework to investigate the local differences of two multivariate data samples, as measured by a statistically significant two-sample test. This framework identifies the locally significant difference regions by computing local test statistics based on the squared difference of two kernel density estimators. The key differences between the data samples are concentrated in these significantly different regions. We illustrate the visualisation and interpretation of local significant differences for simulated data, and their potential in the role of biomarker discovery for biological/biomedical data.

Keywords: asymptotic normality; density difference; multivariate kernel estimator

MSC: 62G07, 62G10

1. Introduction

In the context of two-sample testing, the classical t-test compares the location of two population means. There has been much interest in generalising this test. Amongst the most widely known nonparametric tests for one-dimensional continuous data are the Kolmogorov–Smirnov, Wald–Wolfowitz and Mann–Whitney tests (see the monograph of Gibbons and Chakraborti 2003). Approaches developed by Bickel (1969), Friedman and Rafsky (1979) and Liu and Singh (1993) generalise these to multivariate data. Though generally, these have not been met with the same wide acceptance as their univariate antecedents.

Since the t-test is a comparison of the two normal density functions with a common variance, it is plausible that a testing procedure based on more general density estimators will be more flexible. The generalisation to normal densities with unequal variances (the so-called Behrens–Fisher problem) has been provided by Nel and van der Merwe (1986). Generalising this idea further by replacing a parametric density with a nonparametric density estimator has lead to a vast body of work which we do not attempt to review comprehensively here. Since we propose a kernel density estimator-based method, we cite only a few references for tests based on these: the L_2 discrepancy

*Email: tarn.duong@upmc.fr

between density functions of Anderson, Hall and Titterington (1994), the empirical likelihood ratio test statistic of Cao and Van Keilegom (2006), the L_2 discrepancy between characteristic functions of Alba Fernández, Jiménez Gamero and Muñoz García (2008), the common area test statistic of Martínez-Camblor, De Uña-Álvarez and Corral (2008) and the density ratio/relative density approach of Molanes-Lopéz and Cao (2008). We take as our starting point the L_2 discrepancy of Anderson et al. (1994) due to its amenability for mathematical analysis. In this sub-class, subsequent contributions include Li (1999), Baringhaus and Franz (2004), Borgwardt et al. (2006) and Duong, Goud and Schauer (2012). We adopt the approach of Duong et al. (2012) which, unlike the much of the preceding work, relies on the asymptotic results to compute the sampling distribution of the test statistic, rather than on resampling methods. These authors believe that the reliance on resampling schemes inhibited the wider user of kernel-based testing, outside the statistical computing community, and especially amongst experimental scientists. From a mathematical point of view, asymptotic methods provide insight into the behaviour of the test statistic which is not always apparent from finite sample resampling, for example, the role of the smoothing parameters.

Nonparametric density estimates typically allow good estimation over the sample space. When they are used in a two-sample comparison, they are usually condensed into a global criterion, for example, via integration or various norms, leading to a single binary decision whether the two density estimates are equal or not as evaluated over the sample space. This type of global binary result can leave much of the local structure concealed. In this manuscript, we examine the problem of finding the regions of the sample space where the two data sets are the locally most different. In a chi-squared test of the homogeneity of two samples, if an overall statistically significant result is obtained, then post hoc tests can be applied to the sample counts within each partition class to establish which of these are main contributors to overall statistical significance (see Cox and Key 1993). Applying these post hoc chi-squared tests to determine the significantly different partition classes of discretised density functions was first posited by Roederer and Hardy (2001) and recently refined by Duong, Koch and Wand (2009). The novelty we introduce is to use smooth kernel density estimates in place of discretised density estimates, to take advantage of the well-known statistical properties of the latter over the former (see Scott 1992; Simonoff 1996 for an overview). The analysis of the global difference of the kernel density estimates was examined in Hall and Wand (1988) for discriminant analysis. Sugiyama et al. (2012) and references therein provide a review of recent developments of density differences in other situations. We extend the analysis of density differences to a local inferential framework.

In Section 2, we generalise the global hypothesis testing framework to local squared differences and provide an algorithm to construct locally significantly different regions where the sample densities are most different. In the last section, we provide graphical visualisations and simulation studies of these locally significantly different regions on simulated and real data.

2. Local significant difference regions

To test the null hypothesis of global equality $H_0 : f_1 \equiv f_2$ of two density functions f_1 and f_2, we require a discrepancy measure between the two density functions. The L_p norm is a popular candidate, with L_1, L_2 and L_∞ the most commonly used. Allen (1997), Louani (2000) and Biau and Györfi (2005) examine the L_1 case, with the latter also concerned with the L_∞ supremum norm, whereas Anderson et al. (1994) investigate an L_2 criterion $T \equiv T(f_1,f_2) = \int_{\mathbb{R}^d} [f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})]^2 \, \mathrm{d}\boldsymbol{x}$. It is shown by Duong et al. (2012) that a direct plug-in estimator for the global test statistic T of the form $\int_{\mathbb{R}^d} [\hat{f}_1(\boldsymbol{x}) - \hat{f}_2(\boldsymbol{x})]^2 \, \mathrm{d}\boldsymbol{x}$ is undesirable since it (a) requires numerical integration and (b) lacks closed-form distributional results. On the other hand, it is useful to consider a plug-in local test statistic.

