SUPPLEMENT TO ESTIMATING MULTIVARIATE LATENT-STRUCTURE MODELS

By Stéphane Bonhomme, Koen Jochmans and Jean-Marc Robin

University of Chicago; Sciences Po, Paris; and Sciences Po, Paris and University College, London

This version: August 19, 2015

CONTENTS

S.1. Mixture representation of a hidden Markov model. We show that Equation (2.5) in the text holds. We normalize the support of Y_i to $\{1, 2, \ldots, \kappa\}$ for notational simplicity.

By definition,

$$\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_r) = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1r} \\ p_{21} & p_{22} & \cdots & p_{2r} \\ \vdots & \vdots & \cdots & \vdots \\ p_{\kappa 1} & p_{\kappa 2} & \cdots & p_{\kappa r} \end{pmatrix}$$

is the matrix whose columns are the (stationary) emission distributions for the different latent states $(1, 2, \ldots, r)$. For example, $p_{11} = \Pr(Y_i = 1 | Z_i = 1)$ and

$$p_{kj} = \Pr(Y_i = k | Z_i = j),$$

in general.

Also,

$$\boldsymbol{K}' = \begin{pmatrix} K_{11} & K_{12} & \cdots & K_{1r} \\ K_{21} & K_{22} & \cdots & K_{2r} \\ \vdots & \vdots & \cdots & \vdots \\ K_{r1} & K_{r2} & \cdots & K_{rr} \end{pmatrix}$$

is the matrix whose jth column gives the probabilities of going from the jth state in period $i-1$ to each of the r states in period i; by stationarity, these probabilities do not depend on i.

The Markovian assumption implies the following conditional-independence restrictions:

(A) Y_i and Z_{i-1} are independent given Z_i,
(B) Y_i and Z_{i+1} are independent given Z_i,
(C) Y_{i-1}, Y_i and Y_{i+1} are independent given Z_i.

By (A) the matrix $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{K}'$ has entries

$$b_{kj} = \sum_{c=1}^{r} p_{kc} \, K_{cj} = \Pr(Y_i = k | Z_{i-1} = j),$$

where k ranges over all κ values Y_i can take and j ranges over all r latent states. The time-reversed decomposition $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Pi}\boldsymbol{K}\boldsymbol{\Pi}^{-1}$ is derived in the same way, using (B).

Now, to verify (2.5) first note that

$$\Pr(Y_i = k) = \sum_{j=1}^{r} p_{kj}\, \pi_j$$

for any i and k. Therefore, the marginal distribution of Y_i is simply the $\kappa \times 1$ vector $\boldsymbol{P\pi}$. Similarly, the joint distribution of $(Y_i, Y_i{+}1)$ is the two-way table

$$\boldsymbol{P\Psi P}' = \boldsymbol{P\Pi K P}' = \boldsymbol{P\Pi}(\boldsymbol{PK}')' = \boldsymbol{P\Pi B}',$$

where $\boldsymbol{\Psi} = \boldsymbol{\Pi K}$ denotes the $r \times r$ contingency table of (Z_i, Z_{i+1}); so $\boldsymbol{\Psi}(j_1, j_2) = \Pr(Z_i = j_1, Z_{i+1} = j_2)$. Note that this indeed is a bivariate mixture representation and that

$$\sum_{j=1}^{r} \pi_j\, (\boldsymbol{p}_j \otimes \boldsymbol{b}_j)$$

is an alternative way of writing it. Lastly, we may consider the probability distribution involving all three outcomes. Let $\boldsymbol{D}_k = \mathrm{diag}(p_{k1}, p_{k2}, \dots, p_{kr})$. By (C), the three-way table $\mathbb{P}$ is composed of the κ matrices

$$\mathbb{P}(:,:,k) = \boldsymbol{P\Pi K D}_k \boldsymbol{K P}' = (\boldsymbol{P\Pi K \Pi}^{-1})\, \boldsymbol{\Pi D}_k(\boldsymbol{PK}')' = \boldsymbol{A\Pi D}_k \boldsymbol{B}'.$$

Because the matrices $\boldsymbol{D}_k$ are the rows of $\boldsymbol{P}$ and each of $\boldsymbol{A}, \boldsymbol{B}$, and $\boldsymbol{P}$ are probability distributions, we indeed have the factorization as a finite mixture

$$\mathbb{P} = \sum_{j=1}^{r} \pi_j\, (\boldsymbol{a}_j \otimes \boldsymbol{p}_j \otimes \boldsymbol{b}_j),$$

as claimed.

S.2. Using marginalizations when no submodels are available. We adapt the conclusions and the proof of Theorem 1 to the situation where submodels are not observable. Recall that the array can be written as the collection of matrices

$$\boldsymbol{A}_k = \boldsymbol{X}_1 \boldsymbol{\Pi D}_k \boldsymbol{X}_2',$$

where $\boldsymbol{D}_k = \mathrm{diag}_k\, \boldsymbol{X}_3$. The marginalization of the array (toward direction 3) is the matrix

$$\overline{\boldsymbol{A}}_0 = \boldsymbol{X}_1 \boldsymbol{\Pi} \overline{\boldsymbol{D}} \boldsymbol{X}_2', \qquad \overline{\boldsymbol{D}} = \sum_k \boldsymbol{D}_k.$$

Assuming that $\overline{\boldsymbol{D}}$ has no zeros on its diagonal, $\overline{\boldsymbol{A}}_0$ has rank r and associated singular-value decomposition

$$\overline{\boldsymbol{A}}_0 = \overline{\boldsymbol{U}}\,\overline{\boldsymbol{S}}\,\overline{\boldsymbol{V}}',$$

say. From this we can construct matrices $\overline{\boldsymbol{W}}_1$ and $\overline{\boldsymbol{W}}_2$ as before. Joint diagonalization of the matrices

$$\overline{\boldsymbol{W}}_1 \boldsymbol{A}_k \overline{\boldsymbol{W}}_2' = \overline{\boldsymbol{W}}_1 \boldsymbol{X}_1 \boldsymbol{\Pi}\boldsymbol{D}_k \boldsymbol{X}_2' \overline{\boldsymbol{W}}_2' = \overline{\boldsymbol{Q}}\,(\overline{\boldsymbol{D}}^{-1}\boldsymbol{D}_k)\,\overline{\boldsymbol{Q}}^{-1}$$

then yields the eigenvalues $\overline{\boldsymbol{D}}^{-1}\boldsymbol{D}_k$, and so the matrix

$$\overline{\boldsymbol{X}}_3 = \boldsymbol{X}_3\,\overline{\boldsymbol{D}}^{-1}$$

(up to a permutation of its columns). By construction, each column of $\overline{\boldsymbol{X}}_3$ sums to one while, in general, the columns of $\boldsymbol{X}_3$ (and, hence, the entries of $\overline{\boldsymbol{D}}$) are unrestricted. So, in the absence of submodels, our approach yields a scaled version of $\boldsymbol{X}_3$.

S.3. Omitted proofs for the theorems in Section 3.

Proof of Theorem 2. Theorem 1 can be applied to each direction of the three-way array $\mathbb{X}$. This yields the $\boldsymbol{X}_i$ up to permutation of their columns. However, as each $\boldsymbol{X}_i$ is recovered from a different simultaneous-diagonalization problem, the ordering of the columns of the $\boldsymbol{X}_i$ so obtained need not be the same. Hence, it remains to be shown that we can unravel the orderings. More precisely, application of Theorem 1 for each i identifies, say,

$$\boldsymbol{X}_1, \qquad \boldsymbol{Y}_2 = \boldsymbol{X}_2\boldsymbol{\Delta}_2, \qquad \boldsymbol{Y}_3 = \boldsymbol{X}_3\boldsymbol{\Delta}_3,$$

where $\boldsymbol{\Delta}_2$ and $\boldsymbol{\Delta}_3$ are two permutation matrices.

Now, given $\boldsymbol{X}_1$ and the lower-dimensional submodels $\mathbb{X}_{\{1,2\}}$ and $\mathbb{X}_{\{1,3\}}$, we observe the projection coefficients

$$\boldsymbol{M}_2 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\,\mathbb{X}_{\{1,2\}} = \boldsymbol{\Pi}\boldsymbol{X}_2' = \boldsymbol{\Pi}\boldsymbol{\Delta}_2\boldsymbol{Y}_2',$$
$$\boldsymbol{M}_3 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\,\mathbb{X}_{\{1,3\}} = \boldsymbol{\Pi}\boldsymbol{X}_3' = \boldsymbol{\Pi}\boldsymbol{\Delta}_3\boldsymbol{Y}_3',$$

where the first transition holds by the structure of the lower-dimensional submodels and the second transition follows from the fact that permutation matrices are orthogonal. Also,

$$\mathbb{X}_{\{2,3\}} = \boldsymbol{X}_2\boldsymbol{\Pi}\boldsymbol{X}_3' = \boldsymbol{Y}_2\boldsymbol{\Delta}_2'\boldsymbol{\Pi}\boldsymbol{\Delta}_3\boldsymbol{Y}_3' = \boldsymbol{M}_2'\boldsymbol{\Delta}_3\boldsymbol{Y}_3', \quad \mathbb{X}_{\{2,3\}}' = \boldsymbol{M}_3'\boldsymbol{\Delta}_2\boldsymbol{Y}_2'.$$

The latter two equations can be solved for the permutation matrices, yielding

$$\begin{aligned}\boldsymbol{\Delta}_2 &= (\boldsymbol{M}_3\boldsymbol{M}_3')^{-1}\boldsymbol{M}_3\ \mathbb{X}_{\{2,3\}}'\ \mathbf{Y}_2(\mathbf{Y}_2'\mathbf{Y}_2)^{-1},\\ \boldsymbol{\Delta}_3 &= (\boldsymbol{M}_2\boldsymbol{M}_2')^{-1}\boldsymbol{M}_2\ \mathbb{X}_{\{2,3\}}\ \mathbf{Y}_3(\mathbf{Y}_3'\mathbf{Y}_3)^{-1}.\end{aligned} \tag{S.1}$$

This concludes the proof. □

S.4. Application of Theorems 1–3 to motivating examples. We specialize our generic identification results to our motivating examples from Section 2.

S.4.1. *Latent-class models.* First reconsider the finite-mixture model with discrete outcomes and a known number of components r in (2.1), that is, the q-way table

$$\mathbb{P} = \sum_{j=1}^{r} \pi_j \bigotimes_{i=1}^{q} \boldsymbol{p}_{ij}.$$

Let $\boldsymbol{P}_i = (\boldsymbol{p}_{i1}, \boldsymbol{p}_{i2}, \ldots, \boldsymbol{p}_{ir})$ and $\boldsymbol{\Pi} = \operatorname{diag}(\pi_1, \pi_2, \ldots, \pi_r)$. The following theorem concerns identification of these parameters.

