Journal of the Korean Statistical Society 45 (2016) 33–50

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Non-parametric smoothed estimation of multivariate cumulative distribution and survival functions, and receiver operating characteristic curves

Tarn Duong *

Sorbonne Universities, University Pierre and Marie Curie—Paris 6, Theoretical and Applied Statistics Laboratory (LSTA), UR 1, F-75005, Paris, France

ARTICLE INFO

Article history:
Received 14 November 2014
Accepted 29 June 2015
Available online 19 August 2015

AMS 2000 subject classifications:
primary 62G05
secondary 62H10

Keywords:
Asymptotic mean integrated squared error
Diagnostic test
Optimal bandwidth matrix selection
Quantile
ROC curve

ABSTRACT

A unified framework to analyse multivariate kernel estimators of distribution and survival functions is introduced, before turning our attention to receiver operating characteristic (ROC) curves. These are well-established visual analytic tools for univariate data samples, though their generalisation to multivariate data has been limited. Since non-parametric multivariate kernel smoothing methods possess excellent visualisation properties, they serve as a solid basis for their estimation. With optimal data-based bandwidth matrix selectors, we demonstrate that they possess suitable properties for exploratory data analysis of simulated and experimental data.

1. Introduction

A basic problem in multivariate data analysis is estimating cumulative distribution functions, though there has been a relative paucity of their analysis as compared to density functions. The former are however important in a wide range of data analytic situations. We set up a unified framework to treat kernel estimators of the multivariate distribution functions and the closely related survival functions, since kernel estimators are widely used in non-parametric data smoothing, see Simonoff (1996) and Wand and Jones (1995) for an overview. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d$ be a d-variate random variable with distribution F and density f. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$, then we define the cumulative distribution of $\boldsymbol{X}$ to be

$$F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x}) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_d \leq x_d) = \int_{-\infty}^{\boldsymbol{x}} f(\boldsymbol{w})\, d\boldsymbol{w}$$

where $\int_{-\infty}^{\boldsymbol{x}} d\boldsymbol{w}$ is an abbreviation of $\int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} dw_1 \cdots dw_d$. The survival function is defined as $\bar{F}(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} > \boldsymbol{x})$, complementary to the cumulative distribution function $F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x})$. The usual relation $\bar{F}(x) = 1 - F(x)$ for univariate data does not hold for multivariate data since the hyper-rectangles $\{\boldsymbol{w} \in \mathbb{R}^d : \boldsymbol{w} \leq \boldsymbol{x}\} \cup \{\boldsymbol{w} \in \mathbb{R}^d : \boldsymbol{w} > \boldsymbol{x}\} \neq \mathbb{R}^d$ in general.

These functions are brought together in the analysis of receiver operating characteristic (ROC) curves. ROC curves were introduced in the context of signal detection, e.g. Peterson, Birdsall, and Fox (1954), though they have been subsequently

* Correspondence to: Sorbonne Paris City, University Paris-North—Paris 13, Computer Science Laboratory (LIPN), CNRS UMR 7030, F-93430, Villetaneuse, France.
E-mail address: duong@lipn.univ-paris13.fr.

http://dx.doi.org/10.1016/j.jkss.2015.06.002
1226-3192/

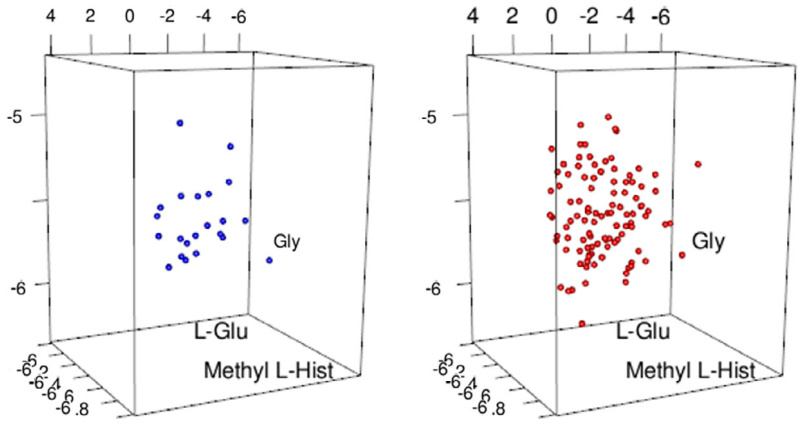

Fig. 1. Scatterplots for Spinal Muscular Atrophy (SMA) data set. The variables are the negative log concentrations (μM) of Glycine (Gly), L-glutamic (L-Glu) and 1-methyl-L-histidine (methyl L-Hist). (Left) Scatterplot for 22 age-matched healthy control children. (Right) Scatterplot for 108 SMA afflicted children.

widely adopted in many different contexts, whenever the values of a second population tend to be greater than those from the first one. The standard definition of ROC curves allows only for scalar valued diagnostic variables whereas the ability to handle multivariate data would be beneficial in many circumstances. Following Handcock and Morris (1998), Hsieh and Turnbull (1996) and Lloyd (1998), instead of comparing the vector of diagnostic variables to a threshold component-wise directly, we apply the survival function $\bar{F}_{\boldsymbol{X}_1}$ from the first population as a pre-transformation. This leads to us to define a multivariate ROC curve as the graph

$$\{(\mathbb{P}(\bar{F}_{\boldsymbol{X}_1}(\boldsymbol{X}_1) > \bar{F}_{\boldsymbol{X}_1}(\boldsymbol{x})), \mathbb{P}(\bar{F}_{\boldsymbol{X}_1}(\boldsymbol{X}_2) > \bar{F}_{\boldsymbol{X}_1}(\boldsymbol{x}))) : \boldsymbol{x} \in \mathbb{R}^d\}.$$

We use this definition rather than the seemingly more straightforward generalisation from the univariate case $\{(\mathbb{P}(\boldsymbol{X}_1 > \boldsymbol{x}), \mathbb{P}(\boldsymbol{X}_2 > \boldsymbol{x})) : \boldsymbol{x} \in \mathbb{R}^d\}$ which is not a well-defined multivariate function, whereas the above ROC curve is monotonic by construction since it is a quantile–quantile plot. This approach is an alternative to current methods such as the dimension reduction via a weighted vector norm of Pepe and Thompson (2000) and Su and Liu (1993) or the singular value decomposition combined with likelihood ratios of Pfeiffer and Bura (2008); or the logistic regression modelling of Pepe (1998). The reader interested in a more comprehensive review is invited to consult Shapiro (1999) and the references contained therein. One of the main advantages of our proposed approach is that it does not require parametric assumptions on the underlying random variables or on the transformation, and so is an ideal candidate within a non-parametric smoothing framework.

A data set for which a multivariate ROC curve analysis would be beneficial is the Pilot Study of Biomarkers for Spinal Muscular Atrophy (BforSMA), available from http://neuinfo.org/smabiomarkers. The full database contains a large variety of measurements taken from a cohort of 130 children aged between 2 and 12 years, with 108 children with genetically confirmed Spinal Muscular Atrophy (SMA) and 22 aged-matched healthy controls. We take a subset of these data identified in Finkel et al. (2012), as potential biomarkers for SMA, namely the (negative log) concentrations of the amino acids Glycine (Gly), L-glutamic (L-Glu) and 1-methyl-L-histidine (methyl L-Hist). The 3-dimensional scatterplots in Fig. 1 give a visual impression that the point cloud of SMA patients have generally higher biomarker values than the control patients, and a ROC curve analysis would visualise and quantify the joint diagnostic efficacy of this biomarker combination.

Our goal is to develop fully multivariate kernel estimators of ROC curves the construction of the ROC curve via a novel combination of kernel estimators of cumulative distribution and survival functions. In Section 2, we set up a framework for the squared error analysis of kernel estimators of multivariate cumulative distribution and survival functions, and most crucially the development of data-based optimal bandwidth selectors. This supporting section is the basis for the data-based implementation for the estimators of ROC curves in Section 3. In Section 4, we demonstrate the efficacy of these kernel estimators for simulated and experimental data. The last section is a discussion, and Appendix contains the mathematical proofs of the results stated in the main text.

2. Cumulative distribution and survival functions

In this section, we introduce a class of plug-in estimators of the cumulative distribution and survival functions. The usual kernel estimator of the cumulative distribution function F is

$$\hat{F}(\boldsymbol{x};\ \mathbf{H}) = n^{-1} \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i) \tag{1}$$

where $\mathcal{K}(\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}} K(\boldsymbol{w})\, d\boldsymbol{w}$ for a multivariate kernel function K, the scaled integrated kernel is $\mathcal{K}_{\mathbf{H}}(\boldsymbol{x}) = \mathcal{K}(\mathbf{H}^{-1/2}\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}} |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\boldsymbol{w})\, d\boldsymbol{w}$, and $\mathbf{H}$ is the bandwidth matrix. This form as the integral of the classical kernel density estimator

was proposed by Watson and Leadbetter (1964) in the framework of hazard function estimation. Subsequent analysis in the univariate case of these kernel distribution estimators in their own right includes Azzalini (1981), Nadaraya (1964), Winter (1973) and Yamato (1973), with Molanes-López and Cao (2008) considering data-based selection of the bandwidth. The first appearance of multivariate kernel distribution estimators is Jin and Shao (1999) who considered the bandwidth matrix as a scalar multiple of the identity matrix. This was later extended to diagonal bandwidth matrices by Liu and Yang (2008). Eq. (1) contains the most general form of a fixed bandwidth kernel estimator since $\mathbf{H}$ is allowed to be any positive definite symmetric $d \times d$ matrix, which we call an unconstrained bandwidth matrix. The novel contribution in this section is the development of data-based selectors of these unconstrained matrices.

The kernel estimator of the survival function is analogously defined as

$$\hat{\bar{F}}(\boldsymbol{x};\ \mathbf{H}) = n^{-1} \sum_{i=1}^{n} \bar{\mathcal{K}}_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i) = n^{-1} \sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_i - \boldsymbol{x}) \tag{2}$$

where $\bar{\mathcal{K}}(\boldsymbol{x}) = \int_{\boldsymbol{x}}^{\infty} K(\boldsymbol{w})\, d\boldsymbol{w} = \int_{-\infty}^{-\boldsymbol{x}} K(\boldsymbol{w})\, d\boldsymbol{w} = \mathcal{K}(-\boldsymbol{x})$ for K a symmetric kernel. The construction of $\hat{\bar{F}}$ involves the reversal of the difference $\boldsymbol{x} - \boldsymbol{X}_i$ in $\hat{F}$, so many of the properties of the latter are correspondingly transferred so we focus on developing theoretical arguments on the former.

The results in this paper rely on the following assumptions. They do not form a minimal set, but they serve as convenient starting point to develop our understanding of these kernel estimators.

(A1) For the distribution function F, all partial derivatives up to order 2 are bounded, continuous and square integrable, and of order 4 exist; and F does not depend on n.
(A2) The kernel K is a positive, symmetric, square integrable density function such that $\int_{\mathbb{R}^d} \boldsymbol{x}\boldsymbol{x}^T K(\boldsymbol{x})\, d\boldsymbol{x} = m_2(K)\mathbf{I}_d$ for some real number $m_2(K)$ with $\mathbf{I}_d$ the $d \times d$ identity matrix.
(A3) The integrated kernel $\mathcal{K}$ is such that $\boldsymbol{m}_1(K\mathcal{K}) = \int_{\mathbb{R}^d} \boldsymbol{x} K(\boldsymbol{x})\mathcal{K}(\boldsymbol{x})\, d\boldsymbol{x}$ has a finite L_2 norm.
(A4) The bandwidth matrix $\mathbf{H} = \mathbf{H}(n)$ forms a sequence of symmetric and positive-definite matrices such that every element of $\mathbf{H} \to 0$ as $n \to \infty$.

Theorem 1. *Suppose that the conditions* (A1)–(A4) *hold. As* $n \to \infty$, *the mean integrated squared error (MISE) is*

$$\operatorname{MISE} \hat{F}(\cdot;\ \mathbf{H}) = \int_{\mathbb{R}^d} \operatorname{MSE} \hat{F}(\boldsymbol{x};\ \mathbf{H})\, d\boldsymbol{x} = \{\operatorname{AMISE} \hat{F}(\cdot;\ \mathbf{H})\}\{1 + o(1)\}$$

where the asymptotic MISE is

$$\operatorname{AMISE} \hat{F}(\cdot;\ \mathbf{H}) = n^{-1}V_1(F) - 2n^{-1}\boldsymbol{m}_1(K\mathcal{K})^T \mathbf{H}^{1/2}\mathbf{1}_d - \tfrac{1}{4}m_2(K)^2(\operatorname{vec}^T \mathbf{H}^2)\boldsymbol{\psi}_2,$$

$V_1(F) = \int_{\mathbb{R}^d} F(\boldsymbol{x})(1 - F(\boldsymbol{x}))\, d\boldsymbol{x}$, $\boldsymbol{\psi}_2 = \int_{\mathbb{R}^d} \operatorname{vec} \mathrm{D}^2 f(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}$, D^2 *is the Hessian matrix operator of second order mixed partial differentials, and* $\mathbf{1}_d$ *is the d-vector of all ones. Equivalent expressions for the survival function estimator are obtained where* $\hat{\bar{F}}, \bar{F}$ *replace* $\hat{F}, F$.

As already indicated by various authors, e.g. Azzalini (1981) and Reiss (1981), the asymptotic variance of $\hat{F}$ is $n^{-1}V_1(F) - 2n^{-1}\boldsymbol{m}_1(K\mathcal{K})^T \mathbf{H}^{1/2}\mathbf{1}_d$ which is smaller than the variance of the empirical cumulative distribution estimator, $n^{-1}V_1(F)$, as $\boldsymbol{m}_1(K\mathcal{K})^T \mathbf{H}^{1/2}\mathbf{1}_d > 0$ since K is a non-negative density function.

As is usual for kernel estimators, the selection of the bandwidth is an important factor in their performance. For the AMISE from Theorem 1, we observe that letting $\mathbf{H}$ to be large leads to decreasing variance, though at the expense of inflating the squared bias, recalling that $(\operatorname{vec}^T \mathbf{H}^2)\boldsymbol{\psi}_2 < 0$. This is usually known as oversmoothing. On the other hand, letting $\mathbf{H}$ to be small leads to decreasing bias with correspondingly inflated variance, which is typical of undersmoothing. An optimal selector is a bandwidth which finds a trade-off between under- and over-smoothing: the oracle optimal selector is $\mathbf{H}_{\mathrm{MISE}} = \operatorname{argmin}_{\mathbf{H} \in \mathcal{F}} \operatorname{MISE} \hat{F}(\cdot;\ \mathbf{H})$ where $\mathcal{F}$ is the space of all positive definite symmetric $d \times d$ matrices.

Theorem 2. *Suppose that conditions* (A1)–(A4) *hold. As* $n \to \infty$, *the oracle optimal bandwidth matrix* $\mathbf{H}_{\mathrm{MISE}} = O(n^{-2/3})\mathbf{J}_d$ *where* $\mathbf{J}_d$ *is the* $d \times d$ *matrix of all ones. The minimal MISE is* $\inf_{\mathbf{H} \in \mathcal{F}} \operatorname{MISE}[\hat{F}(\cdot;\ \mathbf{H})] = n^{-1}V_1(F) + O(n^{-4/3})$.

This order of $\mathbf{H}_{\mathrm{MISE}}$ agrees with univariate results for the unweighted MISE of Azzalini (1981) and the weighted MISE of Altman and Léger (1995), i.e. $\int_{\mathbb{R}} \{\operatorname{MSE} \hat{F}(x; h)\} w(x) f(x)\, dx$ for an arbitrary weighting function w. Thus the choice of the weighting function does not affect the asymptotic order of the optimal bandwidth matrix. Furthermore, this order for the unconstrained bandwidth $\mathbf{H}_{\mathrm{MISE}}$ agrees with the multivariate case of Liu and Yang (2008) for constrained bandwidth matrices. We note that Liu and Yang (2008) exhibit more general results suitable for dependent samples and for higher order kernels: condition (A2) implies that our results hold only for second order kernels. The minimal MISE rate obtained by Altman and Léger (1995) is order $n^{-8/9}$. This is slower than what we obtain because they consider estimation of weighted functionals of higher order than the second order functional $\boldsymbol{\psi}_2$ which we consider. Our case is a special case where the weighting function w is inversely proportional to the density f, affording a faster rate of convergence.

The oracle optimal bandwidth $\mathbf{H}_{\mathrm{MISE}}$ is mathematically intractable in general: more tractable is its asymptotic proxy $\mathbf{H}_{\mathrm{AMISE}} = \operatorname{argmin}_{\mathbf{H}\in\mathcal{F}} \mathrm{AMISE}\, \hat{F}(\cdot;\, \mathbf{H})$. The order $n^{-4/3}$ convergence of $\mathbf{H}_{\mathrm{AMISE}}$ to $\mathbf{H}_{\mathrm{MISE}}$ (whose proof is omitted for brevity) implies that the AMISE is an appropriate proxy in place of the MISE as the discrepancy of $\mathbf{H}_{\mathrm{MISE}}$ and $\mathbf{H}_{\mathrm{AMISE}}$ is asymptotically smaller than $\mathbf{H}_{\mathrm{MISE}}$. The tractability of $\mathbf{H}_{\mathrm{AMISE}}$ allows it to be estimated more easily, and we say that $\hat{\mathbf{H}} = \operatorname{argmin}_{\mathbf{H}\in\mathcal{F}} \widehat{\mathrm{AMISE}}\, \hat{F}(\cdot;\, \mathbf{H})$ is a data-based estimator of $\mathbf{H}_{\mathrm{AMISE}}$, where $\widehat{\mathrm{AMISE}}$ is an estimator of AMISE. Following Duong and Hazelton (2005), the relative rate of convergence to $\hat{\mathbf{H}}$ to $\mathbf{H}_{\mathrm{AMISE}}$ is $O_P(n^{-\alpha})$ if $\hat{\mathbf{H}} - \mathbf{H}_{\mathrm{AMISE}} = O_P(n^{-\alpha}\mathbf{J}_{d^2})\mathrm{vec}\, \mathbf{H}_{\mathrm{AMISE}}$. Furthermore it follows that the rate of $\hat{\mathbf{H}}$ to $\mathbf{H}_{\mathrm{MISE}}$ will remain $n^{-\alpha}$ whenever $\alpha < 4/3$, from Theorem 2.