Let $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_{n_1}\}$ and $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \dots, \boldsymbol{Y}_{n_2}\}$ be d-variate random samples from their respective common densities f_1 and f_2. The kernel density estimates of f_1 and f_2 are

$$\hat{f}_1(\boldsymbol{x}; \mathbf{H}_1) = n_1^{-1} \sum_{i=1}^{n_1} K_{\mathbf{H}_1}(\boldsymbol{x} - \boldsymbol{X}_i), \quad \hat{f}_2(\boldsymbol{x}; \mathbf{H}_2) = n_2^{-1} \sum_{j=1}^{n_2} K_{\mathbf{H}_2}(\boldsymbol{x} - \boldsymbol{Y}_j),$$

where K is the kernel function with $K_{\mathbf{H}_\ell}(\boldsymbol{x}) = |\mathbf{H}_\ell|^{-1/2} K(\mathbf{H}_\ell^{-1/2}\boldsymbol{x})$ and $\mathbf{H}_\ell$ is a bandwidth matrix, for $\ell = 1, 2$. At a non-random point $\boldsymbol{x}$, we consider the local hypothesis $H_0(\boldsymbol{x}) : f_1(\boldsymbol{x}) = f_2(\boldsymbol{x})$ using the analogous local test statistic $\hat{U}(\boldsymbol{x}) = [\hat{f}_1(\boldsymbol{x}; \mathbf{H}_1) - \hat{f}_2(\boldsymbol{x}; \mathbf{H}_2)]^2$. Investigating local differences between multivariate data point clouds in terms of the difference of density functions was suggested by Duong et al. (2009) who applied chi-squared test to discretised data. Our proposed test bypasses this discretisation, thus avoiding any source of inaccuracy induced here. The key result of our testing procedure is the asymptotic chi-squared distribution of $\hat{U}(\boldsymbol{x})$, presented in the following theorem.

THEOREM 2.1 *Suppose that the following conditions hold. For $\ell = 1, 2$,*

(F) *the target densities f_ℓ are bounded and continuous*;
(H) *the bandwidths $\mathbf{H}_\ell = \mathbf{H}_\ell(n_\ell)$ are a sequence of symmetric positive definite matrices such that all elements of $\mathbf{H}_\ell \to 0$ and $n_\ell^{-1}|\mathbf{H}_\ell|^{-1/2} \to 0$ as $n_\ell \to \infty$*;
(K) *the kernel K is a symmetric, square integrable probability density function and such that $\int_{\mathbb{R}^d} \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} K(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = m_2(K)\mathbf{I}_d$ for some real number $m_2(K)$ and $\mathbf{I}_d$ is the $d \times d$ identity matrix, and $R(K) = \int_{\mathbb{R}^d} K(\boldsymbol{x})^2\, \mathrm{d}\boldsymbol{x}$.*

Further suppose that local null hypothesis $H_0(\boldsymbol{x}) : f_1(\boldsymbol{x}) = f_2(\boldsymbol{x}) = f(\boldsymbol{x})$ holds for a non-random point $\boldsymbol{x}$. Then,

$$\sigma_U^{-2}(\boldsymbol{x})\hat{U}(\boldsymbol{x}) \xrightarrow{d} \chi_1^2,$$

where $\sigma_U^2(\boldsymbol{x}) = (n_1^{-1}|\mathbf{H}_1|^{-1/2} + n_2^{-1}|\mathbf{H}_1|^{-1/2})R(K)f(\boldsymbol{x})$ and χ_1^2 is a chi-squared distribution with 1 degree of freedom.

Proof The classical result of Parzen (1962), relying on that kernel density estimators are local means, shows their asymptotic normality via appropriate central limit theorems. Thus, we can write that $[n_\ell^{-1}|\mathbf{H}_\ell|^{-1/2}R(K)f_\ell(\boldsymbol{x})]^{-1/2}[\hat{f}_\ell(\boldsymbol{x}; \mathbf{H}_\ell) - f_\ell(\boldsymbol{x})] \xrightarrow{d} N(0, 1)$, as $n_\ell \to \infty$, using expressions for the mean and variance of $\hat{f}_\ell$ from Wand (1992). Under the null hypothesis $f_1(\boldsymbol{x}) = f_2(\boldsymbol{x}) = f(\boldsymbol{x})$, the difference $\hat{U}^{1/2} = \hat{f}_1 - \hat{f}_2$ follows

$$[(n_1^{-1}|\mathbf{H}_1|^{-1/2} + n_2^{-1}|\mathbf{H}_2|^{-1/2})R(K)f(\boldsymbol{x})]^{-1/2}\hat{U}(\boldsymbol{x})^{1/2} \xrightarrow{d} N(0, 1).$$

■

The remaining unknown in the asymptotic formula is the parameter σ_U^2. The bandwidth matrices are consistently estimated using the plug-in selectors $\hat{\mathbf{H}}_1, \hat{\mathbf{H}}_2$ of Duong and Hazelton (2003). Then, $\hat{\sigma}_U^2(\boldsymbol{x}) = R(K)[n_1^{-1}|\hat{\mathbf{H}}_1|^{-1/2}\hat{f}_1(\boldsymbol{x}; \hat{\mathbf{H}}_1) + n_2^{-1}|\hat{\mathbf{H}}_2|^{-1/2}\hat{f}_2(\boldsymbol{x}; \hat{\mathbf{H}}_2)]$.

What is left is the adjustment for multiple correlated testing. We adapt the approach used by Duong, Cowling, Koch and Wand (2008) in the context of one-sample significance testing, relying on the Hochberg (1988) adjustment to control the family-wise level of significance α. Our proposed algorithm to find locally different regions based on the set of local hypothesis tests $\{H_0(\boldsymbol{x}) : f_1(\boldsymbol{x}) = f_2(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d\}$ is as follows:

(1) For all estimation points $\boldsymbol{x}_j, j = 1, \dots, m$, compute the test statistic $X_j^2 = \hat{\sigma}_U^{-2}(\boldsymbol{x}_j)\hat{U}(\boldsymbol{x}_j)^2$ and the corresponding p-value $P_j = \mathbb{P}(X_j^2 \geq \chi_1^2)$. A common choice of the estimation points is a

fixed uniform grid of points, for example, $(d, m) = (1, 401), (2, 151^2), (3, 51^3)$ as typical grid sizes used in kernel density estimation (Wand 1994).

