Nonparametric spectral-based estimation of latent structures

Stéphane Bonhomme (Chicago), Koen Jochmans (Sciences Po) and J.-M. Robin (Sciences Po and UCL)

May 27, 2014
Economist like unobserved heterogeneity and dynamic factor models.

Usually discrete mixtures of parametric distributions (derived from theory)

For identification and also estimation, it is useful to consider discrete mixtures of nonparametric models.

This paper proposes a simple estimation procedure for discrete mixtures and hidden Markov models of nonparametric distribution components.
The question of identification in latent structures is the topic of a very recent and active literature.

Nonparametric identification from univariate/cross-sectional data typically fails. (Some exceptions for location models)

Multivariate data (panel data) can present a powerful identification source.

1. Finite mixtures/latent-class models: Hall and Zhou (2003); Allman et al. (2009)
2. (Dynamic) discrete-choice models: Magnac and Thesmar (2002); Kasahara and Shimotsu (2009)
We propose a new constructive identification argument... that delivers a least square-type estimator for mixture weights... allowing for asymptotic distributional theory.
Let \((y_1, \ldots, y_q)\) be \(q\) discrete variables with \(\text{supp}(y_i) = \{1, \ldots, \kappa_i\}\).

There exists a latent variable \(x \in \{1, \ldots, r\}\) with \(\pi_j \equiv \Pr\{x = j\}\).

Let \(p_{ij} \in [0, 1]^\kappa_i\) denote the vector of conditional probability masses of \(y_i\) given \(x = j\):

\[
p_{ij}(k) \equiv \Pr\{y_i = k | x = j\}, \quad k = 1, \ldots, \kappa_i
\]
Unconditional distribution for DMs

- The unconditional joint PDF of \((y_1, \ldots, y_q)\) is

\[
P(y_1, \ldots, y_q) = \sum_{j=1}^{r} \pi_j p_{1j}(y_1) p_{2j}(y_2) \cdots p_{qj}(y_q)
\]

- The set of values \(P(y_1, \ldots, y_q)\) for all \((y_1, \ldots, y_q)\) defines a \(q\)-dimensional array

\[
P = \sum_{j=1}^{r} \pi_j p_{1j} \otimes p_{2j} \otimes \cdots \otimes p_{qj}
\]

- \(\otimes\) is the Kronecker product
Hidden Markov models

- There are $q$ discrete latent variables ($x_1, \ldots, x_q$) for $q$ measurements ($y_1, \ldots, y_q$).

- **Stationarity:**
  
  \[
  \Pr\{x_i = j\} = \pi_j, \quad i = 1, \ldots, q
  \]
  \[
  \Pr\{x_{i+1} | x_i\} = K(x_i, x_{i+1}), \quad i = 1, \ldots, q - 1
  \]
  \[
  \Pr\{y_i = k | x_i = j\} = p_{j}(k), \quad k = 1, \ldots, \kappa
  \]

- **Conditional independence:** measurements $y_1, \ldots, y_q$ are independent conditional on $(x_1, \ldots, x_q)$. 
The unconditional joint PDF of \((y_1, \ldots, y_3)\) is

\[
\mathbb{P}(y_1, y_2, y_3) = \sum_{j_1=1}^{r} \left\{ \pi_{j_1} p_{j_1}(y_1) \sum_{j_2=1}^{r} \left[ K(j_1, j_2) p_{j_2}(y_2) \sum_{j_3=1}^{r} K(j_2, j_3) p_{j_3}(y_3) \right] \right\}
\]

\[
= \sum_{j_2=1}^{r} \left\{ \sum_{j_1=1}^{r} p_{j_1}(y_1) \pi_{j_1} K(j_1, j_2) \right\} p_{j_2}(y_2) \left[ \sum_{j_3=1}^{r} K(j_2, j_3) p_{j_3}(y_3) \right]
\]
Let $P = [p_1, \ldots, p_r] \in \mathbb{R}^{k \times r}$ and $\Pi = \text{diag}(\pi_1, \ldots, \pi_r)$.