THEOREM S.1 (Identification of finite mixtures). *The matrices $\{\boldsymbol{P}_i\}$ and $\boldsymbol{\Pi}$ in the finite mixture model in (2.1) are all identified if $\operatorname{rank} \boldsymbol{P}_i = r$ for all i, $\pi_j > 0$ for all j, and $q \geq 3$.*

PROOF. To show Theorem S.1 it suffices again to set $q = 3$. The proof is then a direct application of our identification result. Theorem 2 yields the matrices of component distributions $\{\boldsymbol{P}_i\}$ and Theorem 3 yields the vector of mixing proportions $\boldsymbol{\pi}$, all up to a common permutation matrix. □

Theorem S.1 requires that $\kappa_i \geq r$ for all $i = 1, 2, \ldots, q$, that is, that all distributions p_{ij} have more than r points of support but applies as soon as $q = 3$. The rank conditions can be weakened when $q > 3$. Our approach to proving Theorem S.1 is a constructive version of the proof of Theorem 4 in Allman, Matias and Rhodes [2009].

S.4.2. *Hidden Markov models.* Now turn to the hidden Markov model in (2.5) with a known number of latent states, r. In this model, the parameters of interest are the $\kappa \times r$ matrix of emission distributions $\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_r)$, the stationary distribution of the r latent states $\boldsymbol{\pi}$, and the $r \times r$ matrix of transition probabilities $\boldsymbol{K}$. The next theorem gives sufficient conditions for identification.

Theorem S.2 (Identification of hidden Markov models). *The matrices* $\boldsymbol{P}, \boldsymbol{K}$, *and* $\boldsymbol{\Pi}$ *in the hidden Markov model are all identified if* $\operatorname{rank} \boldsymbol{P} = r$ *and* $\operatorname{rank} \boldsymbol{K} = r$, *and* $\pi_j > 0$ *for all* j *provided* $q \geq 3$.

Proof. Set $q = 3$. Then the contingency table of three measurements factors as

$$\mathbb{P} = \sum_{j=1}^{r} \pi_j \, (\boldsymbol{a}_j \otimes \boldsymbol{p}_j \otimes \boldsymbol{b}_j);$$

see (2.5). Moreover, this states that appropriate conditioning allows to write the hidden Markov models as a finite-mixture model of the form in (2.1). Furthermore, the rank conditions on $\boldsymbol{P}$ and $\boldsymbol{K}$ imply that both $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{K}'$ and $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Pi}\boldsymbol{K}\boldsymbol{\Pi}^{-1}$ also have full column rank r. Theorem S.1 immediately yields identification of the matrix of emission distributions $\boldsymbol{P}$ and of the stationary distribution $\boldsymbol{\pi}$.

Finally, Theorem S.1 also provides the matrix $\boldsymbol{B}$ and, because the model implies that $\boldsymbol{B} = \boldsymbol{P}\boldsymbol{K}'$,

$$\boldsymbol{K} = \boldsymbol{B}'\boldsymbol{P}(\boldsymbol{P}'\boldsymbol{P})^{-1}.$$

The hidden Markov model is overidentified. Indeed, besides $\boldsymbol{B}$ we also have the matrix $\boldsymbol{A}$, which yields the same type of restrictions on the matrix $\boldsymbol{K}$ as does $\boldsymbol{B}$. □

Theorem S.2 states the same identification requirements as Theorem 2.1 of Gassiat, Cleynen and Robin [2013], but the method of proof followed here is constructive.

S.5. Omitted proofs for the theorems in Section 4.

Proof of Theorem 4. For a generic matrix $\boldsymbol{M} = (\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots, \boldsymbol{M}_\kappa)$ of dimension $r \times r\kappa$, denote the objective function as

$$L(\boldsymbol{Q}, \boldsymbol{M}) = \sum_{k=1}^{\kappa} \| \mathrm{off}(\boldsymbol{Q}^{-1}\boldsymbol{M}_k\boldsymbol{Q}) \|_F^2.$$

Write $\boldsymbol{Q}_{-i,-j}$ for the principal minors of $\boldsymbol{Q}$. Because for any $\boldsymbol{Q} \in \mathscr{Q}$ we have

$$[\boldsymbol{Q}^{-1}]_{i,j} = (-1)^{i+j} \det \boldsymbol{Q}_{-i,-j}$$

and $\det \boldsymbol{A}$ is a polynomial function of $\boldsymbol{A}$, the function $L(\boldsymbol{Q}, \boldsymbol{M})$ is continuous in each of its arguments. Let

$$L_0(\boldsymbol{Q}) = L(\boldsymbol{Q}, \boldsymbol{C}), \qquad L_n(\boldsymbol{Q}) = L(\boldsymbol{Q}, \widehat{\boldsymbol{C}}).$$

Note that $\mathscr{Q}_0 = \arg\inf_{\boldsymbol{Q} \in \mathscr{Q}} L_0(\boldsymbol{Q})$ is the equivalence class containing all $\boldsymbol{Q} \in \mathscr{Q}$ that are equal to $\boldsymbol{Q}_0$ up to a permutation and direction of their columns.

Because $\|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_F = o_p(1)$ and $L(\boldsymbol{Q}, \boldsymbol{M})$ is continuous in $\boldsymbol{M}$, for all $\boldsymbol{Q} \in \mathscr{Q}$,

$$L_n(\boldsymbol{Q}) \xrightarrow{p} L_0(\boldsymbol{Q})$$

by the continuous-mapping theorem. Further, by the same argument, for all $\boldsymbol{Q}, \boldsymbol{Q}' \in \mathscr{Q}$,

$$|L_n(\boldsymbol{Q}) - L_n(\boldsymbol{Q}')| \leq O_p(1) \, \|\boldsymbol{Q} - \boldsymbol{Q}'\|_F,$$

because $L_n(\boldsymbol{Q})$ is polynomial in $\boldsymbol{Q}$, and thus Lipschitz continuous. With $\mathscr{Q}$ compact it follows that $L_n(\boldsymbol{Q})$ is stochastically equicontinuous (see, for example, Newey and McFadden 1994, Lemma 2.9), and so

$$\sup_{\boldsymbol{Q} \in \mathscr{Q}} |L_n(\boldsymbol{Q}) - L_0(\boldsymbol{Q})| = o_p(1).$$

Then, for any open subset $\mathscr{O}$ of $\mathscr{Q}$ containing $\mathscr{Q}_0$, with complement $\mathscr{O}^c$, it holds that

$$L_0(\widehat{\boldsymbol{Q}}) < \inf_{\boldsymbol{Q} \in \mathscr{O}^c} L_0(\boldsymbol{Q})$$

with probability approaching one. Hence, we have $\lim_{n \to \infty} \Pr(\widehat{\boldsymbol{Q}} \in \mathscr{O}) = 1$ (Newey and McFadden 1994, Theorem 2.1). □

PROOF OF THEOREM 5. For the proof of Theorem 5 it is convenient to work with a different yet equivalent normalization on $\boldsymbol{Q}_0$. More precisely, the set

$$\mathscr{Q} = \{\boldsymbol{Q} : \det \boldsymbol{Q} = 1, \|\boldsymbol{q}_j\|_F = c \text{ for } j = 1, 2, \ldots, r \text{ and } c \leq m\}$$

is in a simple one-to-one correspondence with the set

$$\mathscr{Q}' = \{\boldsymbol{Q} : \det \boldsymbol{Q} = c^{-r}, \|\boldsymbol{q}_j\|_F = 1 \text{ for } j = 1, 2, \ldots, r \text{ and } c \leq m\}.$$

The latter is easier to work with here because constraining columns to have unit norm is easier than requiring the determinant of the matrix to equal unity. In any case, the asymptotic distribution of $\widehat{\boldsymbol{Q}}$ turns out not to depend on c.

We first derive the first-order conditions to the constrained minimization problem that defines $\widehat{\boldsymbol{Q}}$. Given these, we can then proceed using standard arguments to derive the asymptotic distribution of the joint approximate diagonalizer.

Lagrangian and first-order conditions. It is useful to reformulate the joint approximate-diagonalization problem as

$$\min_{\boldsymbol{Q},\boldsymbol{R}} \sum_{k=1}^{\kappa} \left\| \operatorname{off}(\boldsymbol{R}\,\widehat{\boldsymbol{C}}_k \boldsymbol{Q})\right\|_F^2, \qquad \text{s.t. } \boldsymbol{R}\boldsymbol{Q} = \boldsymbol{I}_r, \quad \|\boldsymbol{q}_j\|_F = 1 \ \forall j.$$

With $[\boldsymbol{Q}]_{i,j}$ and $[\boldsymbol{R}]_{i,j}$ denoting the (i,j)th entries of matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$, respectively, the Lagrangian for this constrained minimization problem with respect to $(\boldsymbol{Q},\boldsymbol{R})$ is

$$\begin{aligned} L(\boldsymbol{Q},\boldsymbol{R}) = &\sum_{k=1}^{\kappa} \left\| \operatorname{off}(\boldsymbol{R}\,\widehat{\boldsymbol{C}}_k \boldsymbol{Q})\right\|_F^2 \\ &+ \sum_{i,j=1}^{r} \lambda_{ij} \left(\sum_{\ell=1}^{r} [\boldsymbol{R}]_{i,\ell} [\boldsymbol{Q}]_{\ell,j} - \delta_{ij}\right) + \sum_{j=1}^{r} \gamma_j (\boldsymbol{q}_j'\boldsymbol{q}_j - 1), \end{aligned}$$

for Lagrange multipliers $[\boldsymbol{\Lambda}]_{i,j} = \lambda_{ij}$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_r)'$ associated with each of the constraints, and δ_{ij} denoting Kronecker's delta.

