Biometrika (2016), **xx**, x, *pp*. 1–14

Printed in Great Britain

Supplementary material for Density estimation in the two-sample problem with likelihood ratio ordering

BY TAO YU

Department of Statistics and Applied Probability, National University of Singapore, 117546, Singapore

stayt@nus.edu.sg

PENGFEI LI

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada

pengfei.li@uwaterloo.ca

JING QIN

National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, U.S.A.

jingqin@niaid.nih.gov

1. INTRODUCTION

This is a supplementary document to the corresponding paper submitted to *Biometrika*. It has eight sections. §2 and §3 restate the regularity conditions and theorems given in the main article. §4 gives the details of the technical proofs, and §5 describes some details of the numerical implementation of our method. §6 presents an alternative method suggested by an anonymous referee and a numerical comparisons of this method and our method. §7 compares the performance of the receiver operating characteristic curve estimates based on our method, the empirical cumulative distribution function, the tilde estimates, and the f and g estimates given in §6. §8 gives more discussion based on all the simulation results in §5 of the main paper.

2. TECHNICAL CONDITIONS

We restate Conditions 1–4 given in the main article, which are needed in our technical derivations. They are not necessarily the weakest. Note that $f_0(x)$, $g_0(x)$, $\theta_0(x)$, and $\psi_0(x)$ are the true values of $f(x)$, $g(x)$, $\theta(x)$, and $\psi(x)$.

Condition 1. In the kernel estimates $\widetilde{f}_n(x)$ and $\widetilde{g}_m(x)$, the kernel $K(\cdot)$ has bounded support and is continuous in $\mathcal{R}$.

Note: Under this condition, $\widehat{\psi}_{n,m}(x)$ has bounded support and is continuous in $x \in \mathcal{R}$. Without loss of generality, we assume that the support is $[a_n, b_n]$.

Condition 2. Assume that $\widetilde{f}_n(x)$ and $\widetilde{g}_m(x)$ satisfy $\int_{\mathcal{R}} |\widetilde{f}_n(x) - f_0(x)|dx = O(\alpha_n)$, $\int_{\mathcal{R}} |\widetilde{g}_m(x) - g_0(x)|dx = O(\alpha_n)$ almost surely, with $\alpha_n \to 0$ as $n \to \infty$.

Condition 3. For some $a < b$, $\widetilde{f}_n(x)$ and $\widetilde{g}_m(x)$ satisfy $\sup_{x\in[a,b]} |\widetilde{f}_n(x) - f_0(x)| = o(1)$, $\sup_{x\in[a,b]} |\widetilde{f}_n'(x) - f_0'(x)| = o(1)$, $\sup_{x\in[a,b]} |\widetilde{g}_m(x) - g_0(x)| = o(1)$, and $\sup_{x\in[a,b]} |\widetilde{g}_m'(x) - g_0'(x)| = o(1)$ almost surely, as $n \to \infty$.

Condition 4. Assume that $f_0(x)$, $g_0(x)$, $f_0'(x)$, and $g_0'(x)$ are all bounded on $[a, b]$. Furthermore, there exists a constant $\delta > 0$ such that $\inf_{x\in[a,b]} \theta_0'(x) > \delta$ and $\inf_{x\in[a,b]} \psi_0(x) > \delta$.

In Conditions 2 and 3, we have assumed the asymptotic properties of $\widetilde{f}_n(x)$ and $\widetilde{g}_m(x)$. Theoretical developments of these properties based on different regularity conditions are widely available in the literature. We list below one set of these conditions; they may not be the weakest possible.

Based on Theorem 4.25 in Eggermont & LaRiccia (2001), Condition 2 is satisfied with $\alpha_n = n^{-2/5}$ if

(i) $K(x)$ is a bounded density and symmetric about 0, $K'(x)$ exists, and $K' \in L_1(\mathcal{R})$, where $L_p(\mathcal{R}) = \{h(x) : \int_{\mathcal{R}} |h(x)|^p dx < \infty\}$;
(ii) $h \propto n^{-1/5}$;
(iii) $f_0(x)$ and $g_0(x)$ are both second-order differentiable and satisfy $f_0, g_0, f_0'', g_0'' \in L_1(\mathcal{R})$, $\int_{\mathcal{R}} |x|^p f_0(x)dx < \infty$, and $\int_{\mathcal{R}} |x|^p g_0(x)dx < \infty$ for some $p > 1$.

Based on Theorems A and C in Silverman (1978), Condition 3 is satisfied if

(i) for every $x \in \mathcal{R}$, the derivative of the kernel function $K'(x)$ exists and is Lipschitz continuous. Further, $\int_{-\infty}^{\infty} |K'(x)|dx < \infty$, $\int_{-\infty}^{\infty} \big|x \log |x|\big|^{1/2} |dK(x)| < \infty$, and $\int_{-\infty}^{\infty} \big|x \log |x|\big|^{1/2} |dK'(x)| < \infty$. The Fourier transform of $K(x)$ is not identically one in any neighbourhood of zero;
(ii) $h \to 0$ and $n^{-1}h^{-3} \log 1/h \to 0$ as $n \to \infty$;
(iii) $f_0'(x)$ and $g_0'(x)$ exist and are uniformly continuous.

3. THEOREMS AND COROLLARIES

Recall that we have the following constraint conditions for densities $f(\cdot)$ and $g(\cdot)$:

$$f(x) \geq 0,\ g(x) \geq 0, \int_{\mathcal{R}} f(x)dx = \int_{\mathcal{R}} g(x) = 1, \tag{1}$$
$$f(x)/g(x) \text{ is nondecreasing in } x. \tag{2}$$

We have also used the following definitions of $\theta(\cdot)$ and $\psi(\cdot)$:

$$\theta(x) = \frac{\lambda f(x)}{\lambda f(x) + (1-\lambda)g(x)}, \qquad \psi(x) = \lambda f(x) + (1-\lambda)g(x), \tag{3}$$

which gives

$$f(x) = \theta(x)\psi(x)/\lambda, \qquad g(x) = \{1 - \theta(x)\}\psi(x)/(1-\lambda). \tag{4}$$

Furthermore, we have the following Proposition in the main article.