Hence the 3-dimensional array

$$
\mathbb{P} = \sum_{j=1}^{r} (P\Pi K)_j \otimes p_j \otimes \left( PK^\top \right)_j
$$

where $M_j$ denotes the $j$th column of matrix $M$.

If $q > 3$ one can select all consecutive triples or regroup observations into 3 consecutive clusters:

$$(y_1, \ldots, y_{k-1}), y_k, (y_{k+1}, \ldots, y_q).$$
Consider a $\kappa_1 \times \kappa_2 \times \kappa_3$ array $P = \sum_{j=1}^{r} p_{1j} \otimes p_{2j} \otimes p_{3j}$

Let $P_i = [p_{i1}, \ldots, p_{ir}] \in \mathbb{R}^{\kappa_i \times r}$, $i = 1, 2, 3$

Let $r_i = \max\{k : \text{all collections of } k \text{ columns of } P_i \text{ are independent}\}$ (the Kruskal-rank of $P_i$).

Note that if $P \in \mathbb{R}^{\kappa \times r}$ has rank $r$ it also has Kruskal-rank $r$.

If $r_1 + r_2 + r_3 \geq 2r + 2$ then $P$ uniquely determines the matrices $P_i$ up to simultaneous column-permutation and common column-scaling.
Allman et al. use Kruskal’s result to give conditions for the identification of discrete mixtures of discrete and continuous nonparametric distributions, hidden Markov models and some stochastic graphs.
Discrete mixtures
Allman, Matias and Rhodes (AoS, 2009)

- Kruskal’s theorem applies with

\[ \mathbb{P} = \sum_{j=1}^{r} \pi_j p_{1j} \otimes p_{2j} \otimes p_{3j} \]

\[ P_1 = [\pi_1 p_{11}, \ldots, \pi_r p_{1r}], \quad P_i = [p_{i1}, \ldots, p_{ir}], \quad i > 1 \]

- (Corollary 2) Since \( \text{sum}(P_1, 1) = [\pi_1, \ldots, \pi_r] \) and \( \text{sum}(P_i, 1) = [1, \ldots, 1], \quad i > 1 \), then, if \( r_1 + r_2 + r_3 \geq 2r + 2 \), group-probabilities \( \pi_j \) and conditional probabilities \( p_{ij} \) are identified up to labeling.

- (Theorem 8) Holds for continuous mixture components if the component densities are linearly independent \( (r_1 = r_2 = r_3 = r) \).
The parameters of an HMM with \( r \) hidden states and \( \kappa \) observable states are generically identifiable from the marginal distribution of \( 2k + 1 \) consecutive variables provided \( k \) satisfies

\[
\binom{k + \kappa - 1}{\kappa - 1} \geq r
\]

Note that \( \binom{k + \kappa - 1}{\kappa - 1} = \kappa \) for \( k = 1 \) (3 measurements) and \( \binom{k + \kappa - 1}{\kappa - 1} = k + 1 \) for \( \kappa = 2 \) (binary outcomes).
They use Allman et al.’s result to prove the following result.

Assume that $r$ is known, $P = [p_1, \ldots, p_r]$ is full column rank, and $K$ has full rank. Then $K$ and $P$ are identifiable from the distribution of 3 consecutive observations $(y_1, y_2, y_3)$ up to label swapping of the hidden states.

Estimation by penalized ML or EM algorithm.
There exists few constructive identification procedures.

De Lathauwer (SIAM, 2006) applies to the case where one outcome (say $y_1$) is such that $P_1$ is full column rank.

However it provides identification only up to relabeling AND scaling.

Group probabilities $\pi_j$ are thus not identified.

We propose one such constructive identification that works both for DMs and HMMs, inspired from ICA or blind deconvolution.
\[ P = \sum_{j=1}^{r} \pi_j p_{1j} \otimes p_{2j} \otimes p_{3j} \]

Let \( \Pi = \text{diag}(\pi_1, \ldots, \pi_r) \), and
\[ P_i = [p_{i1}, \ldots, p_{ir}] \in \mathbb{R}^{k_i \times r}, i = 1, 2, 3. \]

Assume \( \text{rank}(P_i) = r \) and \( \pi_j > 0 \).
DMs

- \( P_1 \Pi P_2^\top = \sum_{j=1}^{r} \pi_j p_{1j} p_{2j}^\top = \sum_{j=1}^{r} \pi_j p_{1j} \otimes p_{2j} \) is the matrix containing probabilities \( \mathbb{P}(y_1, y_2) \). Observable.