Application of the chain rule readily gives

$$\text{(S.2)} \qquad \frac{\partial L(\boldsymbol{Q},\boldsymbol{R})}{\partial \boldsymbol{Q}} = 2 \sum_{k=1}^{\kappa} \widehat{\boldsymbol{C}}_k' \boldsymbol{R}' \operatorname{off}(\boldsymbol{R}\,\widehat{\boldsymbol{C}}_k \boldsymbol{Q}) + \boldsymbol{R}'\boldsymbol{\Lambda} + 2\boldsymbol{Q}\boldsymbol{\Gamma},$$

$$\text{(S.3)} \qquad \frac{\partial L(\boldsymbol{Q},\boldsymbol{R})}{\partial \boldsymbol{R}} = 2 \sum_{k=1}^{\kappa} \operatorname{off}(\boldsymbol{R}\,\widehat{\boldsymbol{C}}_k \boldsymbol{Q})\, \boldsymbol{Q}'\widehat{\boldsymbol{C}}_k' + \boldsymbol{\Lambda}\boldsymbol{Q}',$$

with $\boldsymbol{\Gamma} = \operatorname{diag}(\boldsymbol{\gamma})$. Substitute $\boldsymbol{R} = \boldsymbol{Q}^{-1}$ in (S.3) and solve $\frac{\partial L(\boldsymbol{Q},\boldsymbol{R})}{\partial \boldsymbol{R}} = \mathbf{0}$ for $\boldsymbol{\Lambda}$ to get

$$\boldsymbol{\Lambda} = -2\sum_{k=1}^{\kappa} \operatorname{off}(\boldsymbol{Q}^{-1}\widehat{\boldsymbol{C}}_k\boldsymbol{Q})\,(\boldsymbol{Q}^{-1}\widehat{\boldsymbol{C}}_k\boldsymbol{Q})'.$$

Next, substitute this value for $\boldsymbol{\Lambda}$ and $\boldsymbol{R} = \boldsymbol{Q}^{-1}$ in (S.2) and premultiply with $\boldsymbol{Q}'$ to get

$$\sum_{k=1}^{\kappa}(\boldsymbol{Q}^{-1}\widehat{\boldsymbol{C}}_k\boldsymbol{Q})' \operatorname{off}(\boldsymbol{Q}^{-1}\widehat{\boldsymbol{C}}_k\boldsymbol{Q}) - \operatorname{off}(\boldsymbol{Q}^{-1}\widehat{\boldsymbol{C}}_k\boldsymbol{Q})(\boldsymbol{Q}^{-1}\widehat{\boldsymbol{C}}_k\boldsymbol{Q})' + \boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{\Gamma} = \mathbf{0}.$$

Force the columns of $\boldsymbol{Q}$ to have unit Euclidean norm, so that $\operatorname{diag}(\boldsymbol{Q}'\boldsymbol{Q}) = \boldsymbol{I}_r$, to see that

$$\boldsymbol{\Gamma} = -\operatorname{diag}\left(\sum_{k=1}^{\kappa}(\boldsymbol{Q}^{-1}\widehat{\boldsymbol{C}}_k\boldsymbol{Q})' \operatorname{off}(\boldsymbol{Q}^{-1}\widehat{\boldsymbol{C}}_k\boldsymbol{Q}) - \operatorname{off}(\boldsymbol{Q}^{-1}\widehat{\boldsymbol{C}}_k\boldsymbol{Q})(\boldsymbol{Q}^{-1}\widehat{\boldsymbol{C}}_k\boldsymbol{Q})'\right),$$

as $\boldsymbol{\Gamma} = \operatorname{diag}(\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{\Gamma})$ because $\boldsymbol{\Gamma}$ is diagonal. Then the first-order condition for $\boldsymbol{Q}$ of our constrained minimization problem is obtained on plugging this expression back in to (S.2). To write it compactly, let

$$\Delta(\boldsymbol{M}) = \boldsymbol{M}' \operatorname{off}(\boldsymbol{M}) - \operatorname{off}(\boldsymbol{M})\boldsymbol{M}'$$

for any matrix $\boldsymbol{M}$ and

$$S(\boldsymbol{Q},\boldsymbol{M}) = \sum_{k=1}^{\kappa}(\boldsymbol{Q}')^{-1}\Delta(\boldsymbol{Q}^{-1}\boldsymbol{M}_k\boldsymbol{Q}) - \boldsymbol{Q}\operatorname{diag}(\Delta(\boldsymbol{Q}^{-1}\boldsymbol{M}_k\boldsymbol{Q}))$$

for any $\boldsymbol{M} = (\boldsymbol{M}_1, \boldsymbol{M}_2, \ldots, \boldsymbol{M}_\kappa)$. Then

$$S(\boldsymbol{Q}, \widehat{\boldsymbol{C}}) = \mathbf{0}$$

is the score equation defining $\widehat{\boldsymbol{Q}}$.

Expansion of first-order conditions. With $S(\boldsymbol{Q},\boldsymbol{M})$ polynomial in each of its arguments, an expansion around $\boldsymbol{Q}_0$ and $\boldsymbol{C}$ gives

$$\text{(S.4)}\quad \left.\frac{dS(\boldsymbol{Q}_0,\boldsymbol{M})}{d\boldsymbol{M}}\right|_{\boldsymbol{M}=\boldsymbol{C}} \operatorname{vec}(\widehat{\boldsymbol{M}}-\boldsymbol{M}) + \left.\frac{dS(\boldsymbol{Q},\boldsymbol{C})}{d\boldsymbol{Q}}\right|_{\boldsymbol{Q}=\boldsymbol{Q}_0} \operatorname{vec}(\widehat{\boldsymbol{Q}}-\boldsymbol{Q}_0) = o_p(n^{-1/2}),$$

where

$$\frac{dS(\boldsymbol{Q},\boldsymbol{M})}{d\boldsymbol{M}} = \frac{\partial \operatorname{vec} S(\boldsymbol{Q},\boldsymbol{M})}{\partial \operatorname{vec}(\boldsymbol{M})'}, \qquad \frac{dS(\boldsymbol{Q},\boldsymbol{M})}{d\boldsymbol{Q}} = \frac{\partial \operatorname{vec} S(\boldsymbol{Q},\boldsymbol{M})}{\partial \operatorname{vec}(\boldsymbol{Q})'}.$$

To derive the asymptotic distribution of $\widehat{\boldsymbol{Q}}$ we need to calculate both these derivatives, and evaluate at true values.

Start with the derivative with respect to $\boldsymbol{Q}$. First observe that

$$\frac{d\boldsymbol{Q}^{-1}\boldsymbol{M}\boldsymbol{Q}}{d\boldsymbol{Q}} = \left(\boldsymbol{I}_r \otimes \boldsymbol{Q}^{-1}\boldsymbol{M}\right) - \left((\boldsymbol{Q}^{-1}\boldsymbol{M}\boldsymbol{Q})' \otimes \boldsymbol{Q}^{-1}\right). \tag{S.5}$$

Furthermore, $\operatorname{vec}(\operatorname{off}\boldsymbol{M}) = \operatorname{vec}(\boldsymbol{M} - \operatorname{diag}\boldsymbol{M}) = (\boldsymbol{I}_{r^2} - \boldsymbol{S}_r)\operatorname{vec}(\boldsymbol{M})$ and, by an application of the chain rule,

$$\begin{aligned}\frac{d\Delta(\boldsymbol{Q}^{-1}\boldsymbol{M}\boldsymbol{Q})}{d\boldsymbol{Q}} &= \left\{\operatorname{off}(\boldsymbol{Q}^{-1}\boldsymbol{M}\boldsymbol{Q})' \ominus \operatorname{off}(\boldsymbol{Q}^{-1}\boldsymbol{M}\boldsymbol{Q})\right\} \frac{d\boldsymbol{Q}^{-1}\boldsymbol{M}\boldsymbol{Q}}{d\boldsymbol{Q}} \\ &\quad - \left\{(\boldsymbol{Q}^{-1}\boldsymbol{M}\boldsymbol{Q}) \ominus (\boldsymbol{Q}^{-1}\boldsymbol{M}\boldsymbol{Q})'\right\}\{\boldsymbol{I}_{r^2} - \boldsymbol{S}_r\}\frac{d\boldsymbol{Q}^{-1}\boldsymbol{M}\boldsymbol{Q}}{d\boldsymbol{Q}}.\end{aligned} \tag{S.6}$$

Therefore, combining (S.5) and (S.6), and using that $\boldsymbol{D}_k = \boldsymbol{Q}_0^{-1}\boldsymbol{C}_k\boldsymbol{Q}_0$ and $\operatorname{off}(\boldsymbol{D}_k) = \boldsymbol{0}$, we have

$$\begin{aligned}\left.\frac{d\Delta(\boldsymbol{Q}^{-1}\boldsymbol{C}_k\boldsymbol{Q})}{d\boldsymbol{Q}}\right|_{\boldsymbol{Q}=\boldsymbol{Q}_0} &= (\boldsymbol{D}_k \ominus \boldsymbol{D}_k)(\boldsymbol{I}_{r^2} - \boldsymbol{S}_r)(\boldsymbol{D}_k \ominus \boldsymbol{D}_k)\left(\boldsymbol{I}_r \otimes \boldsymbol{Q}_0^{-1}\right) \\ &= (\boldsymbol{D}_k \ominus \boldsymbol{D}_k)^2\left(\boldsymbol{I}_r \otimes \boldsymbol{Q}_0^{-1}\right)\end{aligned}$$

for all k, where the last transition follows from the fact that $\boldsymbol{S}_r(\boldsymbol{D}_k \ominus \boldsymbol{D}_k) = 0$ because $\boldsymbol{S}_r$ selects only the $\{(ir+(i+1), ir+(i+1))\}_{i=0}^{r-1}$ entries of the $r^2 \times r^2$ matrix $\boldsymbol{D}_k \ominus \boldsymbol{D}_k$, and these are equal to zero. Then

$$\left.\frac{dS(\boldsymbol{Q}, \boldsymbol{C})}{d\boldsymbol{Q}}\right|_{\boldsymbol{Q}=\boldsymbol{Q}_0} = \left(\boldsymbol{I}_r \otimes \boldsymbol{Q}_0^{-1}\right)'\left\{\sum_{k=1}^{\kappa}(\boldsymbol{D}_k \ominus \boldsymbol{D}_k)^2\right\}\left(\boldsymbol{I}_r \otimes \boldsymbol{Q}_0^{-1}\right) \tag{S.7}$$

follows readily.