PROPOSITION 1. *With the reparameterization in* (3)*, $f(\cdot)$ and $g(\cdot)$ satisfy constraints* (1) *and* (2) *if and only if $\theta(\cdot)$ and $\psi(\cdot)$ satisfy*

(a) $\psi(x) \geq 0$ and $\int_{\mathcal{R}} \psi(x)dx = 1$;

(b) $\theta(x) \in [0, 1]$ *and* $\theta(x)$ *is a nondecreasing function of* x*;*
(c) $\int_{\mathcal{R}} \psi(x)\theta(x)dx = \lambda$.

THEOREM 1. *Assume Condition 1. Let*

$$\{\widehat{\theta}_{n,m}(x), \widehat{\psi}_{n,m}(x)\} = \arg\max_{\theta,\,\psi} l_{n,m}(\theta, \psi),$$

subject to (a) and (b) in Proposition 1*. We have*

(a) $\widehat{\psi}_{n,m}(x) = \lambda\widetilde{f}_n(x) + (1-\lambda)\widetilde{g}_m(x)$*;*
(b) $\widehat{\theta}_{n,m}(x) = \arg\min_\theta \int_{\mathcal{R}} \left\{\widetilde{\theta}_{n,m}(x) - \theta(x)\right\}^2 \widehat{\psi}_{n,m}(x)dx$ *subject to (b) in Proposition* 1*, where*

$$\widetilde{\theta}_{n,m}(x) = \frac{\lambda\widetilde{f}(x)}{\lambda\widetilde{f}(x) + (1-\lambda)\widetilde{g}(x)};$$

(c) $\widehat{\psi}_{n,m}(x)$ *and* $\widehat{\theta}_{n,m}(x)$ *satisfy (c) in Proposition* 1.

Let $\widetilde{\theta}(x)$, $w(x) > 0$ be functions defined on $[a, b]$, and let $\widehat{\theta}$ be the solution of the weighted continuous isotonic regression, represented by

$$\widehat{\theta}(x) = \arg\min_{\theta\in\mathcal{F}} \int_a^b \{\widetilde{\theta}(x) - \theta(x)\}^2 w(x)dx, \tag{5}$$

where $\mathcal{F}$ denotes the set of nondecreasing functions defined on $[a, b]$.

THEOREM 2. *Consider* $\widehat{\theta}(x)$*,* $\widetilde{\theta}(x)$*, and* $w(x)$ *defined above. Assume that both* $\widetilde{\theta}(x)$ *and* $w(x)$ *are continuous functions defined on* $[a, b]$*, and* $w(x) > 0$ *for every* $x \in [a, b]$*. Let* $H(t) > 0$ *be a convex function defined on* $(-\infty, \infty)$*, and let* $\eta(\cdot)$ *be an arbitrary nondecreasing function on* $[a, b]$*. Then we have*

$$\int_a^b H\{\widehat{\theta}(x) - \eta(x)\}w(x)dx \le \int_a^b H\{\widetilde{\theta}(x) - \eta(x)\}w(x)dx. \tag{6}$$

Let $\theta_0(x) = \lambda f_0(x)/\{\lambda f_0(x) + (1-\lambda)g_0(x)\}$ be the true value of $\theta(x)$, where $f_0(x)$ and $g_0(x)$ are the true values of $f(x)$ and $g(x)$. We assume $f_0(x)/g_0(x)$ is a nondecreasing function of x, and so is $\theta_0(x)$. The following corollary results from an application of Theorem 2.

COROLLARY 1. *Assume Condition 1. Consider* $\widehat{\theta}_{n,m}(x)$ *and* $\widehat{\psi}_{n,m}(x)$ *given in §2 of the main article. Then for any* $p \ge 1$*,*

$$\int_{\mathcal{R}} |\widehat{\theta}_{n,m}(x) - \theta_0(x)|^p \widehat{\psi}_{n,m}(x)dx \le \int_{\mathcal{R}} |\widetilde{\theta}_{n,m}(x) - \theta_0(x)|^p \widehat{\psi}_{n,m}(x)dx. \tag{7}$$

THEOREM 3. *Assume Conditions 1 and 2. Let* $f_0(x)$ *and* $g_0(x)$ *be the true values of* $f(x)$ *and* $g(x)$*. We have*

$$\int_{\mathcal{R}} |\widehat{f}_{n,m}(x) - f_0(x)|dx = O(\alpha_n), \tag{8}$$

$$\int_{\mathcal{R}} |\widehat{g}_{n,m}(x) - g_0(x)|dx = O(\alpha_n), \tag{9}$$

almost surely, where α_n *is the almost sure convergence rate of the integrated* L_1 *error of* $\widetilde{f}_n(x)$ *and* $\widetilde{g}_m(x)$ *assumed in Condition 2.*

THEOREM 4. *Assume Conditions 1, 3, and 4. Let $[a,b] \subset \mathcal{R}$ be any interval satisfying Conditions 3 and 4. Then (Chow & Teicher, 1997),*

$$\mathrm{pr}\Big\{\widehat{\theta}_{n,m}(x) \neq \widetilde{\theta}_{n,m}(x) \textit{ for some } x \in (a,b),\ \textit{infinitely often}\Big\} = 0. \tag{10}$$

COROLLARY 2. *Under Conditions 1, 3, and 4,*

$$\mathrm{pr}\Big\{\widehat{f}_{n,m}(x) \neq \widetilde{f}_{n}(x) \textit{ for some } x \in (a,b),\ \textit{infinitely often}\Big\} = 0, \tag{11}$$

$$\mathrm{pr}\Big\{\widehat{g}_{n,m}(x) \neq \widetilde{g}_{m}(x) \textit{ for some } x \in (a,b),\ \textit{infinitely often}\Big\} = 0. \tag{12}$$

THEOREM 5. *Assume Conditions 1 and 2. For any $t \in (0,1)$, we have*

$$\widehat{R}(t) - R_0(t) = O(\alpha_n),$$

almost surely, where α_n is given in Condition 2, and $R_0(t)$ is the true value of $R(t)$.

