Supplementary Material to: Hidden Markov Models with state-dependent mixtures: Minimal representation, model testing and applications to clustering

Hajo Holzmann and Florian Schwaiger

Fakultät für Mathematik und Informatik, Philipps-Universität Marburg, Germany

1 Methodology for HMMs with state-dependent mixtures: Additional Proofs

1.1 Markov chains under dependence structure restrictions

Let $(S_t)_t$ be a k-state Markov chain with ergodic transition probability matrix (t.p.m.) $\mathbf{\Gamma} = (\gamma_{i,j})_{i,j=1,\dots,k}$ having the stationary distribution $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)$. In the following we always assume $\pi_j > 0$ for $j = 1, \dots, k$.
For a (disjoint) partition $\mathcal{G} = \{G_1, \dots, G_r\}$ of the state space (into non-empty sets), we let $G(j)$ be the function which maps a state $j \in \{1, \dots, k\}$ onto its group, i.e. for $j \in G_l$ we have $G(j) = G_l$. If $P\big(S_{t-1} = i,\ S_t \in G(j)\big) > 0$) we have the general formula

$$\gamma_{i,j} = P\big(S_t \in G(j) \big| S_{t-1} = i\big) \cdot P\big(S_t = j \big| S_{t-1} = i,\ S_t \in G(j)\big). \tag{1}$$

Define the reduced t.p.m. $\lambda_{\mathcal{G}}(\Gamma)$ by

$$\big(\lambda_{\mathcal{G}}(\mathbf{\Gamma})\big)_{i,j} = P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) \cdot P\big(S_t = j \big| S_t \in G(j)\big), \qquad i, j = 1, \dots, k. \tag{2}$$

Lemma 1. *The matrix $\lambda_{\mathcal{G}}(\mathbf{\Gamma})$ is a t.p.m., and the following statements are equivalent.*

1. *We have*
$$\lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}. \tag{3}$$

2. *For $i, j = 1, \dots, k$ it holds*
$$P\big(S_t \in G(j) \big| S_{t-1} = i\big) = P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big)$$
and if $P\big(S_t \in G(j), S_{t-1} = i\big) > 0$ also
$$P\big(S_t = j \big| S_t \in G(j), S_{t-1} = i\big) = P\big(S_t = j \big| S_t \in G(j)\big).$$

3. *There exists a t.p.m.* $(\nu_{l,m})_{l,m} \in \mathbb{R}^{r\times r}$ *and* $(p_1,\dots,p_k) \in \mathbb{R}^k$*, with* $p_j \geq 0$*,* $\sum_{g\in G_l} p_g = 1$*,* $l = 1,\dots,r$*, such that*

$$\gamma_{i,j} = \nu_{a(i),a(j)} \cdot p_j, \qquad i,j = 1,\dots,k$$

where $a : \{1,\dots,k\} \to \{1,\dots,r\}$ *and* $a(g) = l :\Leftrightarrow g \in G_l$*.*

Proof of Lemma 1. $\lambda_{\mathcal{G}}(\mathbf{\Gamma})$ is a t.p.m. since

$$\sum_{j=1}^{k} \big(\lambda_{\mathcal{G}}(\mathbf{\Gamma})\big)_{i,j} = \sum_{l=1}^{r} \sum_{g\in G_l} P\big(S_t \in G_l \big| S_{t-1} \in G(i)\big) \cdot P\big(S_t = g \big| S_t \in G_l\big) = 1.$$

In order to validate that 1. implies 2., note

$$\gamma_{i,j} = P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) \cdot P\big(S_t = j \big| S_t \in G(j)\big), \qquad i,j = 1,\dots,k,$$

implies $\gamma_{i,j} = \gamma_{h,j}$ for all $h \in G(i)$, i.e. under 1. $\mathbf{\Gamma}$ has equal rows with indices in the same group of partition $\mathcal{G}$. Thus,

$$P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) = \sum_{g\in G(j)} \sum_{h\in G(i)} \left(\frac{P\big(S_{t-1} = h\big)}{P\big(S_{t-1} \in G(i)\big)} \cdot \gamma_{h,g}\right)$$

$$= \sum_{g\in G(j)} \gamma_{i,g} \sum_{h\in G(i)} \left(\frac{P\big(S_{t-1} = h\big)}{P\big(S_{t-1} \in G(i)\big)}\right) = \sum_{g\in G(j)} \gamma_{i,g} = P\big(S_t \in G(j) \big| S_{t-1} = i\big),$$

which gives the first claim of 2. If further $P\big(S_t \in G(j), S_{t-1} = i\big) > 0$ then $\gamma_{i,g} > 0$ for at least one $g \in G(j)$ and thus $P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) > 0$. Further, due to 1. and (1)

$$\begin{aligned} &P\big(S_t \in G(j) \big| S_{t-1} = i\big) \cdot P\big(S_t = j \big| S_{t-1} = i,\ S_t \in G(j)\big) \\ &= P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) \cdot P\big(S_t = j \big| S_t \in G(j)\big) \end{aligned}$$

and hence also the second claim of 2. follows.

