Nonparametric estimation of finite mixtures from repeated measurements
Supplementary material

Stéphane Bonhomme
University of Chicago
Koen Jochmans†
Sciences Po, Paris
Jean-Marc Robin
Sciences Po, Paris and University College London
[Revised July 22, 2014]

Table of contents

Notation

Throughout the supplementary material we use the notation $\overline{z}_N$ to indicate the sample average of a generic sequence $z_1, z_2, \ldots, z_N$.

†*Address for correspondence:* Sciences Po, Department of Economics, 28 rue des Saints-Pères, 75007 Paris, France. *E-mail:* koen.jochmans@sciencespo.fr.

1. Lemmas

The following lemmas provide distribution theory for the estimators of the transformation matrix W and the matrix of joint eigenvectors U.

LEMMA S.1. $\|\widehat{A}-A\| = o_P(1)$ *and* $\sqrt{N}\,\mathrm{vec}(\widehat{A}-A) \xrightarrow{L} \mathcal{N}(0, \mathbb{E}[\upsilon_n\upsilon_n'])$.

PROOF. Because $\widehat{A}$ is a sample average over $\chi(x_{nm_1})\chi(x_{nm_2})'$ and $A = \mathbb{E}[\chi(x_{m_1})\chi(x_{m_2})']$ exists, consistency follows by the law of large numbers. Further, $\mathrm{var}[\chi_{i_1}(x_{m_1})\chi_{i_2}(x_{m_2})]$ exists for any i_1, i_2 by Assumption 2 because

$$\mathbb{E}[\chi_{i_1}(x_{m_1})^2\chi_{i_2}(x_{m_2})^2] = \sum_{k=1}^{K} \mathbb{E}_k[\chi_{i_1}(x_{m_1})^2]\; \mathbb{E}_k[\chi_{i_2}(x_{m_2})^2]\,\omega_k = O(1),$$

where we have used the conditional independence of the measurements within groups. We have also used the fact that $f_k \leq \omega_k^{-1} f$, so $\mathbb{E}_k[\chi_i(x_m)^2] \leq \omega_k^{-1}\,\mathbb{E}[\chi_i(x_m)^2]$, which is finite. As the data constitute a random sample, the asymptotic-normality claim follows from the Lindeberg-Lévy version of the central limit theorem. □

LEMMA S.2. $\|\widehat{W}-W\| = o_P(1)$ *and* $\sqrt{N}\,\mathrm{vec}(\widehat{W}-W) \xrightarrow{L} \mathcal{N}(0, J_W\,\mathbb{E}[\upsilon_n\upsilon_n']J_W')$

PROOF. Matrix A is real and symmetric, has rank K, and has K distinct non-zero eigenvalues $\lambda \equiv (\lambda_1, \lambda_2, \ldots, \lambda_K)'$ by Assumption 3. By combining Lemma S.1, Theorem 4.2 in Eaton and Tyler (1991), and Theorem 1 in Magnus (1985), it then follows that the estimator $\widehat{\lambda}$ constructed from an eigen decomposition of $\widehat{A}$ has the asymptotically-linear representation

$$\sqrt{N}(\widehat{\lambda}-\lambda) = (V_K \overset{\mathrm{col}}{\otimes} V_K)'\,\sqrt{N}\,\mathrm{vec}(\widehat{A}-A) + o_P(1). \tag{S.1.1}$$

Let $\Gamma \equiv \Lambda_K^{-1/2}$ and let $\widehat{\Gamma}$ denote its estimator. Λ_K and λ are linked as $\mathrm{vec}\,\Lambda_K = (\mathrm{I}_K \overset{\mathrm{col}}{\otimes} \mathrm{I}_K)\,\lambda$ and the transformation from Λ_K to Γ is continuous. Therefore, together with an application of the delta method, (S.1.1) implies that

$$\sqrt{N}\,\mathrm{vec}(\widehat{\Gamma}-\Gamma) = -\frac{1}{2}(\mathrm{I}_K \overset{\mathrm{col}}{\otimes} \mathrm{I}_K)\,\Gamma^3\,(V_K \overset{\mathrm{col}}{\otimes} V_K)'\,\sqrt{N}\,\mathrm{vec}(\widehat{A}-A) + o_P(1).$$

This expression can be simplified by distributing the term Γ^3 over the columnwise Kronecker products and substituting $W = \Gamma V_K'$ to arrive at

$$\sqrt{N}\,\mathrm{vec}(\widehat{\Gamma}-\Gamma) = -\frac{1}{2}(\Gamma \overset{\mathrm{col}}{\otimes} \mathrm{I}_K)\;(W \overset{\mathrm{row}}{\otimes} W)\,\sqrt{N}\,\mathrm{vec}(\widehat{A}-A) + o_P(1). \tag{S.1.2}$$

Further, it is known (see, e.g., Anderson 1963) that the estimated eigenvectors, $\widehat{V}_K$, satisfy

$$\sqrt{N}\,\mathrm{vec}(\widehat{V}_K - V_K) = -(\mathrm{I}_K \otimes V)\,(\Lambda_K \ominus \Lambda)^+\,\left(V_K \otimes V\right)'\,\sqrt{N}\,\mathrm{vec}(\widehat{A}-A) + o_P(1). \tag{S.1.3}$$