4. TECHNICAL DETAILS

Proof of Theorem 1

Substituting (8) into (5) in the main article, we immediately have

$$l_{n,m}(\theta,\psi) = l_{n,m,1}(\theta) + l_{n,m,2}(\psi) + \text{constant},$$

where

$$l_{n,m,1}(\theta) = \int_{\mathcal{R}} \lambda \widetilde{f}_n(x) \log\theta(x)dx + \int_{\mathcal{R}} (1-\lambda)\widetilde{g}_m(x)\log\{1-\theta(x)\}dx, \tag{13}$$

$$l_{n,m,2}(\psi) = \int_{\mathcal{R}} \{\lambda\widetilde{f}_n(x) + (1-\lambda)\widetilde{g}_m(x)\}\log\psi(x)dx.$$

Note that $l_{n,m,1}(\theta)$ depends only on θ; likewise, $l_{n,m,2}(\psi)$ depends only on ψ. Therefore,

$$\widehat{\theta}_{n,m}(x) = \arg\max_{\theta(x)} l_{n,m,1}(\theta) \tag{14}$$

subject to Part (b) in Proposition 1; and

$$\widehat{\psi}_{n,m}(x) = \arg\max_{\psi(x)} l_{n,m,2}(\psi) \tag{15}$$

subject to Part (a) in Proposition 1. We consider the optimization problems in (14) and (15) separately. The solution of (15) is available in the literature; see Eggermont & LaRiccia (2001), page 122. It is uniquely given by

$$\widehat{\psi}_{n,m}(x) = \lambda\widetilde{f}_n(x) + (1-\lambda)\widetilde{g}_m(x),$$

which proves the claim of Part (a).

For (14), based on Condition 1, we can rewrite $l_{n,m,1}(\theta)$ in (13) as

$$l_{n,m,1}(\theta) = \int_{a_n}^{b_n} \left[\frac{\lambda\widetilde{f}_n(x)}{\widehat{\psi}_{n,m}(x)}\log\theta(x) + \frac{(1-\lambda)\widetilde{g}_m(x)}{\widehat{\psi}_{n,m}(x)}\log\{1-\theta(x)\}\right]\widehat{\psi}_{n,m}(x)dx.$$

By Example 2 in Groeneboom & Jongbloed (2010), the solution of the optimization problem (14) is equivalent to the minimizer of

$$\int_{a_n}^{b_n} \left\{ \frac{\lambda \widetilde{f}_n(x)}{\widehat{\psi}_{n,m}(x)} - \theta(x) \right\}^2 \widehat{\psi}_{n,m}(x)dx = \int_{\mathcal{R}} \left\{ \widetilde{\theta}(x) - \theta(x) \right\}^2 \widehat{\psi}_{n,m}(x)dx$$

subject to Part (b) in Proposition 1. Noting that $\widetilde{\theta}(x) = \lambda \widetilde{f}_n(x)/\widehat{\psi}_{n,m}(x)$, we obtain Part (b).

Finally, we show Part (c). Based on Part (b) and Equation (6) of Groeneboom & Jongbloed (2010), we immediately have

$$\int_{a_n}^{b_n} \left\{ \widehat{\theta}_{n,m}(x) - \widetilde{\theta}_{n,m}(x) \right\} \widehat{\psi}_{n,m}(x)dx = 0,$$

which implies that

$$\int_{\mathcal{R}} \widehat{\theta}_{n,m}(x)\widehat{\psi}_{n,m}(x)dx = \int_{\mathcal{R}} \lambda \widetilde{f}_n(t)dt = \lambda,$$

giving Part (c).

Proof of Theorem 2

We need the following lemma in our proof of Theorem 2. It is essentially based on the results in Groeneboom & Jongbloed (2010).

LEMMA 1. *Assume the conditions in Theorem* 2*.*

(a) For any nondecreasing function $\tau(x)$ on $[a, b]$ such that $\int_a^b \tau^2(x)w(x)dx < \infty$, we have

$$\int_a^b \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}\tau(x)w(x)dx \leq 0. \tag{16}$$

(b) Let $L_c = \{x : \widehat{\theta}(x) = c\}$ be the level set. If L_c is nonempty, then L_c is either a singleton in which $\widehat{\theta}(x) = \widetilde{\theta}(x)$, or a closed interval with a positive length, namely the level interval. The number of level intervals with a positive length is at most countable. Furthermore,

$$\int_{L_c} \left\{ \widetilde{\theta}(x) - c \right\} w(x)dx = 0. \tag{17}$$

Proof: To show Part (a), let $L\{\widetilde{\theta}(x), \theta(x)\} = \int_a^b \{\widetilde{\theta}(x) - \theta(x)\}^2 w(x)dx$. Referring to the definition of $\widehat{\theta}(x)$ in (5), we have

$$\frac{L\{\widetilde{\theta}(x), \widehat{\theta}(x)\} - L\{\widetilde{\theta}(x), \widehat{\theta}(x) + t\tau(x)\}}{t} \leq 0, \tag{18}$$

for an arbitrary $t \geq 0$. On the other hand,

$$\begin{aligned} \frac{L\{\widetilde{\theta}(x), \widehat{\theta}(x)\} - L\{\widetilde{\theta}(x), \widehat{\theta}(x) + t\tau(x)\}}{t} &= 2\int_a^b \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}\tau(x)w(x)dx \\ &\quad -t\int_a^b \tau^2(x)w(x)dx. \end{aligned} \tag{19}$$

Combining (18) and (19) leads to (16) by making $t \to 0+$.

Part (b) and the proof are adapted from Lemma 2 in Groeneboom & Jongbloed (2010).

We now move to the proof of Theorem 2. Since $H(\cdot)$ is convex,

$$H(v) \geq H(u) + (v-u)h(u),$$

where $h(\cdot)$ is any determination of the derivative of $H(\cdot)$. Setting $v = \widetilde{\theta}(x) - \eta(x)$ and $u = \widehat{\theta}(x) - \eta(x)$, we obtain

$$\begin{aligned}\int_a^b H\{\widetilde{\theta}(x) - \eta(x)\}w(x)dx \geq & \int_a^b H\{\widehat{\theta}(x) - \eta(x)\}w(x)dx \\ & + \int_a^b \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}h\Big\{\widehat{\theta}(x) - \eta(x)\Big\}w(x)dx.\end{aligned}$$

To prove (6), it suffices to show that

$$\int_a^b \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}h\Big\{\widehat{\theta}(x) - \eta(x)\Big\}w(x)dx \geq 0. \tag{20}$$