Now, assume 2. to hold, then for $l,m = 1,\dots,r$ set $\nu_{l,m} = P(S_t \in G_m | S_{t-1} \in G_l)$ and $p_j = P(S_t = j | S_t \in G(j))$. Note $a(\cdot)$ is constant on each group of the partition, $\boldsymbol{N} = (\nu_{l,m})_{l,m} \in \mathbb{R}^{r\times r}$ defines a t.p.m. and $\boldsymbol{p} = (p_1,\dots,p_k)$ has the desired property of 3.. At first, if $P\big(S_{t-1} = i,\ S_t \in G(j)\big) = 0$ we have $\gamma_{i,j} = 0$ and also $P\big(S_t \in G(j) \big| S_{t-1} = i\big) = 0$. Thus, due to 2. also $P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) = 0$ and hence 3. holds in this case. If otherwise $P\big(S_{t-1} = i,\ S_t \in G(j)\big) > 0$ due to (1) and the validity of both statements in 2.

$$\gamma_{i,j} = P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) \cdot P\big(S_t = j \big| S_t \in G(j)\big) = \nu_{a(i),a(j)} \cdot p_j.$$

Finally, assume 3. to hold. Hence, for $i,j = 1,\dots,k$

$$P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) = \sum_{g\in G(j)} \sum_{h\in G(i)} \left(\frac{P\big(S_{t-1} = h\big)}{P\big(S_{t-1} \in G(i)\big)} \cdot \nu_{a(h),a(g)} \cdot p_g\right) = \nu_{a(i),a(j)},$$

$$P\big(S_t \in G(j) \big| S_{t-1} = i\big) = \sum_{g\in G(j)} \big(\nu_{a(h),a(g)} \cdot p_g\big) = \nu_{a(i),a(j)}.$$

If $P\big(S_{t-1} = i,\ S_t \in G(j)\big) = 0$, again $\gamma_{i,g} = 0$ for all $g \in G(j)$. Due to 3. we further have for $h \in G(i)$ and $g \in G(j)$ $\gamma_{h,g} = \nu_{a(i),a(g)} \cdot p_g = \gamma_{i,g} = 0$. Thus also $\nu_{a(i),a(j)} = 0$ and hence $\gamma_{i,j} = \big(\lambda_{\mathcal{G}}(\mathbf{\Gamma})\big)_{i,j} = 0$. If otherwise $P\big(S_{t-1} = i,\ S_t \in G(j)\big) > 0$ due to (1) we directly get $p_j = P\big(S_t = j \big| S_{t-1} = i,\ S_t \in G(j)\big)$, and since p_j is independent of i, $p_j = P\big(S_t = j \big| S_t \in G(j)\big)$, which finally gives

$$\gamma_{i,j} = \nu_{a(i),a(j)} \cdot p_j = P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) \cdot P\big(S_t = j \big| S_t \in G(j)\big) = \big(\lambda_{\mathcal{G}}(\mathbf{\Gamma})\big)_{i,j}.$$

□

1.2 Representations of HMMs with state-dependent mixtures

Let $(X_t, S_t)_t$ be a k-state HMM with state space $\{1, \dots, k\}$, state dependent densities $f_j(x) = f_{X_t|S_t=j}(x),\ j = 1, \dots, k$, t.p.m. $\mathbf{\Gamma}$ and stationary distribution $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)$.

Definition 1 (Reducing states to mixture components in an HMM)**.** Let $\mathcal{G} = \{G_1, \dots, G_r\}$ be a partition of $\{1, \dots, k\}$. Call *reducing states to mixture components with respect to* $\mathcal{G}$ the mapping of the HMM $(X_t, S_t)_t$ onto the new HMM $(X_t^{(\mathcal{G})}, S_t^{(\mathcal{G})})_t$, the distribution of which is determined by the t.p.m. $\mathbf{\Gamma}^{(\mathcal{G})}$,

$$\big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l,m} := P\big(S_t \in G_m \big| S_{t-1} \in G_l\big), \qquad l, m = 1, \dots, r$$

of the Markov chain $(S_t^{(\mathcal{G})})_t$ (on the state space $\{1, \dots, r\}$), and the state-dependent densities

$$f_l^{(\mathcal{G})}(x) := f_{X_t^{(\mathcal{G})}|S_t^{(\mathcal{G})}=l}(x) := f_{X_t|S_t \in G_j}(x), \qquad x \in \mathbb{R}^d,\ l = 1, \dots, r.$$

of the observable process $(X_t^{(\mathcal{G})})_t$. ◇

Theorem 1. *The distribution of the observable process* $\big(X_t^{(\mathcal{G})}\big)_t$ *after reducing states to mixture components w.r.t. the partition* $\mathcal{G}$ *is the same as that of an HMM with t.p.m.* $\lambda_{\mathcal{G}}(\mathbf{\Gamma})$ *(on the original state space* $\{1, \dots, k\}$*) and state-dependent densities* $f_j(x)$, $j = 1, \dots, k$. *In particular, if* $\lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$ *we have that* $(X_t)_t \overset{(d)}{=} \big(X_t^{(\mathcal{G})}\big)_t$.

Proof of Theorem 1. Denote by $(Y_t^{(\mathcal{G})}, T_t^{(\mathcal{G})})_t$ the HMM with Markov chain $(T_t^{(\mathcal{G})})_t$ having t.p.m. $\lambda_{\mathcal{G}}(\mathbf{\Gamma})$ and observable process $(Y_t^{(\mathcal{G})})_t$ with state dependent densities $f_j(x)$, $j = 1, \dots, k$.

Proving $(Y_t^{(\mathcal{G})})_t \overset{(d)}{=} (X_t^{(\mathcal{G})})_t$ yields the claim, since then under $\lambda_{\mathcal{G}}(\mathbf{\Gamma}) = \mathbf{\Gamma}$ directly $(X_t)_t \overset{(d)}{=} (Y_t^{(\mathcal{G})})_t$ (t.p.m.'s and state dependent densities coincide) and thus $(X_t)_t \overset{(d)}{=} \big(X_t^{(\mathcal{G})}\big)_t$.