Let $\mathcal{K}$ be the $IK \times IK$ commutation matrix (Magnus and Neudecker 1979); recall that $\mathrm{vec}(\widehat{V}_K' - V_K') = \mathcal{K}\,\mathrm{vec}(\widehat{V}_K - V_K)$. A linearization of $\widehat{W}-W$ gives

$$\sqrt{N}\,\mathrm{vec}(\widehat{W}-W) = \left(\mathrm{I}_I \otimes \Gamma\right)\,\mathcal{K}\,\sqrt{N}\,\mathrm{vec}(\widehat{V}_K - V_K) + \left(V_K \otimes \mathrm{I}_K\,\right)\sqrt{N}\,\mathrm{vec}(\widehat{\Gamma}-\Gamma) + o_P(1).$$

Elementary properties of the commutation matrix, and (S.1.2) and (S.1.3), then imply that

$$\sqrt{N}\,\mathrm{vec}(\widehat{W}-W) = J_W\,\sqrt{N}\,\mathrm{vec}(\widehat{A}-A) + o_P(1).$$

The asymptotic-normality statement in Lemma S.1 then yields asymptotic normality. This concludes the proof. □

LEMMA S.3. $\|\widehat{U}-U\| = o_P(1)$ *and* $\sqrt{N}\,\mathrm{vec}(\widehat{U}-U) \xrightarrow{L} \mathcal{N}(0, J_U\,\mathbb{E}[\psi_n\psi_n']J_U')$.

PROOF. It suffices to show that the input matrices for the objective function in (3.4) are $\sqrt{N}$-consistent and jointly asymptotically normal, and to verify their influence-function representation. The desired result will then follow from an application of Theorem 5 in Bonhomme and Robin (2009). Note that Lemma S.2 already provides the asymptotic distribution of $\sqrt{N}\,\mathrm{vec}(\widehat{W}-W)$, and that

$$\sqrt{N}\,\mathrm{vec}(\widehat{W}'-W') = \mathcal{K}\,\sqrt{N}\,\mathrm{vec}(\widehat{W}-W) \xrightarrow{L} \mathcal{N}(0, J_{W'}\,\mathbb{E}[\upsilon_n\upsilon_n']J_{W'}').$$

Also, following the same steps as in the proof to Lemma S.1, the asymptotic normality of

$$\mathrm{vec}\left(\mathrm{horzcat}[\widehat{A}_i] - \mathrm{horzcat}[A_i]\right) = \overline{\Upsilon}_N$$

is immediate. It follows that the estimators $\widehat{W}\widehat{A}_i\widehat{W}'$ are jointly asymptotically-normal estimators of the WA_iW'. By a linearization, their joint influence-function representation is found to be

$$\mathrm{vec}\left(\mathrm{horzcat}[\widehat{W}\widehat{A}_i\widehat{W}'] - \mathrm{horzcat}[WA_iW']\right) = \overline{\psi}_N + o_P(N^{-1/2}),$$

where

$$\psi_n = \underbrace{\mathrm{vertcat}[WA_i\otimes \mathrm{I}_K]\,J_W\,\upsilon_n}_{\text{contribution of } W} + \underbrace{\mathrm{I}_I\otimes(W\otimes W)\,\Upsilon_n}_{\text{contribution of the } A_i} + \underbrace{\mathrm{vertcat}[\mathrm{I}_K\otimes WA_i]\,J_{W'}\,\upsilon_n}_{\text{contribution of } W'},$$

which indeed agrees with the definition of ψ_n given in the main text. Now, because the sample average $\overline{\psi}_N$ satisfies the conditions for the Lindeberg-Lévy version of the central limit theorem, the proof is complete. □

2. Proofs

Proof of Theorem 2. Recall that $\widehat{\theta}-\theta_0 = (\widehat{\theta}-\widetilde{\theta}) + (\widetilde{\theta}-\theta_0)$ where the first term captures the estimation noise in the weight function, that is,

$$\widehat{\theta}-\widetilde{\theta} = \frac{1}{N}\frac{(M-3)!}{M!}\sum_{n=1}^{N}\sum_{(m_1,m_2,m_3)}\left\{\widehat{\tau}_k(x_{nm_1},x_{nm_2}) - \tau_k(x_{nm_1},x_{nm_2})\right\}\varphi(x_{nm_3}),$$

and the second term is the sampling-error representation of $\widetilde{\theta}$, being

$$\widetilde{\theta}-\theta_0 = \frac{1}{N}\frac{(M-3)!}{M!}\sum_{n=1}^{N}\sum_{(m_1,m_2,m_3)}\tau_k(x_{nm_1},x_{nm_2})\varphi(x_{nm_3}) - \mathbb{E}[\tau_k(x_{m_1},x_{m_2})\varphi(x_{m_3})].$$

The asymptotic distribution of $\sqrt{N}(\widetilde{\theta}-\theta_0)$ is easy to characterize. Therefore, to obtain the asymptotic distribution of $\sqrt{N}(\widehat{\theta}-\theta_0)$, the main task is to derive the impact of estimating the weight function τ_k.

Let $\eta_k \equiv W' u_k$ and let $\widehat{\eta}_k$ be its plug-in estimator. Note that, using (2.5),

$$\widehat{\tau}_k(x_{nm_1}, x_{nm_2}) - \tau_k(x_{nm_1}, x_{nm_2}) = \widehat{\eta}_k' X_{nm_1,nm_2} \widehat{\eta}_k - \eta_k' X_{nm_1,nm_2} \eta_k,$$

where $X_{nm_1,nm_2} \equiv \chi(x_{nm_1})\chi(x_{nm_2})'$. By consequence of Lemmas S.2 and S.3, we have that

$$\widehat{\tau}_k(x_{nm_1}, x_{nm_2}) - \tau_k(x_{nm_1}, x_{nm_2}) = \eta_k'(X_{nm_1,nm_2} + X_{nm_1,nm_2}')(\widehat{\eta}_k - \eta_k) + o_P(N^{-1/2}).$$