Now turn to the derivative with respect to $\boldsymbol{M}$. Proceeding in the same way as before, now using that

$$\frac{d\boldsymbol{Q}^{-1}\boldsymbol{M}\boldsymbol{Q}}{d\boldsymbol{M}} = \boldsymbol{Q}' \otimes \boldsymbol{Q}^{-1},$$

we obtain

$$\left.\frac{d\Delta(\boldsymbol{Q}_0^{-1}\boldsymbol{M}\boldsymbol{Q}_0)}{d\boldsymbol{M}}\right|_{\boldsymbol{M}=\boldsymbol{C}_k} = -(\boldsymbol{D}_k \ominus \boldsymbol{D}_k)(\boldsymbol{Q}_0' \otimes \boldsymbol{Q}_0^{-1}).$$

This yields

$$\left.\frac{dS(\boldsymbol{Q}_0,\boldsymbol{M})}{d\boldsymbol{M}_k}\right|_{\boldsymbol{M}_k=\boldsymbol{C}_k} = -(\boldsymbol{I}_r\otimes\boldsymbol{Q}_0^{-1})'(\boldsymbol{D}_k\ominus\boldsymbol{D}_k)(\boldsymbol{Q}_0'\otimes\boldsymbol{I}_r)(\boldsymbol{I}_r\otimes\boldsymbol{Q}_0^{-1})$$

for each k, and concatenating these matrices gives

$$\left.\frac{dS(\boldsymbol{Q}_0,\boldsymbol{M})}{d\boldsymbol{M}}\right|_{\boldsymbol{M}=\boldsymbol{C}} = -(\boldsymbol{I}_r\otimes\boldsymbol{Q}_0^{-1})'\,\boldsymbol{T}\,(\boldsymbol{I}_\kappa\otimes\boldsymbol{Q}_0'\otimes\boldsymbol{Q}_0^{-1}). \tag{S.8}$$

Combining (S.4) with (S.7) and (S.8) then yields

$$\mathrm{vec}(\widehat{\boldsymbol{Q}}-\boldsymbol{Q}_0) = \boldsymbol{G}\;\mathrm{vec}(\widehat{\boldsymbol{C}}-\boldsymbol{C}) + o_p(n^{-1/2}),$$

with matrix $\boldsymbol{G}$ as defined in the main text. This completes the proof of the theorem. □

Proof of Theorem 6. Because we have $\|\widehat{\boldsymbol{C}}_k-\boldsymbol{C}_k\| = O_p(n^{-1/2})$ and $\|\widehat{\boldsymbol{Q}}-\boldsymbol{Q}_0\| = O_p(n^{-1/2})$, a linearization of

$$\widehat{\boldsymbol{D}}_k-\boldsymbol{D}_k = \mathrm{diag}(\widehat{\boldsymbol{Q}}^{-1}\widehat{\boldsymbol{C}}_k\widehat{\boldsymbol{Q}}-\boldsymbol{Q}_0^{-1}\boldsymbol{C}_k\boldsymbol{Q}_0)$$

up to $o_p(n^{-1/2})$ will yield the result. Moreover, the term inside the diagonal operator equals

$$(\widehat{\boldsymbol{Q}}-\boldsymbol{Q}_0)^{-1}\boldsymbol{C}_k\boldsymbol{Q}_0+\boldsymbol{Q}_0^{-1}(\widehat{\boldsymbol{C}}_k-\boldsymbol{C}_k)\boldsymbol{Q}_0+\boldsymbol{Q}_0^{-1}\boldsymbol{C}_k(\widehat{\boldsymbol{Q}}-\boldsymbol{Q}_0)+o_p(n^{-1/2}).$$

Because matrix inversion is a continuous transformation, the delta method can further be applied to yield

$$\mathrm{vec}\left((\widehat{\boldsymbol{Q}}-\boldsymbol{Q}_0)^{-1}\boldsymbol{C}_k\boldsymbol{Q}_0\right) = -\left(\boldsymbol{D}_k\otimes\boldsymbol{Q}_0^{-1}\right)\mathrm{vec}\left(\widehat{\boldsymbol{Q}}-\boldsymbol{Q}_0\right)+o_p(n^{-1/2}).$$

The remaining right-hand side terms are already linear in the estimators $\widehat{\boldsymbol{Q}}$ and $\widehat{\boldsymbol{C}}_k$. Then, using that $\boldsymbol{Q}_0^{-1}\boldsymbol{C}_k = \boldsymbol{D}_k\boldsymbol{Q}_0^{-1}$, $\mathrm{vec}(\widehat{\boldsymbol{Q}}^{-1}\widehat{\boldsymbol{C}}_k\,\widehat{\boldsymbol{Q}}-\boldsymbol{Q}_0^{-1}\boldsymbol{C}_k\boldsymbol{Q}_0)$ equals

$$(\boldsymbol{Q}_0'\otimes\boldsymbol{Q}_0^{-1})\,\mathrm{vec}(\widehat{\boldsymbol{C}}_k-\boldsymbol{C}_k)-(\boldsymbol{D}_k\ominus\boldsymbol{D}_k)(\boldsymbol{I}_r\otimes\boldsymbol{Q}_0^{-1})\,\mathrm{vec}(\widehat{\boldsymbol{Q}}-\boldsymbol{Q}_0),$$

up to $o_p(n^{-1/2})$. Now,

$$\mathrm{vec}(\widehat{\boldsymbol{D}}_k-\boldsymbol{D}_k) = \boldsymbol{S}_r\,\mathrm{vec}(\widehat{\boldsymbol{Q}}^{-1}\widehat{\boldsymbol{C}}_k\,\widehat{\boldsymbol{Q}}-\boldsymbol{Q}_0^{-1}\boldsymbol{C}_k\boldsymbol{Q}_0),$$

thus implying asymptotic linearity of the estimated eigenvalues for each k. Concatenating the influence functions gives

$$\mathrm{vec}(\widehat{\boldsymbol{D}} - \boldsymbol{D}) = \big(\boldsymbol{I}_\kappa \otimes (\boldsymbol{S}_r(\boldsymbol{Q}_0' \otimes \boldsymbol{Q}_0^{-1}))\big)\, \mathrm{vec}(\widehat{\boldsymbol{C}} - \boldsymbol{C}) + o_p(n^{-1/2}).$$

This proves the theorem because, indeed,

$$\boldsymbol{I}_\kappa \otimes (\boldsymbol{S}_r(\boldsymbol{Q}_0' \otimes \boldsymbol{Q}_0^{-1})) = (\boldsymbol{I}_\kappa \otimes \boldsymbol{S}_r)\,(\boldsymbol{I}_\kappa \otimes \boldsymbol{Q}_0' \otimes \boldsymbol{Q}_0^{-1}) = \boldsymbol{H},$$

as claimed. □

S.6. Omitted proofs for the theorems in Section 5. We provide detailed derivations for the orthogonal-series estimator under the conditions collected under **A.1**–**A.2** and **C.1**–**C.3**.

The corresponding results under Assumptions **B.1** and **C.1**–**C.3** follow by the same arguments. The bounds on Hermite polynomials in Walter [1977] allow to bypass the integrability requirements in **A.2** used in the proofs below. Further, when establishing asymptotic normality for Hermite series based on the proof of Theorem 8 given below, Theorem 3 in Liebsher [1990] can be used to replace Theorem 2.2.3 in Viollaz [1989] when used to justify (S.13).

Proof of Theorem 7. For the proof it suffices to consider the case with $q = 3$. Without loss of generality we fix $i = 3$ throughout. As in the proof of Theorem 1,

$$\boldsymbol{A}_0 = \boldsymbol{B}_1 \boldsymbol{\Pi} \boldsymbol{B}_2', \qquad \boldsymbol{A}_k = \boldsymbol{B}_1 \boldsymbol{\Pi} \boldsymbol{D}_k \boldsymbol{B}_2', \qquad \boldsymbol{D}_k = \mathrm{diag}_k\, \boldsymbol{B}_3. \tag{S.9}$$

The Fourier coefficients are then estimated by solving the sample version of

$$\boldsymbol{C}_k = \boldsymbol{W}_1 \boldsymbol{A}_k \boldsymbol{W}_2' = \boldsymbol{Q} \boldsymbol{D}_k \boldsymbol{Q}^{-1}. \tag{S.10}$$

The proof consists of two steps. We first derive integrated squared-error and uniform convergence rates for the infeasible estimator that assumes that the matrices $\boldsymbol{Q}$ and $\boldsymbol{W}_1, \boldsymbol{W}_2$ are observable without noise. That is, for the estimator

$$\tilde{f}_{ij} = \sum_{k=1}^{\varkappa} \tilde{b}_{ijk}\, \varphi_k \tag{S.11}$$

where the $\widetilde{b}_{ijk}$ are constructed from $\widetilde{\boldsymbol{D}}_k = \operatorname{diag}[(\boldsymbol{Q}^{-1}\boldsymbol{W}_1)\,\widehat{\boldsymbol{A}}_k\,(\boldsymbol{W}_2'\boldsymbol{Q})]$. We then show that the additional noise in our (feasible) estimator,

$$\hat{f}_{ij} = \sum_{k=1}^{\varkappa} \hat{b}_{ijk}\,\varphi_k$$

that is, the one that uses $\widehat{\boldsymbol{D}}_k = \operatorname{diag}[(\widehat{\boldsymbol{Q}}^{-1}\widehat{\boldsymbol{W}}_1)\,\widehat{\boldsymbol{A}}_k\,(\widehat{\boldsymbol{W}_2'\boldsymbol{Q}})]$, is asymptotically negligible. We will write $\boldsymbol{\beta}_{ij} = (b_{ij1}, b_{ij2}, \ldots, b_{ij\varkappa})'$ and denote its feasible and infeasible estimator by $\widehat{\boldsymbol{\beta}}_{ij}$ and $\widetilde{\boldsymbol{\beta}}_{ij}$, respectively.