The remainder is to prove $(Y_t^{(\mathcal{G})})_t \overset{(d)}{=} (X_t^{(\mathcal{G})})_t$: At first, reducing dependence information does not change the stationary distribution of the Markov chain, i.e. for $j \in \{1, \dots k\}$, we have

$$P\big(T_t^{(\mathcal{G})} = j\big) = P\big(S_t = j\big).$$

To show this, let $\boldsymbol{\pi}$ denote the stationary distribution of $(S_t)_t$, we have

$$\begin{aligned}
&\sum_{i=1}^{k} \pi_i \, P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) = \sum_{l=1}^{r} \sum_{g \in G_l} \pi_g \, P\big(S_t \in G(j) \big| S_{t-1} \in G_l\big) \\
=&\sum_{l=1}^{r} \Big\{ P\big(S_t \in G(j) \big| S_{t-1} \in G_l\big) \, P\big(S_t \in G_l\big)\Big\} = \sum_{l=1}^{r} \Big\{ P\big(S_t \in G(j),\ S_{t-1} \in G_l\big)\Big\} \\
=&P\big(S_t \in G(j)\big),
\end{aligned}$$

and thus

$$\begin{aligned}
\boldsymbol{\pi} \cdot \big(\lambda_{\mathcal{G}}(\boldsymbol{\Gamma})\big)_{\cdot,j} &= \sum_{i=1}^{k} \Big\{ \pi_i \, P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big) \, P\big(S_t = j \big| S_t \in G(j)\big)\Big\} \\
=&P\big(S_t = j \big| S_t \in G(j)\big) \, \sum_{i=1}^{k} \Big\{ \pi_i \, P\big(S_t \in G(j) \big| S_{t-1} \in G(i)\big)\Big\} = P(S_t = j).
\end{aligned}$$

Further, as mentioned before, for the MC $(S_t^{(\mathcal{G})})_t$ it holds

$$P\big(S_t^{(\mathcal{G})} = l\big) = P\big(S_t \in G_l\big), \qquad l = 1, \ldots, r$$

$$\big(\boldsymbol{\Gamma}^{(\mathcal{G})}\big)_{l,m} = P\big(S_t^{(\mathcal{G})} = l \big| S_{t-1}^{(\mathcal{G})} = m\big) = P\big(S_t \in G_l \big| S_{t-1} \in G_m\big), \qquad l, m = 1, \ldots, r.$$

Further, since $f_j(x) = f_{X_t|S_t=j}(x)$ denotes the state dependent density of the original HMM, the state dependent densities of the transformed HMMs are given for $x \in \mathbb{R}^d$ by

$$f_l^{(\mathcal{G})}(x) = f_{X_t^{(\mathcal{G})}|S_t^{(\mathcal{G})}=l}(x) = \sum_{g \in G_l} P\big(S_t = g | S_t \in G_l\big) \cdot f_g(x), \qquad l = 1, \ldots, r$$

and

$$f_j(x) = f_{Y_t^{(\mathcal{G})}|T_t^{(\mathcal{G})}=j}(x), \qquad j = 1, \ldots, k.$$

Assuming the Markov chain to start in its stationary distribution given by the t.p.m., the density of the observable process $(Y_t^{(\mathcal{G})})_{t=1,\ldots,n}$ of the reduced model is given by

$$\begin{aligned}
&f_{(Y_1^{(\mathcal{G})},\ldots,Y_n^{(\mathcal{G})})}(x_1, \ldots, x_n) \\
&= \sum_{j_1,\ldots,j_n=1}^{k} \Big(P\big(T_1^{(\mathcal{G})} = j_1\big) \cdot f_{j_1}(x_1) \cdot \prod_{t=2}^{n} P\big(T_t^{(\mathcal{G})} = j_t \big| T_{t-1}^{(\mathcal{G})} = j_{t-1}\big) \cdot f_{j_t}(x_t)\Big) \\
&= \sum_{l_1,\ldots,l_n=1}^{r} \sum_{g_1 \in G_{l_1}} \cdots \sum_{g_n \in G_{l_n}} \Big(P\big(S_1 = g_1\big) \cdot f_{g_1}(x_1) \cdot \prod_{t=2}^{n} \big(\lambda_{\mathcal{G}}(\boldsymbol{\Gamma})\big)_{g_{t-1},g_t} \cdot f_{g_t}(x_t)\Big)
\end{aligned} \tag{4}$$

Since here $g_{t-1} \in G_{l_{t-1}}$ and $g_t \in G_{l_t}$,

$$\big(\lambda_{\mathcal{G}}(\boldsymbol{\Gamma})\big)_{g_{t-1},g_t} = \big(\boldsymbol{\Gamma}^{(\mathcal{G})}\big)_{l_{t-1},l_t} \cdot P\big(S_t = g_t | S_t \in G_{l_t}\big).$$

Therefore

$$
\begin{aligned}
&f_{(Y_1^{(\mathcal{G})},\dots,Y_n^{(\mathcal{G})})}(x_1,\dots,x_n)\\
&= \sum_{l_1,\dots,l_n=1}^{r} \sum_{g_1\in G_{l_1}} \cdots \sum_{g_n\in G_{l_n}} \Big(P\big(S_1=g_1\big)\cdot f_{g_1}(x_1)\cdot \prod_{t=2}^{n} \big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l_{t-1},l_t}\cdot P\big(S_t=g_t|S_t\in G_{l_t}\big)\cdot f_{g_t}(x_t)\Big)
\end{aligned}
$$

and hence in the latter sum only $P(S_n = g_n|S_n \in G_{l_n})\cdot f_{g_n}(x)$ depends on g_n, i.e. everything else can be factorized. Iterating this procedure over g_t gives