Different data-based selectors arise from different estimators of the AMISE. For consistent estimation, the key quantity is the integrated functional $\boldsymbol{\psi}_2$. Denote by $\mathsf{D} = \partial/\partial\boldsymbol{x} = (\partial/\partial x_1, \ldots, \partial/\partial x_d)$ the first derivative (gradient) operator. If the usual convention $(\partial/\partial x_i)(\partial/\partial x_j) = \partial^2/(\partial x_i \partial x_j)$ is taken into account, the rth order differential operator is rth fold Kronecker product of D with itself, $\mathsf{D}^{\otimes r}$, see Holmquist (1996). For a general r, let $\boldsymbol{\psi}_{2r} = \mathbb{E}\mathsf{D}^{\otimes 2r} f(\boldsymbol{X}) = \int_{\mathbb{R}^d} \mathsf{D}^{\otimes 2r} f(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}$. The usual kernel estimator is

$$\hat{\boldsymbol{\psi}}_{2r}(\mathbf{G}) = n^{-1} \sum_{i=1}^{n} \mathsf{D}^{\otimes r} L_{\mathbf{G}}(\boldsymbol{X}_i) = n^{-2} \sum_{i,j=1}^{n} \mathsf{D}^{\otimes r} L_{\mathbf{G}}(\boldsymbol{X}_i - \boldsymbol{X}_j)$$

where L and $\mathbf{G}$ are possibly different kernel and bandwidth from K and $\mathbf{H}$. The plug-in estimator of AMISE is obtained by replacing $\boldsymbol{\psi}_2$ by $\hat{\boldsymbol{\psi}}_2(\mathbf{G})$,

$$\mathrm{PI}(\mathbf{H};\, \mathbf{G}) = n^{-1} V_1(F) - 2n^{-1} \boldsymbol{m}_1(K\mathcal{K})\, \mathbf{H}^{1/2} \mathbf{1}_d - \tfrac{1}{4} m_2(K)^2 (\mathrm{vec}^T\, \mathbf{H})^{\otimes 2} \hat{\boldsymbol{\psi}}_2(\mathbf{G}). \tag{3}$$

The plug-in bandwidth selector is

$$\hat{\mathbf{H}}_{\mathrm{PI}} = \operatorname*{argmin}_{\mathbf{H}\in\mathcal{F}} \{\mathrm{PI}(\mathbf{H};\, \mathbf{G}) - n^{-1} V_1(F)\} \tag{4}$$

where we have subtracted $n^{-1}V_1(F)$ since it does not involve the bandwidth. Thus the problem becomes the selection of this pilot matrix $\mathbf{G}$.

Let the optimal bandwidth be $\mathbf{G}_{\mathrm{PI}} = \operatorname{argmin}_{\mathbf{G}\in\mathcal{F}} \mathrm{MSE}\hat{\boldsymbol{\psi}}_2(\mathbf{G})$ where $\mathrm{MSE}\, \hat{\boldsymbol{\psi}}_2(\mathbf{G}) = \mathbb{E}\|\hat{\boldsymbol{\psi}}_2(\mathbf{G}) - \boldsymbol{\psi}_2\|^2$. According to Chacón and Duong (2010, Theorem 2), $\mathbf{G}_{\mathrm{PI}}$ is order $n^{-2/(d+4)}$. Furthermore, these authors outline a multi-stage selection algorithm improves on the zero-stage selectors (normal scale selectors) of Altman and Léger (1995), Hall and Hyndman (2003) and Liu and Yang (2008), and the 1-stage selector of Zhou and Harezlak (2002) and generalises the multi-stage univariate plug-in selectors of Polansky and Baker (2000) for $d > 1$.

With the PI criterion consistently estimated, the next theorem establishes the convergence in probability of $\hat{\mathbf{H}}_{\mathrm{PI}}$ to the oracle selector $\mathbf{H}_{\mathrm{MISE}}$.

Theorem 3. *Suppose that the conditions* (A1)–(A4) *hold. Suppose that* (A1) *also holds for kernel L. The relative convergence rate of* $\hat{\mathbf{H}}_{\mathrm{PI}}$ *to* $\mathbf{H}_{\mathrm{MISE}}$ *is* $n^{-2/(d+4)}$.

This reduces for $d = 1$ to the $n^{-2/5}$ rate for a 1-stage plug-in selector of Polansky and Baker (2000). We are not able to reproduce their $n^{-1/2}$ rate for 2-stage selectors for $d > 1$ since this relies on a bias annihilation argument that is not possible for multivariate MSE expressions, see Chacón and Duong (2010). Furthermore, we observe that $\mathrm{AMISE}\, \hat{F}(\cdot;\, \mathbf{H}) - n^{-1}V_1(F) = \mathrm{AMISE}\, \hat{\bar{F}}(\cdot;\, \mathbf{H}) - n^{-1}V_1(\bar{F})$, implying that an AMISE-optimal selector for $\hat{F}$ is thus also optimal for $\hat{\bar{F}}$. These agree with the univariate case where it is more obvious since $\hat{\bar{F}}(x;\, h) = 1 - \hat{F}(x;\, h)$, see Berg and Politis (2006).

We do not consider cross validation methods, the main competitors to plug-in methods for data-based bandwidth selection, here since this would involve lengthy algebraic manipulations and leave them for future work for the purposes of brevity.

To conclude this section, we compare the potential gain in using an unconstrained bandwidth matrix over diagonal bandwidth matrices, as measured by the asymptotic relative error $\mathrm{ARE}(\mathcal{F} : \mathcal{D}) = \mathrm{AMISE}\, \{\hat{F}(\cdot;\, \mathbf{H} \in \mathcal{F})\}/\mathrm{AMISE}\, \{\hat{F}(\cdot;\, \mathbf{H} \in \mathcal{D})\}$. The diagonal selector is similar except that the optimisation ranges over the class of all positive definite diagonal matrices $\mathcal{D}$ instead of over all positive definite matrices $\mathcal{F}$.

Theorem 4. *Suppose that* $F = \Phi_\rho$ *is a bivariate normal distribution with variance* $[1, \rho;\, \rho, 1]$, *and* $K = \phi$ *is the normal kernel. These imply that* (A1)–(A3) *hold. Further suppose that the condition* (A4) *holds. The optimal matrix in* $\mathcal{F}$ *has the form* $\mathbf{H} = [h^2, h_{12};\, h_{12}, h^2]$ *and in* $\mathcal{D}$, *it has* $\mathbf{H} = [h^2, 0;\, 0, h^2]$. *The asymptotic relative error, as a function of the correlation coefficient* ρ, *is*

$$\mathrm{ARE}(\mathcal{F} : \mathcal{D};\, \rho) = \left[V_2(\rho) - 2^{1/2}\pi^{-1/2} n^{-1} \frac{h^2 + (h^4 - h_{12}^2)^{1/2} + h_{12}}{[h^2 + (h^4 - h_{12}^2)^{1/2}]^{1/2}} + \tfrac{1}{4}(4\pi)^{-d/2}(1-\rho^2)^{-1/2}(h_{12}^2 - 2\rho h^2 h_{12})\right] \Big/ \left[V_2(\rho) - 2\pi^{-1/2} n^{-1} h\right]$$

where $V_2(\rho) = n^{-1}V_1(\Phi_\rho) + \frac{1}{4}(4\pi)^{-d/2}(1-\rho^2)^{-1/2} h^4$.

The term $V_2(\rho)$ in Theorem 4 contains expressions which are common to the both matrix classes, so we focus our attention on the other terms to investigate when an unconstrained matrix leads to a decrease in the ARE. Without lost of generality, assume that $\rho > 0$. We can show that if we have $h_{12} > 0$, i.e. the unconstrained matrix also has positive correlation, then the terms immediately following $V_2(\rho)$ in the numerator and denominator respectively satisfy $2^{1/2}[h^2+(h^4-h_{12}^2)^{1/2}+h_{12}]/[h^2+(h^4-h_{12}^2)^{1/2}]^{1/2} > 2h$ which implies a reduction in the ARE. Furthermore, as ρ increases to 1, the third expression in the numerator $(h_{12}^2 - 2\rho h^2 h_{12})$ becomes increasingly negative, again implying a reduction in the ARE. The improvements in terms of these asymptotic relative errors are more modest than those observed for kernel estimators of density functions (Wand & Jones, 1993) and of derivatives of density functions (Chacón, Duong, & Wand, 2011). This is due to that the $O(n^{-1})$ dominant term in the AMISE of distribution estimators does not depend on the bandwidth matrix. The bandwidth modifies the AMISE only in the secondary terms of $O(n^{-1}\mathbf{H}^{1/2} + \mathbf{H}^2)$, unlike the case for density and density derivatives where the bandwidth modifies the dominant terms in their respective AMISE.

3. Receiver operating characteristic curves

As noted in the introduction, the univariate ROC curve is $\{(\bar{F}_{X_1}(x), \bar{F}_{X_2}(x)) : x \in \mathbb{R}\}$, whereas the straightforward multivariate generalisation $\{(\bar{F}_{\boldsymbol{X}_1}(\boldsymbol{x}), \bar{F}_{\boldsymbol{X}_2}(\boldsymbol{x})) : \boldsymbol{x} \in \mathbb{R}^d\}$ does not result in a well-defined function, motivating us to define a ROC curve as $\{(\mathbb{P}(\bar{F}_{\boldsymbol{X}_1}(\boldsymbol{X}_1) > \bar{F}_{\boldsymbol{X}_1}(\boldsymbol{x})), \mathbb{P}(\bar{F}_{\boldsymbol{X}_1}(\boldsymbol{X}_2) > \bar{F}_{\boldsymbol{X}_1}(\boldsymbol{x}))) : \boldsymbol{x} \in \mathbb{R}^d\}$. For an alternative definition which is more amenable for estimation, let $Y_j = \bar{F}_{\boldsymbol{X}_1}(\boldsymbol{X}_j),\ j = 1, 2$, then a multivariate ROC curve is

$$\{(F_{Y_1}(z), F_{Y_2}(z)) : z = \bar{F}_{\boldsymbol{X}_1}(\boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{R}^d\} \tag{5}$$

where F_{Y_j} is the cumulative distribution of Y_j, as

$$F_{Y_1}(z) = \mathbb{P}(Y_1 \le z) = \mathbb{P}(\bar{F}_{\boldsymbol{X}_1}(\boldsymbol{X}_1) > \bar{F}_{\boldsymbol{X}_1}(\boldsymbol{x}))$$
$$F_{Y_2}(z) = \mathbb{P}(Y_2 \le z) = \mathbb{P}(\bar{F}_{\boldsymbol{X}_1}(\boldsymbol{X}_2) > \bar{F}_{\boldsymbol{X}_1}(\boldsymbol{x}))$$

since $\bar{F}_{\boldsymbol{X}_1}$ is monotonically decreasing. The transformations $Y_j = \bar{F}_{\boldsymbol{X}_1}(\boldsymbol{X}_j)$ offer an alternative to the other dimension reduction transformations, e.g. rank 1 singular value decomposition or the linear scalar projection of Su and Liu (1993). One of their main advantages is that it does not require parametric assumptions on the underlying random variables $\boldsymbol{X}_1, \boldsymbol{X}_2$ or on the transformation.

We have two random samples $\mathcal{X}_1 = \{\boldsymbol{X}_{1,1}, \ldots, \boldsymbol{X}_{1,n_1}\}$ and $\mathcal{X}_2 = \{\boldsymbol{X}_{2,1}, \ldots, \boldsymbol{X}_{2,n_2}\}$, with each $\mathcal{X}_j$ sample being drawn from a common distribution function $F_{\boldsymbol{X}_j}, j = 1, 2$. As a visual tool, a ROC curve is well-suited to being cast in a kernel estimation framework, as first proposed by Zou, Hall, and Shapiro (1997). With kernel estimators for cumulative distribution and survival functions established, we are ready to define a kernel estimator of the components of a ROC curve based on Eq. (5) as

$$\hat{F}_{\hat{Y}_j}(z;\ \mathbf{H}_1, h_2) = n_j^{-1} \sum_{i=1}^{n_j} \mathcal{L}_{h_2}(z - \hat{Y}_{j,i}) \tag{6}$$

for a univariate kernel L with corresponding integrated kernel $\mathcal{L}$, and $\hat{Y}_{j,i} = \hat{\bar{F}}_{\boldsymbol{X}_1}(\boldsymbol{X}_{j,i};\ \mathbf{H}_1),\ i = 1, \ldots, n_j$. Recall that for univariate data X_1, X_2 that $F_{Y_1} \sim \text{Unif}[0, 1]$ always so no estimation is required, which implies that the multivariate problem is more involved since both F_{Y_1}, F_{Y_2} must be estimated.

We begin by focusing on estimating the true positive rate F_{Y_2}. A main result concerning a separable form of the mean squared error of $\hat{F}_{\hat{Y}_2}$ is stated next. This separability will greatly facilitate automatic bandwidth selection.

Theorem 5. *Suppose that the condition* (A1) *holds for the distribution* F_{Y_2}, (A2)–(A3) *for the kernels* K, L *and* $\mathcal{K}$, $\mathcal{L}$, *and* (A4) *for the bandwidths* $\mathbf{H}_1, h_2$. *The MISE of the estimator of the true positive rate* $\hat{F}_{\hat{Y}_2}(\cdot; \mathbf{H}_1, h_2)$ *has the conditionally separable form, as* $n_1, n_2 \to \infty$,

$$\text{MISE}\,[\hat{F}_{\hat{Y}_2}(\cdot;\ \mathbf{H}_1, h_2)] = \text{MISE}[\mathbb{E}\hat{F}_{\hat{Y}_2}(\cdot;\ \mathbf{H}_1)|\mathcal{X}_1] + \mathbb{E}[\text{MISE}\,\hat{F}_{\hat{Y}_2}(\cdot; h_2)|\mathcal{X}_1]$$

where

$$\text{MISE}\,[\mathbb{E}\hat{F}_{\hat{Y}_2}(\cdot;\ \mathbf{H}_1)|\mathcal{X}_1] = \int_{\mathbb{R}^d} \text{MSE}[\hat{F}_{\boldsymbol{X}_1}(\boldsymbol{x};\ \mathbf{H}_1)] f_{\boldsymbol{X}_2}(\boldsymbol{x})^2 / f_{\boldsymbol{X}_1}(\boldsymbol{x}) \{1 + o(1)\}\, d\boldsymbol{x}$$
$$\mathbb{E}[\text{MISE}\,\hat{F}_{\hat{Y}_2}(z; h_2)|\mathcal{X}_1] = \left\{n_2^{-1} V_1(F_{Y_2}) - 2n_2^{-1} m_1(L\mathcal{L}) h_2 - \tfrac{1}{4} m_2(L)^2 h_2^4 \psi_{Y_2,2}\right\}\{1 + o(1)\},$$

$\psi_{Y_2,2} = \int_0^1 f_{Y_2}''(z) f_{Y_2}(z)\, dz$ *and* f_{Y_2} *is the density of* F_{Y_2}.

By leaving out the contribution of the weighting function $f_{\boldsymbol{X}_2}(\boldsymbol{x})^2 / f_{\boldsymbol{X}_1}(\boldsymbol{x})$, our optimal selector based on $\mathcal{X}_1$ is defined as

$$\mathbf{H}_{1,\text{AMISE}} = \underset{\mathbf{H} \in \mathcal{F}}{\text{argmin}}\ \text{AMISE}\,\hat{F}_{\boldsymbol{X}_1}(\cdot;\ \mathbf{H}), \tag{7}$$

allowing us to use, e.g. the plug-in data-based selector $\hat{\mathbf{H}}_{\mathrm{PI}}$ developed in Eq. (4), to estimate it. For h_2, the optimal bandwidth has the explicit formula

$$h_{2,\mathrm{AMISE}} = \operatorname*{argmin}_{h>0} \mathrm{AMISE}\, \hat{F}_Y(\cdot; h) = \left[\frac{2m_1(L\mathcal{L})}{-m_2(L)^2\psi_{Y_2,2}}\right]^{1/3} n_2^{-1/3}. \tag{8}$$

The data-based plug-in estimation of $h_{2,\mathrm{AMISE}}$ of Polansky and Baker (2000) relies on the analysis of the univariate functionals of the type $\psi_{Y_2,2}$ in Hall and Marron (1987).

The next theorem asserts that this sequential bandwidth selection strategy leads to the same minimal MISE order as with a more complex joint bandwidth selection $\operatorname{argmin}_{\mathbf{H}_1\in\mathcal{F},h_2>0}\mathrm{MISE}\,[\hat{F}_{\hat{Y}_2}(\cdot;\ \mathbf{H}_1, h_2)]$.

Theorem 6. *Suppose that the condition* (A1) *holds for the distribution* F_{Y_2}, (A2)–(A3) *for the kernels* K, L *and* $\mathcal{K}$, $\mathcal{L}$, *and* (A4) *for the bandwidths* $\mathbf{H}_1, h_2$. *The MISE rate obtained by the sequential unweighted selectors* $\mathbf{H}_{1,\mathrm{AMISE}}, h_{2,\mathrm{AMISE}}$ *in Eqs.* (7)–(8), *is asymptotically the same order as the minimal MISE as* $n_1, n_2 \to \infty$,

$$\left\{\inf_{\mathbf{H}_1\in\mathcal{F},h_2>0} \mathrm{MISE}\,[\hat{F}_{\hat{Y}_2}(\cdot;\ \mathbf{H}_1, h_2)]\right\} - \mathrm{MISE}\,[\hat{F}_{\hat{Y}_2}(\cdot;\ \mathbf{H}_{1,\mathrm{AMISE}}, h_{2,\mathrm{AMISE}})] = O(n_1^{-4/3} + n_2^{-4/3}).$$

The last step is to verify that the bandwidth choices in Eqs. (7)–(8) also ensure that $\hat{F}_{\hat{Y}_1}$ in Eq. (6) remains a consistent estimator of the false positive rate F_{Y_1}. We further require the following assumptions.