We begin by showing that

$$\|\widetilde{\boldsymbol{\beta}}_{ij} - \boldsymbol{\beta}_{ij}\|_F = O_p(\sqrt{\varkappa/n}).$$

The convergence rates for $\tilde{f}_{ij}$ will then follow easily. Write $a_{k_1k_2k}$ for the (k_1, k_2)th entry of $\boldsymbol{A}_k$ and let $\hat{a}_{k_1k_2k}$ be its estimator. Note that

$$\hat{a}_{k_1k_2k} = \frac{1}{n}\sum_{m=1}^{n} \varphi_{k_1}(Y_{1m})\rho(Y_{1m})\varphi_{k_2}(Y_{2m})\rho(Y_{2m})\varphi_k(Y_{3m})\rho(Y_{3m})$$

is an unbiased estimator of $a_{k_1k_2k}$. Hence, for any k,

$$\begin{aligned} E\big\|\widehat{\boldsymbol{A}}_k - \boldsymbol{A}_k\big\|_F^2 &= \sum_{k_1=1}^{\kappa_1}\sum_{k_2=1}^{\kappa_2} E\big[(\hat{a}_{k_1k_2k} - a_{k_1k_2k})^2\big] \\ &= \sum_{k_1=1}^{\kappa_1}\sum_{k_2=1}^{\kappa_2} \frac{E\big[\varphi_{k_1}(Y_1)^2\rho(Y_1)^2\varphi_{k_2}(Y_2)^2\rho(Y_2)^2\varphi_k(Y_3)^2\rho(Y_3)^2\big] - a_{k_1k_2k}^2}{n} \\ &\leq \sum_{k_1=1}^{\kappa_1}\sum_{k_2=1}^{\kappa_2} \frac{\sum_{j=1}^{r}\prod_{i'=1}^{q}\pi_j\,\big(\int_{-1}^{1}\psi(y)^2\rho(y)^2 f_{i'j}(y)\,\mathrm{d}y\big) - a_{k_1k_2k}^2}{n}. \end{aligned}$$

As the $\psi^2\rho^2 f_{i'j}$ are integrable and the Fourier coefficients $a_{k_1k_2k}$ are square summable, we have that $E\|\widehat{\boldsymbol{A}}_k - \boldsymbol{A}_k\|_F^2 = O(1/n)$ uniformly in k. Therefore, $\sum_{k=1}^{\varkappa}\|\widehat{\boldsymbol{A}}_k - \boldsymbol{A}_k\|_F^2 = O_p(\varkappa/n)$ follows from Markov's inequality, and so also

$$\begin{aligned} \|\widetilde{\boldsymbol{\beta}}_{ij} - \boldsymbol{\beta}_{ij}\|_F^2 &\leq \sum_{k=1}^{\varkappa}\big\|\widetilde{\boldsymbol{D}}_k - \boldsymbol{D}_k\big\|_F^2 \\ &\leq \big\|\boldsymbol{Q}^{-1}\boldsymbol{W}_1 \otimes \boldsymbol{Q}'\boldsymbol{W}_2\big\|_F^2\;\sum_{k=1}^{\varkappa}\big\|\widehat{\boldsymbol{A}}_k - \boldsymbol{A}_k\big\|_F^2 = O_p(\varkappa/n) \end{aligned}$$

follows by the Cauchy-Schwarz inequality. This establishes the rate result on the Fourier coefficients sought for. Now turn to the convergence rates for $\widetilde{f}_{ij}$. By orthonormality of the φ_k,

$$\begin{aligned}\big\|\tilde{f}_{ij} - f_{ij}\big\|_2^2 &= \big\|\tilde{f}_{ij} - \operatorname{Proj}_\varkappa f_{ij}\big\|_2^2 + \big\|\operatorname{Proj}_\varkappa f_{ij} - f_{ij}\big\|_2^2 \\ &= \|\widetilde{\boldsymbol{\beta}}_{ij} - \boldsymbol{\beta}_{ij}\|_F^2 + \big\|\operatorname{Proj}_\varkappa f_{ij} - f_{ij}\big\|_2^2.\end{aligned}$$

The first right-hand side term is known to be $O_p(\varkappa/n)$ from above. For the second right-hand side term, by Assumption **C.2**,

$$\big\|\operatorname{Proj}_\varkappa f_{ij} - f_{ij}\big\|_2^2 \leq \int_{-1}^{1} \big\|\operatorname{Proj}_\varkappa f_{ij} - f_{ij}\big\|_\infty^2 \, \rho(y)\,\mathrm{d}y = O(\varkappa^{-2\beta})$$

because ρ is integrable. This establishes the integrated squared-error rate for $\widetilde{f}_{ij}$. To obtain the uniform convergence rate, use the triangle inequality to see that

$$\big\|\tilde{f}_{ij} - f_{ij}\big\|_\infty \leq \big\|\tilde{f}_{ij} - \operatorname{Proj}_\varkappa f_{ij}\big\|_\infty + \big\|\operatorname{Proj}_\varkappa f_{ij} - f_{ij}\big\|_\infty.$$

By the Cauchy-Schwarz inequality in the first step and by the uniform bound on the norm of the basis functions and the convergence rate of $\|\widetilde{\boldsymbol{\beta}}_{ij} - \boldsymbol{\beta}_{ij}\|_F$ in the second, the first right-hand side term satisfies

$$\big\|\tilde{f}_{ij} - \operatorname{Proj}_\varkappa f_{ij}\big\|_\infty \leq \Big\|\sqrt{\boldsymbol{\varphi}_\varkappa' \boldsymbol{\varphi}_\varkappa}\Big\|_\infty \, \|\widetilde{\boldsymbol{\beta}}_{ij} - \boldsymbol{\beta}_{ij}\|_F = O\big(\zeta_\varkappa\big)\; O_p\big(\sqrt{\varkappa/n}\big).$$

By Assumption **C.2**, $\|\operatorname{Proj}_\varkappa f_{ij} - f_{ij}\|_\infty = O(\varkappa^{-\beta})$. This yields the uniform convergence rate.

To extend the results to the feasible density estimator $\hat{f}_{ij}$ we first show that the presence of estimation noise in $\boldsymbol{Q}$ and $(\boldsymbol{W}_1, \boldsymbol{W}_2)$ implies that

$$\|\widehat{\boldsymbol{\beta}}_{ij} - \widetilde{\boldsymbol{\beta}}_{ij}\|_F = O_p(n^{-1/2}) + O_p(\sqrt{\varkappa}/n). \tag{S.12}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}\big\|\widehat{\boldsymbol{\beta}}_{ij} - \widetilde{\boldsymbol{\beta}}_{ij}\big\|_F^2 &\leq \sum_{k=1}^{\varkappa} \big\|\widehat{\boldsymbol{D}}_k - \widetilde{\boldsymbol{D}}_k\big\|_F^2 \\ &\leq \big\|\widehat{\boldsymbol{Q}}^{-1}\widehat{\boldsymbol{W}}_1 \otimes \widehat{\boldsymbol{Q}}'\widehat{\boldsymbol{W}}_2 - \boldsymbol{Q}^{-1}\boldsymbol{W}_1 \otimes \boldsymbol{Q}'\boldsymbol{W}_2\big\|_F^2 \; \sum_{k=1}^{\varkappa} \big\|\widehat{\boldsymbol{A}}_k\big\|_F^2.\end{aligned}$$

Because both $\widehat{\boldsymbol{Q}}$ and $(\widehat{\boldsymbol{W}}_1, \widehat{\boldsymbol{W}}_2)$ are $\sqrt{n}$-consistent,

$$\big\|\widehat{\boldsymbol{Q}}^{-1}\widehat{\boldsymbol{W}}_1 \otimes \widehat{\boldsymbol{Q}}'\widehat{\boldsymbol{W}}_2 - \boldsymbol{Q}^{-1}\boldsymbol{W}_1 \otimes \boldsymbol{Q}'\boldsymbol{W}_2\big\|_F^2 = O_p(1/n).$$

Also, from above, we have that

$$\sum_{k=1}^{\varkappa} \big\|\widehat{\boldsymbol{A}}_k\big\|_F^2 \leq 2\sum_{k=1}^{\varkappa} \big\|\boldsymbol{A}_k\big\|_F^2 + 2\sum_{k=1}^{\varkappa} \big\|\widehat{\boldsymbol{A}}_k - \boldsymbol{A}_k\big\|_F^2 = O(1) + O_p(\varkappa/n).$$

Together, these results imply (S.12). Next,

$$\|\hat{f}_{ij} - f_{ij}\|_2^2 \leq 2\|\widehat{\boldsymbol{\beta}}_{ij} - \widetilde{\boldsymbol{\beta}}_{ij}\|_F^2 + 2\|\tilde{f}_{ij} - f_{ij}\|_2^2.$$

From above, the first right-hand side term is $O_p(1/n) + O_p(\varkappa/n^2)$ while the second right-hand side term is $O_p(\varkappa/n + \varkappa^{-2\beta})$. Therefore, the difference between $\widehat{\boldsymbol{\beta}}_{ij}$ and $\widetilde{\boldsymbol{\beta}}_{ij}$ has an asymptotically-negligible impact on the density estimator, and

$$\|\hat{f}_{ij} - f_{ij}\|_2^2 = O_p(\varkappa/n + \varkappa^{-2\beta}).$$

For the uniform-convergence result, similarly, the triangle inequality gives the bound

$$\|\hat{f}_{ij} - f_{ij}\|_\infty \leq \|\hat{f}_{ij} - \tilde{f}_{ij}\|_\infty + \|\tilde{f}_{ij} - f_{ij}\|_\infty.$$

Again,

$$\|\hat{f}_{ij} - \tilde{f}_{ij}\|_\infty \leq \left\|\sqrt{\boldsymbol{\varphi}_\varkappa'\boldsymbol{\varphi}_\varkappa}\right\|_\infty \big\|\widehat{\boldsymbol{\beta}}_{ij} - \widetilde{\boldsymbol{\beta}}_{ij}\big\|_F = O_p(\zeta_\varkappa/\sqrt{n}) + O_p(\zeta_\varkappa\sqrt{\varkappa}/n),$$

which is of a smaller stochastic order than is $\|\tilde{f}_{ij} - f_{ij}\|_\infty$. This concludes the proof. □

PROOF OF THEOREM 8. We proceed in two steps. First, we derive the asymptotic distribution of the infeasible estimator $\tilde{f}_{ij}$ defined in (S.11) above. Next we show that $\hat{f}_{ij} - \tilde{f}_{ij}$ is asymptotically negligible. Like in the proof of Theorem 7 it again suffices to consider the case with $q = 3$. Throughout, we fix $i = 3$ and omit i as subscript when there is no risk of confusion.

To analyze $\tilde{f}_{ij}$ we will make extensive use of the fact that it can be written as

$$\tilde{f}_{ij}(y) = n^{-1}\sum_{m=1}^{n} \boldsymbol{e}_j'\boldsymbol{\Omega}_m\boldsymbol{e}_j \sum_{k=1}^{\varkappa} \varphi_k(Y_{im})\varphi_k(y)\rho(Y_{im})$$

where we define the $\kappa_1 \times \kappa_2$ matrix

$$\boldsymbol{\Omega}_m = \boldsymbol{Q}^{-1}(\boldsymbol{W}_1\,(\boldsymbol{\varphi}_{\kappa_1}(Y_{1m})\rho(Y_{1m}) \otimes \boldsymbol{\varphi}_{\kappa_2}(Y_{2m})\rho(Y_{2m}))\,\boldsymbol{W}_2')\,\boldsymbol{Q}.$$

This follows in the same way as did (5.2) in the main text. Indeed, on replacing $\boldsymbol{\Omega}_m$ by its sample counterpart $\widehat{\boldsymbol{\Omega}}_m$, we obtain $\hat{f}_{ij}(y)$. We will also use the notational shorthand

$$\phi_{\varkappa}(a,b) = \boldsymbol{\varphi}_{\varkappa}(a)'\boldsymbol{\varphi}_{\varkappa}(b) = \sum_{k=1}^{\varkappa} \varphi_k(a)\varphi_k(b),$$

for the Christoffel-Darboux kernel associated with the polynomial system $\{\varphi_1, \varphi_2, \ldots, \varphi_{\varkappa}\}$.