$$
\begin{aligned}
&f_{(Y_1^{(\mathcal{G})},\dots,Y_n^{(\mathcal{G})})}(x_1,\dots,x_n)\\
&= \sum_{l_1,\dots,l_n=1}^{r} \Big(\sum_{g_1\in G_{l_1}} P\big(S_1=g_1\big)\cdot f_{g_1}(x_1)\cdot \prod_{t=2}^{n} \Big\{ \big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l_{t-1},l_t}\cdot \sum_{g_t\in G_{l_t}} P(S_t=g_t|S_t\in G_{l_t})\cdot f_{g_t}(x_t)\Big\}\Big)\\
&= \sum_{l_1,\dots,l_n=1}^{r} \Big(P\big(S_1^{(\mathcal{G})}=l_1\big)\cdot f_{l_1}^{(\mathcal{G})}(x_1)\cdot \prod_{t=2}^{n} \Big\{ \big(\mathbf{\Gamma}^{(\mathcal{G})}\big)_{l_{t-1},l_t}\cdot f_{l_t}^{(\mathcal{G})}(x_t)\Big\}\Big)\\
&= f_{(X_1^{(\mathcal{G})},\dots,X_n^{(\mathcal{G})})}(x_1,\dots,x_n).
\end{aligned}
$$

□

2 Clustering serially-dependent observations: Additional estimation results

All transition probability matrices are given in percent.

1. Five-state normal HMM with three independence clusters

As in Section 3.3 of Holzmann and Schwaiger (2014), we consider the following five-state bivariate normal HMM

$$\begin{gathered}\boldsymbol{\mu}_1 = \begin{pmatrix}2.5\\1.5\end{pmatrix}^T \quad \boldsymbol{\mu}_2 = \begin{pmatrix}3.5\\2\end{pmatrix}^T \quad \boldsymbol{\mu}_3 = \begin{pmatrix}2\\7\end{pmatrix}^T \quad \boldsymbol{\mu}_4 = \begin{pmatrix}3\\0.5\end{pmatrix}^T \quad \boldsymbol{\mu}_5 = \begin{pmatrix}2.5\\6\end{pmatrix}^T \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix}0.30 & 0.18\\0.18 & 0.30\end{pmatrix}\\ \boldsymbol{\Sigma}_2 = \begin{pmatrix}0.30 & -0.18\\-0.18 & 0.30\end{pmatrix} \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix}0.48 & -0.42\\-0.42 & 0.48\end{pmatrix} \quad \boldsymbol{\Sigma}_4 = \begin{pmatrix}1.20 & 0.27\\0.27 & 1.20\end{pmatrix} \quad \boldsymbol{\Sigma}_5 = \begin{pmatrix}0.5 & 0.4\\0.4 & 0.5\end{pmatrix},\\ \boldsymbol{\Gamma} = \begin{pmatrix}25.50 & 17.00 & 42.50 & 10.00 & 5.00\\25.50 & 17.00 & 42.50 & 10.00 & 5.00\\25.50 & 17.00 & 42.50 & 10.00 & 5.00\\6.00 & 4.00 & 10.00 & 70.00 & 10.00\\4.50 & 3.00 & 7.50 & 10.00 & 75.00\end{pmatrix}\end{gathered} \tag{5}$$

from which we simulate a series of length 1000.

The estimated BIC-optimal unrestricted five-state HMM is given by

$$\begin{gathered}\tilde{\boldsymbol{\mu}}_1 = \begin{pmatrix}2.51\\1.51\end{pmatrix}^T \quad \tilde{\boldsymbol{\mu}}_2 = \begin{pmatrix}3.57\\2.01\end{pmatrix}^T \quad \tilde{\boldsymbol{\mu}}_3 = \begin{pmatrix}2.03\\6.98\end{pmatrix}^T \quad \tilde{\boldsymbol{\mu}}_4 = \begin{pmatrix}2.94\\0.40\end{pmatrix}^T \quad \tilde{\boldsymbol{\mu}}_5 = \begin{pmatrix}2.48\\6.02\end{pmatrix}^T \quad \tilde{\boldsymbol{\Sigma}}_1 = \begin{pmatrix}0.20 & 0.14\\0.14 & 0.32\end{pmatrix}\\ \tilde{\boldsymbol{\Sigma}}_2 = \begin{pmatrix}0.29 & -0.23\\-0.23 & 0.32\end{pmatrix} \quad \tilde{\boldsymbol{\Sigma}}_3 = \begin{pmatrix}0.43 & -0.37\\-0.37 & 0.42\end{pmatrix} \quad \tilde{\boldsymbol{\Sigma}}_4 = \begin{pmatrix}1.09 & 0.13\\0.13 & 1.04\end{pmatrix} \quad \tilde{\boldsymbol{\Sigma}}_5 = \begin{pmatrix}0.56 & 0.42\\0.42 & 0.49\end{pmatrix},\\ \tilde{\boldsymbol{\Gamma}} = \begin{pmatrix}25.01 & 20.21 & 40.10 & 12.42 & 2.26\\21.95 & 19.54 & 46.37 & 9.72 & 2.43\\29.31 & 13.69 & 40.86 & 13.98 & 2.16\\6.04 & 3.14 & 10.62 & 70.27 & 9.94\\9.02 & 0.00 & 7.11 & 8.93 & 74.93\end{pmatrix}.\end{gathered}$$