Based on Part (b) of Lemma 1, there exist at most a countable number of intervals $[l_k, u_k]$ and constants c_k, $k = 1, 2, \ldots$, such that $\widehat{\theta}(x) = c_k$ when $x \in [l_k, u_k]$; $\widehat{\theta}(x) = \widetilde{\theta}(x)$ otherwise. Therefore,

$$\int_a^b \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}h\Big\{\widehat{\theta}(x) - \eta(x)\Big\}w(x)dx = \sum_{k=1}^{\infty} \int_{l_k}^{u_k} \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}h\Big\{c_k - \eta(x)\Big\}w(x)dx. \tag{21}$$

Next, we verify that for each k,

$$\int_{l_k}^{u_k} \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}h\Big\{c_k - \eta(x)\Big\}w(x)dx \geq 0. \tag{22}$$

Define

$$\tau_k(x) = \begin{cases} -h\Big\{c_k - \eta(l_k)\Big\}, & x < l_k \\ -h\Big\{c_k - \eta(x)\Big\}, & x \in [l_k, u_k] \\ -h\Big\{c_k - \eta(u_k)\Big\}, & x > u_k \end{cases}. \tag{23}$$

Clearly, since both $h(\cdot)$ and $\eta(\cdot)$ are nondecreasing functions, so is $\tau_k(x)$. Furthermore, it is easy to verify that $\int_a^b \tau_k^2(x)w(x) < \infty$. Applying Part (a) of Lemma 1, we have

$$\int_a^b \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}\tau_k(x)w(x)dx \leq 0.$$

Let

$$\begin{aligned} J_{1k} &= h\{c_k - \eta(l_k)\} \int_a^{l_k} \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}w(x)dx, \\ J_{2k} &= h\{c_k - \eta(u_k)\} \int_{u_k}^{b} \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}w(x)dx, \\ J_{3k} &= \int_{l_k}^{u_k} \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}h\{c_k - \eta(x)\}w(x)dx. \end{aligned} \tag{24}$$

With the definition of $\tau_k(x)$ in (23), we get

$$\int_a^b \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}\tau_k(x)w(x)dx = -(J_{1k} + J_{2k} + J_{3k}) \leq 0,$$

which implies that

$$J_{1k} + J_{2k} + J_{3k} \geq 0. \tag{25}$$

For J_{1k},

$$\begin{aligned} J_{1k} &= h\{c_k - \eta(l_k)\} \int_a^{l_k} \{\widetilde{\theta}(x) - \widehat{\theta}(x)\}w(x)dx & (26)\\ &= h\{c_k - \eta(l_k)\} \sum_{\{j: u_j \leq l_k\}} \int_{l_j}^{u_j} \{\widetilde{\theta}(x) - c_j\}w(x)dx & (27)\\ &= 0, & (28)\end{aligned}$$

where from (26) to (27) we have applied Part (b) of Lemma 1; from (27) to (28) we have used (17). Following the same procedure, we have

$$J_{2k} = 0. \tag{29}$$

(25), (28), and (29) lead to $J_{3k} \geq 0$, which with the definition of J_{3k} in (24) leads to (22). This together with (21) implies (20) and consequently the claim of this theorem.

Proof of Corollary 1

Based on Part (b) of Theorem 1 and Condition 1, we have

$$\begin{aligned} \widehat{\theta}_{n,m}(x) &= \arg\min_{\theta} \int_{\mathcal{R}} \left\{\widetilde{\theta}_{n,m}(x) - \theta(x)\right\}^2 \widehat{\psi}_{n,m}(x)dx \\ &= \arg\min_{\theta} \int_{a_n}^{b_n} \left\{\widetilde{\theta}_{n,m}(x) - \theta(x)\right\}^2 \widehat{\psi}_{n,m}(x)dx \end{aligned}$$

subject to (b) in Proposition 1. Applying Theorem 4, we immediately have

$$\int_{a_n}^{b_n} |\widehat{\theta}_{n,m}(x) - \theta_0(x)|^p \widehat{\psi}_{n,m}(x)dx \leq \int_{a_n}^{b_n} |\widetilde{\theta}_{n,m}(x) - \theta_0(x)|^p \widehat{\psi}_{n,m}(x)dx,$$

which is equivalent to (7) with Condition 1.

Proof of Theorem 3

Applying Corollary 1 by setting $p = 1$, we have

$$\int_{\mathcal{R}} |\widehat{\theta}_{n,m}(x) - \theta_0(x)|\widehat{\psi}_{n,m}(x)dx \leq \int_{\mathcal{R}} |\widetilde{\theta}_{n,m}(x) - \theta_0(x)|\widehat{\psi}_{n,m}(x)dx. \tag{30}$$

Since $\widehat{\theta}_{n,m}(x) = \lambda \widehat{f}_{n,m}(x)/\widehat{\psi}_{n,m}(x)$ and $\theta_0(x) = \lambda f_0(x)/\psi_0(x)$, we have

$$\begin{aligned}\int_{\mathcal{R}} |\widehat{\theta}_{n,m}(x) - \theta_0(x)|\widehat{\psi}_{n,m}(x)dx &\geq \lambda \int_{\mathcal{R}} |\widehat{f}_{m,n}(x) - f_0(x)|dx \\ &\quad -\lambda \int_{\mathcal{R}} |\widehat{\psi}_{n,m}(x)/\psi_0(x) - 1|f_0(x)dx, \\ \int_{\mathcal{R}} |\widetilde{\theta}_{n,m}(x) - \theta_0(x)|\widehat{\psi}_{n,m}(x)dx &\leq \lambda \int_{\mathcal{R}} |\widetilde{f}_n(x) - f_0(x)|dx \\ &\quad +\lambda \int_{\mathcal{R}} |\widehat{\psi}_{n,m}(x)/\psi_0(x) - 1|f_0(x)dx. \qquad (31)\end{aligned}$$

Combining (30) and (31) leads to

$$\begin{aligned}\int_{\mathcal{R}} |\widehat{f}_{m,n}(x) - f_0(x)|dx &\leq \int_{\mathcal{R}} |\widetilde{f}_n(x) - f_0(x)|dx + 2\int_{\mathcal{R}} |\widehat{\psi}_{n,m}(x)/\psi_0(x) - 1|f_0(x)dx \\ &= \int_{\mathcal{R}} |\widetilde{f}_n(x) - f_0(x)|dx + 2\int_{\mathcal{R}} |\widehat{\psi}_{n,m}(x) - \psi_0(x)|\theta_0(x)/\lambda dx \\ &\leq \int_{\mathcal{R}} |\widetilde{f}_n(x) - f_0(x)|dx + \frac{2}{\lambda}\int_{\mathcal{R}} |\widehat{\psi}_{n,m}(x) - \psi_0(x)|dx \\ &= O(\alpha_n),\end{aligned}$$

giving (8).