(2) Apply the Hochberg (1988) multiple adjustment to these m local hypothesis tests. Sort these p-values into ascending order $P_{(1)}, \ldots, P_{(m)}$. Define $j^* = \text{argmax}_{1 \le j \le m}\{P_{(j)} \le \alpha/(m - j + 1)\}$. The rejection region is $\{\boldsymbol{x}_j : P_{(j)} \le j^*, 1 \le j \le m\}$.

(3) For all points $\boldsymbol{x}_j$ in the rejection region,

  (a) if $\hat{f}_1(\boldsymbol{x}_j; \mathbf{H}_1) > \hat{f}_2(\boldsymbol{x}_j; \mathbf{H}_2)$, then conclude that $f_1(\boldsymbol{x}_j) > f_2(\boldsymbol{x}_j)$;

  (b) if $\hat{f}_1(\boldsymbol{x}_j; \mathbf{H}_1) < \hat{f}_2(\boldsymbol{x}_j; \mathbf{H}_2)$, then conclude that $f_1(\boldsymbol{x}_j) < f_2(\boldsymbol{x}_j)$.

3. Numerical study

3.1. *Bivariate simulated data*

For a simulation study, we compare pairs of bivariate mixture normal densities, taken from Duong et al. (2012) as a testing ground for the finite sample performance of our proposed test. The first pair $N((-1/2, 0), \mathbf{I}_2)$ and $N((1/2, 0), \mathbf{I}_2)$ are two single normal densities with identity variance and whose means are separated by a distance of 1, so this can be treated as a base case. Pair #2 are both bimodal densities, $1/2N((1, -1), \boldsymbol{\Sigma}) + 1/2N((-1, 1), \boldsymbol{\Sigma})$ and $1/2N((1, -1), \boldsymbol{\Sigma}) + 1/2N((-1, 1), \mathbf{I}_2)$ where $\boldsymbol{\Sigma} = [4/9\ 4/15; 4/15\ 4/9]$. The lower right components are almost the same but the upper left components are different so this is potentially a challenging case to distinguish between two finite samples. Both densities in pair #3, $N((0, 0), \mathbf{I}_2)$ and $1/2N((0, 0), \mathbf{I}_2) + 1/10N((0, 0), 1/16\mathbf{I}_2) + 1/10N((-1, -1), 1/16\mathbf{I}_2) + 1/10N((-1, 1), 1/16\mathbf{I}_2) + 1/10N((1, -1), 1/16\mathbf{I}_2) + 1/10N((1, 1), 1/16\mathbf{I}_2)$, have (approximately) zero mean and identity variance, though with different internal structure, so would most likely benefit from a density-based test. The contour plots and difference regions of these test densities are displayed in Figure 1. Our aim is not to estimate these entire difference regions, but the most influential subsets of them.

For each pair, we take two random samples of size $n = n_1 = n_2 = 1000$ and $10,000$ from density #1, and two random samples from each of density #1 and density #2. We compute the kernel-based global p-values and locally significantly different regions, using the functionality provided in the `ks` library (Duong 2007) in the `R` statistical programming language.

For the locally significant regions, we fix the level of significance at $\alpha = 0.05$. In Figure 2, the dark grey region is where density #2 is significantly greater than density #1, light grey is where density #1 is significantly greater density #2 and otherwise is where the densities are equal. As expected, the regions are uniformly larger for the larger sample size. For Pair #1, the locally significant regions are centred around the modes of the two individual densities. For Pair #2, for $n = 1000$ only small locally significant regions appear, whereas for $n = 10,000$, these regions more clearly show the differences in the upper left modes. For Pair #3, the dark grey regions are more compactly delimited than the light grey regions, indicating that for this pair, it is easier to distinguish where density #2 is more abundant at its five local modes than vice versa.

In Figure 2, for comparison, we computed the locally significant regions based on the method of Duong et al. (2009) who used the PRIM (Patient Induction Rule Method) to partition the data sets for a chi-squared test. First, the sample of density #1 are labelled 1, and density #2 are labelled -1. The PRIM algorithm builds rectangular regions where the sample mean of the labels exceeds given thresholds, which are 0.3 and -0.3 in our case, based on thresholds computed by a data-based algorithm provided by Duong et al. (2009). The minimum size of the rectangles is set to be 5% of the combined sample size for $n = 1000$ and 2.5% for $n = 10,000$. The algorithm searches for rectangles where one density tends to outnumber the other. Compared to the kernel-based difference regions, PRIM regions are located in the same general area of the sample space, but