Under the independence restriction $\mathcal{G}^* = \big\{\{1,2,3\},\{4\},\{5\}\big\}$, which has been found by backward selection, the estimated HMM is given by

$$\begin{gathered}\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix}2.48\\1.47\end{pmatrix}^T \quad \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix}3.53\\2.03\end{pmatrix}^T \quad \hat{\boldsymbol{\mu}}_3 = \begin{pmatrix}2.03\\6.98\end{pmatrix}^T \quad \hat{\boldsymbol{\mu}}_4 = \begin{pmatrix}2.94\\0.39\end{pmatrix}^T \quad \hat{\boldsymbol{\mu}}_5 = \begin{pmatrix}2.48\\6.02\end{pmatrix}^T \quad \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix}0.19 & 0.13\\0.13 & 0.30\end{pmatrix}\\ \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix}0.30 & -0.22\\-0.22 & 0.31\end{pmatrix} \quad \hat{\boldsymbol{\Sigma}}_3 = \begin{pmatrix}0.43 & -0.37\\-0.37 & 0.42\end{pmatrix} \quad \hat{\boldsymbol{\Sigma}}_4 = \begin{pmatrix}1.09 & 0.12\\0.12 & 1.03\end{pmatrix} \quad \hat{\boldsymbol{\Sigma}}_5 = \begin{pmatrix}0.56 & 0.43\\0.43 & 0.49\end{pmatrix},\\ \hat{\boldsymbol{\Gamma}} = \begin{pmatrix}26.00 & 17.25 & 41.71 & 12.82 & 2.22\\26.00 & 17.25 & 41.71 & 12.82 & 2.22\\26.00 & 17.25 & 41.71 & 12.82 & 2.22\\6.36 & 4.22 & 10.20 & 69.28 & 9.94\\4.72 & 3.13 & 7.57 & 9.67 & 74.91\end{pmatrix}.\end{gathered}$$

2. Two-state skew-normal HMM

Second, we consider the following two-state bivariate skew-normal HMM :

$$\begin{aligned}
&\boldsymbol{\Sigma}_1 = \begin{pmatrix} 4.80 & -0.48 \\ -0.48 & 1.20 \end{pmatrix},\ \boldsymbol{\Sigma}_2 = \begin{pmatrix} 4.0 & -0.4 \\ -0.4 & 1.0 \end{pmatrix}, \qquad \boldsymbol{\Gamma} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix} \\
&\boldsymbol{\alpha}_1 = \begin{pmatrix} 14 & -6 \end{pmatrix},\ \boldsymbol{\alpha}_2 = \begin{pmatrix} 14 & 0 \end{pmatrix}, \\
&\boldsymbol{\mu}_1 = \begin{pmatrix} -5.0 & 3.3 \end{pmatrix},\ \boldsymbol{\mu}_2 = \begin{pmatrix} -1.5 & 6.0 \end{pmatrix},
\end{aligned} \tag{6}$$

from which we simulate a series of length 5000.

The estimated BIC-optimal unrestricted five-state HMM is given by

$$\tilde{\boldsymbol{\mu}}_1 = \begin{pmatrix} -4.52 \\ 3.16 \end{pmatrix}^T \tilde{\boldsymbol{\mu}}_2 = \begin{pmatrix} -3.52 \\ 2.87 \end{pmatrix}^T \tilde{\boldsymbol{\mu}}_3 = \begin{pmatrix} -2.00 \\ 2.35 \end{pmatrix}^T \tilde{\boldsymbol{\mu}}_4 = \begin{pmatrix} -0.88 \\ 5.94 \end{pmatrix}^T \tilde{\boldsymbol{\mu}}_5 = \begin{pmatrix} 0.69 \\ 5.73 \end{pmatrix}^T \tilde{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 0.76 & 0.73 \\ 0.73 & 0.88 \end{pmatrix}$$
$$\tilde{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 1.08 & 0.62 \\ 0.62 & 0.89 \end{pmatrix} \tilde{\boldsymbol{\Sigma}}_3 = \begin{pmatrix} 2.57 & 0.35 \\ 0.35 & 0.84 \end{pmatrix} \tilde{\boldsymbol{\Sigma}}_4 = \begin{pmatrix} 0.16 & 0.01 \\ 0.01 & 0.94 \end{pmatrix} \tilde{\boldsymbol{\Sigma}}_5 = \begin{pmatrix} 1.33 & -0.08 \\ -0.08 & 1.03 \end{pmatrix},$$
$$\hat{\boldsymbol{\Gamma}}_{Nor} = \begin{pmatrix} 25.92 & 39.45 & 23.92 & 7.37 & 3.35 \\ 23.35 & 44.16 & 21.30 & 2.43 & 8.75 \\ 20.72 & 44.38 & 25.37 & 4.12 & 5.41 \\ 2.89 & 4.32 & 2.50 & 33.76 & 56.53 \\ 3.75 & 2.85 & 3.45 & 36.20 & 53.75 \end{pmatrix}.$$

Under the independence restriction $\mathcal{G}^* = \big\{\{1,2,3\},\{4,5\}\big\}$, which has been found by backward selection, the estimated HMM is given by