- SVD on \( P_1 \Pi P_2^\top \), which has rank \( r \), allows to construct \( U \) and \( V \) such that

\[
U_{r \times \kappa_1} P_1 \Pi P_2^\top V_{\kappa_2 \times r}^\top = I_r \Rightarrow (VP_2)^\top = (UP_1 \Pi)^{-1} \equiv Q^{-1}_{r \times r}
\]

- \( \mathbb{P}(\cdot, \cdot, k) = \sum_{j=1}^{r} \pi_j p_{1j} \otimes p_{2j} \otimes p_{3j}(k) = P_1 \Pi D_{3k} P_2^\top \), with \( D_{3k} = \text{diag} [p_{31}(k), \ldots, p_{3r}(k)] \), is the matrix containing probabilities \( \mathbb{P}(y_1, y_2, k) \) (for any \( y_1, y_2 \) and \( y_3 = k \)). Also observable.

- \( W_k = U\mathbb{P}(\cdot, \cdot, k) V^\top = QD_{3k} Q^{-1} \) (whitening)

- \( P_3 \) identified by the eigenvalues of matrices \( W_1, \ldots, W_{\kappa_3} \)

- Repeat for \( P_1 \) and \( P_2 \).

- \( \pi = [\pi_1; \ldots; \pi_r] \) identified from \( \mathbb{P}(y_i) = \sum_{j=1}^{r} \pi_j p_{ij}(y_i) = P \pi \)
DMs

- \( P_1 \Pi P_2^\top = \sum_{j=1}^r \pi_j \rho_{1j} \rho_{2j}^\top = \sum_{j=1}^r \pi_j \rho_{1j} \otimes \rho_{2j} \) is the matrix containing probabilities \( \mathbb{P}(y_1, y_2) \). Observable.

- SVD on \( P_1 \Pi P_2^\top \), which has rank \( r \), allows to construct \( U \) and \( V \) such that

\[
U_{r \times k_1} P_1 \Pi P_2^\top V_{k_2 \times r}^\top = I_r \Rightarrow (VP_2)^\top = (UP_1 \Pi)^{-1} \equiv Q^{-1}
\]

- \( \mathbb{P}(\cdot, \cdot, k) = \sum_{j=1}^r \pi_j \rho_{1j} \otimes \rho_{2j} \otimes \rho_{3j}(k) = P_1 \Pi D_{3k} P_2^\top \), with \( D_{3k} = \text{diag}[\rho_{31}(k), \ldots, \rho_{3r}(k)] \), is the matrix containing probabilities \( \mathbb{P}(y_1, y_2, k) \) (for any \( y_1, y_2 \) and \( y_3 = k \)). Also observable.

- \( W_k = U \mathbb{P}(\cdot, \cdot, k) V^\top = QD_{3k} Q^{-1} \) (whitening)

- \( P_3 \) identified by the eigenvalues of matrices \( W_1, \ldots, W_{k_3} \)

- Repeat for \( P_1 \) and \( P_2 \).

- \( \pi = [\pi_1; \ldots; \pi_r] \) identified from \( \mathbb{P}(y_i) = \sum_{j=1}^r \pi_j \rho_{ij}(y_i) = P \pi \)
\[ P_1 \Pi P_2^\top = \sum_{j=1}^r \pi_j p_{1j} p_{2j}^\top = \sum_{j=1}^r \pi_j p_{1j} \otimes p_{2j} \] is the matrix containing probabilities \( \mathbb{P}(y_1, y_2) \). Observable.