It remains to show (9). Given the definition of $g_0(\cdot)$ and $\widehat{g}_{n,m}(x)$ in (4) and (9) in the main article,

$$\begin{aligned}&\int_{\mathcal{R}} |\widehat{g}_{n,m}(x) - g_0(x)|dx \\ &= \int_{\mathcal{R}} \Big|\{1 - \widehat{\theta}_{n,m}(x)\}\widehat{\psi}_{n,m}(x)/(1-\lambda) - \{1-\theta_0(x)\}\psi_0(x)/(1-\lambda)\Big|dx \\ &\leq \frac{1}{1-\lambda}\int_{\mathcal{R}} |\widehat{\psi}_{n,m}(x) - \psi_0(x)|dx + \frac{1}{1-\lambda}\int_{\mathcal{R}} |\widehat{f}_{n,m}(x) - f_0(x)|dx = O(\alpha_n).\end{aligned}$$

Proof of Theorem 4

We need the following lemma in the proof of Theorem 4.

LEMMA 2. *Let $\theta_1(x)$ and $\theta_2(x)$ be arbitrary functions defined on $x \in \mathcal{R}$. Further, $\theta_1(x)$ is continuous and $\theta_2(x)$ is nondecreasing. For any arbitrary $(a, b) \subset \mathcal{R}$, if*

$$\int_a^b |\theta_1(x) - \theta_2(x)| = 0 \qquad (32)$$

then $\theta_1(x) = \theta_2(x)$ for every $x \in (a, b)$.

Proof: If there exists an $x_0 \in (a, b)$ such that $\theta_1(x_0) \neq \theta_2(x_0)$, without loss of generality, we assume $\theta_1(x_0) > \theta_2(x_0)$. Let $c_0 = \theta_1(x_0) - \theta_2(x_0) > 0$. By the continuity of $\theta_1(x)$, there exists a $\delta_1 \in (0, x_0 - a)$ such that $\theta_1(x) > \theta_2(x_0) + c_0/2$ for any $x \in [x_0 - \delta_1, x_0]$. Because $\theta_2(x)$ is nondecreasing, we immediately have $\theta_1(x) > \theta_2(x) + c_0/2$ for any $x \in [x_0 - \delta_1, x_0]$. As a consequence, we have

$$\int_a^b |\theta_1(x) - \theta_2(x)|dx \geq \int_{x_0-\delta_1}^{x_0} |\theta_1(x) - \theta_2(x)| \geq \delta_1 c_0/2 > 0,$$

which contradicts (32). This completes our proof of this lemma.

We now move to the proof of Theorem 4. Using Conditions 3 and 4, it is straightforward to verify that

$$\sup_{x\in[a,b]} \left| \widetilde{f}_n'(x)\widetilde{g}_m(x) - \widetilde{f}_n(x)\widetilde{g}_m'(x) - f_0'(x)g_0(x) + f_0(x)g_0'(x) \right| = o(1), \tag{33}$$

almost surely as $n \to \infty$. On the other hand, based on Condition 4, we have

$$\inf_{x\in[a,b]} \left\{ f_0'(x)g_0(x) - f_0(x)g_0'(x) \right\} = \inf_{x\in[a,b]} \theta_0'(x)\psi_0^2(x)/\{\lambda(1-\lambda)\} > \delta_1, \tag{34}$$

where $\delta_1 = \delta^3/\{\lambda(1-\lambda)\} > 0$. Combining (33) and (34), we have

$$\inf_{x\in[a,b]} \left\{ \widetilde{f}_n'(x)\widetilde{g}_m(x) - \widetilde{f}_n(x)\widetilde{g}_m'(x) \right\} \geq \delta_1/2,$$

almost surely as $n \to \infty$. Therefore,

$$\inf_{x\in[a,b]} \widetilde{\theta}_{n,m}'(x) = \inf_{x\in[a,b]} \frac{\widetilde{f}_n'(x)\widetilde{g}_m(x) - \widetilde{f}_n(x)\widetilde{g}_m'(x)}{\widehat{\psi}_{n,m}^2(x)} \geq 0,$$

almost surely as $n \to \infty$, indicating that $\widetilde{\theta}_{n,m}(x)$ is nondecreasing, which together with Theorem 2 leads to

$$\int_a^b |\widehat{\theta}_{n,m}(x) - \widetilde{\theta}_{n,m}(x)|\widehat{\psi}_{n,m}(x)dx \leq 0, \tag{35}$$

almost surely as $n \to \infty$. Furthermore, by Condition 3, $\sup_{x\in[a,b]} |\widehat{\psi}_{n,m}(x) - \psi_0(x)| < \delta/2$ almost surely as $n \to \infty$. Hence, since from Condition 4 $\inf_{x\in[a,b]} \psi_0(x) > \delta$, we have $\inf_{x\in[a,b]} \widehat{\psi}_{n,m}(x) > \delta/2$ almost surely as $n \to \infty$. This combined with (35) implies

$$\int_a^b |\widehat{\theta}_{n,m}(x) - \widetilde{\theta}_{n,m}(x)|dx \leq 0,$$

almost surely as $n \to \infty$. This together with the continuity of $\widetilde{\theta}_{n,m}(x)$, the monotonicity of $\widehat{\theta}_{n,m}(x)$, and Lemma 2 leads to (10).

Proof of Corollary 2

Based on Condition 1, for any $x \in (a,b)$, we have $\widehat{\psi}_{n,m}(x) \neq 0$. Since $\widehat{\theta}_{n,m}(x) = \lambda\widehat{f}_{n,m}(x)/\widehat{\psi}_{n,m}(x)$ and $\widetilde{\theta}_{n,m}(x) = \lambda\widetilde{f}_n(x)/\widehat{\psi}_{n,m}(x)$, we get $\widehat{f}_{n,m}(x) = \widetilde{f}_n(x)$ if and only if $\widehat{\theta}_{n,m}(x) = \widetilde{\theta}_{n,m}(x)$. Hence,

$$\begin{aligned} &\mathrm{pr}\big\{\widehat{f}_{n,m}(x) \neq \widetilde{f}_n(x) \text{ for some } x \in (a,b), \text{infinitely often}\big\} \\ &= \mathrm{pr}\big\{\widehat{\theta}_{n,m}(x) \neq \widetilde{\theta}_{n,m}(x) \text{ for some } x \in (a,b), \text{infinitely often}\big\} = 0, \end{aligned}$$

giving (11). Similarly, we can show (12).