Also, by iterating expectations,

$$\mathbb{E}[\varphi(x_{m_3})\, \eta_k'(X_{m_1,m_2} + X_{m_1,m_2}')] = \mathbb{E}\left[\varphi(x_{m_3})\, \frac{\partial \eta_k'\, A(x_{m_3})\, \eta_k}{\partial \eta_k'}\right] = \vartheta_0.$$

Hence, by a standard law of large numbers,

$$\sqrt{N}(\widehat{\theta}-\widetilde{\theta}) \quad = \quad \vartheta_0 \sqrt{N}(\widehat{\eta}_k - \eta_k) + o_P(1).$$

Further, again by Lemmas S.2 and S.3, we have

$$\sqrt{N}(\widehat{\eta}_k - \eta_k) = \frac{1}{\sqrt{N}} \sum_{n=1}^N (e_k' \otimes \mathrm{I}_I)\, \iota_n + o_P(1).$$

Combining the results yields the influence-function representation in the main text and concludes the proof. □

Proof of Proposition 1. Given Theorem 2 and the regularity conditions in Assumption 4, the result is standard; see, e.g., Hansen (1982). □

Proof of Proposition 2. The estimation noise in the weights can be ignored throughout the proof. Indeed, with

$$\widetilde{f}_k(x) \equiv \frac{1}{N} \frac{(M-3)!}{M!} \sum_{n=1}^N \sum_{(m_1,m_2,m_3)} \frac{\tau_k(x_{nm_1}, x_{nm_2})}{h}\, \kappa\left(\frac{x_{nm_3}-x}{h}\right),$$

Lemmas S.2 and S.3 imply that $\|\widehat{f}_k - \widetilde{f}_k\|_\infty = O_P(N^{-1/2})$, which is asymptotically negligible. Therefore, it suffices to show that the infeasible estimator $\widetilde{f}_k$ satisfies the conclusions of Proposition 2.

We first show consistency, and start by characterizing the bias and variance of $\widetilde{f}_k(x)$. First note that, by Theorem 1,

$$\mathbb{E}[\widetilde{f}_k(x)] = \mathbb{E}\left[\frac{\tau_k(x_{m_1}, x_{m_2})}{h}\, \kappa\left(\frac{x_{m_3}-x}{h}\right)\right] = \mathbb{E}_k\left[\frac{1}{h}\, \kappa\left(\frac{x_{m_3}-x}{h}\right)\right].$$

Standard arguments and Assumption 5 then give

$$\mathbb{E}_k \left[\frac{1}{h}\, \kappa \left(\frac{x_{m_3} - x}{h} \right) \right] = \int_{-\infty}^{+\infty} \frac{f_k(z)}{h}\, \kappa \left(\frac{z-x}{h} \right) \mathrm{d}z \; = \; f_k(x) + h^2 \, \mu_f + o(h^2).$$

where, recall, $\mu_f = \frac{1}{2} f_k''(x) \int_{-\infty}^{+\infty} u^2 \kappa(u)\, \mathrm{d}u$. Furthermore, because the observations are independent and identically distributed,

$$\mathrm{var}[\widetilde{f}_k(x)] = \frac{\mathrm{var}[\beta_n]}{Nh}, \quad \beta_n \equiv \frac{(M-3)!}{M!} \sum_{(m_1,m_2,m_3)} \frac{\tau_k(x_{nm_1}, x_{nm_2})}{\sqrt{h}} \kappa \left(\frac{x_{nm_3} - x}{h} \right).$$

To analyze the numerator, note that, by Assumptions 5 and 6,

$$\mathbb{E}[\beta_n] = \mathbb{E}_k \left[\frac{1}{\sqrt{h}}\, \kappa \left(\frac{x_m - x}{h} \right) \right] = \sqrt{h} \int_{-\infty}^{+\infty} f_k(x + hu)\, \kappa(u)\, \mathrm{d}u = O(\sqrt{h}).$$

Also, by essentially the same argument, for any pair of triples (m_1, m_2, m_3) and (m_1', m_2', m_3') with $m_3 \neq m_3'$,

$$\mathbb{E} \left[\frac{\tau_k(x_{m_1}, x_{m_2})}{\sqrt{h}} \frac{\tau_k(x_{m_1'}, x_{m_2'})}{\sqrt{h}}\, \kappa \left(\frac{x_{m_3} - x}{h} \right) \kappa \left(\frac{x_{m_3'} - x}{h} \right) \right] = O(h).$$

Hence,

$$\mathrm{var}[\beta_n] = M \left(\frac{(M-3)!}{M!} \right)^2 \mathbb{E} \left[q_k(x_m)\, \frac{1}{h} \kappa \left(\frac{x_m - x}{h} \right)^2 \right] + O(h) \to \mathscr{V}_f,$$

where the last step follows from a bounded-convergence argument. The mean-squared error of $\widetilde{f}_k(x)$ is thus

$$\mathrm{mse}[\widetilde{f}_k(x)] = O\left(h^4\right) + O \left(\frac{1}{Nh} \right),$$

and $|\widetilde{f}_k(x) - f_k(x)| = o_P(1)$ follows.