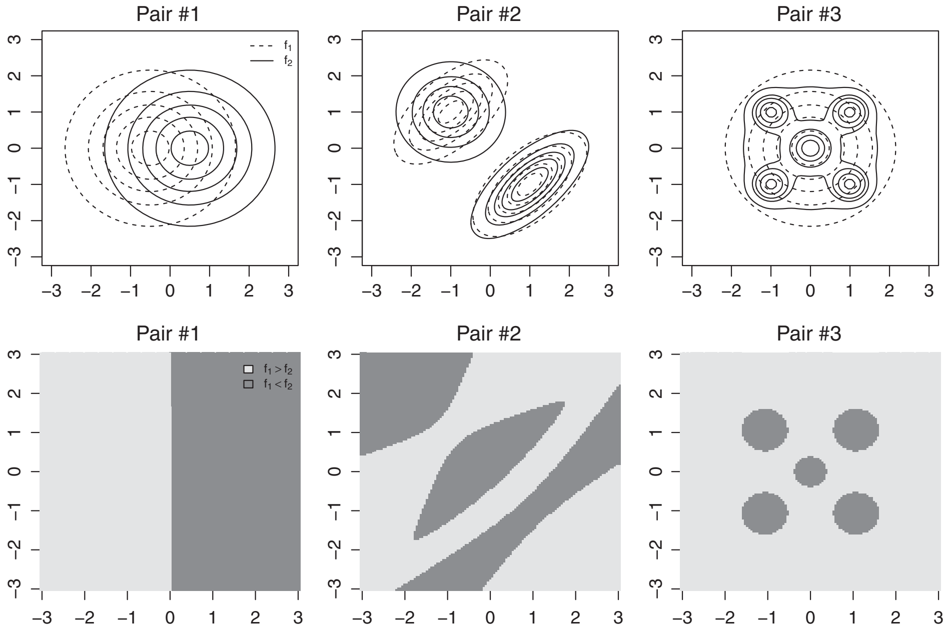

Figure 1. Bivariate target density pairs. The first row are contour plots. Dashed lines: density #1. Solid lines: density #2. In the second row, the difference regions are coloured as follows: light grey: density #1 $>$ density #2; dark grey: density #1 $<$ density #2.

the latter, due to their rectangular construction, do not follow the structure of the data as the former. The PRIM algorithm results in some dark grey regions in the lower right corner which are spuriously significant since visual inspection of the true densities indicates little difference between the two densities. Thus, kernel-based difference regions are preferred here.

To check the performance of Hochberg's (1988) procedure to control Type I and Type II errors for finite samples, we compare the samples drawn from the second density in each of the target density pairs, providing a range of density shapes, for a nominal level of significance of $\alpha = 0.05$. The first sample is drawn from the density, and the second sample is drawn from the same density translated by μ in the x-axis, for $\mu = 0.0, 0.1, \ldots, 0.8$. For the sample sizes $n = 1000, 10,000$, we calculate the any power rate, that is, the proportion of the $N = 100$ trials where at least one $\boldsymbol{x}$ in the grid defined on $[-3, 3] \times [-3, 3]$ leads to a rejection of the local null hypothesis $H_0(\boldsymbol{x})$. For $\mu = 0$, this gives an empirical estimate of the level of significance $\hat{\alpha}$. Hochberg's (1988) procedure guarantees a family-wise error rate at all testing points to be no greater than α rather than exactly α so our results $\hat{\alpha} = 0.00$ for all cases, except $\hat{\alpha} = 0.02$ for Pair #2, $n = 10,000$, are consistent with this. This is the first column in Table 1. These furthermore indicate that the proposed testing procedure is more conservative than indicated by the nominal level of significance $\alpha = 0.05$. In the remaining columns, we examine the case where the alternative hypothesis ($\mu = 0.1, 0.2, \ldots, 0.8$) holds to compute the estimated powers of the test $1 - \hat{\beta}$. For $n = 1000$, the test has low power until a higher value of μ, whereas for $n = 10,000$, the test correctly rejects the null hypothesis for a smaller separation μ.

Another important property is the behaviour of the estimated level of significance and any power rate of the test as functions of the bandwidth. To verify these, we take the case of comparing samples from the second density from the target density pairs against the same density, and against the same density translated by $(0.6, 0)$ as this separation of two normal mixture distributions leads to excellent power, so it will allow us to examine how the power changes as the bandwidth changes.