(A5) For the random variable $Y_1 = F_{\boldsymbol{X}_1}(\boldsymbol{X}_1)$, for its density function f_{Y_1}, all derivatives up to order 2 are bounded, continuous and square integrable, and of order 4 exist, and the expected value of Y_1 is $\mu_{Y_1} < \infty$.

(A6) For the random variable $\hat{Y}_1 = \hat{F}_{\boldsymbol{X}_1}(\boldsymbol{X}_1;\ \mathbf{H}_1)$, the random variable $\boldsymbol{X}_1$ does not coincide with any of the $\boldsymbol{X}_{1,1}, \ldots, \boldsymbol{X}_{1,n_1}$.

Theorem 7. *Suppose that the conditions for Theorem 4 hold, and further suppose that* (A5)–(A6) *hold. As* $n_1, n_2 \to \infty$, *the MISE of the estimator of the false positive rate* $\hat{F}_{\hat{Y}_1}$, *when the optimal selectors in Eqs.* (7)–(8) *are used, is*

$$\begin{aligned}\mathrm{MISE}\,[\hat{F}_{\hat{Y}_1}(\cdot;\ \mathbf{H}_{1,\mathrm{AMISE}}, h_{2,\mathrm{AMISE}})] &= n_1^{-1}\left[V_1(F_{Y_1}) + \psi_{Y_1,0}\mu_{Y_1}(1-\mu_{Y_1}) + \tfrac{1}{3} - 2\mu_{Y_1}V_0(F_{Y_1})\right] \\ &\quad + O(n_1^{-4/3} + n_1^{-1}n_2^{-1/3} + n_2^{-4/3})\end{aligned}$$

where $\psi_{Y_1,0} = \int_0^1 f_{Y_1}(z)^2\, dz$ *and* $V_0(F_{Y_1}) = \int_0^1 F_{Y_1}(z)\, dz$.

This MISE rate implies that consistency of $\hat{F}_{\hat{Y}_1}$ is maintained. Taking Theorems 5–7 together, the pairs $\{(\hat{F}_{\hat{Y}_1}(z;\ \mathbf{H}_1, h_2), \hat{F}_{\hat{Y}_2}(z;\ \mathbf{H}_1, h_2)) : z \in [0, 1]\}$ are MISE consistent estimators of the target ROC curve $\{(F_{Y_1}(z), F_{Y_2}(z)) : z \in [0, 1]\}$.

Given that $\hat{Y}_{j,i}, j = 1, 2, i = 1, \ldots, n_j$ are supported on the unit interval, we base our estimation on $\hat{Y}'_{j,i} = G^{-1}(\hat{Y}_{j,i})$, for a known univariate distribution function G with infinite support, to avoid the potential bias problems at the end points. So we have $F_{\hat{Y}_j}(z) = F_{\hat{Y}'_j}(G^{-1}(z))$. This allows us to avoid boundary bias when estimating $h_{2,\mathrm{AMISE}}$ for $\hat{Y}'_{2,1}, \ldots, \hat{Y}'_{2,n_2}$ and thus $\hat{F}_{\hat{Y}'_2}(G^{-1}(z); h_2) = n_2^{-1}\sum_{i=1}^{n_2}\mathcal{L}_{h_2}(G^{-1}(z) - G^{-1}(\hat{Y}_i))$. The ROC curve estimate is defined as

$$\hat{F}_{\hat{Y}_j}(z; h_2) = n_j^{-1}\sum_{i=1}^{n_j}\mathcal{L}_{h_2}(G^{-1}(z) - G^{-1}(\hat{Y}_{j,i})). \tag{9}$$

This sequential bandwidth approach recalls the univariate proposals posited by Hall and Hyndman (2003), Lloyd and Yong (1999) and Zhou and Harezlak (2002). Lloyd and Yong (1999) gave the bandwidth order as $n^{-1/3}$ but did not specify the constants. Zhou and Harezlak (2002) investigated several data-based selectors for kernel distribution estimators, but did not establish their convergence for ROC curves. Hall and Hyndman (2003) proposed selectors based on the weighted MISE, $\int_0^1 \mathrm{MSE}[\tilde{R}(z; h_1, h_2)]f_{X_1}(F_{X_1}^{-1}(z))\, dz = \int_{\mathbb{R}} \mathrm{MSE}[\hat{F}_{X_1}(x; h_1)]f_{X_2}(x)^2\, dx + \int_{\mathbb{R}} \mathrm{MSE}[\hat{F}_{X_2}(x; h_2)]f_{X_1}(x)^2\, dx$, where $\tilde{R}(z; h_1, h_2) = 1 - \hat{F}_{X_2}(\hat{F}_{X_1}^{-1}(1-z; h_1); h_2)$ is their ROC curve estimator, to avoid problems that may arise from the division of f_{X_1} in an unweighted $\int_0^1 \mathrm{MSE}[\tilde{R}(z; h_1, h_2)]\, dz$. This leads to other challenges since these weighted kernel estimators are difficult to analyse mathematically. Hall and Hyndman (2003) circumvented this difficulty by relying on normal scale approximations of these quantities, even though these normal scale estimators are not consistent if the true densities are non-normal. So these authors optimised a mathematically precise weighted MISE criterion but introduced simplifying assumptions into the computation. Results, from e.g. Altman and Léger (1995), Hall and Hyndman (2003) and Sarda (1993), demonstrate that using a weighting function which is independent of the data sample does not affect the asymptotic order of the MISE.

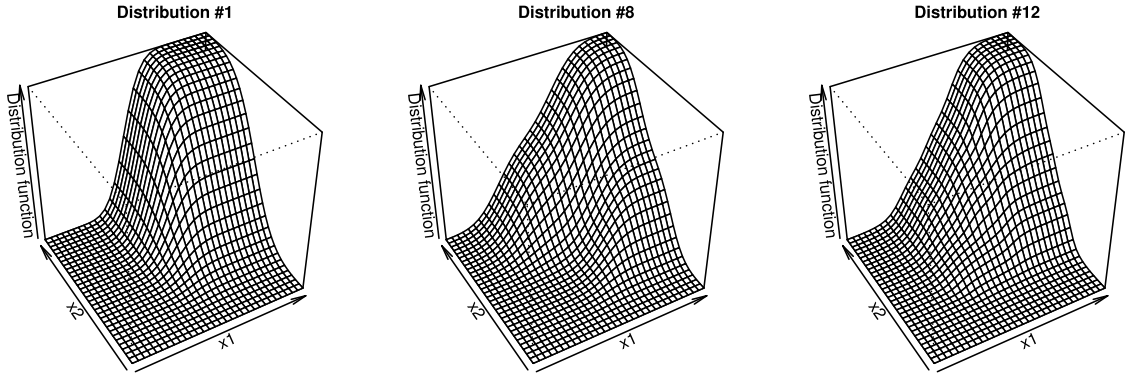

Fig. 2. Perspective plots of three bivariate normal mixture distributions.

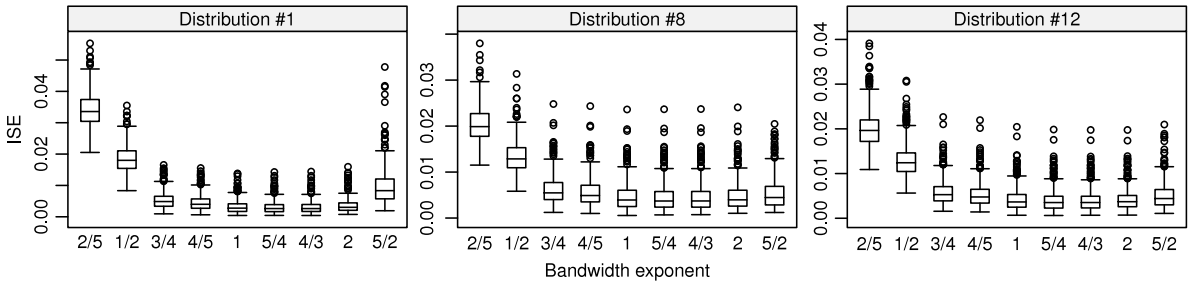

Fig. 3. Box plots of ISEs of kernel estimators of bivariate normal mixture distributions for $N = 400$ simulation trials of sample size $n = 1000$, as a function of exponents of the optimal bandwidth matrix. Within each panel are the ISE box plots for the kernel estimator $\hat{F}$ with bandwidths $\hat{\mathbf{H}}_{\mathrm{PI}}^{a}$, for $a = 2/5, 1/2, 3/4, 4/5, 1, 5/4, 4/3, 2, 5/2$.

4. Numerical results

4.1. Cumulative distribution functions

We examine the finite sample properties of our proposed kernel estimators of cumulative distribution functions for the twelve bivariate normal mixture distributions from Chacón (2009). Since they give similar results, we present only three of them, namely Distribution #1: $N((0,0),[1/4,0;0,1])$; Distribution #8: $1/2N((1,-1),[4/9,14/25;14/25,4/9]) + 1/2N((1,-1),4/9\mathbf{I}_2)$ and Distribution #12: $1/2N((0,0),\mathbf{I}_2) + 3/40N((0,0),1/16[1,-0.9;-0.9,1]) + 1/5N((1,1),1/4[1,-0.9;-0.9,1]) + 3/40N((-1,1),1/8\mathbf{I}_2) + 3/40N((-1,-1),1/8[1,-0.9;-0.9,1]) + 3/40N((1,-1),1/16\mathbf{I}_2)$, whose perspective plots are given in Fig. 2.

For a data sample, we compute the plug-in selector $\hat{\mathbf{H}}_{\mathrm{PI}}$ in Eq. (4), as implemented in the ks package (Duong, 2007) in the R statistical programming environment. This 2-stage selector, using the normal kernel $K = \phi$, is:

1. Begin with $\hat{\boldsymbol{\psi}}_4^{\mathrm{NS}}(\mathbf{S}) = \mathrm{D}^{\otimes 4}\phi_{2\mathbf{S}}(\mathbf{0})$ where $\mathbf{S}$ is the sample variance of $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$.
2. Compute the pilot bandwidth $\hat{\mathbf{G}}_{\mathrm{PI}} = \operatorname{argmin}_{\mathbf{G}\in\mathcal{F}} \|n^{-1}|\mathbf{G}|^{-1/2}(\mathbf{G}^{-1/2}\otimes\mathbf{G}^{-1/2})\mathrm{D}^{\otimes 2}\phi(\mathbf{0}) + \frac{1}{2}(\mathrm{vec}^T\mathbf{G}\otimes\mathbf{I}_{d^2})\hat{\boldsymbol{\psi}}_4^{\mathrm{NS}}(\mathbf{S})\|^2$.
3. Compute the integrated density functional $\hat{\boldsymbol{\psi}}_2(\hat{\mathbf{G}}_{\mathrm{PI}}) = n^{-2}\sum_{i,j=1}^{n}\mathrm{D}^{\otimes 2}\phi_{\hat{\mathbf{G}}_{\mathrm{PI}}}(\boldsymbol{X}_i - \boldsymbol{X}_j)$.
4. Compute the plug-in bandwidth $\hat{\mathbf{H}}_{\mathrm{PI}} = \operatorname{argmin}_{\mathbf{H}\in\mathcal{F}}\,[-2(4\pi)^{-1/2}n^{-1}\mathrm{tr}(\mathbf{H}^{1/2}\mathbf{J}_d) - \frac{1}{4}(\mathrm{vec}^T\,\mathbf{H})^{\otimes 2}\hat{\boldsymbol{\psi}}_2(\hat{\mathbf{G}}_{\mathrm{PI}})]$.

The optimisations are carried using a quasi-Newton BFGS routine in the R base software. This algorithm closely follows the bandwidth selectors for density derivative estimation in Chacón et al. (2011). These authors also outline how to compute efficiently the derivatives of the normal density function. Kernel estimators on the same data sample are then recomputed with $\hat{\mathbf{H}}_{\mathrm{PI}}^{a}$ for $a = 2/5, 1/2, 3/4, 4/5, 5/4, 4/3, 2, 5/2$. Since $|\hat{\mathbf{H}}_{\mathrm{PI}}| < 1$, then values of $a < 1$ lead to over-smoothing, and $a > 1$ to undersmoothing. The performance measure is the integrated squared error $\mathrm{ISE}(\mathbf{H}) = \int_{\mathbb{R}^2}[\hat{F}(\boldsymbol{x};\mathbf{H}) - F(\boldsymbol{x})]^2\,d\boldsymbol{x}$ which is approximated by a Riemann sum, since it does not have a closed form like the ISE of normal mixture densities. The box plots of these ISEs for $N = 400$ trials of sample size $n = 1000$ is shown in Fig. 3. The proposed $\hat{\mathbf{H}}_{\mathrm{PI}}$ falls in the interval of the exponents ($a = 1, 5/4, 4/3$) which yield the smallest ISEs, as expected from Theorem 3. Large amounts of undersmoothing ($a = 5/2$) or oversmoothing ($a = 2/5$) produce estimates with inflated ISEs, though we note that the undersmoothing is less serious than oversmoothing.

We compare the empirical gain in using an unconstrained bandwidth matrix over diagonal bandwidth matrices, as measured by the empirical $\mathrm{ARE}(\mathcal{F} : \mathcal{D}) = \mathrm{ISE}\{\hat{F}(\cdot;\hat{\mathbf{H}}_{\mathrm{PI}})\}/\mathrm{ISE}\{\hat{F}(\cdot;\hat{\mathbf{H}}_{\mathrm{PI},\mathcal{D}})\}$. The diagonal selector $\hat{\mathbf{H}}_{\mathrm{PI},\mathcal{D}}$ is similar to $\hat{\mathbf{H}}_{\mathrm{PI}}$

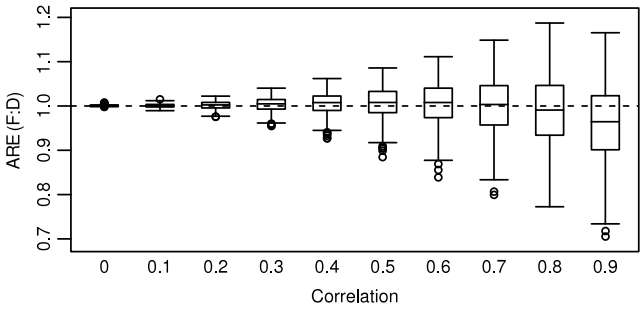

Fig. 4. Box plots of AREs of kernel estimators of bivariate correlated normal distributions for $N = 400$ simulation trials of sample size $n = 1000$, as a function of the correlation coefficient. The ARE$(\mathscr{F} : \mathscr{D})$ is estimated by the ratio of ISE $\hat{F}(\cdot; \hat{\mathbf{H}}_{\mathrm{PI}})$ for an unconstrained plug-in matrix $\hat{\mathbf{H}}_{\mathrm{PI}} \in \mathscr{F}$, divided by ISE $\hat{F}(\cdot; \hat{\mathbf{H}}_{\mathrm{PI},\mathscr{D}})$ obtained with a diagonal plug-in matrix $\hat{\mathbf{H}}_{\mathrm{PI},\mathscr{D}} \in \mathscr{D}$.

except that the optimisation ranges over the class of all positive definite diagonal matrices $\mathscr{D}$. We draw $N = 400$ samples of $n = 1000$ from a bivariate normal with mean zero and variance $1/4[1, \rho; \rho, 1]$, for $\rho \in (0, 0.1, 0.2, \ldots, 0.9)$. Fig. 4 is the ARE plots as estimated by the ratios of the corresponding ISEs, as a function of the correlation coefficient ρ. In Fig. 4, the ISE is not inflated using an unconstrained bandwidth matrix when a diagonal matrix is optimal (low values of correlations), and is most beneficial for ISE reduction, as expected from the ARE results in Theorem 4, for high correlations ($\rho > 0.7$), though the parametrisation of the bandwidth matrix has only a limited effect in ISE reduction for kernel distribution estimators.

4.2. Receiver operating characteristic curves

The computational algorithm for our proposed multivariate ROC curve estimator is:

1. The two data samples are $\mathscr{X}_1 = \{\boldsymbol{X}_{1,1}, \ldots, \boldsymbol{X}_{1,n_1}\}$ and $\mathscr{X}_2 = \{\boldsymbol{X}_{2,1}, \ldots, \boldsymbol{X}_{2,n_2}\}$. From the first sample $\mathscr{X}_1$, compute the kernel estimator in Eq. (2) of its survival function $\hat{\bar{F}}_{\boldsymbol{X}_1}(\cdot; \hat{\mathbf{H}}_1)$, using a data-based plug-in selector, $\hat{\mathbf{H}}_1$ in Eq. (7), using the algorithm for $\hat{\mathbf{H}}_{\mathrm{PI}}$ in Section 4.1.
2. Create the univariate auxiliary random variables $\hat{Y}_{j,i} = \hat{\bar{F}}_{\boldsymbol{X}_1}(\boldsymbol{X}_{j,i}; \hat{\mathbf{H}}_1),\ i = 1, \ldots, n_j,\ j = 1, 2$. Compute the kernel estimators in Eq. (9) of their cumulative distribution functions, using the scalar plug-in selector $\hat{h}_2$ in Eq. (8), using the algorithm in Polansky and Baker (2000), and $G = \Phi$. The quantile–quantile plot of these two distributions $\hat{F}_{\hat{Y}_1}$ vs. $\hat{F}_{\hat{Y}_2}$ is the required kernel estimator of the ROC curve which compares $\mathscr{X}_1$ and $\mathscr{X}_2$.