We first show that $\tilde{f}_{ij}(y)$ is an unbiased estimator of the projection $\mathrm{Proj}_{\varkappa} f_{ij}(y)$ or, equivalently, that the associated estimator of the Fourier coefficients is unbiased. Because the outcomes are independent conditional on realizations of Z,

$$\begin{aligned} E\big[\tilde{f}_{ij}(y)\big] &= E\big[\boldsymbol{e}_j'\boldsymbol{\Omega}_m\boldsymbol{e}_j\,\phi_{\varkappa}(Y_{3m},y)\rho(Y_{3m})\big] \\ &= \sum_{j'=1}^{r} E\big[\boldsymbol{e}_j'\boldsymbol{\Omega}_m\boldsymbol{e}_j\big|Z=j'\big]\,E\big[\phi_{\varkappa}(Y_{3m},y)\rho(Y_{3m})\big|Z=j'\big]\ \pi_{j'}. \end{aligned}$$

As is well-known, for each $j' \in \{1,2,\ldots,r\}$, the Christoffel-Darboux kernel satisfies

$$E\big[\phi_{\varkappa}(Y_{im},y)\rho(Y_{im})\big|Z=j'\big] = \mathrm{Proj}_{\varkappa} f_{ij'}(y).$$

Furthermore, because, again by using conditional independence, we also have

$$E[\boldsymbol{\varphi}_{\kappa_1}(Y_{1m})\rho(Y_{1m}) \otimes \boldsymbol{\varphi}_{\kappa_2}(Y_{2m})\rho(Y_{2m})|Z=j'] = \boldsymbol{b}_{1j'} \otimes \boldsymbol{b}_{2j'},$$

we find that

$$E\big[\boldsymbol{e}_j'\boldsymbol{\Omega}_m\boldsymbol{e}_j\big|Z=j'\big] = \boldsymbol{e}_j'\boldsymbol{Q}^{-1}\boldsymbol{W}_1\,(\boldsymbol{b}_{1j'} \otimes \boldsymbol{b}_{2j'})\,\boldsymbol{W}_2'\boldsymbol{Q}\boldsymbol{e}_j = \pi_j^{-1}\delta_{jj'},$$

where, recall, $\delta_{jj'}$ is Kronecker's delta. The last transition follows by the equalities in (S.9)–(S.10) as they imply that $\boldsymbol{Q}^{-1}\boldsymbol{W}_1\boldsymbol{B}_1\boldsymbol{B}_2'\boldsymbol{W}_2'\boldsymbol{Q} = \boldsymbol{\Pi}^{-1}$. Therefore,

$$E\big[\tilde{f}_{ij}(y)\big] = \sum_{j'=1}^{r} \delta_{jj'}\,\frac{\pi_{j'}}{\pi_j}\ \mathrm{Proj}_{\varkappa} f_{ij'}(y) = \mathrm{Proj}_{\varkappa} f_{ij}(y),$$

as claimed.

Centering the estimator $\tilde{f}_{ij}(y)$ around its expectation gives

$$\tilde{f}_{ij}(y) - \operatorname{Proj}_{\varkappa} f_{ij}(y) = n^{-1} \sum_{m=1}^{n} \psi_m,$$

where we let

$$\psi_m = \boldsymbol{e}_j' \boldsymbol{\Omega}_m \boldsymbol{e}_j \, \phi_{\varkappa}(Y_{3m}, y) \rho(Y_{3m}) - \operatorname{Proj}_{\varkappa} f_{ij}(y).$$

Because we have that $|\operatorname{Proj}_{\varkappa} f_{ij}(y) - f_{ij}(y)| \leq \|\operatorname{Proj}_{\varkappa} f_{ij} - f_{ij}\|_{\infty} = O(\varkappa^{-\beta})$ by Assumption **C.2** and we require that $\sqrt{n}\varkappa^{-\beta} \to 0$, the bias induced by truncating the projection is asymptotically negligible. It thus suffices to derive the limit distribution of the sample average of the ψ_m. For this we verify that the conditions of Lyapunov central limit theorem for triangular arrays are satisfied. We have already demonstrated that $E[\psi_m] = 0$ and so, if we can show that

$$\text{(i) } E\left[\frac{\psi_m^2}{\phi_{\varkappa}(y,y)}\right] = O(1); \qquad \text{(ii) } E\left[\left(\frac{\psi_m^2}{\operatorname{var}[\psi_m]}\right)^2\right] = o(n),$$

the result will be proven.

To show Condition (i), first use conditional independence of the measurements to see that

$$E\left[\psi_m^2\right] = \sum_{j'=1}^{r} E\left[(\boldsymbol{e}_j' \boldsymbol{\Omega}_m \boldsymbol{e}_j)^2 | Z = j'\right] \; E\left[\phi_{\varkappa}(Y_{3m}, y)^2 \rho(Y_{3m})^2 | Z = j'\right] \; \pi_{j'}.$$

Again exploiting conditional independence, a direct calculation shows that, for each j',

$$\mathbb{E}[(\boldsymbol{e}_j' \boldsymbol{\Omega}_m \boldsymbol{e}_j)^2 | Z = j'] = O(1),$$

with boundedness following from the fact that the $\boldsymbol{Q}^{-1}\boldsymbol{W}_1$ and $\boldsymbol{W}_2'\boldsymbol{Q}$ are $O(1)$ and that, for each $k_1 \in \{1, 2, \ldots, \kappa_1\}$ and $k_2 \in \{1, 2, \ldots, \kappa_2\}$ we have that

$$E[\varphi_{k_1}(Y_{i'm})\rho(Y_{i'm})\varphi_{k_2}(Y_{i'm})\rho(Y_{i'm}) | Z = j'] \leq \int_{-1}^{1} \psi(u)^2 \rho(u)^2 f_{i'j'}(u) \mathrm{d}u,$$

which is $O(1)$ for all i', j' because $\psi^2\rho^2 f_{i'j'}$ is integrable. Next, under our maintained conditions, we can apply Theorem 2.2.3 in Viollaz [1989] to get

$$\frac{E\left[\phi_{\varkappa}(Y_{3m}, y)^2 \rho(Y_{3m})^2 | Z = j'\right]}{\phi_{\varkappa}(y,y)} \longrightarrow f_{ij'}(y)\rho(y). \tag{S.13}$$

which exists for all j'. Finally, as

$$\mathrm{var}[\psi_m] = E[\psi_m^2] - (\mathrm{Proj}_\varkappa \, f_{ij}(y))^2 = E[\psi_m^2] + o(\phi_\varkappa(y,y)),$$

we have that $\mathrm{var}[\psi_m]/\phi_\varkappa(y,y)$ tends to a positive constant, and so Condition (i) is satisfied.

To verify Condition (ii), introduce

$$\boldsymbol{X}_{mk} = (\boldsymbol{\varphi}_{\kappa_1}(Y_{1m})\rho(Y_{1m}) \otimes \boldsymbol{\varphi}_{\kappa_2}(Y_{2m})\rho(Y_{2m}))\, \varphi_k(Y_{3m})\rho(Y_{3m}),$$

which is a $\kappa_1 \times \kappa_2$ matrix. As

$$\widehat{\boldsymbol{A}}_k = n^{-1} \sum_{m=1}^{n} \boldsymbol{X}_{mk},$$

we have

$$\begin{aligned}\psi_m &= \sum_{k=1}^{\varkappa} \big(\boldsymbol{e}_j' \boldsymbol{Q}^{-1} \boldsymbol{W}_1 (\boldsymbol{X}_{mk} - \boldsymbol{A}_k) \boldsymbol{W}_2' \boldsymbol{Q} \boldsymbol{e}_j\big) \varphi_k(y) \\ &= \sum_{k=1}^{\varkappa} \mathrm{trace}\,\big(\big(\boldsymbol{W}_2' \boldsymbol{Q} \boldsymbol{e}_j \boldsymbol{e}_j' \boldsymbol{Q}^{-1} \boldsymbol{W}_1\big)\, (\boldsymbol{X}_{mk} - \boldsymbol{A}_k) \,\big) \varphi_k(y).\end{aligned}$$

By repeatedly applying the Cauchy-Schwarz inequality to this expression we then establish

$$\psi_m^2 \leq \left\| \boldsymbol{W}_2' \boldsymbol{Q} \boldsymbol{e}_j \boldsymbol{e}_j' \boldsymbol{Q}^{-1} \boldsymbol{W}_1 \right\|^2 \, \sum_{k=1}^{\varkappa} \| \boldsymbol{X}_{mk} - \boldsymbol{A}_k \|^2 \; \phi_\varkappa(y,y).$$

Because $\| \boldsymbol{W}_2' \boldsymbol{Q} \boldsymbol{e}_j \boldsymbol{e}_j' \boldsymbol{Q}^{-1} \boldsymbol{W}_1 \| = O(1)$ and $\mathrm{var}[\psi_m] \asymp \phi_\varkappa(y,y)$ we then obtain

$$\begin{aligned} E\left[\left(\frac{\psi_m^2}{\mathrm{var}[\psi_m]} \right)^2 \right] &\leq O(1) \; E\Big[\Big(\sum_{k=1}^{\varkappa} \| \boldsymbol{X}_{mk} - \boldsymbol{A}_k \|^2 \Big)^2 \Big] \\ &\leq O(\varkappa) \; \sum_{k=1}^{\varkappa} E \, \| \boldsymbol{X}_{mk} - \boldsymbol{A}_k \|^4 , \end{aligned}$$

where the last transition follows by the Cauchy-Schwarz inequality. Then, using similar arguments as in the proof of Theorem 7, it is straightforward to show that $E \, \| \boldsymbol{X}_{mk} - \boldsymbol{A}_k \|^4 = O(1)$ uniformly in k, because $\pi^4 \rho^4 f_{ij}$ is integrable. Therefore,

$$\frac{1}{n} E \left[\left(\frac{\psi_m^2}{\mathrm{var}[\psi_m]} \right)^2 \right] = O\left(\frac{\varkappa^2}{n} \right),$$

which converges to zero as $n \to \infty$. This shows that Condition (ii) is satisfied. With the requirements of the central limit theorem verified, it follows that

$$\frac{1}{\sqrt{n}} \sum_{m=1}^{n} \frac{\psi_m}{\sqrt{\operatorname{var}[\psi_m]}} \xrightarrow{d} \mathcal{N}(0,1).$$

It is easy to see that $\hat{\sigma}_{ij}(y)$ in the text is a consistent estimator of $\sqrt{\operatorname{var}[\psi_m]}$ and so the theorem has been proven to hold for $\tilde{f}_{ij}$. It only remains to extend the result to the feasible estimator.