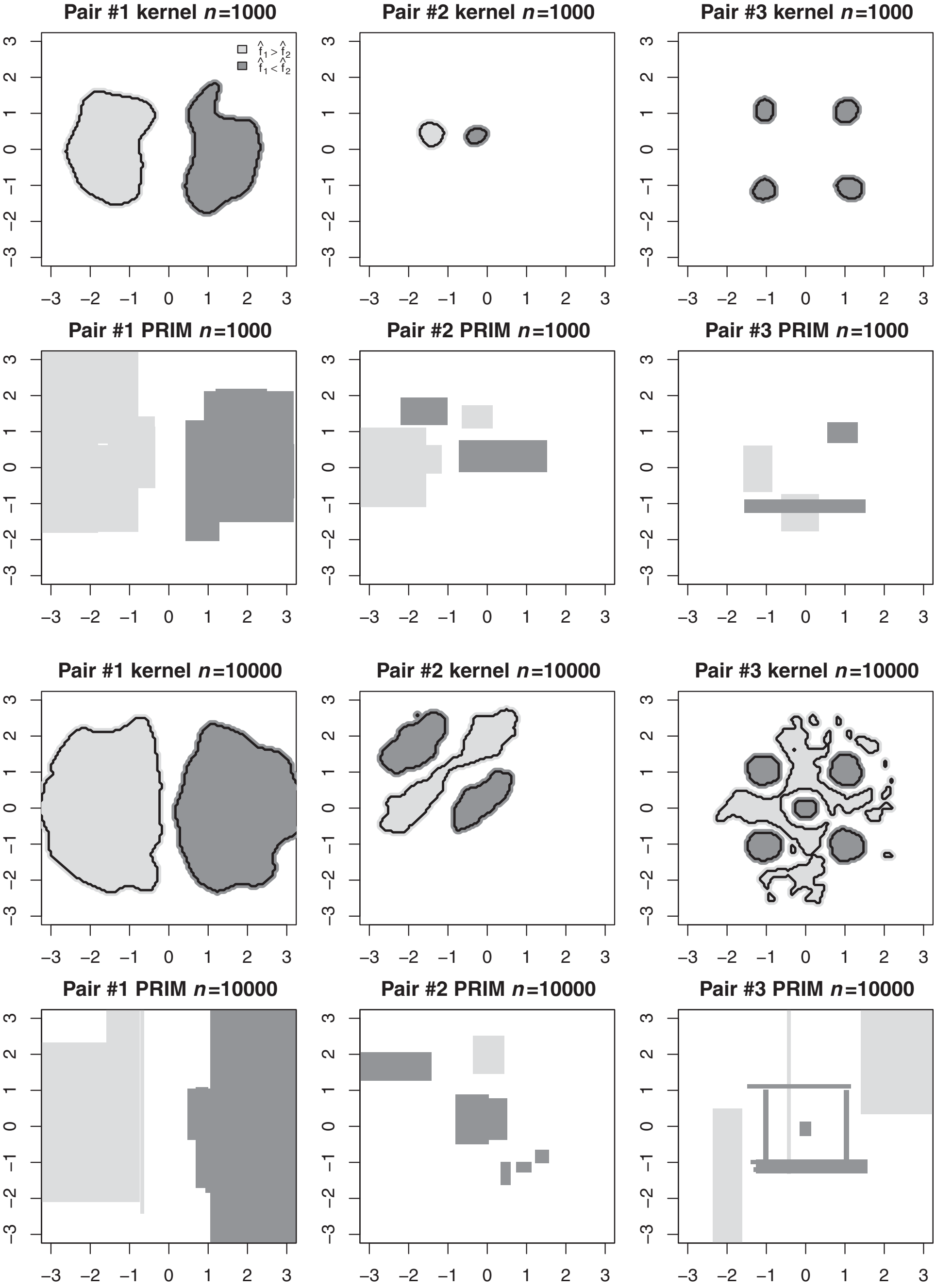

Figure 2. Comparison of kernel-based and PRIM-based locally significant difference regions, at $\alpha = 0.05$ level of significance, for bivariate target density pairs, at sample sizes $n = 1000, 10,000$. The difference regions are coloured as follows: light grey, density #1 > density #2; white, density #1 = density #2; dark grey, density #1 < density #2.

In Table 2, the fifth column ($\gamma = 1.0$) is the base case with the data-based optimal bandwidths $\hat{\mathbf{H}}_1, \hat{\mathbf{H}}_2$. The other columns are the level of significance and power of the tests carried on the same sample data with the bandwidths $\hat{\mathbf{H}}_1^\gamma, \hat{\mathbf{H}}_2^\gamma$, $\gamma = -1.0, -0.5, 0.0, 0.5, 1.5, 2.0, 2.5, 3.0$. Since $|\hat{\mathbf{H}}_1|, |\hat{\mathbf{H}}_2| < 1$, then using $\gamma > 1$ leads to undersmoothing, and $\gamma < 1$ leads to oversmoothing.

Table 1. Empirical level of significance and any power rate for the comparison of samples from the second density from the target density pairs against the same density translated by $(\mu, 0)$, for a nominal level of significance $\alpha = 0.05$.

	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Separation (μ)	$\hat{\alpha}$	$1 - \hat{\beta}$							
Pair #1									
$n = 1000$	0.00	0.00	0.03	0.09	0.46	0.84	1.00	1.00	1.00
$n = 10,000$	0.00	0.07	0.97	1.00	1.00	1.00	1.00	1.00	1.00
Pair #2									
$n = 1000$	0.00	0.00	0.10	0.79	1.00	1.00	1.00	1.00	1.00
$n = 10,000$	0.02	0.66	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Pair #3									
$n = 1000$	0.00	0.00	0.34	0.97	1.00	1.00	1.00	1.00	1.00
$n = 10,000$	0.00	0.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Notes: The two sample sizes considered are $n = 1000, 10,000$. The first column with $\mu = 0$ gives the estimated level of significance $\hat{\alpha}$ and the remaining columns with $\mu > 0$ give the estimated any power rate $(1 - \hat{\beta})$ for these alternative hypotheses.

Table 2. Empirical level of significance and any power rate, as functions of varying bandwidths for a nominal level of significance $\alpha = 0.05$.

Bandwidth exponent (γ)	−1.0	−0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
				Level of significance $(\hat{\alpha})$					
Pair #1									
$n = 1000$	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$n = 10,000$	0.32	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Pair #2									
$n = 1000$	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$n = 10,000$	0.36	0.00	0.00	0.00	0.02	0.01	0.00	0.00	0.00
Pair #3									
$n = 1000$	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$n = 10,000$	0.42	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
				Power $(1 - \hat{\beta})$					
Pair #1									
$n = 1000$	0.13	1.00	1.00	1.00	1.00	0.46	0.00	0.00	0.00
$n = 10,000$	0.35	1.00	1.00	1.00	1.00	1.00	0.81	0.00	0.00
Pair #2									
$n = 1000$	0.12	1.00	1.00	1.00	1.00	1.00	0.01	0.00	0.00
$n = 10,000$	0.30	1.00	1.00	1.00	1.00	1.00	0.33	0.00	0.00
Pair #3									
$n = 1000$	0.22	1.00	1.00	1.00	1.00	0.98	0.00	0.00	0.00
$n = 10,000$	0.41	0.02	1.00	1.00	1.00	1.00	0.00	0.00	0.00

Notes: Each entry in the upper half is the estimated level of significance $\hat{\alpha}$, for the comparison of samples from the second density from the target density pairs against the same density. Each entry in the lower half is the estimated any power rate $(1 - \hat{\beta})$, for the comparison of samples from the second density from the target density pairs against the same density translated by $(0.6, 0)$. The two sample sizes considered are $n = 1000, 10,000$. The fifth column $(\gamma = 1.0)$ is the base case with the optimal bandwidths $\hat{\mathbf{H}}_1, \hat{\mathbf{H}}_2$. The other columns are with the bandwidths $\hat{\mathbf{H}}_1^\gamma, \hat{\mathbf{H}}_2^\gamma$. Since $|\hat{\mathbf{H}}_1|, |\hat{\mathbf{H}}_2| < 1$, then using $\gamma > 1$ leads to undersmoothing and $\gamma < 1$ leads to oversmoothing.