This algorithm has been implemented in the `ks` package. As noted previously, since there is little difference in the performance between unconstrained and diagonal bandwidth matrices for kernel estimators of distribution functions, we have omitted the latter from this section. For each pair of simulated data, we compute the

- bivariate kernel ROC curve with optimal (unconstrained) bandwidths $\hat{R}_{\mathrm{KOPT}}$
- univariate kernel ROC curves of the marginal variables $\hat{R}_{\mathrm{K1}}, \hat{R}_{\mathrm{K2}}$
- univariate kernel ROC curve of the best linear scalar projection of the variables $\hat{R}_{\mathrm{KSL}}$ of Su and Liu (1993). The best scalar linear projection is defined as $Y_{j,i} = (\mathrm{tr}\mathbf{S}_1 + \mathrm{tr}\mathbf{S}_2)^{-1}(\bar{\boldsymbol{X}}_2 - \bar{\boldsymbol{X}}_1)^T \boldsymbol{X}_{j,i}$ where $\bar{\boldsymbol{X}}_j, \mathbf{S}_j$ are the sample mean and variance of $\mathscr{X}_j,\ j = 1, 2,\ i = 1, \ldots, n_j$.

4.2.1. Simulated data for bivariate normal mixtures

To examine the finite sample properties of this computational algorithm, we examine three pairs of bivariate normal mixture densities: Pair #1: $N((0, 0), \mathbf{I}_2)$ vs. $N((2/3, 2/3), \mathbf{I}_2)$, Pair #2: $1/2N((-3/2, 0), 2/3\mathbf{I}_2) + 1/2N((1/2, 0), 2/3\mathbf{I}_2)$ vs. $1/2N((-1, 1/4), 2/3\mathbf{I}_2) + 1/2N((1, 1/2), 2/3\mathbf{I}_2)$, and Pair #3: $1/2N((-7/8, 7/8), 1/4\mathbf{I}_2) + 1/2N((7/8, -7/8), 1/4\mathbf{I}_2)$ vs. $1/2N((-7/8, -7/8), 1/4\mathbf{I}_2) + 1/2N((7/8, 7/8), 1/4\mathbf{I}_2)$. Pair #1 differs only in the means of the two normal variables and so can be considered a base case. Pair #2 differs by a small translation in both variables, though the difference for x_1 is between two marginal bimodal distributions. Pair #3 is where both pairs of marginal densities are the same, whereas the join densities are different, so we expect an important gain from using a bivariate method over univariate one.

We begin with some plots from a representative sample of size $n_1 = n_2 = 1000$ in Fig. 5. Each row corresponds to each pair of target densities. In the first column are the kernel estimates of the marginal density for the first variable x_1, the second column for the second variable x_2, and the third column for the joint density of (x_1, x_2). The density estimates based solely on the x_1 variable are coloured in black, the x_2 variable in blue, and the bivariate (x_1, x_2) variable in red. For the first three columns, the estimates based on the first sample $\mathscr{X}_1$ are represented by solid lines, the second sample $\mathscr{X}_2$ by the dashed lines. In the fourth column, the ROC curve estimate $\hat{R}_{\mathrm{K1}}$ is the black dashed line, $\hat{R}_{\mathrm{K2}}$ is dotted blue line, and $\hat{R}_{\mathrm{KOPT}}$ is the solid red line. The horizontal axis is the false positive rate $\hat{\bar{F}}_{\boldsymbol{X}_1}$ or the complement of the specificity $\hat{F}_{\boldsymbol{X}_1}$ (labelled

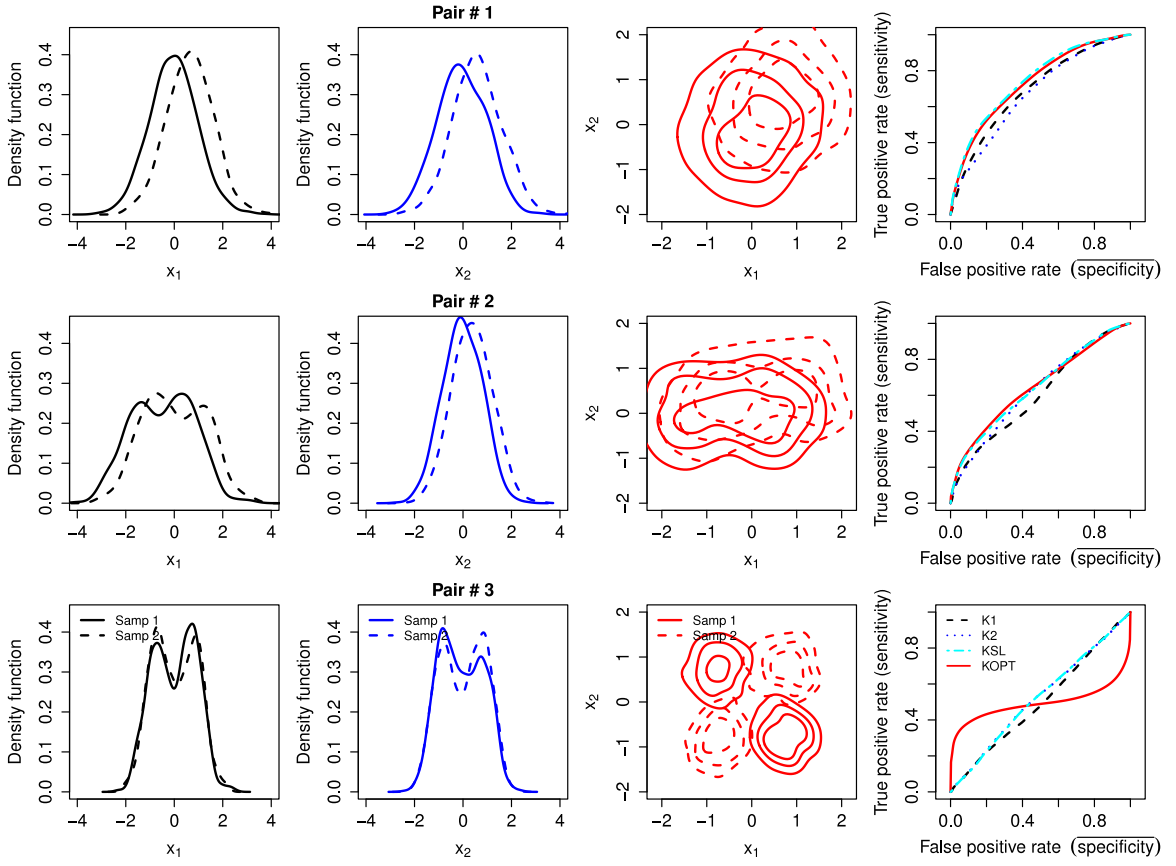

Fig. 5. Density estimates and ROC curves for marginal and joint variables for bivariate normal mixture pairs. Each row corresponds to the results from a representative $n_1 = n_2 = 1000$ sample drawn from each of the target bivariate density pairs. Black represents estimates based on the first variable x_1 only, blue the second variable x_2 only, and red the joint variables (x_1, x_2). The first column is the marginal density estimates of x_1, the second column the marginal density estimates of x_2, the third column the joint density estimates of (x_1, x_2). In these first three columns, the first sample is represented by the solid lines, the second sample by the dashed lines. The fourth column displays the ROC curves: $\hat{R}_{\mathrm{K1}}$ is black dashed, $\hat{R}_{\mathrm{K2}}$ is dotted blue, scalar projected $\hat{R}_{\mathrm{SL}}$ is dot-dashed cyan, and $\hat{R}_{\mathrm{KOPT}}$ is solid red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$\overline{\text{specificity}}$); the vertical axis is the true positive rate $\hat{F}_{\boldsymbol{X}_2}$ or the sensitivity. For Pair #1, the two bivariate based ROC curves KOPT, KSL are uniformly better than the marginal ROC curves K1, K2, as expected. Since both distributions are bivariate normal, the performance of KSL is optimal (see Su and Liu (1993)) and we observe that our kernel based ROC curve exhibits very similar performance. For Pair #2, KSL does less well than KOPT since the latter becomes less optimal when the underlying distributions are non-normal: in this case, bimodality is noticeable in the x_1-variable. KSL remains narrowly better than K2 and uniformly better than K1. For Pair #3, both univariate ROC curves K1, K2 are non-informative. Since each distribution in this pair are well-separated bimodal distributions, KSL chooses a sub-optimal scalar projection which produces a ROC curve that is narrowly better than the non-informative marginal univariate ROC curves. Whereas the bivariate ROC curve KOPT correctly shows that a large difference exists between the samples.

In addition to these representative plots, we perform a quantitative simulation study for $N = 400$ repetitions for each of the normal mixture pairs for sample size $n_1 = n_2 = 100$ and $n_1 = n_2 = 1000$. We compute the Youden index as our summary measure of performance, rather than the ISE or the area under the ROC curve (AUC), see Shapiro (1999) for a discussion of the relative merits of the Youden index over the AUC for continuous random variables. Youden (1950) originally defined his subsequently eponymous index as the maximum of the difference of the true positive rate (sensitivity) and the false positive rate (complement of specificity), assuming that false positives and true positives are equally weighted. For ROC curves, the Youden index is the maximum deviation of the ROC curve from the diagonal line, i.e. a non-informative ROC curve. For interpretation, a set of variables is discriminatory if their ROC curve has a Youden index close to 1, whereas a non-informative ROC curve has a Youden index close to 0.

Since the results for both $n_1 = n_2 = 100$ and $n_1 = n_2 = 1000$ are similar, we show only the box plots for these summary indices in Fig. 6 for the larger sample size. These box plots verify our intuitive observations from the representative samples in Fig. 5. The target Youden index is approximated by the maximum of $|F_{\boldsymbol{X}_1} - F_{\boldsymbol{X}_2}|$ over a grid, since the index is also the supremum norm of the L_1 difference between the two distributions $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |F_{\boldsymbol{X}_1}(\boldsymbol{x}) - F_{\boldsymbol{X}_2}(\boldsymbol{x})|$, see Lloyd and Yong (1999). For Pair #1, the target Youden index is 0.32. As expected, the univariate marginal ROC curves K1, K2 underestimate the Youden

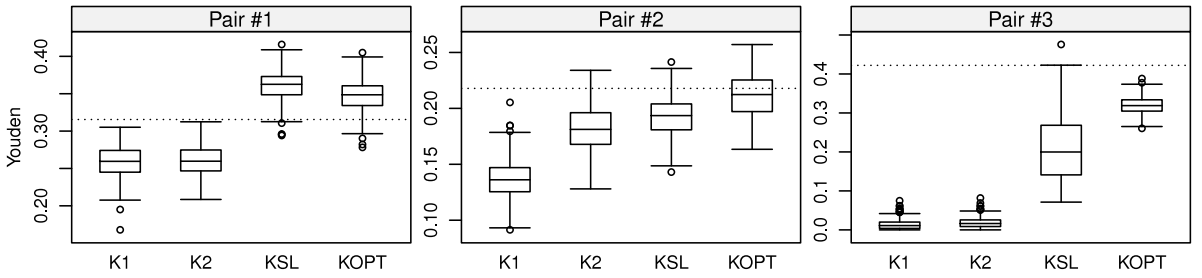

Fig. 6. Box plots for Youden indices of ROC curves of bivariate normal mixture pairs for $N = 400$ simulation trials of sample size $n_1 = n_2 = 1000$. The marginal univariate ROC curves $\hat{R}_{K1}$, $\hat{R}_{K2}$ the scalar projected ROC curve $\hat{R}_{KSL}$, and the bivariate ROC curve $\hat{R}_{KOPT}$. The dotted line is the target Youden index.

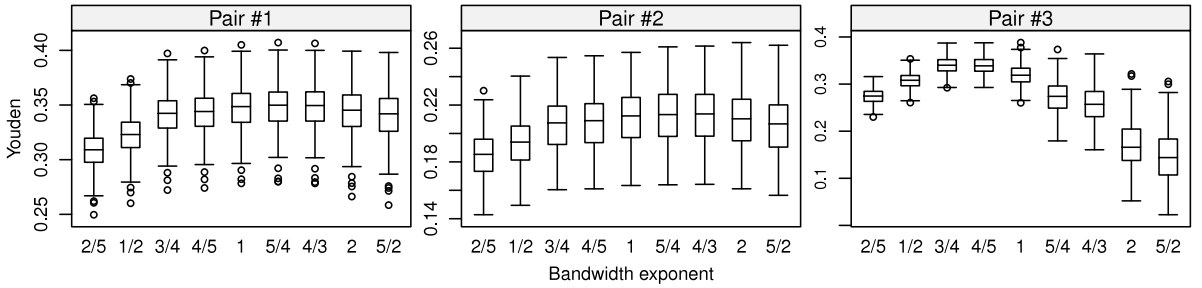

Fig. 7. Box plots for Youden indices of ROC curves of bivariate normal mixture pairs for $N = 400$ simulation trials of sample size $n_1 = n_2 = 1000$, as a function of exponents of the optimal bandwidths. Within each panel are the box plots of the summary indices for KOPT, the kernel estimator $\hat{R}$ with bandwidths $\hat{\mathbf{H}}^a_{PI}$, $\hat{h}^a_{PI}$, for $a = 2/5, 1/2, 3/4, 4/5, 1, 5/4, 4/3, 2, 5/2$.

index, whereas the KSL and KOPT both overestimate it, with the latter with a slightly smaller overestimation and the closest overall to the target Youden value. For Pair #2, the target Youden index is 0.22. The separation of KSL and KOPT from the univariate marginal ROC curves K1, K2 is maintained. KOPT most accurately estimates the Youden index, though both K2 and KOPT give reasonable Youden indices. For Pair #3, the target Youden index is 0.42, implying that the advantage of KOPT is amplified. Overall, KOPT is better than KSL which better than the 1-dimensional marginal ROC curves. This confirms that multivariate non-parametric estimators outperform parametric estimators when the parametric assumptions are not satisfied, and that multivariate estimators can distinguish structure which are hidden by lower dimensional projections.

We proceed with an examination of the behaviour of the optimal plug-in selector KOPT as a function of bandwidth exponents in Fig. 7. The KOPT curve is first computed, i.e. the kernel ROC curve estimator with the plug-in selectors $\hat{\mathbf{H}}_1, \hat{h}_2$ from Eqs. (7)–(8). Kernel estimators on the same data sample are then recomputed with $\hat{\mathbf{H}}^a_1, \hat{h}^a_2$ for $a = 2/5, 1/2, 3/4, 4/5, 5/4, 4/3, 2, 5/2$. Since $|\hat{\mathbf{H}}_1|, \hat{h}_2 < 1$, then values of $a < 1$ lead to over-smoothing, and $a > 1$ to undersmoothing. The proposed KOPT falls inside the interval of exponents ($a = 1, 5/4, 4/3$) with maximal Youden indices for Pairs #1,#2, and the interval of exponents ($a = 3/4, 4/5, 1$) for Pair #3. Although Theorem 6 asserts only the MISE optimality of the KOPT estimator, Fig. 7 empirically shows that this optimality also applies to the other summary measures such as the Youden index.

4.2.2. Joint multivariate biomarker validation

To illustrate our proposed multivariate ROC curve algorithm on experimental data, we return to the Pilot Study of Biomarkers for Spinal Muscular Atrophy data set presented in the introduction. Our goal here is to validate the diagnostic power in terms of ROC curves of these combinations of biomarkers: Glycine (Gly), L-glutamic (L-Glu) and 1-methyl-L-histidine (methyl L-Hist) The marginal univariate kernel density estimates are given in the upper row of Fig. 8, followed by the trivariate density estimates on the lower row. On the lower centre panel are the individual ROC curves for Gly $\hat{R}_{K1}$ in dashed black, L-Glu $\hat{R}_{K2}$ in dotted blue, methyl L-Hist $\hat{R}_{K3}$ in dot-dashed cyan, the scalar projected ROC curve $\hat{R}_{KSL}$ in long dashed magenta, the joint ROC curve $\hat{R}_{KOPT}$ in solid red. The joint trivariate ROC curve has better true positive rates for all false positive rates, except for those in the approximate range [0, 2, 0.4]. This visual improvement is verified by an improvement in the summary indices: $\text{Youden}(\hat{R}_{K1}) = 0.112$, $\text{Youden}(\hat{R}_{K2}) = 0.279$, $\text{Youden}(\hat{R}_{K3}) = 0.195$, $\text{Youden}(\hat{R}_{KSL}) = 0.198$, $\text{Youden}(\hat{R}_{KOPT}) = 0.286$. We demonstrate that while the diagnostic power of the marginal univariate biomarkers singly can already be good, it can be improved on by the ROC curve of the joint trivariate biomarkers.