To show asymptotic normality we verify the conditions for Lyapunov's central limit theorem for triangular arrays. Observe that

$$\sqrt{Nh} \left(\widetilde{f}_k(x) - \mathbb{E}[\widetilde{f}_k(x)] \right) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (\beta_n - \mathbb{E}[\beta_n]).$$

We need to show that (i) $\mathbb{E}[\beta_n] = O(1)$; (ii) $\mathrm{var}[\beta_n] \to \mathscr{V}_f$; and (iii) that the Lyapunov condition

$$\sum_{n=1}^{N} \mathbb{E} \left| \frac{\beta_n - \mathbb{E}[\beta_n]}{\sqrt{N}} \right|^{2+\delta} = o(1)$$

holds for some $\delta > 0$. Conditions (i) and (ii) have already been shown to hold above. Also, because $\chi_i^4 f$ is integrable and the kernel function is bounded,

$$\sum_{n=1}^{N} \mathbb{E} \left| \frac{\beta_n - \mathbb{E}[\beta_n]}{\sqrt{N}} \right|^{4} = O \left(\frac{1}{Nh} \right) = o(1),$$

from which Condition (iii) follows. Therefore, with $Nh^5 \to c$ as N diverges,

$$\sqrt{Nh}\,[\widehat{f}_k(x) - f_k(x)] \overset{L}{\to} \mathcal{N}(\sqrt{c}\,\mu_f, \mathscr{V}_f).$$

The proof is complete. □

3. A reweighting interpretation

The function $d_k(x)$ is a tilt function. Indeed,

$$f_k(x) = d_k(x)\, f(x).$$

For estimation purposes, the formulation in Theorem 1 is more interesting as it prevents the need for nonparametric estimation of the tilt. An alternative view on τ_k is to see it as a weight. More precisely, it is readily verified that

$$\mathbb{E}[\tau_k(x_{m_1}, x_{m_2})] = 1, \qquad \mathbb{E}_{k_1}[\tau_{k_2}(x_{m_1}, x_{m_2})] = \frac{1\{k_1 = k_2\}}{\omega_{k_2}}.$$

By means of conditional independence of the measurements, this implies that

$$\mathbb{E}[\tau_{k_1}(x_{m_1}, x_{m_2})\, \varphi(x_{m_3})] = \sum_{k_2=1}^{K} 1\{k_1 = k_2\}\, \frac{\omega_{k_2}}{\omega_{k_1}}\; \mathbb{E}_{k_2}[\varphi(x_{m_3})] = \mathbb{E}_{k_1}[\varphi(x_{m_3})].$$

So, inclusion of the weight ensures correct classification of observations to latent groups, on average.

4. Estimation of the mixing proportions

Corollary 2 suggests estimating the mixing proportions ω by

$$\widehat{\omega} \equiv (\widehat{B}'\widehat{B})^{-1}\widehat{B}'\widehat{a}, \tag{S.4.1}$$

where $\widehat{a} \equiv (NM)^{-1}\sum_{n=1}^{N}\sum_{m=1}^{M}\chi(x_{nm})$, and

$$\widehat{B} \equiv (\widehat{b}_1, ..., \widehat{b}_K), \qquad \widehat{b}_k \equiv \frac{1}{N}\,\frac{(M-3)!}{M!}\sum_{n=1}^{N}\sum_{(m_1,m_2,m_3)} \widehat{\tau}_k(x_{nm_1}, x_{nm_2})\chi(x_{nm_3}).$$

Given the results in the main text, the asymptotic properties of $\widehat{\omega}$ are easy to derive. First, Theorem 1 implies that

$$B \;=\; \mathbb{E}\left[\begin{pmatrix} \chi_1(x_{m_3}) \\ \chi_2(x_{m_3}) \\ \vdots \\ \chi_I(x_{m_3}) \end{pmatrix}\begin{pmatrix} \tau_1(x_{m_1}, x_{m_2}) \\ \tau_2(x_{m_1}, x_{m_2}) \\ \vdots \\ \tau_K(x_{m_1}, x_{m_2}) \end{pmatrix}'\right] = \mathbb{E}[\chi(x_{m_3})\, T(x_{m_1}, x_{m_2})],$$

say. An application of Theorem 2 with $\varphi(x_m) = \chi(x_m)$ further shows that the plug-in estimator $\widehat{B}$ of B is consistent and asymptotically linear. Moreover, the influence function of $\mathrm{vec}[\widehat{B} - B]$ equals

$$\Theta_0 \, (\mathrm{I}_K \otimes \mathrm{I}_I) \, \iota_n + \frac{(M-3)!}{M!} \sum_{(m_1,m_2,m_3)} \mathrm{vec}[\chi(x_{nm_3}) \, T(x_{nm_1}, x_{nm_2}) - B],$$

where $\Theta_0 \equiv \mathrm{vertcat}[2 \, \mathbb{E}[\chi(x_m) u_k' W A(x_m)]]$. In addition, by Assumption 2, we have that $\sqrt{N}(\widehat{a} - a)$, too, is asymptotically normal.

Now, because $\widehat{B}$ is consistent for B and BB' is full rank, a linearization gives

$$\sqrt{N}(\widehat{\omega} - \omega) = (B'B)^{-1} \sqrt{N}(\widehat{B} - B)' a + (B'B)^{-1} B' \sqrt{N}(\widehat{a} - a) + o_P(1).$$

The asymptotic distribution of $\widehat{\omega}$ then follows readily from an application of the delta method.

5. Tests of rank and selection of I

Our approach requires choosing I so that the $I \times I$ matrix

$$A = A_I = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1I} \\ a_{21} & a_{22} & \cdots & a_{2I} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1} & a_{I2} & \cdots & a_{II} \end{pmatrix}, \qquad a_{i_1 i_2} = \mathbb{E}[\chi_{i_1}(x_{m_1}) \chi_{i_2}(x_{m_2})],$$

has rank K. For any fixed I, the plug-in estimator $\widehat{A}$ is $\sqrt{N}$-consistent and asymptotically normal by Lemma S.1, i.e.,

$$\sqrt{N} \, \mathrm{vec}(\widehat{A} - A) \xrightarrow{L} \mathcal{N}\big(0, \mathbb{E}[\upsilon_n \upsilon_n']\big).$$

This result can be used to construct a test of the rank of A, and thus to test our identifying condition in Assumption 1.