In the upper half of Table 2, for $\gamma = 1, 1.5$, the bandwidths $\hat{\mathbf{H}}_1^\gamma, \hat{\mathbf{H}}_2^\gamma$ exhibit the closest empirical levels of significance to the nominal level $\alpha = 0.05$. Though we note that for all $\gamma > -1$, the empirical levels are less than 0.05, indicating that undersmoothing and mild oversmoothing do not over-estimate the level of significance. Only for the highly oversmoothed case $(\gamma = -1)$ do the empirical levels exceed the nominal level. In the lower half of Table 2, for $\gamma = 0, 0.5, 1, 1.5$, the bandwidths $\hat{\mathbf{H}}_1^\gamma, \hat{\mathbf{H}}_2^\gamma$ exhibit good power. It appears that moderate under- or oversmoothing does not affect power, though the highly undersmoothed $\gamma = 2, 2.5, 3$ and oversmoothed cases $\gamma = -1, -0.5$ result in poor power performance. We recall that the optimal bandwidths used here are computed according to Duong and Hazelton (2003) and minimise the squared error for density estimation, which does not necessarily imply that they would be optimal for local significance testing. Fortuitously, these bandwidth matrices are contained in a range of bandwidth matrices which yield good empirical level of significance and power properties.

3.2. *Univariate simulated data*

The theory presented in the previous section is not restricted to only multivariate data. Upon inspection, it is straightforward to simplify these multivariate results to univariate data. To verify

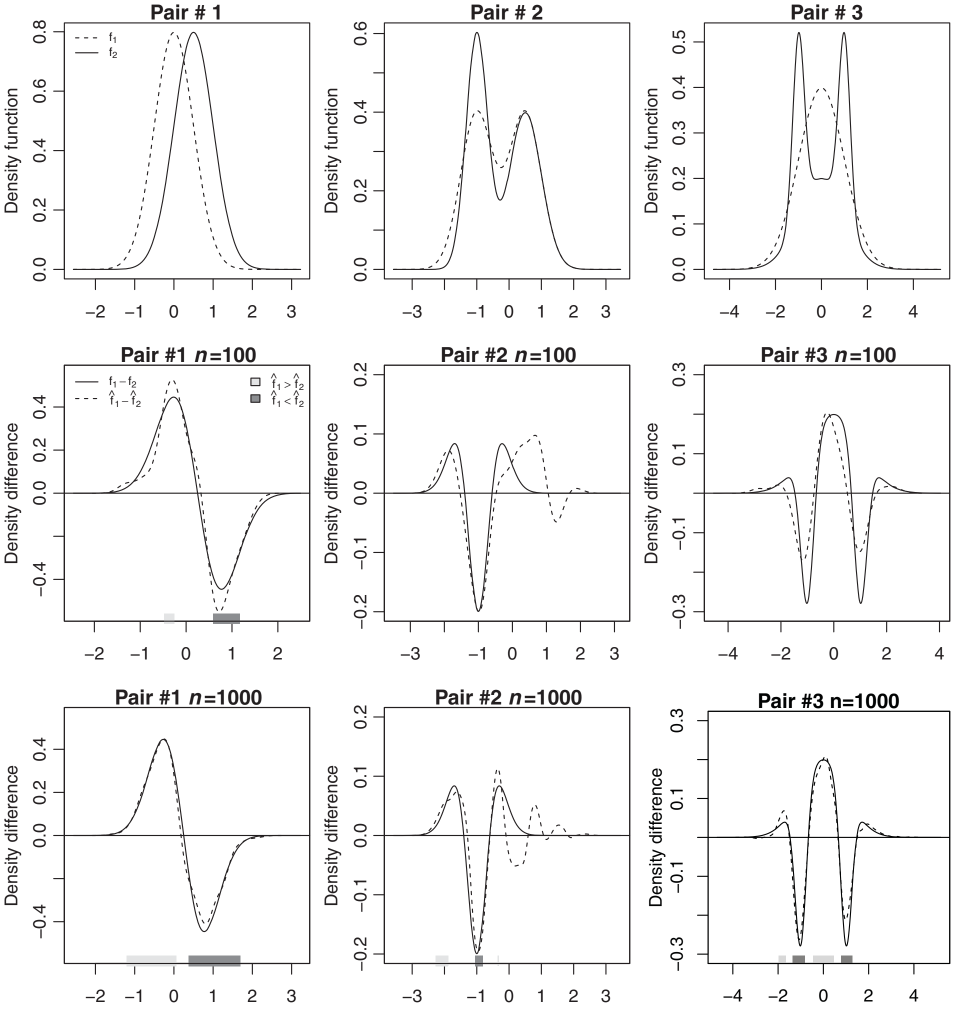

Figure 3. Comparison of kernel-based and PRIM-based locally significant difference regions, at $\alpha = 0.05$ level of significance, for univariate target density pairs at sample sizes $n = 100, 1000$. The first row are contour plots for the target density pairs. Dashed lines: density #1. Solid lines: density #2. In the subsequent rows, the true density differences are the solid lines, and the estimated density differences are the dashed lines. The difference regions (horizontal bars on the horizontal axis) are coloured as follows: light grey, density #1 > density #2; white, density #1 = density #2; dark grey, density #1 < density #2.

the finite sample behaviour, we have taken one-dimensional versions of the bivariate density pairs which we considered earlier. Pair #1 is $N(0, 1/2)$ versus $N(1/2, 1/2)$, two normal shifted densities. Pair #2 is $1/2N(1/2, 1/2) + 1/2N(-1, 1/2)$ versus $1/2N(1/2, 1/2) + 1/2N(-1, 1/3)$, two bimodal densities with one similar component and one differing component. Pair #3 is $N(0, 1)$ versus $1/2N(0, 1) + 1/4N(1, 1/4) + 1/4N(-1, 1/4)$ which have (approximately) overall zero mean and unit variance but with different internal structure. The plots of these density pairs are given in the first row of Figure 3. For sample sizes $n = 100, 1000$, in the subsequent two rows are the true density differences $f_1 - f_2$ (solid lines) and the estimated density differences $\hat{f}_1 - \hat{f}_2$ (dashed lines). The local difference regions are plotted as a rug-like plot: the intervals on the x-axis are coloured in light grey for $\hat{f}_1 > \hat{f}_2$ and dark grey for $\hat{f}_1 < \hat{f}_2$. Overall, we deduce that whilst $n = 100$ is sufficient for reasonable density estimation, it is not sufficient in these cases to estimate reliably local significant difference regions. Therefore, focusing on the latter sample size $n = 1000$, we see that the local significant regions correspond closely to local extrema (peaks and valleys) in the individual densities, as expected. We note that they also appear in the 'shoulders' of the densities where there are no local extrema, for example, the leftmost light grey regions for pairs #2 and #3 where large differences in probability mass are present.