5. Discussion

We have introduced a fully non-parametric estimator of the receiver operating characteristic curve to compare two multivariate samples, based on a combination of kernel estimators of cumulative distribution and survival functions. As is

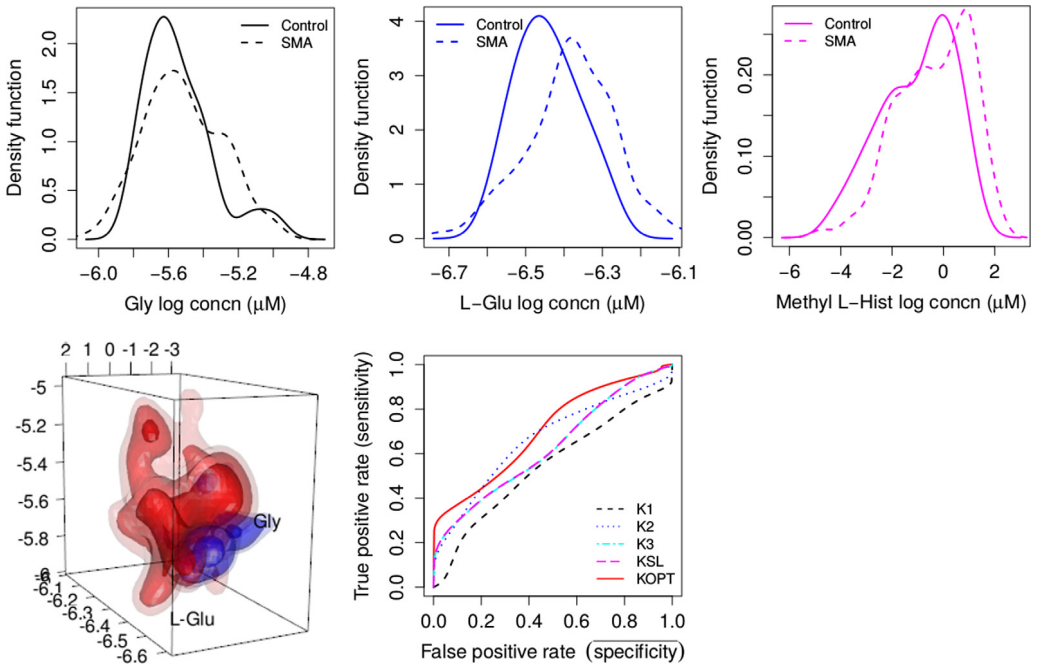

Fig. 8. Joint multivariate biomarker validation for Spinal Muscular Atrophy (SMA). Kernel density estimates of negative log concentrations (μM). The control density estimates are solid lines, and the SMA density estimates are dashed lines. Upper left—Glycine (Gly). Upper centre—L-glutamic (L-Glu). Upper right—1-methyl-L-histidine (methyl L-Hist). Lower right—Kernel density estimate of (Gly, L-Glu, methyl L-Hist). The control density estimate are blue contour shells, and the SMA density estimate are red contour shells. Lower centre—ROC curves for individual and joint biomarkers: Gly $\hat{R}_{\mathrm{K1}}$ is dashed black, L-Glu $\hat{R}_{\mathrm{K2}}$ is dotted blue, methyl L-Hist $\hat{R}_{\mathrm{K3}}$ is dot-dashed cyan, scalar projection $\hat{R}_{\mathrm{KSL}}$ is long dashed magenta, joint (Gly, L-Glu, methyl L-Hist) $\hat{R}_{\mathrm{KOPT}}$ is solid red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

crucial for all kernel estimators, we have supplied consistent, data-based algorithms for optimal selection for all required smoothing parameters. Whilst the performance gain from using maximally general unconstrained bandwidth matrices over their constrained counterparts is modest, one of the main contributions of this manuscript has been to recast their analysis in a mathematical framework which will facilitate the development of kernel estimators of other quantities based on cumulative distribution and survival functions, e.g. the hazard function, the Lorenz curve and the Gini coefficient. We have treated the kernel estimation case, since they are an important case in their own right for low dimensional data analysis, but importantly they also serve as a learning ground for future research. The results developed here are amenable to being extended to other non-parametric estimation techniques which are more suitable for high dimensional data, e.g. splines, wavelets, nearest neighbour, etc. For high dimensional discrimination and variable (subset) selection problems, multivariate ROC curves would be an important addition to the suite of data analytic methods.

Acknowledgements

The author would like to thank José E. Chacón (Universidad de Extremadura, Spain) and Boris Béranger (Université Pierre et Marie Curie—Paris 6, France/University of New South Wales, Australia) for careful reading of the manuscript. The author was partially supported by a "Projet d'investisssment d'avenir" (PIA 2013) grant at the Computer Science Laboratory (LIPN) at the University of Paris-North—Paris 13, France.

Appendix

A.1. Proofs for Section 2

Proof of Theorem 1. Under the regularity guaranteed by (A1)–(A3), the expected value is, by repeatedly applying integration by parts,

$$\mathbb{E}\hat{F}(\boldsymbol{x};\ \mathbf{H}) = \mathbb{E}\mathcal{K}_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X}) = \int_{\mathbb{R}^d} \mathcal{K}(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))f(\boldsymbol{y})\,d\boldsymbol{y} = \int_{\mathbb{R}^d} \mathcal{K}(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))\frac{\partial^d}{\partial y_1\cdots\partial y_d}F(\boldsymbol{y})\,d\boldsymbol{y}$$

$$\vdots$$

$$= (-1)^d \int_{\mathbb{R}^d} \frac{\partial^d}{\partial y_1 \cdots \partial y_d} \mathcal{K}(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))F(\boldsymbol{y})\,d\boldsymbol{y}$$
$$= (-1)^{2d} \int_{\mathbb{R}^d} |\,\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))F(\boldsymbol{y})\,d\boldsymbol{y} = K_{\mathbf{H}} * F(\boldsymbol{x})$$

using that $\mathcal{K}(\infty) = F(\infty) = 1$ and $\mathcal{K}(-\infty) = F(-\infty) = 0$ as $\mathcal{K}, F$ are multivariate distribution functions.

For the variance, we have $\operatorname{Var}\hat{F}(\boldsymbol{x};\ \mathbf{H}) = n^{-1}\mathbb{E}[\mathcal{K}_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})^2] - n^{-1}[\mathbb{E}\mathcal{K}_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})]^2$. The second term is the same as above, so it leaves us to evaluate

$$\mathbb{E}[\mathcal{K}_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{X})^2] = \int_{\mathbb{R}^d} \mathcal{K}(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))^2 f(\boldsymbol{y})\,d\boldsymbol{y} = \int_{\mathbb{R}^d} \mathcal{K}(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))^2 \frac{\partial^d}{\partial y_1 \cdots \partial y_d} F(\boldsymbol{y})\,d\boldsymbol{y}$$
$$= (-1)^d \int_{\mathbb{R}^d} \frac{\partial^d}{\partial y_1 \cdots \partial y_d} \mathcal{K}(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))^2 F(\boldsymbol{y})\,d\boldsymbol{y}$$
$$= \int_{\mathbb{R}^d} 2|\,\mathbf{H}|^{-1/2} \mathcal{K}(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))K(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{y}))F(\boldsymbol{y})\,d\boldsymbol{y}$$
$$= \int_{\mathbb{R}^d} 2\mathcal{K}_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})F(\boldsymbol{y})\,d\boldsymbol{y} = 2(K_{\mathbf{H}}\mathcal{K}_{\mathbf{H}} * F)(\boldsymbol{x}).$$

Obtaining the MSE is straightforward. Obtaining the MISE from the MSE is also straightforward since we can interchange the order of integration according to the conditions (A1)–(A3).

For the asymptotic analysis, we follow the framework established by Nadaraya (1964) for univariate data, later adapted to multivariate data, e.g. Liu and Yang (2008). For the expected value,

$$\mathbb{E}\hat{F}(\boldsymbol{x};\ \mathbf{H}) = \int_{\mathbb{R}^d} K(\boldsymbol{w})F(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{w})\,d\boldsymbol{w}$$
$$= \int_{\mathbb{R}^d} K(\boldsymbol{w})\left[F(\boldsymbol{x}) - \boldsymbol{w}^T\mathbf{H}^{1/2}\mathsf{D}F(\boldsymbol{x}) + \tfrac{1}{2}\boldsymbol{w}^T\mathbf{H}^{1/2}\mathsf{D}^2F(\boldsymbol{x})\mathbf{H}^{1/2}\boldsymbol{w}\right]\{1+o(1)\}\,d\boldsymbol{w}$$
$$= F(\boldsymbol{x}) + \tfrac{1}{2}m_2(K)\operatorname{tr}(\mathbf{H}\,\mathsf{D}^2F(\boldsymbol{x})) + o(\operatorname{tr}\mathbf{H}),$$

using the usual rearrangement $\boldsymbol{w}^T\mathbf{H}^{1/2}\mathsf{D}^2F(\boldsymbol{x})\mathbf{H}^{1/2} = \operatorname{tr}(\boldsymbol{w}\boldsymbol{w}^T\mathbf{H}\,\mathsf{D}^2F(\boldsymbol{x}))$. For the variance, we have

$$\mathbb{E}[\mathcal{K}(\mathbf{H}^{-1/2}(\boldsymbol{x}-\boldsymbol{X}))^2] = \int_{\mathbb{R}^d} 2\mathcal{K}(\boldsymbol{w})K(\boldsymbol{w})F(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{w})\,d\boldsymbol{w}$$
$$= \int_{\mathbb{R}^d} 2\mathcal{K}(\boldsymbol{w})K(\boldsymbol{w})[F(\boldsymbol{x}) - \boldsymbol{w}^T\mathbf{H}^{1/2}\mathsf{D}F(\boldsymbol{x})]\{1+o(1)\}\,d\boldsymbol{w}$$
$$= F(\boldsymbol{x}) - 2\int_{\mathbb{R}^d} \mathcal{K}(\boldsymbol{w})K(\boldsymbol{w})\boldsymbol{w}^T\mathbf{H}^{1/2}\mathsf{D}F(\boldsymbol{x})\{1+o(1)\}\,d\boldsymbol{w}$$

since $\int_{\mathbb{R}} \mathcal{K}(\boldsymbol{w})K(\boldsymbol{w})\,d\boldsymbol{w} = 1/2$. Thus $\operatorname{Var}\hat{F}(\boldsymbol{x};\ \mathbf{H}) = \{n^{-1}F(\boldsymbol{x})(1-F(\boldsymbol{x})) - 2n^{-1}\boldsymbol{m}_1(K\mathcal{K})^T\mathbf{H}^{1/2}\mathsf{D}F(\boldsymbol{x})\}\{1+o(1)\}$.

For further simplification in the MISE, we note that for any x_j,

$$F(\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}} f(\boldsymbol{w})\,d\boldsymbol{w} \le \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{x_j}\cdots\int_{-\infty}^{\infty} f(\boldsymbol{w})\,d\boldsymbol{w} = \int_{-\infty}^{x_j} f_j(w_j)\,dw_j$$

where f_j is the jth marginal density of F. Thus element-wise $\mathsf{D}F(\boldsymbol{x}) \le (f_1(x_1), \ldots, f_j(x_j)) = f(\boldsymbol{x})\mathbf{1}_d\{1+o(1)\}$ as F does not depend on n from (A1). Furthermore, $\mathsf{D}^2F(\boldsymbol{x}) = \mathsf{D}f(\boldsymbol{x})\mathbf{1}_d^T\{1+o(1)\}$. Using these, the squared bias becomes $\operatorname{tr}^2(\mathbf{H}\,\mathsf{D}^2F) = (\operatorname{vec}^T\mathbf{H}\operatorname{vec}\mathsf{D}^2F)(\operatorname{vec}^T\mathsf{D}^2F\operatorname{vec}\mathbf{H}) = \operatorname{vec}^T(\mathbf{H}^T\mathsf{D}f\,\mathbf{1}_d)\operatorname{vec}(\mathbf{1}_d^T\mathbf{H}\,\mathsf{D}f) = (\operatorname{vec}^T\mathbf{H}^2)(\mathsf{D}f\otimes\mathsf{D}f)$. Integrating this,

$$\int_{\mathbb{R}^d} \operatorname{Bias}^2\hat{F}(\boldsymbol{x};\mathbf{H})\,d\boldsymbol{x} = \tfrac{1}{4}m_2(K)^2(\operatorname{vec}^T\mathbf{H}^2)\int_{\mathbb{R}^d}(\mathsf{D}f(\boldsymbol{x})\otimes\mathsf{D}f(\boldsymbol{x}))\{1+o(1)\}\,d\boldsymbol{x}$$
$$= -\tfrac{1}{4}m_2(K)^2(\operatorname{vec}^T\mathbf{H}^2)\int_{\mathbb{R}^d}\mathsf{D}^{\otimes 2}f(\boldsymbol{x})f(\boldsymbol{x})\{1+o(1)\}\,d\boldsymbol{x} = -\tfrac{1}{4}m_2(K)^2(\operatorname{vec}^T\mathbf{H}^2)\boldsymbol{\psi}_2\{1+o(1)\}.$$

Likewise, the integrated variance is

$$\int_{\mathbb{R}^d} \operatorname{Var}\hat{F}(\boldsymbol{x};\mathbf{H})\,d\boldsymbol{x} = \int_{\mathbb{R}^d}\left\{n^{-1}F(\boldsymbol{x})(1-F(\boldsymbol{x})) - 2n^{-1}\boldsymbol{m}_1(K\mathcal{K})^T\mathbf{H}^{1/2}f(\boldsymbol{x})\mathbf{1}_d\right\}\{1+o(1)\}\,d\boldsymbol{x}$$
$$= \{n^{-1}V(F) - 2n^{-1}\boldsymbol{m}_1(K\mathcal{K})^T\mathbf{H}^{1/2}\}\{1+o(1)\},$$

and the AMISE $\hat{F}(\cdot;\mathbf{H})$ result follows immediately.

For the survival function estimator, the first moment of $\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}-\boldsymbol{x})$ is

$$\mathbb{E}\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}-\boldsymbol{x}) = \int_{\mathbb{R}^d} \mathcal{K}_{\mathbf{H}}(\boldsymbol{y}-\boldsymbol{x})f(\boldsymbol{y})\,d\boldsymbol{y} = \int_{\mathbb{R}^d} K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})\bar{F}(\boldsymbol{y})\,d\boldsymbol{y} = K_{\mathbf{H}} * \bar{F}(\boldsymbol{x})$$

and its second moment is

$$\mathbb{E}\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}-\boldsymbol{x})^2 = \int_{\mathbb{R}^d} \mathcal{K}_{\mathbf{H}}(\boldsymbol{y}-\boldsymbol{x})^2 f(\boldsymbol{y})\,d\boldsymbol{y} = -2\int_{\mathbb{R}^d} K_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})\mathcal{K}_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})\bar{F}(\boldsymbol{y})\,d\boldsymbol{y} = 2K_{\mathbf{H}}\mathcal{K}_{\mathbf{H}} * \bar{F}(\boldsymbol{x}),$$

using the symmetry of K. Combining Taylor's expansions of these expressions leads to the appropriate asymptotic bias and variance formulas. For the AMISE, we use $\mathsf{D}\bar{F}(\boldsymbol{x}) = -f(\boldsymbol{x})\mathbf{1}_d\{1+o(1)\}$ and $\mathsf{D}^2\bar{F}(\boldsymbol{x}) = -\mathsf{D}f(\boldsymbol{x})\mathbf{1}_d^T\{1+o(1)\}$, in a similar calculation to that for F above. □

Proof of Theorem 2. Let $\mathsf{D}_{\mathbf{H}} = \partial/(\partial \mathrm{vec}\,\mathbf{H})$ be the vector of differentials with respect to vec $\mathbf{H}$. Then $\mathbf{H}_{\mathrm{MISE}}$ is a solution to $\mathsf{D}_{\mathbf{H}}\mathrm{MISE}\,\hat{F}(\cdot;\,\mathbf{H}) = \mathbf{0}$. Computing derivatives directly is complicated so we adopt the approach of Magnus and Neudecker (1999) of converting differentials to derivatives. For a $d\times d$ symmetric matrix $\mathbf{A}$, $d\mathrm{tr}(\mathbf{H}^2\mathbf{A}) = \mathrm{vec}^T(\mathbf{A}\mathbf{H}+\mathbf{H}\mathbf{A})(d\mathrm{vec}\,\mathbf{H})$, and $d\mathrm{tr}(\mathbf{H}^{1/2}\mathbf{A}) = \frac{1}{2}\mathrm{vec}^T(\mathbf{A}\mathbf{H}^{-1/2}+\mathbf{H}^{-1/2}\mathbf{A})d\,\mathrm{vec}\,\mathbf{H}$. Thus the derivative of the MISE is

$$\begin{aligned}\mathsf{D}_{\mathbf{H}}\mathrm{MISE}\,\hat{F}(\cdot;\,\mathbf{H}) &= \tfrac{1}{4}m_2(K)^2\int_{\mathbb{R}^d}\mathrm{vec}\,(\mathbf{H}\,\mathsf{D}^2F(\boldsymbol{x})+\mathsf{D}^2F(\boldsymbol{x})\,\mathbf{H})\,d\boldsymbol{x}\\ &\quad - n^{-1}\int_{\mathbb{R}^d}(\mathbf{H}^{-1/2}\otimes\boldsymbol{m}_1(K\mathcal{K})\mathsf{D}F(\boldsymbol{x})^T+\boldsymbol{m}_1(K\mathcal{K})\mathsf{D}F(\boldsymbol{x})^T\otimes\mathbf{H}^{-1/2})\,d\boldsymbol{x}.\end{aligned}$$

In general this matrix equation does not have an explicit solution. Nonetheless if we let $\mathbf{H}_{\mathrm{MISE}} = O(n^{-\alpha})\mathbf{J}_d$, where $\mathbf{J}_d$ is a $d\times d$ matrix whose elements are all ones, then matching coefficients requires that $\alpha = 2/3$. Substituting this $\mathbf{H}_{\mathrm{MISE}} = O(n^{-2/3})$ into the MISE formula, we obtain that the minimal MISE is order $n^{-1}V_1(F)+O(n^{-4/3})$. □

Following Duong and Hazelton (2005), the relative rate of convergence to $\hat{\mathbf{H}}$ to $\mathbf{H}_{\mathrm{AMISE}}$ is $O_P(n^{-\alpha})$ if $\hat{\mathbf{H}}-\mathbf{H}_{\mathrm{AMISE}} = O_P(n^{-\alpha}\mathbf{J}_{d^2})\mathrm{vec}\,\mathbf{H}_{\mathrm{AMISE}}$. Since convergence in probability can be difficult to demonstrate directly, these authors show that in their Lemma 1 that this order in probability is implied by

$$\mathrm{MSE}[\mathsf{D}_{\mathbf{H}}(\widehat{\mathrm{AMISE}}-\mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}})] = O(n^{-2\alpha}\mathbf{J}_{d^2})(\mathrm{vec}\,\mathbf{H}_{\mathrm{AMISE}}\mathrm{vec}^T\,\mathbf{H}_{\mathrm{AMISE}})$$

since $\mathrm{MSE}\,\hat{\mathbf{H}} = \mathbb{E}\|\,\mathrm{vec}\,(\hat{\mathbf{H}}-\mathbf{H}_{\mathrm{AMISE}})\|^2 = \mathrm{MSE}[\mathsf{D}_{\mathbf{H}}(\widehat{\mathrm{AMISE}}-\mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}})]\{1+o(1)\}$.