$$\hat{\boldsymbol{\mu}}_1 = \begin{pmatrix} -4.58 \\ 3.13 \end{pmatrix}^T \hat{\boldsymbol{\mu}}_2 = \begin{pmatrix} -3.54 \\ 2.89 \end{pmatrix}^T \hat{\boldsymbol{\mu}}_3 = \begin{pmatrix} -2.02 \\ 2.38 \end{pmatrix}^T \hat{\boldsymbol{\mu}}_4 = \begin{pmatrix} -0.88 \\ 5.93 \end{pmatrix}^T \hat{\boldsymbol{\mu}}_5 = \begin{pmatrix} 0.70 \\ 5.73 \end{pmatrix}^T \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 0.74 & 0.72 \\ 0.72 & 0.88 \end{pmatrix}$$
$$\hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 1.03 & 0.61 \\ 0.61 & 0.91 \end{pmatrix} \hat{\boldsymbol{\Sigma}}_3 = \begin{pmatrix} 2.53 & 0.34 \\ 0.34 & 0.85 \end{pmatrix} \hat{\boldsymbol{\Sigma}}_4 = \begin{pmatrix} 0.17 & 0.01 \\ 0.01 & 0.95 \end{pmatrix} \hat{\boldsymbol{\Sigma}}_5 = \begin{pmatrix} 1.33 & -0.08 \\ -0.08 & 1.02 \end{pmatrix},$$
$$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 89.37 \cdot \begin{pmatrix} 0.26 & 0.47 & 0.27 \\ 0.26 & 0.47 & 0.27 \\ 0.26 & 0.47 & 0.27 \end{pmatrix} & 10.63 \cdot \begin{pmatrix} 0.39 & 0.61 \\ 0.39 & 0.61 \\ 0.39 & 0.61 \end{pmatrix} \\ 09.92 \cdot \begin{pmatrix} 0.26 & 0.47 & 0.27 \\ 0.26 & 0.47 & 0.27 \end{pmatrix} & 90.08 \cdot \begin{pmatrix} 0.39 & 0.61 \\ 0.39 & 0.61 \end{pmatrix} \end{pmatrix}.$$

3 Additional simulation results for the backward selection algorithm to determine the independence partition

Five-state normal HMM from (5)

We consider the five-state normal HMM (5), and perform the backward selection algorithm for series of length of 500. The results are given in Table 1 below, compare to table 2 of the paper.

length T	500			
level/ind. clusters	2	3	4	5
10%	-	726	227	47
5%	-	809	176	15
1%	1	908	90	1

Table 1: Simulation results of backward selection under the normal HMM (5): Absolute frequency of selected sets in independence cluster according to used level. $M = 1.000$ repetitions for length of $T = 500$.

Five-state normal HMM from with distinct t.p.m.

We consider a slightly distinct five-state normal HMM. We keep the state-dependent parameters as in (5), but use instead the t.p.m.

$$\mathbf{\Gamma} = \begin{pmatrix} 3.00 & 2.00 & 5.00 & 85.00 & 5.00 \\ 3.00 & 2.00 & 5.00 & 85.00 & 5.00 \\ 3.00 & 2.00 & 5.00 & 85.00 & 5.00 \\ 6.00 & 4.00 & 10.00 & 70.00 & 10.00 \\ 4.50 & 3.00 & 7.50 & 10.00 & 75.00 \end{pmatrix} \tag{7}$$

The results are given in Tables 2 and 3. Since the states 1 - 3 and 4 of the Markov chain are somewhat similar, they are several instances, in particular for the shorter series, where these states cannot be separated.

length T	500			
level/ind. clusters	2	3	4	5
10%	415	483	81	21
5%	500	436	50	14
1%	650	320	25	5

Table 2: Simulation results of backward selection under the normal HMM with t.p.m. (7): Absolute frequency of selected sets in independence cluster according to used level. $M = 1.000$ repetitions for lenght of $T = 500$.

length T	1.000				2.500				5.000			
level/ind. clusters	2	3	4	5	2	3	4	5	2	3	4	5
10%	130	718	116	36	-	733	146	121	-	428	61	11
5%	170	723	86	21	1	802	113	84	-	455	41	4
1%	288	668	39	5	4	880	48	68	-	471	29	-

Table 3: Simulation results of backward selection under a normal HMM: Absolute frequency of selected sets in independence cluster according to used level. $M = 1.000$ repetitions for lenghts of $T = 1.000$ and 2.500; $M = 500$ repetitions for lenght $T = 5.000$.

4 Clustering results for the algorithm of Volant et al. (2013)

The function `HMMmix` of the R package `HMMmix` published by Volant and Berard on CRAN, implements the procedure of estimating an HMM with normal-mixtures as state-dependent distributions as proposed in Volant et al. (2013). For an application two parameters have to be specified: The number of initial states K, and the number of clusters D, whereby D can be seen as a lower bound for the number of clusters, since also the model with the ICL_S-optimal number of clusters is estimated.

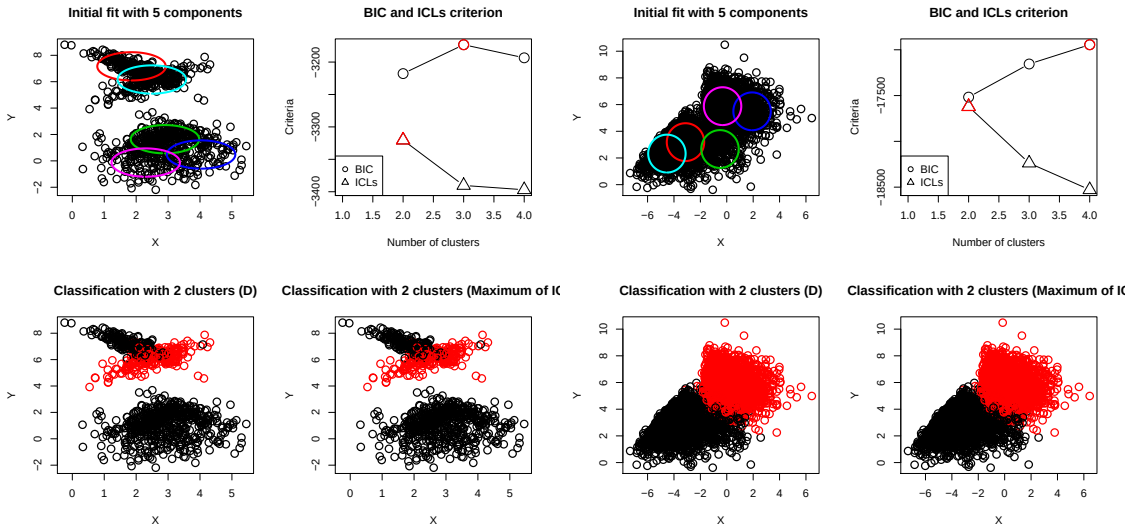
We apply this function to both of our simulated examples with parameters $K = 5$, which is the BIC-optimal number of states in both cases, and $D = 2$ to analyze how many clusters are selected by ICL_S. .