Proof of Theorem 5

Let $F_0(x)$ and $G_0(x)$ be the true values of $F(x)$ and $G(x)$. Note that

$$\sup_x |\widehat{F}(x) - F_0(x)| \leq \int_{\mathcal{R}} |\widehat{f}_{n,m}(t) - f_0(t)|dt = O(\alpha_n),$$

where the final equality comes from Theorem 3. Similarly, we have

$$\sup_x |\widehat{G}(x) - G_0(x)| = O(\alpha_n).$$

With a proof similar to that of Part (b) of Theorem 2 in Chen et al. (2016), we have that for any $t \in (0, 1)$,

$$\widehat{R}(t) - R_0(t) = O(\alpha_n).$$

This completes the proof.

5. IMPLEMENTATION

In this section, we describe some implementation details. First, throughout our numerical studies, we set $K(\cdot)$ to be the Epanechnikov kernel, which leads to the minimum asymptotic mean integrated squared error over all kernel densities; see Epanechnikov (1969). In principle, other kernels can be used; it is well known that the choice of the kernel function has little impact on kernel density estimation.

Second, we propose the following procedure for choosing h_1 and h_2.

Step 1. We choose $h_1^{(0)}$ based on using sample 1 to estimate the classical kernel density $\widetilde{f}_n(x)$, and similarly for $h_2^{(0)}$, so that $h_i^{(0)}$, $i = 1, 2$ are the best bandwidths if we use $\widetilde{f}_n(x)$ and $\widetilde{g}_m(x)$ as our density estimates. Note that methods for choosing a bandwidth for kernel density estimation are readily available in the literature. In our implementation, we have used function `dpik()` in the R package `KernSmooth` implemented by Wand and Ripley, publicly available at `http://CRAN.R-project.org/package=KernSmooth`, to choose the bandwidth $h_i^{(0)}$, $i = 1, 2$, for sample i. This package essentially implements the kernel methods in Wand & Jones (1995).

Step 2. For $t = 1, 2, \ldots$, we compute $\{\widehat{\theta}^{(t-1)}(X_1), \ldots, \widehat{\theta}^{(t-1)}(X_n), \widehat{\theta}^{(t-1)}(Y_1), \widehat{\theta}^{(t-1)}(Y_m)\}$, using the method given at the end of this section and bandwidth $h_i^{(t-1)}$, $i = 1, 2$.

Step 3. Let $\mathcal{S}^{(t)} = \{X_1, \ldots, X_n\} \cup \{Y_j : \widehat{\theta}^{(t-1)}(Y_j) \geq \lambda\}$, and $\overline{\mathcal{S}}^{(t)} = \{Y_1, \ldots, Y_m\} \cup \{X_i : \widehat{\theta}^{(t-1)}(X_i) < \lambda\}$. We choose $h_1^{(t)}$ by treating the observations in $\mathcal{S}^{(t)}$ as if they are sampled from $f(x)$; and similarly for $h_2^{(t)}$.

Step 4. We repeat Steps 2 and 3 until $h_i^{(t-1)}$, $i = 1, 2$, converges.

The philosophy of the above bandwidth selection algorithm is as follows. Note that $\widehat{\theta}^{(t-1)}(x)$ is essentially the estimate of the posterior probability that given $X = x$, the corresponding observation belongs to sample 1. Therefore, $\mathcal{S}^{(t)}$ collects the observations that act as if they are from sample 1; and in the estimation, these observations may have a significant impact on the estimation of $f(x)$. We use these observations to select the bandwidth for $f(x)$.

Third, we consider the evaluation of $\widehat{\theta}_{n,m}(x)$ in Theorem 1. By Theorem 1 in Groeneboom & Jongbloed (2010), $\widehat{\theta}_{n,m}(x)$ is the first derivative of the greatest convex minorant of $\int_{-\infty}^{x} \lambda \widetilde{f}_n(t)dt$ subject to the weight $\int_{-\infty}^{x} \widehat{\psi}_{n,m}(t)dt$. This can be directly evaluated by the function `gcmlcm()` in the R package `fdrtool` implemented by Klaus and Strimmer, publicly available at `http://CRAN.R-project.org/package=fdrtool`.

6. Comparison with an Alternative Method

An anonymous referee suggests an alternative way to formulate the problem. Let $\{Z_1, \ldots, Z_N\}$ denote the set of pooled observations from samples 1 and 2, where $N = n + m$. Let $\Delta_i = I\{Z_i \in (X_1, \ldots, X_n)\}$, where $I\{\cdot\}$ is the indicator function. With the definition of $\theta(\cdot)$ and $\psi(\cdot)$ given by (2), the classical likelihood can be formulated as

$$\begin{aligned}
\widetilde{l}_{n,m}(f, g) &= \sum_{i=1}^{n} \log f(X_i) + \sum_{j=1}^{m} \log g(Y_j) \\
&= \sum_{i=1}^{N} \Delta_i \log \theta(Z_i) + \sum_{i=1}^{N} (1 - \Delta_i) \log \{1 - \theta(Z_i)\} + \sum_{i=1}^{N} \log \psi(Z_i) \\
&= \widetilde{l}_{n,m,1}(\theta) + \widetilde{l}_{n,m,2}(\psi),
\end{aligned}$$

where

$$\widetilde{l}_{n,m,1}(\theta) = \sum_{i=1}^{N} \Delta_i \log \theta(Z_i) + \sum_{i=1}^{N} (1 - \Delta_i) \log \{1 - \theta(Z_i)\}\,, \;\; \widetilde{l}_{n,m,2}(\psi) = \sum_{i=1}^{N} \log \psi(Z_i).$$

Therefore, we can define

$$\widehat{\theta}_{\text{alt}}(x) = \arg\max_{\theta(x)} \widetilde{l}_{n,m,1}(\theta) \tag{36}$$

subject to Part (b) in Proposition 1; and $\psi(x)$ can be estimated by some other method, for example, we can use $\widehat{\psi}_{n,m}(x) = \lambda \widetilde{f}_n(x) + (1 - \lambda)\widetilde{g}_m(x)$. Note that the optimization problem (36) can be readily solved by the classical pool-adjacent-violation-algorithm (Ayer et al., 1955) and active set methods (de Leeuw et al., 2009).