We will show that

$$\left|\hat{f}_{ij}(y) - \tilde{f}_{ij}(y)\right| = o_p\left(\sqrt{\frac{\phi_{\varkappa}(y,y)}{n}}\right), \tag{S.14}$$

which implies that the additional noise in $\hat{f}_{ij}$ is asymptotically negligible. To do so, recall that

$$\hat{f}_{ij}(y) = n^{-1} \sum_{m=1}^{n} \boldsymbol{e}_j' \widehat{\boldsymbol{\Omega}}_m \boldsymbol{e}_j \sum_{k=1}^{\varkappa} \varphi_k(Y_{im})\varphi_k(y)\rho(Y_{im}),$$

where $\widehat{\boldsymbol{\Omega}}_m$ differs from $\boldsymbol{\Omega}_m$ in that it uses $\widehat{\boldsymbol{Q}}$ and $\widehat{\boldsymbol{W}}_1, \widehat{\boldsymbol{W}}_2$ rather than $\boldsymbol{Q}$ and $\boldsymbol{W}_1, \boldsymbol{W}_2$.

Let $\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_r)$ be so that $\boldsymbol{R}\boldsymbol{Q}' = \boldsymbol{I}_r$, so $\boldsymbol{R}' = \boldsymbol{Q}^{-1}$ and also define

$$\boldsymbol{U}_m = \boldsymbol{\varphi}_{\kappa_1}(Y_{1m})\rho(Y_{1m}) \otimes \boldsymbol{\varphi}_{\kappa_2}(Y_{2m})\rho(Y_{2m});$$

let $u_{k_1k_2m}$ be its (k_1,k_2)th entry. Indeed, $\widehat{\boldsymbol{A}}_0 = n^{-1}\sum_{m=1}^{n} \boldsymbol{U}_m$. This allows us to write

$$\boldsymbol{e}_j' \widehat{\boldsymbol{\Omega}}_m \boldsymbol{e}_j - \boldsymbol{e}_j' \boldsymbol{\Omega}_m \boldsymbol{e}_j = \sum_{k_1=1}^{\kappa_1} \sum_{k_2=1}^{\kappa_2} \left(\widehat{\boldsymbol{r}}_j' \widehat{\boldsymbol{w}}_{1k_1} \widehat{\boldsymbol{q}}_j' \widehat{\boldsymbol{w}}_{2k_2} - \boldsymbol{r}_j' \boldsymbol{w}_{1k_1} \boldsymbol{q}_j' \boldsymbol{w}_{2k_2}\right) u_{k_1k_2m},$$

and to get the bound

$$\begin{aligned}\left|\hat{f}_{ij}(y) - \tilde{f}_{ij}(y)\right| \leq \sum_{k_1=1}^{\kappa_1} \sum_{k_2=1}^{\kappa_2} &\left|\widehat{\boldsymbol{r}}_j' \widehat{\boldsymbol{w}}_{1k_1} \widehat{\boldsymbol{q}}_j' \widehat{\boldsymbol{w}}_{2k_2} - \boldsymbol{r}_j' \boldsymbol{w}_{1k_1} \boldsymbol{q}_j' \boldsymbol{w}_{2k_2}\right| \\ &\times \left|\frac{1}{n} \sum_{m=1}^{n} u_{k_1k_2m}\ \phi_{\varkappa}(Y_{3m}, y)\rho(Y_{3m})\right|.\end{aligned}$$

We will handle each of the terms on the right-hand side in turn. First, by Theorem 5 and by Assumption **C.3**,

$$|\widehat{\boldsymbol{r}}_j'\widehat{\boldsymbol{w}}_{1k_1}\widehat{\boldsymbol{q}}_j'\widehat{\boldsymbol{w}}_{2k_2} - \boldsymbol{r}_j'\boldsymbol{w}_{1k_1}\boldsymbol{q}_j'\boldsymbol{w}_{2k_2}| = O_p(1/\sqrt{n}) \tag{S.15}$$

for all (k_1, k_2), and so this first term converges at the parametric rate. For the second term, $E[u_{k_1k_2m}\ \phi_\varkappa(Y_{3m}, y)\rho(Y_{3m})] = O(1)$ is easily verified while

$$\frac{\text{var}[u_{k_1k_2m}\ \phi_\varkappa(Y_{3m}, y)\rho(Y_{3m})]}{\phi_\varkappa(y,y)} = O(1)$$

follows from the same arguments as those that were used to establish the asymptotic distribution of the infeasible density estimator. Therefore, for all (k_1, k_2),

$$\Big|\frac{1}{n}\sum_{m=1}^n u_{k_1k_2m}\ \phi_\varkappa(Y_{3m}, y)\rho(Y_{3m})\Big| = O_p(\sqrt{\phi_\varkappa(y,y)/n}). \tag{S.16}$$

Combining (S.15) and (S.16) then gives

$$|\hat{f}_{ij}(y) - \tilde{f}_{ij}(y)| = O_p\left(\frac{\sqrt{\phi_\varkappa(y,y)}}{n}\right),$$

which implies (S.14). This completes the proof. □

S.7. Cross-validation for the orthogonal-series estimator. We fix indices i, j throughout and consider choosing the integer $\varkappa$—that is, the number of series terms to include in the expansion—in the orthogonal-series estimator

$$\begin{aligned}\hat{f}_{ij}(y) &= \sum_{k=1}^{\varkappa} \hat{b}_{ijk}\,\varphi_k(y) \\ &= n^{-1}\sum_{m=1}^n \boldsymbol{e}_j'\widehat{\boldsymbol{\Omega}}_m\boldsymbol{e}_j \ \sum_{k=1}^{\varkappa}\varphi_k(Y_{im})\varphi_k(y)\rho(Y_{im})\end{aligned}$$

as to minimize the squared L^2_ρ-loss

$$\|\hat{f}_{ij} - f_{ij}\|_2^2 = \int_{\mathscr{Y}} \big(\hat{f}_{ij}(x) - f_{ij}(x)\big)^2\rho(x)\,\mathrm{d}x.$$

Expanding the square and collecting terms that do not depend on $\varkappa$ gives

$$\|\hat{f}_{ij} - f_{ij}\|_2^2 = \int_{\mathscr{Y}} \hat{f}_{ij}(x)^2 \rho(x)\,\mathrm{d}x - 2\int_{\mathscr{Y}} \hat{f}_{ij}(x)\, f_{ij}(x)\rho(x)\,\mathrm{d}x + \text{const.} \tag{S.17}$$

Now,

$$\begin{aligned}\hat{f}_{ij}(x)^2 &= \left(n^{-1}\sum_{m=1}^{n} \boldsymbol{e}_j'\widehat{\boldsymbol{\Omega}}_m\boldsymbol{e}_j \ \sum_{k=1}^{\varkappa}\varphi_k(Y_{im})\varphi_k(x)\rho(Y_{im})\right)^2 \\ &= n^{-2}\sum_{m=1}^{n}\sum_{o=1}^{n} \boldsymbol{e}_j'\widehat{\boldsymbol{\Omega}}_m\boldsymbol{e}_j\boldsymbol{e}_j'\widehat{\boldsymbol{\Omega}}_o\boldsymbol{e}_j \\ &\quad\times \sum_{k_1=1}^{\varkappa}\sum_{k_2=1}^{\varkappa}\varphi_{k_1}(Y_{im})\varphi_{k_2}(Y_{io})\rho(Y_{im})\rho(Y_{io})\,\varphi_{k_1}(x)\varphi_{k_2}(x).\end{aligned}$$

Because the functions $\{\varphi_k\}$ are orthogonal with respect to the function ρ, integrating $\hat{f}_{ij}^2\rho$ gives

$$\int_{\mathscr{Y}} \hat{f}_{ij}(x)^2\rho(x)\,\mathrm{d}x = \sum_{k=1}^{\varkappa}\left(n^{-1}\sum_{m=1}^{n}\boldsymbol{e}_j'\widehat{\boldsymbol{\Omega}}_m\boldsymbol{e}_j\varphi_k(Y_{im})\rho(Y_{im})\right)^2 = \sum_{k=1}^{\varkappa}\hat{b}_{ijk}^2.$$

For the second right-hand side term in (S.17), a small calculation allows to establish that

$$\int_{\mathscr{Y}} \hat{f}_{ij}(x)\, f_{ij}(x)\rho(x)\,\mathrm{d}x = E\left[\boldsymbol{e}_j'\boldsymbol{\Omega}_m\boldsymbol{e}_j\,\hat{f}_{ij}(Y_{im})\rho(Y_{im})\right],$$

where the expectation is taken with respect to the random variables $\boldsymbol{\Omega}_m$ and Y_{im}, and $\boldsymbol{\Omega}_m$ is the population version of $\widehat{\boldsymbol{\Omega}}_m$ as introduced in the proof of Theorem 8. A plug-in estimator is

$$n^{-1}\sum_{m=1}^{n}\boldsymbol{e}_j'\widehat{\boldsymbol{\Omega}}_m\boldsymbol{e}_j\,\hat{f}_{ij}^{-m}(Y_{im})\rho(Y_{im}),$$

where

$$\hat{f}_{ij}^{-o}(Y_{io}) = (n-1)^{-1}\sum_{m\neq o}\boldsymbol{e}_j'\widehat{\boldsymbol{\Omega}}_m\boldsymbol{e}_j \ \sum_{k=1}^{\varkappa}\varphi_k(Y_{im})\varphi_k(Y_{io})\rho(Y_{im})$$

is the leave-one-out density estimator at Y_{io}.

Put together, a feasible version of (S.17) is the cross-validation criterion

$$\sum_{k=1}^{\varkappa} \hat{b}_{ijk}^2 - 2n^{-1} \sum_{m=1}^{n} \boldsymbol{e}_j' \widehat{\boldsymbol{\Omega}}_m \boldsymbol{e}_j \, \hat{f}_{ij}^{-m}(Y_{im}) \rho(Y_{im}),$$

which, on plugging in the functional form for the leave-one-out estimator, gives the minimand stated in the main text.