1. Five-state normal HMM with three independence clusters

For the series of length 1000 from the five-state normal HMM in (5) ICL_S selects two clusters, which is contrary to our result and interpretation of the true model - only states one and two should be combined because both the dependence structure and the shape of the state-dependent distributions are suitable for merging. See Figure 1 (a.) for the output of the function `HMMmix`.

2. Two-state skew-normal HMM

For the series of length 5000 from the two-state skew-normal HMM in (6) the result of ICL_S coincides with our result of two clusters, see Figure 1 (b.)



(a) normal HMM (b) skew-normal HMM

Figure 1: Results of applying the R package HMMmix to the simulated datasets.

5 Clustering results for the algorithm of Baudry et al. (2013) for the skew normal example

For the series of length 5000 from the two-state skew-normal HMM in (6) we also applied applied the algorithm of Baudry et al. (2013) to investigate the relevance of the serial dependence. We used the R package `mclust` in the version 4.2[1], i.e. we at first use a normal mixture model to fit the data and then apply the merging algorithm of Baudry et al. (2010) to combine the components. First we select the number of components via BIC using the function `mclust`, see Figure 2. The result are six components and the corresponding estimated parameters. Then the function `clustCombi` can be used to combine step-wise components to clusters. The entropy- and the normalized-entropy plot both find two clusters, see Figure 3 for the normalized-entropy plot.

Figure 3b shows the typical hard cut between clusters of a maximum a-posteriori analysis using an i.i.d. model. For assigning data points lying in the region between two clusters the information of the serial dependence can be beneficial. In the current example using the maximum a-posteriori clustering of the mixture-model 92 of 5.000 observations are wrongly classified, whereas using the merged HMM 49 are wrongly classified. Thus, both methods perform well, but as expected the merged HMM can recover more observations which lie between the two clusters.

[1] see MCLUST Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation, Technical Report no. 597, Department of Statistics, University of Washington, June 2012

6 An alternative merging algorithm based on the ridgeline

We propose a modification of the merging algorithm in Section 3.2 of Holzmann and Schwaiger (2014) for Gaussian HMMs, which is mode-based and not based on entropies.

Input: The observed series $x_1, \ldots, x_T$ and the parametric family of the state dependent densities $f(\cdot, \theta)$.

Step 1 Select and fit an appropriate finite-state HMM with state dependent densities from $f(\cdot, \theta)$, e.g. by using the BIC (or possibly the AIC). Denote the number of states of the selected HMM by k.

Step 2 Determine the independence partition $\mathcal{G}^* = \{G_1, \ldots, G_r\}$ of the selected HMM with k states using a backward selection based on the p-values of the test in Theorem 3 in Holzmann and Schwaiger (2014), according to a certain significance level (say 0.05 or 0.1).

We let $\hat{\Gamma}$ and $\hat{\theta}_1, \ldots, \hat{\theta}_k$ denote the parameters of the ML-fit under the independence restrictions given by $\mathcal{G}^*$, so that $\hat{\Gamma}$ is a $k \times k$-t.p.m. for which $\lambda_{\mathcal{G}^*}(\hat{\Gamma}) = \hat{\Gamma}$, and we let $(X_t, S_t)_t$ denote a k-state HMM with these parameters.

Step 3 For the states in each element of the independence partition $\mathcal{G}^*$, we apply separately one of the ridgeline merging algorithms of Henning (2010), either the ridgeline unimodal method or the ridgeline ratio method. Thus, only merges within each element of the independence partition are allowed.

Indeed, all local extrema of Gaussian mixtures occur along the ridgeline, see Ray and Lindsay (2005) or Hennig (2010). The ridgeline ratio for a two-component normal mixture is defined as 1 if the mixture is unimodal, and as the ratio between the minimum of the density along the ridgeline between the exterior modes and the value of the density at the minimal mode in case of more than one mode.

1. Five-state normal HMM with three independence clusters

For the series of length 1000 from the five-state normal HMM in (5), a merge based on modality within the independence cluster $\{1, 2, 3\}$ can only be possible for states 1 and 2, all other ridgeline ratios are zero. Figure 4 plots the ridgeline along the corresponding two-component mixture (the weights being $(0.255, 0.17)/(0.255+0.17)$) as well as the value of the mixture density, along the parameter α which parametrizes the ridgeline.

The mixture is bimodal, with ratio between the second largest modes and the minimum between the two modes being 0.95. Thus, states 1 and 2 cannot be merged based on the strict ridgeline unimodal method, but they can be merged based on the ridgeline ratio method with parameter say $r = 0.9$ (Hennig 2010 recommends 0.2).