As a consequence, $f(x)$ and $g(x)$ are estimated by

$$\widehat{f}_{\text{alt}}(x) = \widehat{\theta}_{\text{alt}}(x)\widehat{\psi}_{n,m}(x)/\lambda, \qquad \widehat{g}_{\text{alt}}(x) = \{1 - \widehat{\theta}_{\text{alt}}(x)\}\widehat{\psi}_{n,m}(x)/(1 - \lambda).$$

In this section, we compare the numerical performance of our hat estimates with those of this alternative method. Figure 1 and Table 1 compare the estimates of $f(x)$, $g(x)$, and $\theta(x)$ from the two methods. The simulation setups are the same as those of Fig. 1 and Table 1 in the main article.

Clearly, the alternative method does not need smoothing techniques in the estimation of $\theta(x)$. On the one hand, it is faster than our method, since it does not need to tune a smoothing parameter. On the other hand, ignoring the smoothness condition on the functions may lead to less accurate estimates when these functions are truly smooth. This is well illustrated by Fig. 1 and Table 1, where our method performs better than the alternative. Furthermore, the asymptotic properties of the estimates based on this method have not been derived. We conjecture that these properties can be established using techniques that are popular in the isotonic regression community; we leave this interesting and important topic for future research.

7. Numerical Comparison of the Receiver Operating Characteristic Curve Estimates

In this section, we compare the performance of the receiver operating characteristic curve estimates from the different methods. The receiver operating characteristic curve estimates include that based on our method, i.e., $\widehat{R}(t)$, that based on the empirical cumulative distribution function,

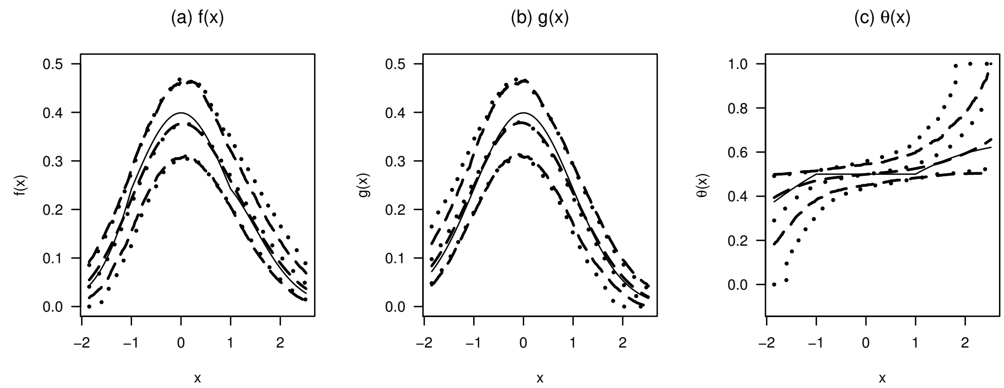

Fig. 1. Density estimation for example 2 with $C_1 = -1, C_2 = 1$. Panels (a)–(c): $f(x)$, $g(x)$, and $\theta(x)$. In each panel: solid lines: true curves; dotted lines: 2.5%, mean, and 97.5% percentiles of the alternative estimates over 1000 repetitions; dashed lines: those of the hat estimates over 1000 repetitions.

Table 1. *Medians of* $\text{RIL}_1(\textit{hat}, \textit{alternative})$ *over 1000 replications (percentages of values of* $\text{RIL}_1(\textit{hat}, \textit{alternative}) > 0$ *over 1000 replications)*

$n \setminus m$		$m = 50$	$m = 100$	$m = 150$	$m = 400$
$n = 50$	f	0.20 (88)	0.26 (91)	0.28 (92)	0.31 (96)
	g	0.20 (85)	0.13 (80)	0.09 (76)	0.02 (64)
	θ	0.33 (90)	0.35 (88)	0.38 (89)	0.43 (85)
$n = 100$	f	0.11 (80)	0.17 (84)	0.20 (90)	0.25 (93)
	g	0.25 (90)	0.16 (86)	0.13 (82)	0.04 (70)
	θ	0.30 (87)	0.32 (89)	0.33 (88)	0.38 (84)
$n = 150$	f	0.08 (76)	0.12 (81)	0.16 (88)	0.21 (92)
	g	0.27 (92)	0.21 (90)	0.15 (84)	0.06 (72)
	θ	0.27 (85)	0.30 (88)	0.33 (91)	0.37 (88)
$n = 400$	f	0.02 (61)	0.03 (64)	0.05 (71)	0.11 (84)
	g	0.28 (95)	0.25 (93)	0.21 (91)	0.13 (83)
	θ	0.24 (77)	0.25 (77)	0.26 (80)	0.29 (89)

i.e., $\widehat{R}_e(t)$, that based on the tilde estimates, i.e., $\widetilde{R}(t)$, and that based on the alternative f and g estimates, i.e., $\widehat{R}_{\text{alt}}(t)$. Here,

$$\widetilde{R}(t) = 1 - \widetilde{F}\{\widetilde{G}^{-1}(1-t)\},$$
$$\widehat{R}_e(t) = 1 - \widehat{F}_e\{\widehat{G}_e^{-1}(1-t)\},$$
$$\widehat{R}_{\text{alt}}(t) = 1 - \widehat{F}_{\text{alt}}\{\widehat{G}_{\text{alt}}^{-1}(1-t)\},$$

where $\widetilde{F}(x) = \int_{-\infty}^{x} \widetilde{f}_n(t)dt$, $\widetilde{G}(x) = \int_{-\infty}^{x} \widetilde{g}_m(t)dt$, $\widehat{F}_e(x) = n^{-1}\sum_{i=1}^{n} I(X_i \le x)$, $\widehat{G}_e(x) = m^{-1}\sum_{i=1}^{m} I(Y_i \le x)$, $\widehat{F}_{\text{alt}}(x) = \int_{-\infty}^{x} \widehat{f}_{\text{alt}}(t)dt$, and $\widehat{G}_{\text{alt}}(x) = \int_{-\infty}^{x} \widehat{g}_{\text{alt}}(t)dt$.