SVD on \( P_1 \Pi P_2^\top \), which has rank \( r \), allows to construct \( U \) and \( V \) such that

\[
U \begin{bmatrix} r & \times & k_1 \end{bmatrix} P_1 \Pi P_2^\top \begin{bmatrix} k_2 & \times & r \end{bmatrix} V^\top = I_r \Rightarrow (VP_2)^\top = (UP_1 \Pi)^{-1} = Q^{-1}
\]

\( \mathbb{P}(:, :, k) = \sum_{j=1}^r \pi_j p_{1j} \otimes p_{2j} \otimes p_{3j}(k) = P_1 \Pi D_{3k} P_2^\top \), with \( D_{3k} = \text{diag}[p_{31}(k), ..., p_{3r}(k)] \), is the matrix containing probabilities \( \mathbb{P}(y_1, y_2, k) \) (for any \( y_1, y_2 \) and \( y_3 = k \)). Also observable.

\( W_k = U \mathbb{P}(:, :, k) V^\top = Q D_{3k} Q^{-1} \) (whitening)

\( P_3 \) identified by the eigenvalues of matrices \( W_1, ..., W_{k_3} \)

Repeat for \( P_1 \) and \( P_2 \).

\( \pi = [\pi_1; ...; \pi_r] \) identified from \( \mathbb{P}(y_i) = \sum_{j=1}^r \pi_j p_{ij}(y_i) = P \pi \)
HMMs

\[ \mathbb{P} = \sum_{j=1}^{r} (P \Pi K)_j \otimes p_j \otimes (PK^\top)_j, \quad \Pi = \text{diag}(\pi_1, \ldots, \pi_r) \]

Assume \( K \) full rank, \( P = [p_1, \ldots, p_r] \in \mathbb{R}^{\kappa \times r} \) full column rank and \( \pi_j > 0 \).
HMMs

- One can put all $P(y_1, y_2, y_3)$ for fixed $y_2 \in \{1, \ldots, \kappa\}$ in the matrix $P(:, k, :) = P \Pi K D_{2k} K P^\top$, $D_{2k} = \text{diag}(p_1(k), \ldots, p_r(k))$

- Note that the matrix $P \Pi K^2 P^\top$ is the matrix containing probabilities $P(y_1, y_3)$.

- SVD on $P \Pi K^2 P^\top$, which has rank $r$, allows to construct $U$ and $V$ such that

$$
U_{r \times \kappa_1} P \Pi K^2 P^\top V_{\kappa_2 \times r}^\top = I_r \iff K P^\top V^\top = (P \Pi K)^{-1} \equiv Q^{-1} \quad (\text{whitening})
$$

- $W_k = U P(:, k, :) V^\top = Q D_k Q^{-1}$

- $P$ identified by the eigenvalues of matrices $W_1, \ldots, W_\kappa$

- $\pi$ identified from $P(y_1) = \sum_{j=1}^r \pi_j p_j(y_1) = P \pi$

- $K$ identified from $P(y_1, y_2) = P \Pi K P^\top$
HMMs

- One can put all $\mathbb{P}(y_1, y_2, y_3)$ for fixed $y_2 \in \{1, \ldots, \kappa\}$ in the matrix $\mathbb{P}(\cdot, k, :) = P \Pi K D_{2k} K P^\top$, $D_{2k} = \text{diag}(p_1(k), \ldots, p_r(k))$

- Note that the matrix $P \Pi K^2 P^\top$ is the matrix containing probabilities $\mathbb{P}(y_1, y_3)$.

- SVD on $P \Pi K^2 P^\top$, which has rank $r$, allows to construct $U$ and $V$ such that

$$U_{r \times \kappa_1} P \Pi K^2 P^\top V_{\kappa_2 \times r}^\top = I_r \Leftrightarrow K P^\top V^\top = (P \Pi K)^{-1} \equiv Q^{-1}_{r \times r}$$

- $W_k = U_{\mathbb{P}(\cdot, k,:)} V^\top = Q D_k Q^{-1}$ (whitening)

- $P$ identified by the eigenvalues of matrices $W_1, \ldots, W_\kappa$

- $\pi$ identified from $\mathbb{P}(y_1) = \sum_{j=1}^{\kappa} \pi_j p_j(y_1) = P \pi$