3.3. *Multivariate real data*

There is vast interest in the identification of biomarkers in biological/biomedical data. A biomarker is the characteristic signature of a disease, mutation, effect of drug treatments, etc. which distinguishes it from the control sample. A commonly used technology for biomarker discovery is flow cytometry where the fluorescence properties of marker proteins inside cells are measured.

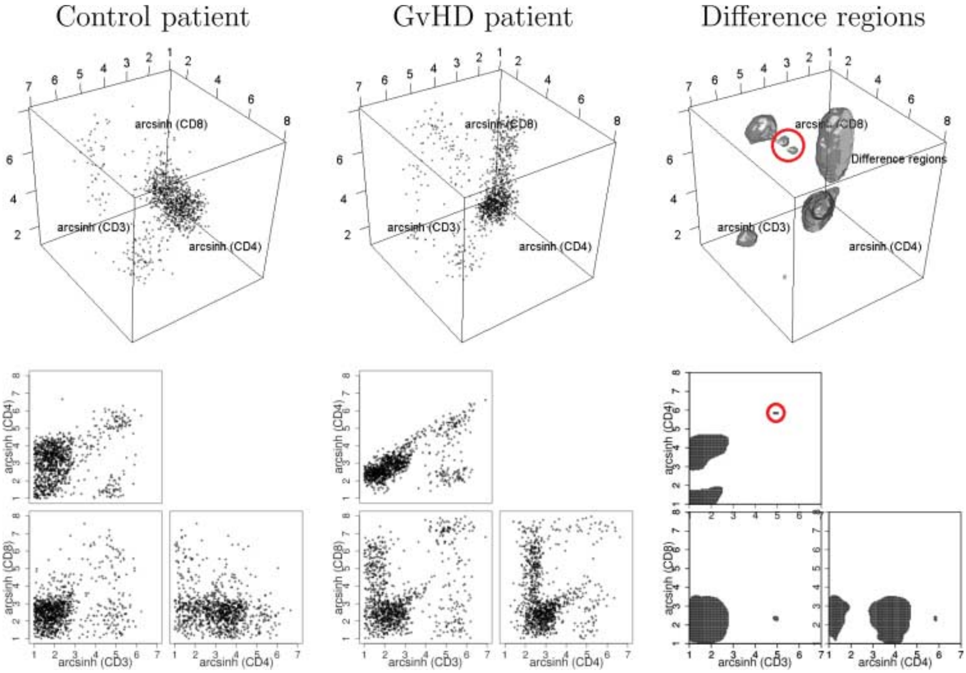

Figure 4. GvHD (graft-versus-host disease) biomarker discovery for (CD3, CD8) flow cytometry fluorescence measurements. Left: scatter plot for a control patient. Centre: scatter plot for a GvHD patient. Right: The $\alpha = 0.05$ local difference regions where GvHD patient levels are significantly higher. In the upper right difference regions (circled) of elevated levels of all (CD3, CD4, CD8) antibodies is a potential biomarker for the disease.

These marker proteins are proxies for the presence of specific cellular structures. We take the measurements from a control patient and a patient with GvHD (graft-versus-host disease), 32 days after each patient received a bone marrow transplant. GvHD occurs when the immune cells in the grafted bone marrow begin to attack the tissues in the host recipient. Since CD3, CD4 and CD8 are marker proteins associated with immune responses, they are good candidates to be GvHD biomarkers. As is usual in flow cytometry data analysis, the fluorescence levels are transformed using an inverse hyperbolic sine and pre-processed to remove the measurements from dead cells. This leaves a control patient sample consisting of $n_1 = 7566$ cells and a GvHD patient sample of $n_2 = 10,142$ cells. A different version of this data set has been already analysed in Chacón, Duong and Wand (2011), who in turn received the data originally from Brinkman et al. (2007). In Figure 4, on the left is the scatter plot of a subsample of 1000 cells from a control patient and in the centre that from a GvHD patient. On the right are the locally significant difference regions. The most interesting region for biomarker discovery is in the upper right (as indicated by the circle) since it is isolated from the 'main' sample. This indicates a relative enrichment of all CD3, CD4 and CD8 antibodies levels concurrently in a GvHD patient with respect to a control patient. This concurrent enrichment in CD3, CD4 and CD8 is a potential biomarker for GvHD. In the long term, drugs which block the action of CD3, CD4 and CD8 could thus potentially play a role in treating this disease.

Acknowledgements

The author acknowledges the financial support from grants at the Curie Institute, Paris, France, and the Institute of Translational Neurosciences (IHU-A-ICM WP2), Paris, France.

References

Alba Fernández, V., Jiménez Gamero, M.D., and Muñoz García, J. (2008), 'A Test for the Two-Sample Problem Based on Empirical Characteristic Functions', *Computational Statistics and Data Analysis*, 52, 3730–3748.

Allen, D.L. (1997), 'Hypothesis Testing Using an L_1-Distance Bootstrap', *American Statistician*, 51, 145–150.

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', *Journal of Multivariate Analysis*, 50, 41–54.