For the simulation examples given in the paper, we establish the receiver operating characteristic curves and compute the medians of $\text{RIL}_1(\widehat{R}, \widetilde{R}^*)$ and the percentages of values of

$\text{RIL}_1(\widehat{R}, \widetilde{R}^*) > 0$ over 1000 replications. Here $\widetilde{R}^*(t)$ denotes $\widehat{R}_e(t)$, $\widetilde{R}(t)$, or $\widehat{R}_{\text{alt}}(t)$. We observe that in our numerical studies, all the medians of the $\text{RIL}_1(\widehat{R}, \widetilde{R}^*)$'s are greater than 0, and the percentages are all greater than 50%, with some close to 100%. These observations indicate that our method leads to more accurate receiver operating characteristic curve estimates than the other methods. To avoid lengthy output, in Table 2, we display only the results for example 2 with $C_1 = -1$ and $C_2 = 1$; the results for the other settings are similar and are omitted.

Table 2. *Medians of* $\text{RIL}_1(\widehat{R}, \widetilde{R}^*)$ *over 1000 replications (percentages of values of* $\text{RIL}_1(\widehat{R}, \widetilde{R}^*) > 0$ *over 1000 replications). Here* $\widehat{R}(t)$ *is the receiver operating characteristic estimate based on our method;* $\widetilde{R}^*(t)$ *denotes* $\widehat{R}_e(t)$, $\widetilde{R}(t)$, *or* $\widehat{R}_{\text{alt}}(t)$.

$n \setminus m$		$m = 50$	$m = 100$	$m = 150$	$m = 400$
$n = 50$	$\text{RIL}_1(\widehat{R}, \widehat{R}_e)$	0.71 (96)	0.74 (97)	0.75 (98)	0.78 (99)
	$\text{RIL}_1(\widehat{R}, \widetilde{R})$	0.22 (66)	0.17 (62)	0.17 (61)	0.18 (58)
	$\text{RIL}_1(\widehat{R}, \widehat{R}_{\text{alt}})$	0.60 (92)	0.59 (93)	0.56 (91)	0.53 (92)
$n = 100$	$\text{RIL}_1(\widehat{R}, \widehat{R}_e)$	0.74 (98)	0.75 (98)	0.76 (98)	0.80 (100)
	$\text{RIL}_1(\widehat{R}, \widetilde{R})$	0.20 (64)	0.24 (66)	0.21 (65)	0.18 (60)
	$\text{RIL}_1(\widehat{R}, \widehat{R}_{\text{alt}})$	0.57 (93)	0.56 (92)	0.55 (90)	0.50 (88)
$n = 150$	$\text{RIL}_1(\widehat{R}, \widehat{R}_e)$	0.75 (99)	0.76 (99)	0.78 (99)	0.81 (100)
	$\text{RIL}_1(\widehat{R}, \widetilde{R})$	0.13 (61)	0.16 (64)	0.15 (63)	0.14 (59)
	$\text{RIL}_1(\widehat{R}, \widehat{R}_{\text{alt}})$	0.56 (92)	0.55 (90)	0.54 (90)	0.51 (90)
$n = 400$	$\text{RIL}_1(\widehat{R}, \widehat{R}_e)$	0.78 (100)	0.81 (100)	0.82 (100)	0.85 (100)
	$\text{RIL}_1(\widehat{R}, \widetilde{R})$	0.10 (56)	0.13 (61)	0.15 (61)	0.12 (60)
	$\text{RIL}_1(\widehat{R}, \widehat{R}_{\text{alt}})$	0.50 (91)	0.51 (90)	0.50 (90)	0.47 (87)

8. More Discussion of the Simulation Studies

We constructed similar plots and tables for the simulated data under all the settings in §5 of the main article. From these plots and tables, we have the following observations.

When both n and m are small, the hat estimates give better accuracy. When both n and m are large, the difference between the hat and tilde estimates is smaller. This complies with our asymptotic theory in §3 of the main article showing that the hat estimates have essentially the same rate of convergence as the tilde estimates.

Density estimates for small samples benefit more from our method than those for larger samples. For example, in Table 1 of the main article, for $n = 100$, when m increases from 50 to 400, both the medians of $\text{RIL}_1(\widehat{f}_{n,m}, \widetilde{f}_n)$ and the corresponding percentages, displayed in parentheses in the table increase. In contrast, the medians for $\text{RIL}_1(\widehat{g}_{n,m}, \widetilde{g}_m)$ decrease.

For the estimates of $\theta(x)$, by comparing the rightmost panel of Fig. 1 in the main article with the corresponding results obtained by replacing the flat area of $\theta(x)$, $[-1, 1]$, with $[0, 2]$, we observe that the efficiency gain of $\widehat{\theta}_{n,m}(x)$ over $\widetilde{\theta}_{n,m}(x)$ is greater in flat areas than in strictly monotonic areas. Furthermore, the improvement of $\widehat{\theta}_{n,m}(x)$ over $\widetilde{\theta}_{n,m}(x)$ is much greater in areas where either $f(x)$ or $g(x)$ is small; see for example $|x| \geq 2$ against $|x| < 2$ in the right panels of Fig. 1. We conjecture that this is because when $f(x)$ or $g(x)$ is small, a sample has a limited number of observations, so the information from that sample is limited. Therefore, appropriately accounting for the monotonicity information greatly benefits estimation of $\theta(x)$.

REFERENCES

AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. & SILVERMAN E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, **26**, 641–647.

CHEN, B., LI, P., QIN, J. & YU, T. (2016). Using a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests. *Journal of the American Statistical Association*, **514**, 861–874.

CHOW Y. S. & TEICHER H. (1997). *Probability Theory: Independence, Interchangeability, Martingales*. New York: Springer.

DE LEEUW, J., HORNIK, K. & PATRICK, M. (2009). Isotone optimization in R: Pool-adjacent-violators-algorithm (PAVA) and active set methods. *Journal of Statistical Software*, **32**, 1–24.

EGGERMONT, P. P. B. & LARICCIA, V. N. (2001). *Maximum Penalized Likelihood Estimation*. New York: Springer.

EPANECHNIKOV, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability and Its Applications*, **14**, 153–158.

GROENEBOOM, P. & JONGBLOED, G. (2010). Generalized continuous isotonic regression. *Statistics and Probability Letters*, **80**, 248–253.

SILVERMAN, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, **6**, 177–184.

WAND, M. P. & JONES, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.