- $K$ identified from $\mathbb{P}(y_1, y_2) = P \Pi K P^\top$
HMMs

- One can put all $\mathbb{P}(y_1, y_2, y_3)$ for fixed $y_2 \in \{1, \ldots, \kappa\}$ in the matrix $\mathbb{P}(\cdot, k, \cdot) = P \Pi K D_{2k} K P^\top$, $D_{2k} = \text{diag}(p_1(k), \ldots, p_r(k))$
- Note that the matrix $P \Pi K^2 P^\top$ is the matrix containing probabilities $\mathbb{P}(y_1, y_3)$.
- SVD on $P \Pi K^2 P^\top$, which has rank $r$, allows to construct $U$ and $V$ such that

$$U_{r \times \kappa_1} P \Pi K^2 P^\top V_{\kappa_2 \times r}^\top = I_r \iff KP^\top V^\top = (P \Pi K)^{-1} \equiv Q^{-1}_{r \times r}$$

- $W_k = U \mathbb{P}(\cdot, k, \cdot) V^\top = Q D_k Q^{-1}$ (whitening)
- $P$ identified by the eigenvalues of matrices $W_1, \ldots, W_\kappa$
- $\pi$ identified from $\mathbb{P}(y_1) = \sum_{j=1}^r \pi_j p_j(y_1) = P \pi$
- $K$ identified from $\mathbb{P}(y_1, y_2) = P \Pi K P^\top$
HMMs

- One can put all $\mathbb{P}(y_1, y_2, y_3)$ for fixed $y_2 \in \{1, \ldots, \kappa\}$ in the matrix $\mathbb{P}(, k, :) = P\Pi K D_{2k} K P^\top$, $D_{2k} = \text{diag}(p_1(k), \ldots, p_r(k))$

- Note that the matrix $P\Pi K^2 P^\top$ is the matrix containing probabilities $\mathbb{P}(y_1, y_3)$.

- SVD on $P\Pi K^2 P^\top$, which has rank $r$, allows to construct $U$ and $V$ such that

\[
U_{r \times \kappa_1} P\Pi K^2 P^\top V_{\kappa_2 \times r}^\top = I_r \iff KP^\top V^\top = (P\Pi K)^{-1} = Q^{-1}_{r \times r}
\]

- $W_k = U\mathbb{P}(, k, :) V^\top = Q D_k Q^{-1}$ (whitening)

- $P$ identified by the eigenvalues of matrices $W_1, \ldots, W_k$

- $\pi$ identified from $\mathbb{P}(y_1) = \sum_{j=1}^r \pi_j p_j(y_1) = P\pi$

- $K$ identified from $\mathbb{P}(y_1, y_2) = P\Pi K P^\top$
HMMs

One can put all $\mathbb{P}(y_1, y_2, y_3)$ for fixed $y_2 \in \{1, \ldots, \kappa\}$ in the matrix $\mathbb{P}(:, k, :) = P \Pi K D_{2k} K P^\top$, $D_{2k} = \text{diag}(p_1(k), \ldots, p_r(k))$

Note that the matrix $P \Pi K^2 P^\top$ is the matrix containing probabilities $\mathbb{P}(y_1, y_3)$.

SVD on $P \Pi K^2 P^\top$, which has rank $r$, allows to construct $U$ and $V$ such that

$$U_{r \times \kappa_1} P \Pi K^2 P^\top V_{\kappa_2 \times r}^\top = I_r \iff K P^\top V^\top = (P \Pi K)^{-1} \equiv Q^{-1}_{r \times r}$$

- $W_k = U \mathbb{P}(:, k, :) V^\top = Q D_k Q^{-1}$ (whitening)
- $P$ identified by the eigenvalues of matrices $W_1, \ldots, W_k$
- $\pi$ identified from $\mathbb{P}(y_1) = \sum_{j=1}^r \pi_j p_j(y_1) = P \pi$
- $K$ identified from $\mathbb{P}(y_1, y_2) = P \Pi K P^\top$
Matrices $W_k$ thus have to be simultaneously diagonalized.