To test the rank of A we suggest using the procedure of Kleibergen and Paap (2006). Although one could use any of a number of alternative available rank tests, their statistic has several attractive features and, therefore, carries our preference. Prime advantages include its non-sensitivity to the ordering of variables, and the fact that its limit distribution under the null is free of nuisance parameters.

To present the test statistic, adapted to the current setting, order the eigenvalues of A in decreasing order, and place the eigenvectors in V accordingly. Fix $k \in \{1, 2, \ldots, I\}$ and let $\Lambda_{I-k} \equiv \Lambda(k+1 : I, k+1 : I)$ denote the lower-right $(I-k) \times (I-k)$ block of Λ. Also, let

$$V_k \equiv V(1 : k, k+1 : I), \qquad V_{I-k} \equiv V(k+1 : I, k+1 : I),$$

using obvious notation for matrix block selection. Now, let

$$\pi_k \equiv (\overline{V}_{I-k} \otimes \overline{V}_{I-k}) \; \mathrm{vec}\, A, \qquad \overline{V}_{I-k} \equiv (V_{I-k} V_{I-k}')^{1/2} V_{I-k}^{-1\prime} [V_k', V_{I-k}'].$$

Under the null H_0 : rank $A = k$,

$$\widehat{r}_k \equiv N\,\widehat{\pi}_k' \widehat{\mathscr{V}}_\pi^{-1} \widehat{\pi}_k \xrightarrow{L} \chi^2\big((I-k)^2\big), \quad \mathscr{V}_\pi \equiv (\overline{V}_{I-k} \otimes \overline{V}_{I-k})\,\mathbb{E}[\upsilon_n \upsilon_n'](\overline{V}_{I-k}' \otimes \overline{V}_{I-k}'), \tag{S.5.1}$$

where $\widehat{\pi}_k$ is the sample analog of π_k, $\widehat{\mathscr{V}}_\pi$ is a consistent estimator of $\mathscr{V}_\pi$, and $\chi^2\big((I-k)^2\big)$ denotes the chi-squared distribution with $(I-k)^2$ degrees of freedom.

The rank statistic in (S.5.1) can be used to test Assumption 1. It further suggests the following algorithm for selecting I in practice.

(i) Initialize the algorithm by setting $I_1 = K$ and move to Step 1.

(ii) Step s of the algorithm is

(iia) Test H_0 : rank $A_{I_s} = K-1$ using (S.5.1) at chosen significance level α.

(iib) If H_0 is not rejected, set $I_{s+1} = I_s + 1$ and continue to Step $s+1$.

(iic) If H_0 is rejected, set $I = I_s$ and stop.

Let $\widehat{I}$ be the resulting estimator, and let I_0 be the minimum I such that rank $A_I = K$. Assume that a known upper bound $\overline{I}$ on I_0 is available. Then $\widehat{I}$ will be weakly consistent for I_0, provided that the significance level α decreases at a suitable rate as a function of the sample size, as in Robin and Smith (2000).

6. Least-squares cross-validation

Recall the nonparametric density estimator

$$\widehat{f}_k(x) = \frac{1}{N}\frac{(M-3)!}{M!}\sum_{n=1}^N \sum_{(m_1,m_2,m_3)} \widehat{\tau}_k(x_{nm_1}, x_{nm_2})\,\frac{1}{h}\,\kappa\left(\frac{x_{nm_3}-x}{h}\right).$$

A popular method to select h in an automated manner is by least-squares cross-validation (Rudemo 1982). Least-squares cross-validation chooses h as to minimize a sample version of the integrated squared error

$$\int \big(\widehat{f}_k(x) - f_k(x)\big)^2 \,\mathrm{d}x.$$

The optimality properties of this selection scheme for standard kernel density estimation are well known (Hall 1983; Stone 1984). Expanding the square under the integral sign and ignoring terms that do not depend on h, the cross-validated bandwidth is the minimizer of

$$\mathrm{CV}(h) \equiv \int \widehat{f}_k(x)^2\,\mathrm{d}x - 2\int \widehat{f}_k(x)\,f_k(x)\,\mathrm{d}x. \tag{S.6.1}$$

To see how least-squares cross-validation can be implemented in our mixture setting we consider each of the right-hand side terms in (S.6.1) separately.

The first term in (S.6.1) is standard. Working out the square, $\int \widehat{f}_k(x)^2\,\mathrm{d}x$ equals

$$\frac{1}{N^2}\left(\frac{(M-3)!}{M!}\right)^2 \sum_{n=1}^N \sum_{n'=1}^N \sum_{(m_1,m_2,m_3)} \sum_{(m_1',m_2',m_3')} \gamma(x_{nm_1}, x_{nm_2}, x_{nm_3}; x_{n'm_1'}, x_{n'm_2'}, x_{n'm_3'}),$$

where $\gamma(x_{nm_1}, x_{nm_2}, x_{nm_3}; x_{n'm'_1}, x_{n'm'_2}, x_{n'm'_3})$ is defined as

$$\widehat{\tau}_k(x_{nm_1}, x_{nm_2})\, \widehat{\tau}_k(x_{n'm'_1}, x_{n'm'_2})\, \frac{1}{h}\, \overline{\kappa}\left(\frac{x_{n'm'_3} - x_{nm_3}}{h}\right)$$

with $\overline{\kappa}(x) \equiv \int \kappa(u)\, \kappa(x-u)\, \mathrm{d}u$. The convolution of the kernel takes a particularly tractable form when κ is taken to be the standard-normal density. Indeed, in this case, $\overline{\kappa}$ is just a normal density with mean zero and variance equal to two.