Proof of Theorem 3. $\mathrm{MSE}[\mathsf{D}_{\mathbf{H}}(\mathrm{PI}-\mathrm{AMISE})(\mathbf{H}_{\mathrm{AMISE}})]$ is a key quantity to compute. From the definitions of PI and AMISE, Eq. (3) and Theorem 2 respectively,

$$(\mathrm{PI}-\mathrm{AMISE})(\mathbf{H}) = \tfrac{1}{4}m_2(K)^2(\mathrm{vec}^T\,\mathbf{H}^2)[\boldsymbol{\psi}_2-\hat{\boldsymbol{\psi}}_2(\mathbf{G})].$$

Since $d((\mathrm{vec}^T\,\mathbf{H}^2)\,\mathrm{vec}\,\mathbf{A}) = \mathrm{vec}^T(\mathbf{A}\mathbf{H}+\mathbf{H}\mathbf{A})(d\mathrm{vec}\,\mathbf{H}) = (\mathrm{vec}^T\mathbf{A})(\mathbf{H}\otimes\mathbf{I}_d+\mathbf{I}_d\otimes\mathbf{H})(d\mathrm{vec}\,\mathbf{H})$, then $\mathsf{D}_{\mathbf{H}}[(\mathrm{PI}-\mathrm{AMISE})](\mathbf{H}) = \frac{1}{4}m_2(K)^2(\mathbf{H}\otimes\mathbf{I}_d+\mathbf{I}_d\otimes\mathbf{H})[\boldsymbol{\psi}_2-\hat{\boldsymbol{\psi}}_2(\mathbf{G})]$. The mean squared error is $\mathrm{MSE}\{\mathsf{D}_{\mathbf{H}}[(\mathrm{PI}-\mathrm{AMISE})]\}(\mathbf{H}) = \frac{1}{16}m_2(K)^4(\mathbf{H}\otimes\mathbf{I}_d+\mathbf{I}_d\otimes\mathbf{H})\mathrm{MSE}[\hat{\boldsymbol{\psi}}_2(\mathbf{G})](\mathbf{H}\otimes\mathbf{I}_d+\mathbf{I}_d\otimes\mathbf{H})$, i.e.

$$\begin{aligned}\mathrm{MSE}\{\mathsf{D}_{\mathbf{H}}[(\mathrm{PI}-\mathrm{AMISE})]\}(\mathbf{H}_{\mathrm{AMISE}}) &= \mathrm{MSE}[\hat{\boldsymbol{\psi}}_2(\mathbf{G})](\mathrm{vec}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vec}^T\,\mathbf{H}_{\mathrm{AMISE}})\\ &= O(n^{-4/(d+4)})(\mathrm{vec}\,\mathbf{H}_{\mathrm{AMISE}})(\mathrm{vec}^T\,\mathbf{H}_{\mathrm{AMISE}})\end{aligned}$$

since Theorem 2 from Chacón and Duong (2010) shows that the infimum of MSE $\hat{\boldsymbol{\psi}}_2(\mathbf{G})$ has order $n^{-4/(d+4)}$. So $\hat{\mathbf{H}}_{\mathrm{PI}}$ converges to $\mathbf{H}_{\mathrm{AMISE}}$ at rate $n^{-2/(d+4)}$. Since this rate dominates $n^{-4/3}$, then $\hat{\mathbf{H}}_{\mathrm{PI}}$ converges to $\mathbf{H}_{\mathrm{MISE}}$ at the same rate. □

Proof of Theorem 4. From Azzalini (1985), the expected value of the skew normal distribution with parameter λ is $\int_{-\infty}^{\infty}2x\phi(x)\Phi(\lambda x)\,dx = \lambda\sqrt{2/[\pi(1+\lambda^2)]}$. By setting $\lambda = 1$, we obtain that $m_1 = m_1(\phi\Phi) = (4\pi)^{-1/2}$. Since the components of $K = \phi_{\mathbf{I}_d}$ are independent of each other then $\boldsymbol{m}_1(\phi_{\mathbf{I}_d}\Phi_{\mathbf{I}_d}) = m_1\mathbf{1}_d$. As usual, $m_2(\phi_{\mathbf{I}_d}) = 1$. From Chacón and Duong (2010, Formula (7)), $\boldsymbol{\psi}_2^{\mathrm{NS}} = -\frac{1}{2}(4\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\mathrm{vec}\boldsymbol{\Sigma}^{-1}$ so the normal scale estimate of the AMISE follows as $\mathrm{AMISE}_{\mathrm{NS}}(\mathbf{H}) = n^{-1}V_1(\Phi_{\boldsymbol{\Sigma}})-2n^{-1}(4\pi)^{-d/2}\mathrm{tr}(\mathbf{H}^{1/2}\mathbf{J}_d)+\frac{1}{8}(4\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\mathrm{tr}(\mathbf{H}^2\boldsymbol{\Sigma}^{-1})$. For the special case of a bivariate normal with a standard correlation matrix $\boldsymbol{\Sigma} = [1,\rho;\rho,1]$, for $\mathbf{H} = [h^2, h_{12}; h_{12}, h^2]$, the matrix square root is $\mathbf{H}^{1/2} = (2h^2+2(h^2-h_{12}^2)^{1/2})^{-1/2}[h^2+(h^4-h_{12}^2)^{1/2}, h_{12}; h_{12}, h^2+(h^4-h_{12}^2)^{1/2}]$ which gives

$$\mathrm{AMISE}_{\mathrm{NS}}(\mathbf{H}) = n^{-1}V_1(\Phi_\rho)-2^{1/2}\pi^{-1/2}n^{-1}\frac{h^2+(h^4-h_{12}^2)^{1/2}+h_{12}}{[h^2+(h^4-h_{12}^2)^{1/2}]^{1/2}}+\tfrac{1}{4}(4\pi)^{-d/2}(1-\rho^2)^{-1/2}(h^2+h_{12}^2-2\rho h^2h_{12}).$$

For $\mathbf{H} = [h^2, 0; 0, h^2]$, we have $\mathrm{AMISE}_{\mathrm{NS}}(\mathbf{H}) = n^{-1}V_1(\Phi_\rho)-2\pi^{-1/2}n^{-1}h+\frac{1}{4}(4\pi)^{-d/2}(1-\rho^2)^{-1/2}h^2$ and the result follows. □

A.2. Proofs for Section 3

The proof of Theorem 5 requires the following Lemmas 1 and 2. Let $\boldsymbol{\eta}$ be a d-dimensional parameter (non-random vector) and $q : \mathbb{R}^d \to \mathbb{R}$ be a multivariate non-random function. If $\hat{\boldsymbol{\eta}}$ and $\hat{q}(\cdot)$ are estimators of $\boldsymbol{\eta}$ and $q(\cdot)$, we study the properties of the 'double plug-in' estimator $\hat{q}(\hat{\boldsymbol{\eta}})$, as well as the case where $\boldsymbol{\eta}$, $\hat{\boldsymbol{\eta}}$ are defined as the solutions of $q(\boldsymbol{\eta}) = \hat{q}(\hat{\boldsymbol{\eta}}) = c$, for a constant c.

Lemma 1. *Suppose that $q : \mathbb{R}^d \to \mathbb{R}$ has continuously differentiable second order partial derivatives, and that $\hat{q}(\cdot)$, $\hat{\boldsymbol{\eta}}$ are estimators of q, $\boldsymbol{\eta}$ based on a random sample of size n which have finite variance.*

(i) *As $n \to \infty$, the mean squared error of $\hat{q}(\hat{\boldsymbol{\eta}})$ is*

$$\operatorname{MSE} \hat{q}(\hat{\boldsymbol{\eta}}) = \{\operatorname{MSE} \hat{q}(\boldsymbol{\eta}) + \mathrm{D}q(\boldsymbol{\eta})^T (\operatorname{MSE} \hat{\boldsymbol{\eta}}) \mathrm{D}q(\boldsymbol{\eta}) + 2 \operatorname{Cov}(\hat{q}(\boldsymbol{\eta}), \hat{\boldsymbol{\eta}}^T \mathrm{D}q(\boldsymbol{\eta}))\}\{1 + o(1)\}.$$

(ii) *Further suppose that $\boldsymbol{\eta}$, $\hat{\boldsymbol{\eta}}$ are solutions to the equations $q(\boldsymbol{\eta}) = \hat{q}(\hat{\boldsymbol{\eta}}) = c$ for a constant c which does not depend on $\boldsymbol{\eta}$, $\hat{\boldsymbol{\eta}}$. As $n \to \infty$, the mean squared error of $\hat{\boldsymbol{\eta}}$ is implicitly defined as*

$$\mathrm{D}q(\boldsymbol{\eta})^T (\operatorname{MSE} \hat{\boldsymbol{\eta}}) \mathrm{D}q(\boldsymbol{\eta}) = \{\operatorname{MSE} \hat{q}(\boldsymbol{\eta})\}\{1 + o(1)\}.$$

Proof of Lemma 1. (i) Following Lloyd (1998) and expanding $\hat{q}(\hat{\boldsymbol{\eta}})$ about $\boldsymbol{\eta}$ yields

$$\begin{aligned}
\hat{q}(\hat{\boldsymbol{\eta}}) &= \{(\hat{q} - q)(\hat{\boldsymbol{\eta}})\} + \{q(\hat{\boldsymbol{\eta}})\} = \{(\hat{q} - q)(\boldsymbol{\eta} + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}))\} + \{q(\boldsymbol{\eta} + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}))\} \\
&= \{(\hat{q} - q)(\boldsymbol{\eta})[1 + o(1)]\} + \{q(\boldsymbol{\eta}) + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T \mathrm{D}q(\boldsymbol{\eta}) + \tfrac{1}{2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T \mathrm{D}^2 q(\boldsymbol{\eta})(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})[1 + o(1)]\} \\
&= q(\boldsymbol{\eta}) + \{(\hat{q} - q)(\boldsymbol{\eta}) + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T \mathrm{D}q(\boldsymbol{\eta}) + \tfrac{1}{2}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T \mathrm{D}^2 q(\boldsymbol{\eta})(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})\}\{1 + o(1)\}.
\end{aligned}$$

Taking expected values yields $\operatorname{Bias} \hat{q}(\hat{\boldsymbol{\eta}}) = \operatorname{Bias} \hat{q}(\boldsymbol{\eta}) + (\operatorname{Bias}^T \hat{\boldsymbol{\eta}}) \mathrm{D}q(\boldsymbol{\eta}) + \frac{1}{2}\operatorname{tr}[(\operatorname{MSE} \hat{\boldsymbol{\eta}}) \mathrm{D}^2 q(\boldsymbol{\eta})]\{1 + o(1)\}$. For the variance, truncating $\hat{q}(\hat{\boldsymbol{\eta}})$ at the linear term and $\mathbb{E}\hat{q}(\hat{\boldsymbol{\eta}})$ at the constant term, we obtain

$$\begin{aligned}
\operatorname{Var} \hat{q}(\hat{\boldsymbol{\eta}}) &= \mathbb{E}[\hat{q}(\hat{\boldsymbol{\eta}}) - \mathbb{E}\hat{q}(\hat{\boldsymbol{\eta}})]^2 = \mathbb{E}[\hat{q}(\boldsymbol{\eta}) + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T \mathrm{D}q(\boldsymbol{\eta}) - \mathbb{E}\hat{q}(\boldsymbol{\eta})]^2\{1 + o(1)\} \\
&= \left\{\mathbb{E}[\hat{q}(\boldsymbol{\eta}) - \mathbb{E}\hat{q}(\boldsymbol{\eta})]^2 + \mathbb{E}[(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T \mathrm{D}q(\boldsymbol{\eta})]^2 + 2\mathbb{E}\{[\hat{q}(\boldsymbol{\eta}) - \mathbb{E}\hat{q}(\boldsymbol{\eta})][(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T \mathrm{D}q(\boldsymbol{\eta})]\}\right\}\{1 + o(1)\} \\
&= \{\operatorname{Var} \hat{q}(\boldsymbol{\eta}) + \mathrm{D}q(\boldsymbol{\eta})^T (\operatorname{MSE} \hat{\boldsymbol{\eta}}) \mathrm{D}q(\boldsymbol{\eta}) + 2 \operatorname{Cov}(\hat{q}(\boldsymbol{\eta}), \hat{\boldsymbol{\eta}}^T \mathrm{D}q(\boldsymbol{\eta}))\}\{1 + o(1)\}.
\end{aligned}$$

To obtain a MSE expression, it is more efficient to compute it directly from $\hat{q}(\hat{\boldsymbol{\eta}}) - q(\boldsymbol{\eta}) = \{(\hat{q} - q)(\boldsymbol{\eta}) + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T \mathrm{D}q(\boldsymbol{\eta})\}\{1 + o(1)\}$ than from the usual squared bias and variance sum. That is

$$\begin{aligned}
\operatorname{MSE} \hat{q}(\hat{\boldsymbol{\eta}}) &= \mathbb{E}[(\hat{q} - q)(\boldsymbol{\eta}) + (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T \mathrm{D}q(\boldsymbol{\eta})]^2\{1 + o(1)\} \\
&= \{\operatorname{MSE} \hat{q}(\boldsymbol{\eta}) + \mathrm{D}q(\boldsymbol{\eta})^T (\operatorname{MSE} \hat{\boldsymbol{\eta}}) \mathrm{D}q(\boldsymbol{\eta}) + 2 \operatorname{Cov}(\hat{q}(\boldsymbol{\eta}), \hat{\boldsymbol{\eta}}^T \mathrm{D}q(\boldsymbol{\eta}))\}\{1 + o(1)\}.
\end{aligned}$$

(ii) For the moments of $\hat{\boldsymbol{\eta}}$, it is straightforward to infer that

$$(\operatorname{Bias}^T \hat{\boldsymbol{\eta}}) \mathrm{D}q(\boldsymbol{\eta}) = \operatorname{Bias} \hat{q}(\hat{\boldsymbol{\eta}}) - \operatorname{Bias} \hat{q}(\boldsymbol{\eta}) - \tfrac{1}{2}\operatorname{tr}[(\operatorname{MSE} \hat{\boldsymbol{\eta}}) \mathrm{D}^2 q(\boldsymbol{\eta})]\{1 + o(1)\}.$$

We rearrange the linear truncation of $\hat{q}(\hat{\boldsymbol{\eta}})$ to isolate $(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T \mathrm{D}q(\boldsymbol{\eta}) = \{\hat{q}(\hat{\boldsymbol{\eta}}) - \hat{q}(\boldsymbol{\eta})\}\{1 + o(1)\}$. Squaring each side and then taking expectations, we obtain

$$\begin{aligned}
\mathrm{D}q(\boldsymbol{\eta})^T (\operatorname{MSE} \hat{\boldsymbol{\eta}}) \mathrm{D}q(\boldsymbol{\eta}) &= \mathbb{E}[\hat{q}(\hat{\boldsymbol{\eta}}) - q(\boldsymbol{\eta}) + q(\boldsymbol{\eta}) - \hat{q}(\boldsymbol{\eta})]^2\{1 + o(1)\} \\
&= \{\operatorname{MSE} \hat{q}(\hat{\boldsymbol{\eta}}) + \operatorname{MSE} \hat{q}(\boldsymbol{\eta}) + 2\mathbb{E}[(\hat{q}(\hat{\boldsymbol{\eta}}) - q(\boldsymbol{\eta}))(\hat{q}(\boldsymbol{\eta}) - q(\boldsymbol{\eta}))]\}\{1 + o(1)\}. \quad \square
\end{aligned}$$

Lemma 1(i) allows us to use the analysis of the simpler single plug-in estimator $\hat{q}(\boldsymbol{\eta})$ instead of those of the more complicated, double plug-in estimator $\hat{q}(\hat{\boldsymbol{\eta}})$. Likewise Lemma 1(ii) implies that the more complicated analysis of the vector estimator $\hat{\boldsymbol{\eta}}$ is asymptotically equivalent to that of the scalar valued estimator $\hat{q}(\boldsymbol{\eta})$.

The proof of Theorem 5 further requires an auxiliary result for kernel estimators of quantile functions, as defined, for example by Azzalini (1981) and Nadaraya (1964). Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$ be the z-quantile of $\boldsymbol{X}$ if $\boldsymbol{\theta} = \boldsymbol{\theta}(z)$ satisfies $F(\boldsymbol{\theta}(z)) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{\theta}(z)) = z, 0 \leq z \leq 1$. This is not the only way to define multivariate quantiles: others have been proposed, see for example Serfling (2002) for a review of approaches based on depth functions, norm minimisation, inversion maps, gradient search and quantile processes. We use this definition as it is intuitively the inverse operation to the cumulative distribution function. With the kernel estimator of the distribution function $\hat{F}$, the implicit definition of the kernel estimator of the z-quantile is $\hat{\boldsymbol{\theta}}(z; \mathbf{H})$ where $\hat{F}(\hat{\boldsymbol{\theta}}(z; \mathbf{H})) = z$. To apply Lemma 1 to obtain the moments of $\hat{\boldsymbol{\theta}}(z; \mathbf{H})$, we set $q = F$, $\boldsymbol{\eta} = \boldsymbol{\theta}(z)$, and the estimators $\hat{q} = \hat{F}(\cdot; \mathbf{H})$, $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\theta}}(z; \mathbf{H})$.

Lemma 2 extends the mean squared error analysis of univariate quantile estimators of Azzalini (1981) and Falk (1984), confirming that their assertions that kernel quantile estimators are asymptotically more efficient than empirical quantiles are equally valid for multivariate quantiles. These authors (amongst others) state closely related results, but only Cheng and Sun (2006) appears to be the exact counterpart to Lemma 2 for $d = 1$. It is straightforward to show that an equivalent result holds for the upper quantiles $\bar{\boldsymbol{\theta}}(z)$, $\hat{\bar{\boldsymbol{\theta}}}(z; \mathbf{H})$ defined as $\bar{F}(\bar{\boldsymbol{\theta}}(z)) = \hat{\bar{F}}(\hat{\bar{\boldsymbol{\theta}}}(z)) = z$.