This cross-validation approach gives an automated way of choosing $\varkappa$ given $\{\kappa_i\}_{i=1}^q$, that is, given the size of the array $\mathbb{B}$ used to compute the joint diagonalizer $\boldsymbol{Q}$. For the large-sample theory of Theorems 7 and 8 to apply to the orthogonal-series estimator we only need these coefficients to be so that the rank conditions in Theorem 1 are satisfied. Bonhomme, Jochmans and Robin [2014] discuss tests for this in greater detail in a slightly different context; their approach and conclusions carry over to the current setting with obvious modification.

In contrast to $\varkappa$, which grows with n, the values of $\kappa_1, \kappa_2, \ldots, \kappa_q$ are held fixed. The optimal choice of κ_i and $\{\kappa_{i'}\}_{i' \neq i}$ is an interesting question that requires additional attention in future research. Regarding κ_i we note that the conclusions of Theorem 7 and Theorem 8 continue to hold if κ_i grows with n, provided that $\kappa_i = o(\varkappa)$ (Theorem 7) and $\kappa_i = o(\phi_{\varkappa}(y,y))$ (Theorem 8). It is not clear, however, whether there are (asymptotic) efficiency gains to be obtained by letting κ_i grow large. Indeed, the asymptotic distribution of the series estimator does not depend on estimation noise in $\widehat{\boldsymbol{Q}}$, and so would not be a function of κ_i. Furthermore, Monte Carlo experiments suggest the density estimator to be fairly insensitive to the number of matrices that are diagonalized.

The small-sample behavior of orthogonal-series estimator as a function of $\{\kappa_{i'}\}_{i' \neq i}$ is a different issue. They influence the pointwise asymptotic distribution of the series estimator, through the dependence of $\boldsymbol{\Omega}_m$ on the $\boldsymbol{\varphi}_{\kappa_{i'}}$.

S.8. Additional simulation results.

S.8.1. *Mixing proportions.* Here we provide results for a Monte Carlo simulation for the minimum-distance estimator of the mixing proportions in the designs of Section 6.1.

The estimator has the form

$$\widehat{\boldsymbol{\pi}}(i) = \left(\widehat{\boldsymbol{B}}_i'\widehat{\boldsymbol{B}}_i\right)^{-1}\widehat{\boldsymbol{B}}_i'\widehat{\boldsymbol{a}}_i,$$

where

$$\widehat{\boldsymbol{B}}_i = (\widehat{\boldsymbol{b}}_{i1}, \widehat{\boldsymbol{b}}_{i2}, \ldots, \widehat{\boldsymbol{b}}_{ir}), \qquad \widehat{\boldsymbol{a}}_i = n^{-1}\sum_{m=1}^{n} \boldsymbol{\varphi}_{\kappa_i}(Y_{im})\rho(Y_{im}),$$

for each outcome $i \in \{1, 2, 3\}$.

Table S.1 contains the mean, the standard deviation, the median, and the interquartile range (computed over the Monte Carlo replications) of this estimator in the mixture of normal densities. The table shows that this estimator performs well for all designs considered. Notably, the bias is small relative to the standard deviation. Inspection of the empirical distribution of the points estimates (not reported) further shows that this distribution is close to normal for all configurations considered.

TABLE S.1
Simulation results for mixing proportions in normal design

		mean		std. dev.		median		interq. range	
π_1	i	π_1	π_2	π_1	π_2	π_1	π_2	π_1	π_2
.10	1	.098	.887	.023	.038	.099	.889	.032	.049
.10	2	.097	.898	.015	.052	.098	.895	.022	.066
.10	3	.101	.896	.027	.062	.099	.895	.020	.074
.20	1	.196	.793	.025	.036	.198	.792	.030	.047
.20	2	.196	.798	.020	.049	.196	.798	.028	.067
.20	3	.198	.802	.019	.051	.198	.798	.026	.069
.30	1	.295	.690	.025	.035	.296	.689	.033	.045
.30	2	.295	.700	.021	.048	.295	.699	.028	.068
.30	3	.296	.702	.020	.052	.297	.702	.026	.068
.40	1	.396	.590	.029	.038	.396	.590	.040	.052
.40	2	.395	.599	.024	.048	.395	.598	.037	.059
.40	3	.398	.599	.023	.052	.398	.599	.034	.067

Table S.2 provides the corresponding results for the design with non-central t distributions. The conclusions of Table S.1 carry over to this design.

S.8.2. *Inference in a hidden Markov model.* Figure S.1 contains results for the orthogonal-series estimator in the hidden Markov model of Section 6.2 for $n \in \{500; 2,000; 2,500; 5000\}$. The results for $n = 500$ and for $n = 5,000$ correspond to those in the main text. The plots clearly show that our

TABLE S.2
Simulation results for mixing proportions in non-central t-design

		mean		std. dev.		median		interq. range	
π_1	i	π_1	π_2	π_1	π_2	π_1	π_2	π_1	π_2
.10	1	.098	.880	.027	.053	.097	.880	.034	.070
.10	2	.096	.894	.016	.065	.095	.895	.019	.088
.10	3	.102	.897	.037	.077	.099	.896	.021	.090
.20	1	.197	.785	.028	.053	.197	.781	.036	.067
.20	2	.193	.804	.022	.066	.194	.799	.027	.083
.20	3	.195	.811	.019	.073	.196	.802	.024	.096
.30	1	.296	.688	.030	.049	.297	.685	.039	.065
.30	2	.293	.696	.024	.061	.293	.692	.032	.076
.30	3	.296	.703	.023	.065	.297	.693	.030	.084
.40	1	.394	.584	.030	.049	.395	.580	.038	.057
.40	2	.390	.597	.025	.057	.389	.591	.037	.071
.40	3	.394	.601	.024	.063	.393	.595	.033	.079

estimated standard error captures well the small-sample variability of the orthogonal-series estimator, even in relatively small samples. Furthermore, the performance of the estimator in terms of bias, as compared to the oracle estimator is remarkable.

The infeasible kernel-density estimator of f_1 and f_2 that we report on as an oracle benchmark assumes that the latent states are observable. Letting

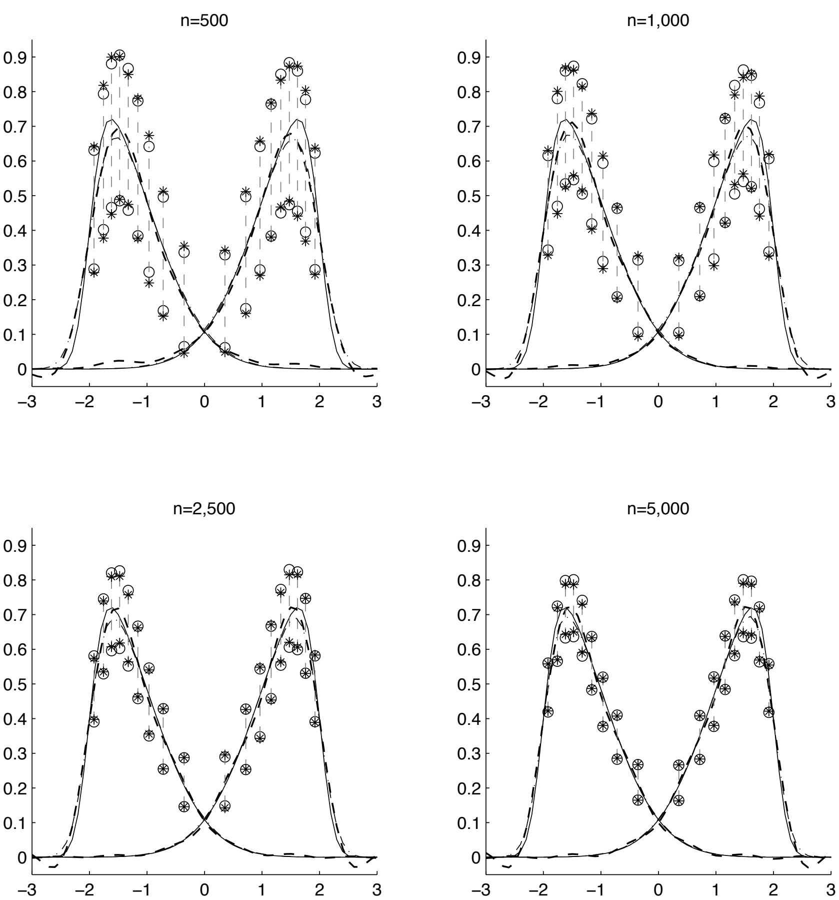
$$n_j = \sum_{m=1}^{n} 1\{Z_{2m} = j\},$$

the estimator of $f_j(y)$ is computed as

$$\frac{1}{n_j h_j} \sum_{m:\, Z_{2m}=j} K\left(\frac{Y_{2m} - y}{h_j}\right),$$

for a chosen kernel function K and a bandwidth h_j. For our simulations here we used a standard-normal kernel and determined the bandwidth via Silverman's rule of thumb (see Silverman 1986), which is a conventional choice in practice.

FIG S.1. *Emission densities in the hidden Markov model*



REFERENCES

ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics* **37** 3099–3132.

BONHOMME, S., JOCHMANS, K. and ROBIN, J. M. (2014). Nonparametric estimation of finite mixtures from repeated measurements. *Journal of the Royal Statistical Society, Series B*, forthcoming.

GASSIAT, E., CLEYNEN, A. and ROBIN, S. (2013). Finite state space non parametric hidden Markov models are in general identifiable. *Statistics and Computing*, forthcoming.

LIEBSHER, E. (1990). Hermite series estimators for probability densities. *Metrica* **37** 321–343.

NEWEY, W. K. and MCFADDEN, D. L. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, **4** 36 2111–2245. Elsevier.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall.

VIOLLAZ, A. J. (1989). Nonparametric estimation of probability density functions based on orthogonal expansions. *Revista Matematica de la Universidad Complutense de Madrid* **2** 41–82.

WALTER, G. G. (1977). Properties of Hermite series estimation of probability density. *Annals of Statistics* **5** 1258–1264.

UNIVERSITY OF CHICAGO
DEPARTMENT OF ECONOMICS
1126 E. 59TH STREET
CHICAGO, IL 60637
U.S.A.
E-MAIL: sbonhomme@uchicago.edu

SCIENCES PO
DEPARTMENT OF ECONOMICS
28 RUE DES SAINTS PÈRES
75007 PARIS
FRANCE
E-MAIL: koen.jochmans@sciencespo.fr
(CORRESPONDING AUTHOR)

SCIENCES PO
DEPARTMENT OF ECONOMICS
28 RUE DES SAINTS PÈRES
75007 PARIS
FRANCE
E-MAIL: jeanmarc.robin@sciencespo.fr