The second term in (S.6.1) appears more complicated because it depends on f_k, which is unknown. However, we can use Theorem 1 to obtain

$$\int \widehat{f}_k(x)\, f_k(x)\, \mathrm{d}x = \mathbb{E}\left[\tau_k(x_{m_1}, x_{m_2}) \widehat{f}_k(x_{m_3})\right],$$

where the expectation is taken with respect to the joint density of $(x_{m_1}, x_{m_2}, x_{m_3})$. Moreover, a plug-in estimator of this expectation is

$$\frac{1}{N}\frac{(M-3)!}{M!}\sum_{n=1}^{N}\sum_{(m_1,m_2,m_3)} \widehat{\tau}_k(x_{nm_1}, x_{nm_2})\, \widehat{f}_k^{-n}(x_{nm_3}),$$

where $\widehat{f}_k^{-n}$ denotes the standard leave-one-out kernel density estimator, that is,

$$\widehat{f}_k^{-n_1}(x) \equiv \frac{1}{N-1}\frac{(M-3)!}{M!}\sum_{n\neq n_1}\sum_{(m_1,m_2,m_3)} \widehat{\tau}_k(x_{nm_1}, x_{nm_2})\, \frac{1}{h}\, \kappa\left(\frac{x_{nm_3} - x}{h}\right).$$

We leave the formal study of the large-sample properties of this cross-validation criterion along the lines of Hall and Marron (1987) to future work.

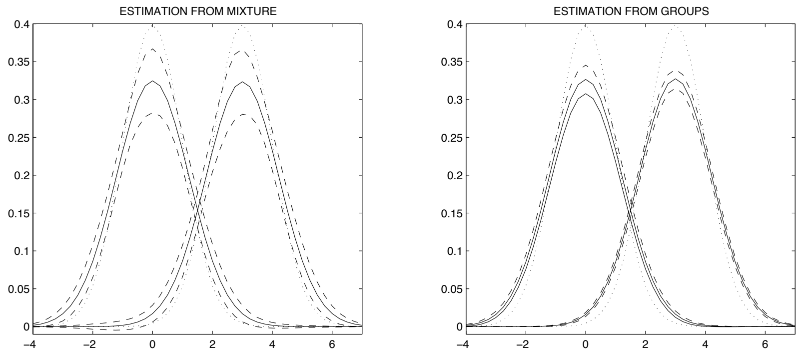
7. Comparison to oracle density estimator

Let g_n be a random variable denoting group membership of observation n. Note that $\Pr[g_n = k] = \omega_k$. In our Monte Carlo experiment we now compare our density estimator $\widehat{f}_k$ to the oracle estimator

$$\widehat{f}_k^{\star}(x) \equiv \frac{\frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M} \frac{1}{h}\,\kappa\left(\frac{x_{nm}-x}{h}\right)\, 1\{g_n = k\}}{\frac{1}{N}\sum_{n=1}^{N} 1\{g_n = k\}},$$

that is, a standard kernel density estimator applied to the subsample of observations drawn from the k-th subpopulation. Figure S.1 contains the average point estimates (full lines), .025 and .975 quantiles of the empirical distributions of the point estimates (dashed lines), and true values (dotted lines) of our estimator $\widehat{f}_k$ (left plot) and the oracle estimator $\widehat{f}_k^{\star}$ just defined (right plot). For both estimators, the bandwidth was chosen via least-squares cross-validation. The plots show that both estimators yield broadly the same density estimate, on average. Thus, our reweighting approach does not suffer from more bias than does the oracle estimator. Not surprisingly, our feasible estimator is somewhat more variable, however.

Figure S.1. Comparison to oracle density estimator. Design from Levine, Hunter, and Chauveau (2011). $N = 500, M = 3$. Statistics obtained over $1,000$ replications.



Average point estimates (full), .025 and .975 quantiles of the empirical distributions of the point estimates (dashed lines), and true values (dotted).

8. Additional simulation results

We present additional evidence on the performance of our estimators in a two-component mixture of Beta distributions on $[-1,1]$. The density function of the Beta distribution on this interval is

$$\mathsf{b}(x;\vartheta_1,\vartheta_2) \equiv \frac{1}{2^{\vartheta_1+\vartheta_2-1}}\frac{1}{\mathsf{B}(\vartheta_1,\vartheta_2)}(1+x)^{\vartheta_1-1}(1-x)^{\vartheta_2-1},$$