Approximate joint diagonalization by least squares:

$$Q = \arg \min_Q \sum_{k=1}^{\kappa_3} \| W_k - QD_k Q^{-1} \|_F^2, \quad D_k \equiv \text{diag}[Q^{-1} W_k Q]$$

Algorithm in Iferroudjene, Abed-Meraim and Belouchrani (Applied Math. and Computation, 2009)

Advantage of LS: asymptotic theory is possible
Continuous outcomes

- Requires discretization
- We use orthogonal polynomials (Chebychev)
Discrete mixtures of continuous distributions

- Conditional PDF of $y_i$ given $= j$:

$$f_{ij}(y) \simeq \sum_{k=1}^{\kappa_i} p_{ij}(k) \varphi_k(y), \quad p_{ij}(k) = \int_{-1}^{1} \varphi_k(u) f_{ij}(u) \, du$$

- $(\varphi_k)$ complete orthonormal set of functions:

$$\int \varphi_k(y) \varphi_\ell(y) \rho(y) \, dy = \delta_{k\ell}$$

- Three observations:

$$f(y_1, y_2, y_3) = \sum_{j=1}^{r} \pi_j f_{1j}(y_1) f_{2j}(y_2) f_{3j}(y_3)$$

$$\simeq \sum_{j=1}^{r} \pi_j \rho_{1j} \otimes \rho_{2j} \otimes \rho_{3j}$$

- Note that sum$(p_{ij}) \neq 1$. Yet the identification algorithm continues to work.
Asymptotic theory

- Standard convergence rates because the weights are root-$n$ consistent
- Extends to hidden Markov models for continuous outcomes
Example: DMs of continuous distributions
Simulation

- We generate data from a heterogenous mixture of beta distributions on $[-1, 1]$
- $r = 2; \ q = 3; \ \pi_1 = \pi_2 = \frac{1}{2}$
- Chebychev polynomials of the first kind for $\phi_i$.
- Orthogonal-series estimators are not bona fide. Adjust estimates ex post via Gajek’s (1986) projection procedure.
\( n = 500 \)
Proportions

<table>
<thead>
<tr>
<th></th>
<th>$n = 500$</th>
<th></th>
<th>$n = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>std</td>
<td>mean</td>
</tr>
<tr>
<td></td>
<td>$\pi_1$</td>
<td>$\pi_2$</td>
<td>$\pi_1$</td>
</tr>
<tr>
<td>$i = 1$</td>
<td>.5133</td>
<td>.4794</td>
<td>.0257</td>
</tr>
<tr>
<td>$i = 2$</td>
<td>.5130</td>
<td>.4854</td>
<td>.0300</td>
</tr>
<tr>
<td>$i = 3$</td>
<td>.4978</td>
<td>.4948</td>
<td>.0319</td>
</tr>
</tbody>
</table>
• Stationary probit model for a binary state variable

\[ s_t = 1\{s_{t-1} \geq \varepsilon_t\}, \quad \varepsilon_t \sim \mathcal{N}(0, 1), \]

and suppose that,

\[ f(y_t|s_t = 0) \sim \text{left-skewed Beta}, \quad f(y_t|s_t = 1) \sim \text{right-skewed Beta} \]

• Steady state gives \( \Pr[s_t = 0] \approx \frac{1}{4} \) and \( K(0, 0) = \frac{1}{2}, K(1, 0) \approx \frac{1}{6} \).

• Most draws are from dominant regime \( (s_t = 1) \).
### State process

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Mean</th>
<th>Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Pr[s_t = 1]$</td>
<td>.7591</td>
<td>.7255</td>
<td>.0755</td>
</tr>
<tr>
<td>$Pr[s_t = 0]$</td>
<td>.2409</td>
<td>.2554</td>
<td>.0786</td>
</tr>
<tr>
<td>$K(0,0)$</td>
<td>.5000</td>
<td>.5731</td>
<td>.3056</td>
</tr>
<tr>
<td>$K(0,1)$</td>
<td>.5000</td>
<td>.3913</td>
<td>.3494</td>
</tr>
<tr>
<td>$K(1,0)$</td>
<td>.1587</td>
<td>.1352</td>
<td>.0587</td>
</tr>
<tr>
<td>$K(1,1)$</td>
<td>.8413</td>
<td>.8500</td>
<td>.0608</td>
</tr>
</tbody>
</table>