Lemma 2. *Suppose that the conditions* (A1)–(A4) *hold. For a non-random value* z, $0 \le z \le 1$, *the expected value, variance and MSE of the kernel quantile estimator* $\hat{\boldsymbol{\theta}}(z;\ \mathbf{H})$ *are implicitly defined by*

$$\begin{aligned}\mathrm{D}F(\boldsymbol{\theta}(z))^T[\mathrm{MSE}\,\hat{\boldsymbol{\theta}}(z;\mathbf{H})]\mathrm{D}F(\boldsymbol{\theta}(z)) &= \left\{n^{-1}z(1-z) - 2n^{-1}\boldsymbol{m}_1(K\mathcal{K})^T\,\mathbf{H}^{1/2}\mathrm{D}F(\boldsymbol{\theta}(z))\right.\\ &\quad\left. +\ \tfrac{1}{4}m_2(K)^2\mathrm{tr}^2(\mathbf{H}\,\mathrm{D}^2F(\boldsymbol{\theta}(z)))\right\}\{1+o(1)\}.\end{aligned}$$

Proof of Lemma 2. Applying Lemma 1 to $q = F$, $\hat{q} = \hat{F}$, and $\boldsymbol{\eta} = \boldsymbol{\theta}(z)$, $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\theta}}(z)$, we have

$$\begin{aligned}(\mathrm{Bias}^T\hat{\boldsymbol{\theta}})\mathrm{D}F(\boldsymbol{\theta}) &= \left\{\mathrm{Bias}\,\hat{F}(\hat{\boldsymbol{\theta}};\ \mathbf{H}) - \mathrm{Bias}\,\hat{F}(\boldsymbol{\theta};\ \mathbf{H}) - \tfrac{1}{2}\mathrm{tr}[(\mathrm{MSE}\,\hat{\boldsymbol{\theta}})\mathrm{D}^2F(\boldsymbol{\theta})]\right\}\{1+o(1)\}\\ &= \left\{-\mathrm{Bias}\,\hat{F}(\boldsymbol{\theta};\ \mathbf{H}) - \tfrac{1}{2}\mathrm{tr}[(\mathrm{MSE}\,\hat{\boldsymbol{\theta}})\mathrm{D}^2F(\boldsymbol{\theta})]\right\}\{1+o(1)\}\end{aligned}$$

since $\hat{F}(\hat{\boldsymbol{\theta}};\ \mathbf{H}) = z$ is a constant, then $\mathrm{Bias}\,\hat{F}(\hat{\boldsymbol{\theta}};\ \mathbf{H})$ is identically zero. From Theorem 1, then

$$(\mathrm{Bias}^T\hat{\boldsymbol{\theta}})\mathrm{D}F(\boldsymbol{\theta}) = \left\{-\tfrac{1}{2}m_2(K)\mathrm{tr}(\mathbf{H}\,\mathrm{D}^2F(\boldsymbol{\theta})) - \tfrac{1}{2}\mathrm{tr}[(\mathrm{MSE}\,\hat{\boldsymbol{\theta}})\mathrm{D}^2F(\boldsymbol{\theta})]\right\}\{1+o(1)\}.$$

To simplify the bias, let $\mathrm{Bias}\,\hat{\boldsymbol{\theta}} = c\boldsymbol{b}\{1+o(1)\}$ where $\boldsymbol{b}$ is a d-vector not involving n, and c is scalar which collects all terms involving n. Then we have $c\boldsymbol{b}^T\mathrm{D}F = -\frac{1}{2}m_2(K)\{\mathrm{tr}[(\mathbf{H}+\mathrm{Var}\,\hat{\boldsymbol{\theta}})\mathrm{D}^2F]+c^2(\boldsymbol{b}^T\mathrm{D}^2F\boldsymbol{b})\}$ or $\frac{1}{2}m_2(K)\boldsymbol{b}^T\mathrm{D}^2F\boldsymbol{b}c^2+\boldsymbol{b}^T\boldsymbol{\xi}c+\frac{1}{2}m_2(K)v = 0$ where $v = \mathrm{tr}((\mathbf{H}+\mathrm{Var}\,\hat{\boldsymbol{\theta}})\mathrm{D}^2F)$. The solution of the quadratic equation is

$$\begin{aligned}c &= \{-\boldsymbol{b}^T\mathrm{D}F + [(\boldsymbol{b}^T\mathrm{D}F)^2 - m_2(K)^2\boldsymbol{b}^T\mathrm{D}^2F\boldsymbol{b}]^{1/2}\}/[m_2(K)\boldsymbol{b}^T\mathrm{D}^2F\boldsymbol{b}]\\ &= \boldsymbol{b}^T\mathrm{D}F/[m_2(K)\boldsymbol{b}^T\mathrm{D}^2F\boldsymbol{b}]\{-1+[1-m_2(K)^2v\boldsymbol{b}^T\mathrm{D}^2F\boldsymbol{b}/(\boldsymbol{b}^T\mathrm{D}F)^2]^{1/2}\}\\ &= \boldsymbol{b}^T\mathrm{D}F/[m_2(K)\boldsymbol{b}^T\mathrm{D}^2F\boldsymbol{b}]\left\{-1+1-\tfrac{1}{2}m_2(K)^2v\boldsymbol{b}^T\mathrm{D}^2F\boldsymbol{b}/(\boldsymbol{b}^T\mathrm{D}F)^2\right\}\{1+o(1)\}\\ &= -\tfrac{1}{2}m_2(K)v/(\boldsymbol{b}^T\mathrm{D}F)\{1+o(1)\}\end{aligned}$$

where the third equality follows from the Taylor expansion $(1-x)^{1/2} = 1-\frac{1}{2}x+o(x)$. This implies that $(\mathrm{Bias}^T\hat{\boldsymbol{\theta}})\mathrm{D}F(\boldsymbol{\theta}) = -\frac{1}{2}m_2(K)\mathrm{tr}[(\mathbf{H}+\mathrm{Var}\,\hat{\boldsymbol{\theta}})\mathrm{D}^2F(\boldsymbol{\theta})]\{1+o(1)\}$, i.e. $\mathrm{Var}\,\hat{\boldsymbol{\theta}}$ has replaced $\mathrm{MSE}\,\hat{\boldsymbol{\theta}}$ in the inner parenthesis.

For the variance, we again apply Lemma 1, since we note that $\hat{F}(\hat{\boldsymbol{\theta}};\ \mathbf{H}) = z$,

$$\begin{aligned}\mathrm{D}F(\boldsymbol{\theta})^T(\mathrm{Var}\,\hat{\boldsymbol{\theta}})\mathrm{D}F(\boldsymbol{\theta}) &= \{\mathrm{Var}\,\hat{F}(\hat{\boldsymbol{\theta}};\ \mathbf{H}) + \mathrm{Var}\,\hat{F}(\boldsymbol{\theta};\ \mathbf{H}) + 2\,\mathrm{Cov}(\hat{F}(\hat{\boldsymbol{\theta}};\ \mathbf{H}),\hat{F}(\boldsymbol{\theta};\ \mathbf{H}))\}\{1+o(1)\}\\ &= \{\mathrm{Var}\,\hat{F}(\boldsymbol{\theta};\ \mathbf{H})\}\{1+o(1)\} = \{n^{-1}z(1-z) - 2n^{-1}\boldsymbol{m}_1(K\mathcal{K})^T\,\mathbf{H}^{1/2}\mathrm{D}F(\boldsymbol{\theta})\}\{1+o(1)\}\end{aligned}$$

and for the mean squared error,

$$\begin{aligned}\mathrm{D}F(\boldsymbol{\theta})^T(\mathrm{MSE}\,\hat{\boldsymbol{\theta}})\mathrm{D}F(\boldsymbol{\theta}) &= \{\mathrm{MSE}\,\hat{F}(\hat{\boldsymbol{\theta}};\ \mathbf{H}) + \mathrm{MSE}\,\hat{F}(\boldsymbol{\theta};\ \mathbf{H}) + 2\,\mathrm{Cov}(\hat{F}(\hat{\boldsymbol{\theta}};\ \mathbf{H}),\hat{F}(\boldsymbol{\theta};\ \mathbf{H}))\}\{1+o(1)\}\\ &= \{\mathrm{MSE}\,\hat{F}(\boldsymbol{\theta};\ \mathbf{H})\}\{1+o(1)\}\\ &= \left\{n^{-1}z(1-z) - 2n^{-1}\boldsymbol{m}_1(K\mathcal{K})^T\,\mathbf{H}^{1/2}\mathrm{D}F(\boldsymbol{\theta}) + \tfrac{1}{4}m_2(K)^2\mathrm{tr}^2(\mathbf{H}\,\mathrm{D}^2F(\boldsymbol{\theta}))\right\}\{1+o(1)\}.\quad\square\end{aligned}$$

For the proofs of Theorems 5–7, we require the definition of the intermediate quantities $Y_i = F_{\boldsymbol{X}}(\boldsymbol{X}_i)$, $i = 1,\ldots,n$, and $\hat{F}_Y(z;h) = n^{-1}\sum_{i=1}^n \mathcal{L}_h(z-Y_i)$.

Proof of Theorem 5. Using the law of total expectation, we have $\mathbb{E}\hat{F}_{\hat{Y}_2}(z) = \mathbb{E}[\mathbb{E}\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1]$ implying that $\mathrm{Bias}\,\hat{F}_{\hat{Y}_2}(z) = \mathbb{E}\hat{F}_{\hat{Y}_2}(z) - F_{Y_2} = \mathbb{E}[\mathbb{E}\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] - F_{Y_2} = \mathbb{E}[\mathbb{E}\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] - \mathbb{E}[\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] + \mathbb{E}[\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] - F_{Y_2} = \mathrm{Bias}(\mathbb{E}\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1) + \mathrm{Bias}(\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1)$, thus $\mathrm{Bias}^2\hat{F}_{\hat{Y}_2}(z) = \mathrm{Bias}^2[\mathbb{E}\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] + \mathrm{Bias}^2[\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1]$. The law of total variance yields $\mathrm{Var}\,\hat{F}_{\hat{Y}_2}(z) = \mathrm{Var}\,[\mathbb{E}\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] + \mathbb{E}\mathrm{Var}[\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1]$. Combining these gives the unconditional MSE as

$$\begin{aligned}\mathrm{MSE}\,\hat{F}_{\hat{Y}_2}(z) &= \mathrm{Var}[\mathbb{E}\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] + \mathbb{E}\mathrm{Var}[\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] + \mathrm{Bias}^2(\mathbb{E}\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1) + \mathbb{E}\mathrm{Bias}^2[\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1]\\ &= \mathrm{MSE}[\mathbb{E}\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] + \mathbb{E}[\mathrm{MSE}\,\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1].\end{aligned}$$

The conditional moments of $\hat{F}_{\hat{Y}_2}$ given $\mathcal{X}_1$ are, from Theorem 1,

$$\begin{aligned}&\mathbb{E}[\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] = \left\{F_{\hat{Y}_2}(z) + \tfrac{1}{2}m_2(L)h_2^2F''_{\hat{Y}_2}(z)\right\}\{1+o(1)\}\\ &\mathrm{Var}[\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] = \{n_2^{-1}F_{\hat{Y}_2}(z)\bar{F}_{\hat{Y}_2}(z) - 2n_2^{-1}h_2m_1(L\mathcal{L})F'_{\hat{Y}_2}(z)\}\{1+o(1)\}\end{aligned}$$

where $F_{\hat{Y}_2}(z) = \bar{F}_{\boldsymbol{X}_2}(\hat{\bar{\boldsymbol{\theta}}}_{\boldsymbol{X}_1}(z))$. The first term in the unconditional MSE $\hat{F}_{\hat{Y}_2}$ is

$$\begin{aligned}
&\mathrm{MSE}[\mathbb{E}(\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1)] = \mathrm{MSE}\, F_{\hat{Y}_2}(z)\{1+o(1)\} = \mathrm{MSE}\, \bar{F}_{\boldsymbol{X}_2}(\hat{\bar{\boldsymbol{\theta}}}_{\boldsymbol{X}_1})\{1+o(1)\}\\
&= \mathsf{D}\bar{F}_{\boldsymbol{X}_2}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1})^T \mathrm{MSE}[\hat{\bar{F}}_{\boldsymbol{X}_1}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1})]\mathsf{D}\bar{F}_{\boldsymbol{X}_2}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1})\{1+o(1)\} = \mathrm{MSE}[\hat{\bar{F}}_{\boldsymbol{X}_1}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1})]f_{\boldsymbol{X}_2}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1})^2/f_{\boldsymbol{X}_1}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1})^2\{1+o(1)\}\\
&= \Big[n_1^{-1}z(1-z) - 2n_1^{-1}\boldsymbol{m}_1(K\mathcal{K})\,\mathbf{H}_1^{1/2}\mathsf{D}F_{\boldsymbol{X}_1}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1}) + \tfrac{1}{4}m_2(K)^2\mathrm{tr}^2(\mathbf{H}_1\mathsf{D}^2F_{\boldsymbol{X}_1}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1}))\Big]f_{\boldsymbol{X}_2}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1})^2/f_{\boldsymbol{X}_1}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1})^2\{1+o(1)\}
\end{aligned}$$

using techniques similar to those in the proof of Lemma 2. For the second term of the unconditional MSE $\hat{F}_{\hat{Y}_2}$, we first evaluate that the expected value of $F_{\hat{Y}_2}$ is $\mathbb{E}F_{\hat{Y}_2}(z) = \mathbb{E}\bar{F}_{\boldsymbol{X}_2}(\hat{\bar{\boldsymbol{\theta}}}_{\boldsymbol{X}_1}(z)) = \bar{F}_{\boldsymbol{X}_2}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1}(z))\{1+o(1)\} = F_{Y_2}(z)\{1+o(1)\}$. Exchanging the order of differentiation and expectation, $\mathbb{E}F'_{\hat{Y}_2}(z) = (\partial/\partial z)\mathbb{E}[F_{\hat{Y}_2}(z)] = F'_{Y_2}(z)\{1+o(1)\}$ and so on for higher order derivatives. Then

$$\begin{aligned}
\mathbb{E}\mathrm{Bias}^2[\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] &= \tfrac{1}{4}m_2(L)^2h_2^4\mathbb{E}[F''_{\hat{Y}_2}(z)]^2\{1+o(1)\} = \tfrac{1}{4}m_2(L)^2h_2^4F''_{Y_2}(z)^2\{1+o(1)\}\\
\mathbb{E}\mathrm{Var}[\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] &= \{n_2^{-1}\mathbb{E}[F_{\hat{Y}_2}(z)(1-F_{\hat{Y}_2}(z))] - n_2^{-1}h_2m_1(L\mathcal{L})\mathbb{E}F'_{\hat{Y}_2}(z)\}\{1+o(1)\}\\
&= \{n_2^{-1}F_{Y_2}(z)(1-F_{Y_2}(z)) - n_2^{-1}h_2m_1(L\mathcal{L})F'_{Y_2}(z)\}\{1+o(1)\}
\end{aligned}$$

and so $\mathbb{E}\mathrm{MSE}[\hat{F}_{\hat{Y}_2}(z)|\mathcal{X}_1] = \{\mathrm{MSE}\,\hat{F}_{Y_2}(z)\}\{1+o(1)\}$. Thus $\mathrm{MSE}\,\hat{F}_{\hat{Y}_2}(z)$ can be decomposed into two (conditionally) independent components

$$\begin{aligned}
\mathrm{MSE}\,\hat{F}_{\hat{Y}_2}(z) = \Big\{&[\mathrm{MSE}\,\hat{F}_{\boldsymbol{X}_1}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1})]f_{\boldsymbol{X}_2}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1})^2/f_{\boldsymbol{X}_1}(\bar{\boldsymbol{\theta}}_{\boldsymbol{X}_1})^2\\
&+ n_2^{-1}F_{Y_2}(z)(1-F_{Y_2}(z)) - 2n_2^{-1}h_2m_1(L\mathcal{L})F'_{Y_2}(z) + \tfrac{1}{4}m_2(L)^2h_2^4F''_{Y_2}(z)^2\Big\}\{1+o(1)\}.
\end{aligned}$$

For univariate data, this MSE expression reduces to well-known expressions, e.g. Lloyd (1998). Integrating the MSE we obtain the result. □

Proof of Theorem 6. Integrating the leading terms of $\mathrm{MSE}\,\hat{F}_{\boldsymbol{X}_1}(\boldsymbol{x};\ \mathbf{H}_1)$ from Theorem 1 and $\mathbb{E}[\mathrm{MISE}\,\hat{F}_{\hat{Y}_2}(\cdot;\ h_2)|\mathcal{X}_1]$ from Theorem 5, we have

$$\begin{aligned}
\mathrm{MISE}\,\hat{F}_{\hat{Y}_2}(\cdot;\ \mathbf{H}_1, h_2) &= \int_0^1 \Big\{n_1^{-1}z(1-z)f_{\boldsymbol{X}_2}(\boldsymbol{x})^2/f_{\boldsymbol{X}_1}(\boldsymbol{x})^2 + n_2^{-1}F_{Y_2}(z)(1-F_{Y_2}(z))\\
&\quad + O(n_1^{-1}\mathrm{tr}\,\mathbf{H}_1^{1/2} + \mathrm{tr}\,\mathbf{H}_1^2 + n_2^{-1}h_2 + h_2^4)\Big\}\,dz\\
&= n_1^{-1}\int_{\mathbb{R}^d}\bar{F}_{\boldsymbol{X}_1}(\boldsymbol{x})(1-\bar{F}_{\boldsymbol{X}_1}(\boldsymbol{x}))f_{\boldsymbol{X}_2}(\boldsymbol{x})^2/f_{\boldsymbol{X}_1}(\boldsymbol{x})\,d\boldsymbol{x} + n_2^{-1}V_1(F_{Y_2})\\
&\quad + O(n_1^{-1}\mathrm{tr}\,\mathbf{H}_1^{1/2} + \mathrm{tr}\,\mathbf{H}_1^2 + n_2^{-1}h_2 + h_2^4).
\end{aligned}$$

For both joint and sequential optimisation, the dominant terms in the minimal MISE remain the two terms not involving the bandwidths. □