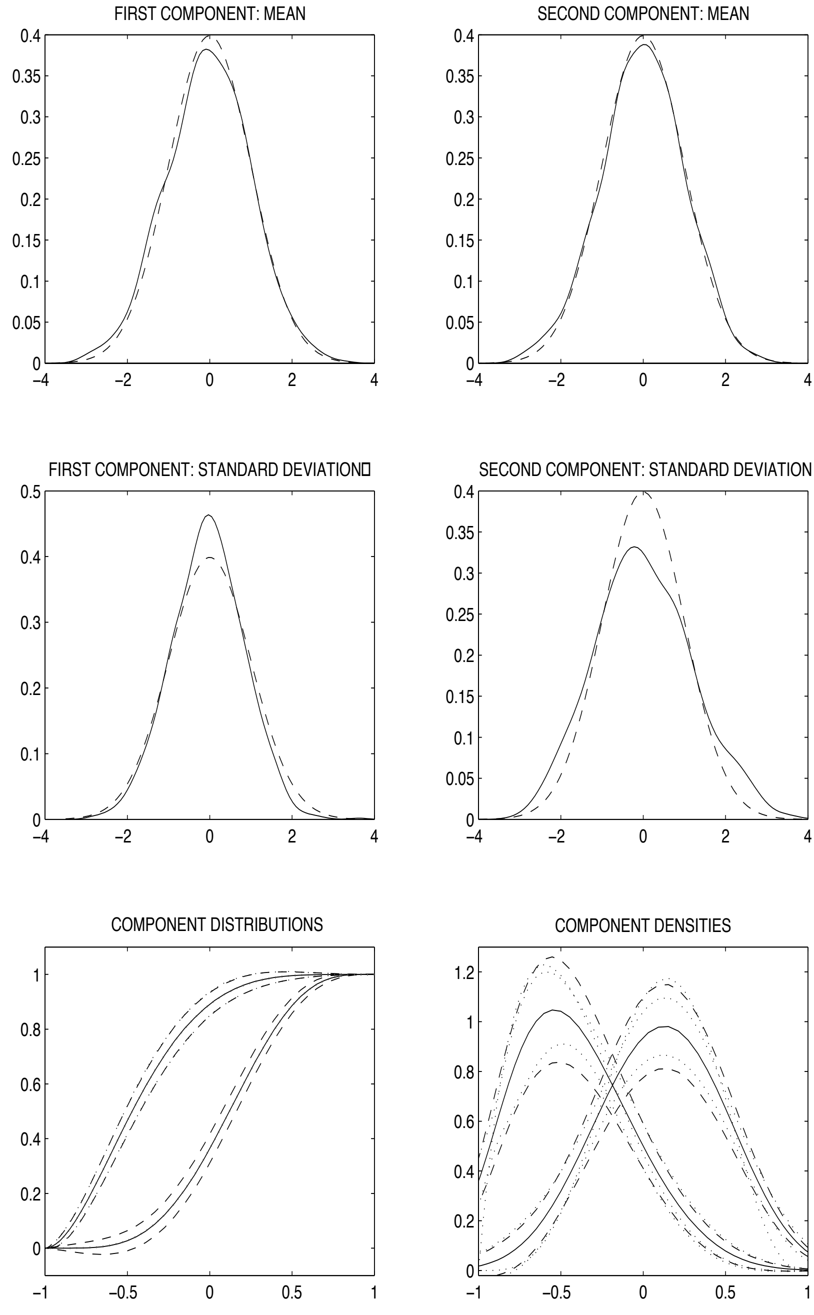
where $\mathsf{B}(\vartheta_1,\vartheta_2) \equiv \int_0^1 x^{\vartheta_1-1}(1-x)^{\vartheta_2-1}\,\mathrm{d}x$, and ϑ_1 and ϑ_2 are positive scale parameters. Its mean and variance are

$$\mu(\vartheta_1,\vartheta_2) \equiv -1 + 2\,\frac{\vartheta_1}{\vartheta_1+\vartheta_2}, \qquad \text{and} \qquad \sigma^2(\vartheta_1,\vartheta_2) \equiv 4\,\frac{\vartheta_1\vartheta_2}{(\vartheta_1+\vartheta_2)^2(\vartheta_1+\vartheta_2+1)}, \qquad \text{(S.8.1)}$$

respectively. We experimented with various parameter values. Here we present results for a design where component densities $f_1(x) = \mathsf{b}(x;2,5)$ and $f_2(x) = \mathsf{b}(x;5,4)$ are mixed with $\omega_1 = .50$ and $\omega_2 = .50$. We use normalized Chebychev polynomials of the first kind for $\chi_1, \chi_2, \ldots, \chi_I$ and consider several choices for I. The sample size is fixed at $N = 500$ and $M = 3$. Implementation of our estimators is as discussed in the main text. All statistics are computed over $1,000$ Monte Carlo replications. All results are collected in Figure S.2 and Table S.1 contain the results.

The upper and middle panels in Figure S.2 provide the empirical densities of the Studentized point estimates (full lines) of the component means (middle panels) and component standard deviations (lower panels), along with a standard-normal density (dashed lines) as a benchmark. These plots show that our asymptotic approximation does well in capturing the small-sample behavior of our estimators. The deviation from normality is somewhat larger for the estimates of the standard deviation than for the estimates of the mean. This is not

Figure S.2. Simulation results for a two-component mixture of Beta distributions on $[-1, 1]$. $N = 500, M = 3$. Statistics obtained over $1,000$ replications.



Upper panels: empirical density functions (full) of Studentized estimates of component means and standard deviations obtained over $1,000$ replications, together with a standard normal density (dashed). Lower panels: average point estimate (full) and .95% confidence band (dashed) for component distributions and densities, together with the functions respective true values (dotted).