Baringhaus, L., and Franz, C. (2004), 'On a New Multivariate Two-Sample Test', *Journal of Multivariate Analysis*, 88, 190–206.

Biau, G., and Györfi, L. (2005), 'On the Asymptotic Properties of a Nonparametric L_1-Test Statistic of Homogeneity', *IEEE Transactions on Information Theory*, 51, 3965–3973.

Bickel, P.J. (1969), 'A Distribution Free Version of the Smirnov Two-Sample Test in the p-Variate Case', *Annals of Mathematics Statistics*, 40, 1–23.

Borgwardt, K., Gretton, A., Rasch, M., Kriegel, H.P., Schölkopf, B., and Smola, A. (2006), 'Integrating Structured Biological Data by Kernel Maximum Mean Discrepancy', *Bioinformatics*, 51, 49–57.

Brinkman, R.R., Gasparetto, M., Lee, S.-J.J., Ribickas, A.J., Perkins, J., Janssen, W., Smiley, R., and Smith, C. (2007), 'High-Content Flow Cytometry and Temporal Data Analysis for Defining a Cellular Signature Graft-Versus-Host Disease', *Biology of Blood and Marrow Transplantation*, 13, 691–700.

Cao, R., and Van Keilegom, I. (2006), 'Empirical Likelihood Ratio Tests for Two-Sample Problems via Non-Parametric Density Estimation', *Canadian Journal of Statistics*, 34, 61–77.

Chacón, J.E., Duong, T., and Wand, M.P. (2011), 'Asymptotics for General Multivariate Kernel Density Derivative Estimators', *Statistica Sinica*, 21, 807–840.

Cox, M.K., and Key, C.H. (1993), 'Post hoc Pair-Wise Comparisons for the Chi-Square Test of Homogeneity of Proportions', *Educational and Psychological Measurement*, 53, 951–962.

Duong, T. (2007), 'ks: Kernel Density Estimation and Kernel Discriminant Analysis for Multivariate Data in R', *Journal of Statistical Software*, 21(7), 1–16.

Duong, T., and Hazelton, M.L. (2003), 'Plug-in Bandwidth Matrices for Bivariate Kernel Density Estimation', *Journal of Nonparametric Statistics*, 15, 17–30.

Duong, T., Cowling, A., Koch, I., and Wand, M.P. (2008), 'Feature Significance for Multivariate Kernel Density Estimation', *Computational Statistics and Data Analysis*, 52, 4225–4242.

Duong, T., Koch, I., and Wand, M.P. (2009), 'Highest Density Difference Region Estimation with Application to Flow Cytometric Data', *Biometrical Journal*, 51, 504–521.

Duong, T., Goud, B., and Schauer, K. (2012), ‘First Closed-Form Density-Based Framework for Automatic Detection of Cellular Morphology Changes’, *Proceedings of the National Academy of Sciences*, 109, 8382–8387.

Friedman, J.H., and Rafsky, L.C. (1979), ‘Multivariate Generalizations of the Wald–Wolfowitz and Smirnov 2-Sample Tests’, *Annals of Statistics*, 7, 697–717.

Gibbons, J.D., and Chakraborti, S. (2003), *Nonparametric Statistical Inference* (4th ed.), New York: Marcel Dekker.

Hall, P., and Wand, M.P. (1988), ‘On Nonparametric Discrimination Using Density Differences’, *Biometrika*, 75, 541–547.

Hochberg, Y. (1988), ‘A Sharper Bonferroni Procedure for Multiple Tests of Significance’, *Biometrika*, 75, 800–802.

Li, Q. (1999), ‘Nonparametric Testing the Similarity of Two Unknown Density Functions: Local Power and Bootstrap Analysis’, *Journal of Nonparametric Statistics*, 11, 189–213.

Liu, R.Y., and Singh, K. (1993), ‘A Quality Index Based on Data Depth and Multivariate Rank-Tests’, *Journal of the American Statistical Association*, 88, 252–260.

Louani, D. (2000), ‘Exact Bahadur Efficiencies for Two-Sample Statistics in Functional Density Estimation’, *Statistics & Decisions*, 18, 389–412.

Martínez-Camblor, P., De Uña-Álvarez, J., and Corral, N. (2008), ‘*k*-Sample Test Based on the Common Area of Kernel Density Estimators’, *Journal of Statistical Planning and Inference*, 138, 4006–4020.

Molanes-López, E.M., and Cao, R. (2008), ‘Plug-in Bandwidth Selector for the Kernel Relative Density Estimator’, *Annals of the Institute of Statistical Mathematics*, 60, 273–300.

Nel, D.G., and van der Merwe, C.A. (1986), ‘A Solution to the Multivariate Behrens–Fisher Problem’, *Communications in Statistics. A. Theory and Methods*, 15, 3719–3735.

Parzen, E. (1962), ‘On Estimation of a Probability Density Function and Mode’, *Annals of Mathematical Statistics*, 33, 1065–1076.

Roederer, M., and Hardy, R.R. (2001), ‘Frequency Difference Gating: A Multivariate Method for Identifying Subsets That Differ Between Samples’, *Cytometry A*, 45, 56–64.

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley & Sons Inc.

Simonoff, J.S. (1996), *Smoothing Methods in Statistics*, New York: Springer-Verlag.

Sugiyama, M., Kanamori, T., Suzuki, T., du Plessis, M.C., Liu, S., and Takeuchi, I. (2012), ‘Density-Difference Estimation’, *Advances in Neural Information Processing Systems*, 25, 692–700.

Wand, M.P. (1992), ‘Error Analysis for General Multivariate Kernel Estimators’, *Journal of Nonparametric Statistics*, 2, 1–15.

Wand, M.P. (1994), ‘Fast Computation of Multivariate Kernel Estimators’, *Journal of Computational and Graphical Statistics*, 3, 433–445.