Proof of Theorem 7. For clarity, we omit the 1 and 2 subscripts on $\boldsymbol{X}$, $\mathbf{H}$, Y, $\hat{Y}$, h, and focus on cumulative distributions $F_{\boldsymbol{X}}$, as the results for survival functions $\bar{F}_{\boldsymbol{X}}$ follow analogously. Applying Lemma 1 with $q = F_Y$, $\boldsymbol{\eta} = Y = F_{\boldsymbol{X}}(\boldsymbol{X})$, and $\hat{q} = \hat{F}_Y$, $\hat{\boldsymbol{\eta}} = \hat{Y}(\mathbf{H}) = \hat{F}_{\boldsymbol{X}}(\boldsymbol{X};\ \mathbf{H})$, then

$$\mathrm{MSE}\,[\hat{F}_{\hat{Y}}(z;\mathbf{H},h)] = \{\mathrm{MSE}\,[\hat{F}_Y(z;h)] + f_Y(z)^2\mathrm{MSE}\,[\hat{Y}(\mathbf{H})] + 2f_Y(z)\,\mathrm{Cov}(\hat{F}_Y(z;h), \hat{Y}(\mathbf{H}))\}\{1+o(1)\}. \tag{A.1}$$

The first term $\mathrm{MSE}\,[\hat{F}_Y(z;h)]$ follows directly from Theorem 1 as

$$\mathrm{MSE}\,[\hat{F}_Y(z;h)] = \Big\{n^{-1}F_Y(z)(1-F_Y(z)) - 2n^{-1}hm_1(L\mathcal{L})f_Y(z) + \tfrac{1}{4}m_2(L)^2h^4f'_Y(z)^2\Big\}\{1+o(1)\}. \tag{A.2}$$

For the second term in Eq. (A.1), by (A6), we assume that $\boldsymbol{X}$ does not coincide with any of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, so $\mathbb{E}\hat{Y}(\mathbf{H}) = \mathbb{E}\hat{F}_{\boldsymbol{X}}(\boldsymbol{X}) = \mathbb{E}[\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)]$. Expanding the right hand side using similar techniques to those in the proof of Theorem 1, we obtain

$$\begin{aligned}
\mathbb{E}[\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)] &= \int_{\mathbb{R}^{2d}}\mathcal{K}_{\mathbf{H}}(\boldsymbol{x}-\boldsymbol{y})f_{\boldsymbol{X}}(\boldsymbol{x})f_{\boldsymbol{X}}(\boldsymbol{y})\,d\boldsymbol{x}d\boldsymbol{y} = \int_{\mathbb{R}^{2d}}K(\boldsymbol{w})F_{\boldsymbol{X}}(\boldsymbol{y} - \mathbf{H}^{1/2}\boldsymbol{w})f_{\boldsymbol{X}}(\boldsymbol{y})\,d\boldsymbol{w}d\boldsymbol{y}\\
&= \int_{\mathbb{R}^d}\Big[F_{\boldsymbol{X}}(\boldsymbol{y}) + \tfrac{1}{2}m_2(K)\mathrm{tr}(\mathbf{H}\,\mathsf{D}^2F_{\boldsymbol{X}}(\boldsymbol{y}))\Big]f_{\boldsymbol{X}}(\boldsymbol{y})\{1+o(1)\}\,d\boldsymbol{y}\\
&= \mathbb{E}Y + \tfrac{1}{2}m_2(K)\int_{\mathbb{R}^d}\big[\mathrm{tr}(\mathbf{H}\,\mathsf{D}^2F_{\boldsymbol{X}}(\boldsymbol{y}))\big]f_{\boldsymbol{X}}(\boldsymbol{y})\{1+o(1)\}\,d\boldsymbol{y} = \mu_Y\{1+o(1)\}
\end{aligned}$$

using that $\mathsf{D}^2 F(\boldsymbol{x}) = \mathsf{D} f_{\boldsymbol{X}}(\boldsymbol{x})\{1 + o(1)\}$ and $\int_{\mathbb{R}^d} \mathsf{D} f_{\boldsymbol{X}}(\boldsymbol{y}) f_{\boldsymbol{X}}(\boldsymbol{y})\, d\boldsymbol{y} = \boldsymbol{0}$. Likewise, the variance $\operatorname{Var} \hat{Y}(\mathbf{H}) = n^{-1}[\operatorname{Var} \mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)]\{1 + o(1)\}$, of which the remaining expression to evaluate is

$$\begin{aligned}
\mathbb{E}[\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)^2] &= \int_{\mathbb{R}^{2d}} \mathcal{K}_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{y})^2 f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{y})\, d\boldsymbol{x} d\boldsymbol{y} = \int_{\mathbb{R}^{2d}} 2\mathcal{K}(\boldsymbol{w}) K(\boldsymbol{w}) F_{\boldsymbol{X}}(\boldsymbol{y} - \mathbf{H}^{1/2}\boldsymbol{w}) f_{\boldsymbol{X}}(\boldsymbol{y})\, d\boldsymbol{w} d\boldsymbol{y} \\
&= \int_{\mathbb{R}^d} \Big[F_{\boldsymbol{X}}(\boldsymbol{y}) - \boldsymbol{m}_1(\mathcal{K}K)^T \mathbf{H}^{1/2} \mathsf{D} F_{\boldsymbol{X}}(\boldsymbol{y}) \Big] f_{\boldsymbol{X}}(\boldsymbol{y})\{1 + o(1)\}\, d\boldsymbol{y} \\
&= \mathbb{E}Y - \psi_{\boldsymbol{X},0} \boldsymbol{m}_1(\mathcal{K}K)^T \mathbf{H}^{1/2} \mathbf{1}_d \{1 + o(1)\}\, d\boldsymbol{y}
\end{aligned}$$

that is, $\operatorname{Var} \hat{Y}(\mathbf{H}) = \big\{ n^{-1}\mu_Y(1 - \mu_Y) - n^{-1}\psi_{\boldsymbol{X},0} \boldsymbol{m}_1(\mathcal{K}K)^T \mathbf{H}^{1/2} \mathbf{1}_d \big\}\{1 + o(1)\}$. Thus

$$\operatorname{MSE}[\hat{Y}(\mathbf{H})] = \Big\{ n^{-1}\mu_Y(1 - \mu_Y) - n^{-1}\psi_{\boldsymbol{X},0} \boldsymbol{m}_1(\mathcal{K}K)^T \mathbf{H}^{1/2} \mathbf{1}_d \Big\}\{1 + o(1)\}. \tag{A.3}$$

For the covariance term in Eq. (A.1), $\operatorname{Cov}(\hat{F}_Y(z; h), \hat{Y}(\mathbf{H}))$, we require

$$\begin{aligned}
\mathbb{E}[\hat{F}_Y(z; h)\hat{Y}(\mathbf{H})] &= n^{-2}\mathbb{E}\Big[\sum_{i,j=1}^{n} \mathcal{L}_h(z - Y_j) \mathcal{K}_{\mathbf{H}}(\boldsymbol{X} - \boldsymbol{X}_i) \Big] \\
&= n^{-1}\mathbb{E}[\mathcal{L}_h(z - Y_2)\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)] + (1 - n^{-1})[\mathbb{E}\mathcal{L}_h(z - Y_3)][\mathbb{E}\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)]
\end{aligned}$$

where $Y_\ell = \bar{F}_{\boldsymbol{X}}(\boldsymbol{X}_2)$, $\ell = 2, 3$, i.e.

$$\begin{aligned}
\operatorname{Cov}(\hat{F}_Y(z; h), \hat{Y}(\mathbf{H})) &= n^{-1}\mathbb{E}[\mathcal{L}_h(z - Y_2)\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)] - n^{-1}\mathbb{E}[\hat{F}_Y(z; h)]\mathbb{E}[\hat{Y}(\mathbf{H})] \\
&= n^{-1}\mathbb{E}[\mathcal{L}_h(z - Y_2)\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)] - n^{-1}\mu_Y F_Y(z) + O(n^{-1}(h^2 + \operatorname{tr}\mathbf{H})).
\end{aligned}$$

The remaining unevaluated term is

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_h(z - Y_2)\mathcal{K}_{\mathbf{H}}(\boldsymbol{X}_1 - \boldsymbol{X}_2)] &= \int_{\mathbb{R}^{2d}} \mathcal{L}_h(z - F_{\boldsymbol{X}}(\boldsymbol{x})) \mathcal{K}_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{y}) f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{y})\, d\boldsymbol{x} d\boldsymbol{y} \\
&= \int_{\mathbb{R}^{2d}} \mathcal{L}_h(z - F_{\boldsymbol{X}}(\boldsymbol{x})) K(\boldsymbol{w}) f_{\boldsymbol{X}}(\boldsymbol{x}) F_{\boldsymbol{X}}(\boldsymbol{x} - \mathbf{H}^{1/2}\boldsymbol{w})\, d\boldsymbol{w} d\boldsymbol{x} \\
&= \int_{\mathbb{R}^d} \mathcal{L}_h(z - F_{\boldsymbol{X}}(\boldsymbol{x})) f_{\boldsymbol{X}}(\boldsymbol{x}) F_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x} + O(\operatorname{tr}\mathbf{H}).
\end{aligned}$$

Since $F_{\boldsymbol{X}}$ and $\mathcal{K}_h$ are both monotonic distribution functions, there exists a bandwidth matrix $\mathbf{H}^*$ of the same asymptotic order as $\mathbf{H}$ such that $\mathcal{L}_h(z - F_{\boldsymbol{X}}(\boldsymbol{x})) = \mathcal{K}_{\mathbf{H}^*}(\boldsymbol{y} - \boldsymbol{x})$ for $z = F_{\boldsymbol{X}}(\boldsymbol{y})$. Thus

$$\begin{aligned}
&\int_{\mathbb{R}^d} \mathcal{L}_h(z - F_{\boldsymbol{X}}(\boldsymbol{x})) f_{\boldsymbol{X}}(\boldsymbol{x}) F_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int_{\mathbb{R}^d} \mathcal{K}_{\mathbf{H}^*}(\boldsymbol{y} - \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) F_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x} = \mathcal{K}_{\mathbf{H}^*}(\boldsymbol{y} - \boldsymbol{x}) \tfrac{1}{2} F_{\boldsymbol{X}}(\boldsymbol{x})^2 \Big|_{\boldsymbol{x} = -\infty}^{\boldsymbol{x} = \infty} + \int_{\mathbb{R}^d} K_{\mathbf{H}^*}(\boldsymbol{y} - \boldsymbol{x}) \tfrac{1}{2} F_{\boldsymbol{X}}(\boldsymbol{x})^2\, d\boldsymbol{x} \\
&= \tfrac{1}{2} \int_{\mathbb{R}^d} K(\boldsymbol{w}) F_{\boldsymbol{X}}(\boldsymbol{y} - \mathbf{H}^{*1/2}\boldsymbol{w})^2\, d\boldsymbol{w} = \tfrac{1}{2} F_{\boldsymbol{X}}(\boldsymbol{y})^2 + O(\operatorname{tr}\mathbf{H}^*) = \tfrac{1}{2} z^2 + O(\operatorname{tr}\mathbf{H}^*)
\end{aligned}$$

and

$$\operatorname{Cov}(\hat{F}_Y(z; h), \hat{Y}(\mathbf{H})) = n^{-1}\Big[\tfrac{1}{2} z^2 - (\mathbb{E}Y) F_Y(z) \Big] + O(n^{-1}(h^2 + \operatorname{tr}\mathbf{H})). \tag{A.4}$$

The remainder terms of order $n^{-1}(h^2 + \operatorname{tr}\mathbf{H})$ can be neglected as they are asymptotically dominated by the order $n^{-1}\operatorname{tr}\mathbf{H}^{1/2}$ terms in the variance in Eq. (A.3). Combining Eqs. (A.2)–(A.4) gives the desired MSE result. Upon integration,

$$\begin{aligned}
\operatorname{MISE}[\hat{F}_{\hat{Y}}(\cdot;\, \mathbf{H}, h)] = \int_0^1 \Big\{ & n^{-1}\big[F_{Y_1}(z)(1 - F_Y(z)) + f_Y(z)^2 \mu_Y(1 - \mu_Y) + z^2 - 2\mu_Y F_Y(z) \big] \\
& - 2n^{-1} h m_1(L\mathcal{L}) f_Y(z) + \tfrac{1}{4} m_2(L)^2 h^4 f_Y'(z)^2 - n^{-1}\psi_{\boldsymbol{X},0} \boldsymbol{m}_1(\mathcal{K}K)^T \mathbf{H}^{1/2} \mathbf{1}_d \Big\}\{1 + o(1)\}\, dz,
\end{aligned}$$

and the AMISE result follows since $\psi_{Y,0} = \int_0^1 f_Y'(z)^2\, dz$, $\psi_{Y,2} = \int_0^1 f_Y''(z) f_Y(z)\, dz$. □

References

Altman, N., & Léger, C. (1995). Bandwidth selection for kernel distribution function. *Journal of Statistical Planning and Inference*, *46*, 195–214.
Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, *68*, 326–328.
Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics, Theory and Applications*, *12*, 171–178.
Berg, A., & Politis, D. (2006). CDF and survival function estimation with infinite-order kernels. *Electronic Journal of Statistics*, *3*, 1436–1454.
Chacón, J. E. (2009). Data-driven choice of the smoothing parametrization for kernel density estimators. *Canadian Journal of Statistics*, *37*, 249–255.
Chacón, J. E., & Duong, T. (2010). Multivariate plug-in bandwidth selection with unconstrained bandwidth matrices. *TEST*, *19*, 375–398.
Chacón, J. E., Duong, T., & Wand, M. P. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, *21*, 807–840.
Cheng, M.-Y., & Sun, S. (2006). Bandwidth selection for kernel quantile estimation. *Journal of the Chinese Statistical Association*, *44*, 271–295.
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., & Hazelton, M. L. (2005). Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation. *Journal of Multivariate Analysis*, *93*, 417–433.
Falk, M. (1984). Relative deficiency of kernel type estimators of quantiles. *Annals of Statistics*, *12*, 261–268.
Finkel, R. S., Crawford, T. O., Swoboda, K. J., Kaufmann, P., Juhasz, P., Li, X. et al., On behalf of the Pilot Study of Biomarkers for Spinal Muscular Atrophy (BforSMA) Trial Group, (2012). Candidate proteins, metabolites and transcripts in the biomarkers for spinal muscular atrophy (BforSMA) clinical study. *PLoS One*, *7*, e35462.
Hall, P., & Hyndman, R. J. (2003). Improved methods for bandwidth selection when estimating ROC curves. *Statistics & Probability Letters*, *64*, 181–189.
Hall, P., & Marron, J. S. (1987). Estimation of integrated squared density derivatives. *Statistics & Probability Letters*, *6*, 109–115.
Handcock, M. S., & Morris, M. (1998). Relative distribution methods. *Sociological Methodology*, *28*, 53–97.
Holmquist, B. (1996). The *d*-variate vector Hermite polynomial of order *k*. *Linear Algebra and its Applications*, *237–238*, 155–190.
Hsieh, F., & Turnbull, B. W. (1996). Nonparametric methods for evaluatin diagnostic tests. *Statistica Sinica*, *6*, 47–62.
Jin, Z., & Shao, Y. (1999). On kernel estimation of a multivariate distribution function. *Statistics & Probability Letters*, *41*, 163–168.
Liu, R., & Yang, L. (2008). Kernel estimation of multivariate cumulative distribution function. *Journal of Nonparametric Statistics*, *20*, 661–677.
Lloyd, C. J. (1998). Using smoothed receiver operating curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, *93*, 1356–1364.
Lloyd, C. J., & Yong, Z. (1999). Kernel estimators of the ROC curve are better than empirical. *Statistics & Probability Letters*, *44*, 221–228.
Magnus, J. R., & Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics* (Revised ed.). Chichester: John Wiley and Sons.
Molanes-López, E. M., & Cao, R. (2008). Plug-in bandwidth selector for the kernel relative density estimator. *Annals of the Institute of Statistical Mathematics*, *60*, 273–300.
Nadaraya, E. (1964). Some new estimates for distribution functions. *Theory of Probability and its Applications*, 9, 497–500.
Pepe, M. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, *54*, 124–135.
Pepe, M. S., & Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*, *1*, 123–140.
Peterson, W., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *IRE Transactions on Information Theory*, *4*, 171–212.
Pfeiffer, R. M., & Bura, E. (2008). A model free approach to combining biomarkers. *Biometricl Journal*, *50*, 558–570.
Polansky, A. M., & Baker, E. R. (2000). Multistage plug-in bandwidth selection for kernel distribution function estimates. *Journal of Statistical Computation and Simulation*, *65*, 63–80.
Reiss, R.-D. (1981). Non-parametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, *8*, 116–119.
Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions. *Journal of Statistical Planning and Inference*, *35*, 65–75.
Serfling, R. (2002). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, *56*, 214–232.
Shapiro, D. (1999). The interpretation of diagnostic tests. *Statistical Methods in Medical Research*, *8*, 113–134.
Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer-Verlag.
Su, J., & Liu, J. (1993). Linear combinations of multiple diagonstic markers. *Journal of the American Statistical Association*, *88*, 1350–1355.
Wand, M. P., & Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, *88*, 520–528.
Wand, M. P., & Jones, M. C. (1995). *Kernel smoothing*. London: Chapman and Hall/CRC.
Watson, G. S., & Leadbetter, M. R. (1964). Hazard analysis. II. *Sankhyā. Series A*, *26*, 101–116.
Winter, B. B. (1973). Strong uniform consistency of integrals of density estimators. *The Canadian Journal of Statistics*, *1*, 247–253.
Yamato, H. (1973). Uniform convergence of an estimator of a distribution function. *Bulletin of Mathematical Statistics*, *15*, 69–78.
Youden, W. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32–35.
Zhou, X., & Harezlak, J. (2002). Comparison of bandwidth selection methods for kernel smoothing of ROC curves. *Statistics in Medicine*, *21*, 2045–2055.
Zou, K. H., Hall, W. J., & Shapiro, D. E. (1997). Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, *16*, 2143–2156.