Table S.1. Simulation results for a two-component mixture of Beta distributions on $[-1, 1]$. $N = 500, M = 3$. Statistics obtained over $1,000$ replications.

		component functionals							
θ_0	I	BIAS		SD		SE/SD		CR(95%)	
μ	5	.000	.000	.022	.025	.987	.996	.943	.948
σ	5	−.001	−.004	.016	.029	1.138	.794	.979	.891
μ	10	−.001	−.002	.021	.022	1.029	1.031	.951	.961
σ	10	−.002	−.001	.015	.024	1.184	.856	.980	.919
		component distributions at x							
x	I	BIAS		SD		SE/SD		CR(95%)	
−.50	5	.000	.000	.027	.019	1.024	1.014	.958	.949
0	5	.000	.000	.023	.027	1.001	.980	.948	.942
.50	5	.000	.000	.007	.014	.989	1.001	.939	.946
−.50	10	.001	.001	.031	.022	1.019	1.018	.952	.960
0	10	.000	.002	.022	.029	1.020	.985	.953	.945
.50	10	−.001	.000	.005	.015	1.111	1.009	.970	.953

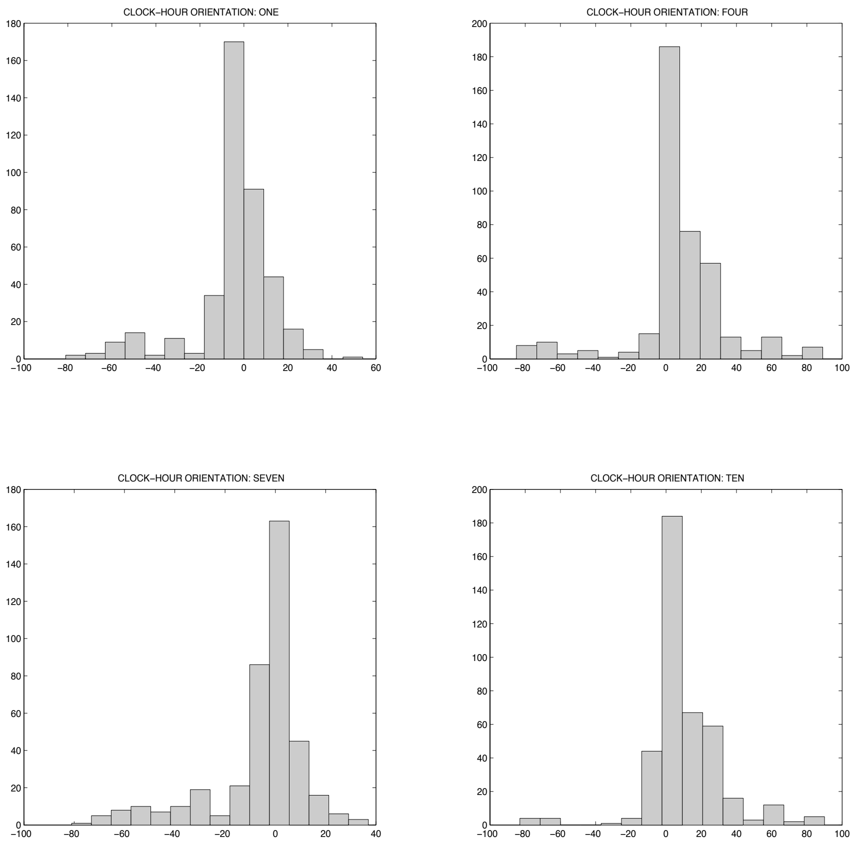
surprising given that σ is a nonlinear functional and that its estimator relies on second-order moments. Table S.1 further gives the ratio of the (mean of the) estimated standard errors to the empirical standard deviation, as well as coverage rates of 95%-confidence intervals. The ratios are fairly close to one, providing further evidence of the usefulness of our asymptotic theory. As a consequence, the coverage of the confidence intervals is close to 95%. The coverage rates are somewhat more accurate for the mean compared to the standard deviation, which is in line with the above discussion.

The lower left plot in Figure S.2 shows the average point estimate (solid lines) and average 95% confidence band (dashed lines) for the component distributions. The true distributions are also plotted (dotted line) but they are difficult to see as they essentially coincide with the mean of the point estimates. That is, our estimator is virtually unbiased. Furthermore, the .025 and .975 quantiles of the point estimates, too, would be almost indistinguishable from the reported average confidence band, and so they are not reported on the figure. This shows that our method yields accurate inference on the component cumulative distribution functions. The bottom panel in Table S.1 confirms this. It shows inference results for the component distributions at $x \in \{-.50, 0, .50\}$. The estimates show small biases, and the confidence intervals have very accurate coverage. The lower right plot in Figure S.2 provides the corresponding results for the density estimator.

9. Additional results for the empirical application

Figure S.3 contains histograms of each of the individual measurements of the water-level dataset.

Figure S.3. Histograms of the individual measurements in the water-level data of Thomas, Lohaus, and Brainerd (1993).



References

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics 34*, 122–148.

Bonhomme, S. and J.-M. Robin (2009). Consistent noisy independent component analysis. *Journal of Econometrics 149*, 12–25.

Eaton, M. L. and D. E. Tyler (1991). On Wielandt's inequality and its applications. *Annals of Statistics 19*, 260–271.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Annals of Statistics 11*, 1156–1174.

Hall, P. and J. S. Marron (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probability Theory and Related Fields 74*, 567–581.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*, 1029–1054.

Kleibergen, F. and R. Paap (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics 133*, 97–126.

Levine, M., D. R. Hunter, and D. Chauveau (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika 98*, 403–416.

Magnus, J. R. (1985). On differentiating eigenvalues and eigenvectors. *Econometric Theory 1*, 179–191.

Magnus, J. R. and H. Neudecker (1979). The commutation matrix: Some properties and applications. *Annals of Statistics 7*, 381–394.

Robin, J.-M. and R. J. Smith (2000). Tests of rank. *Econometric Theory 16*, 151–175.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics 9*, 65–78.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics 12*, 1285–1297.

Thomas, H., A. Lohaus, and C. J. Brainerd (1993). Modeling growth and individual differences in spatial tasks. *Monographs of the Society for Research in Child Development 58*, 1–191.