2. Two-state skew-normal HMM

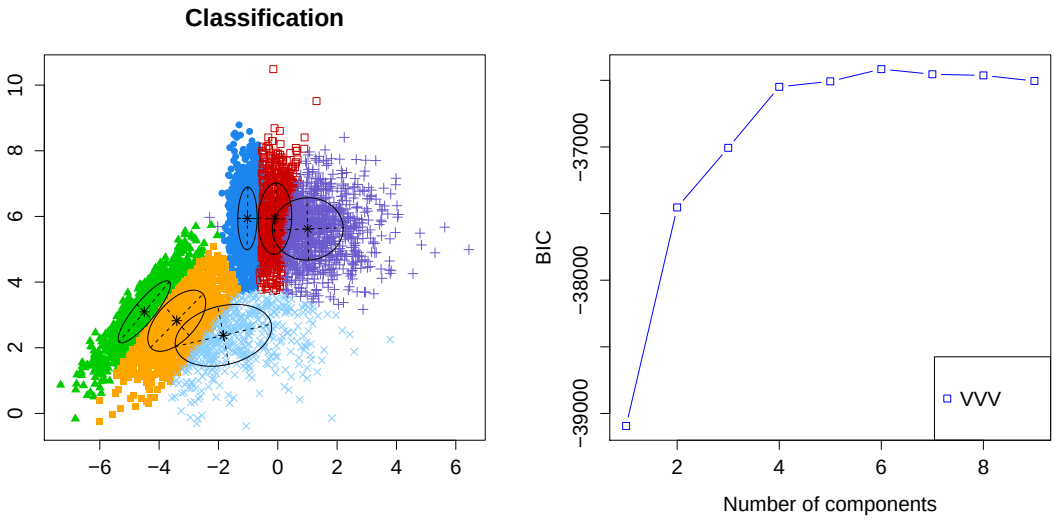
For the series of length 5000 from the two-state skew-normal HMM in (6), we obtain the five-state HMM with independence partition $\mathcal{G} = \{\{1, 2, 3, \}, \{4, 5\}\}$. First, consider the states 4 and 5. The resulting two-component normal mixture is bimodal, with a ridgeline ratio of 0.82. For the states $\{1, 2, 3, \}$, the two-component mixture of states 2 and 3 (with

weights $\pi = 0.27/(0.27 + 0.47$ for states 2 and $1 - \pi$ for state 3) is unimodal (the others being bimodal), so that these two states are merged in the first step. Next, we form the mean vector and the covariance matrix of the resulting two-component normal mixture of states 2 and 3, which are given by (vectors are taken as column vectors)

$$\mu_{mix} = \pi\,\mu_2 + (1-\pi)\,\mu_3, \qquad \Sigma_{mix} = \pi\big(\Sigma_2 + \mu_2\,\mu_2^T\big) + (1-\pi)\,\big(\Sigma_3 + \mu_3\,\mu_3^T\big) - \mu_{mix}\mu_{mix}^T.$$

Then the ridgeline is formed for the parameters (μ_1, Σ_1), $(\mu_{mix}, \Sigma_{mix})$ and weights $(0.26, 0.74)$, yielding a unimodal density.

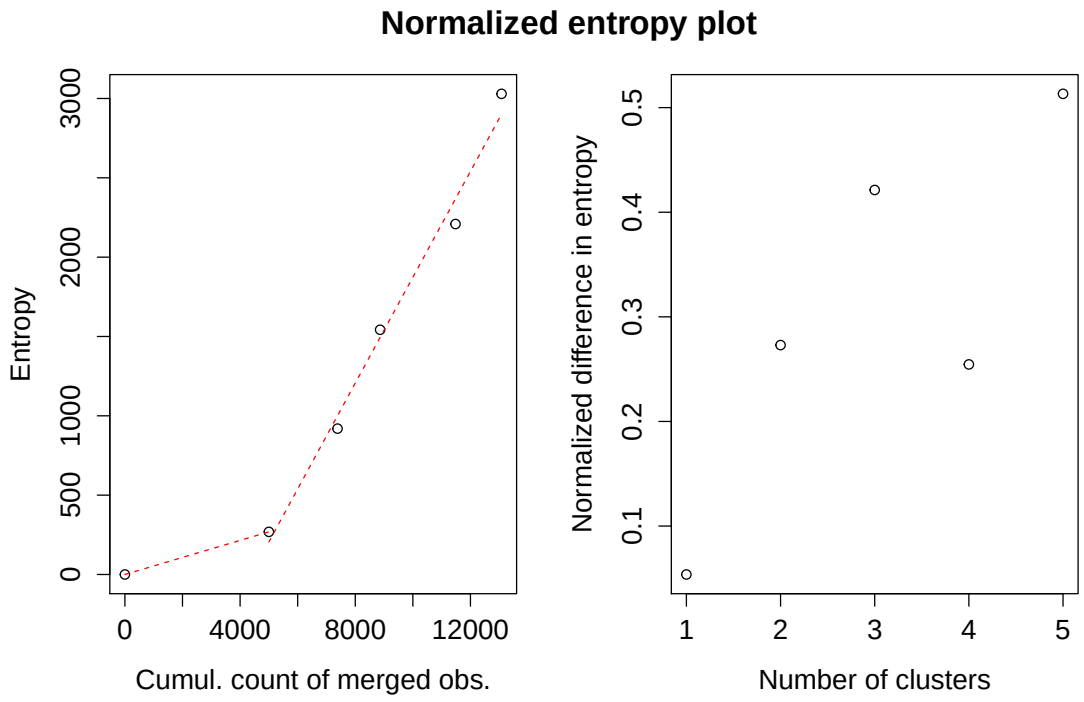
Thus, the ridgeline ratio method with tuning parameter 0.8 yields two clusters. The ridgeline unimodal method seems not to work that well, it retains too many clusters, corresponding to very small (artifical) modes of the estimated density.



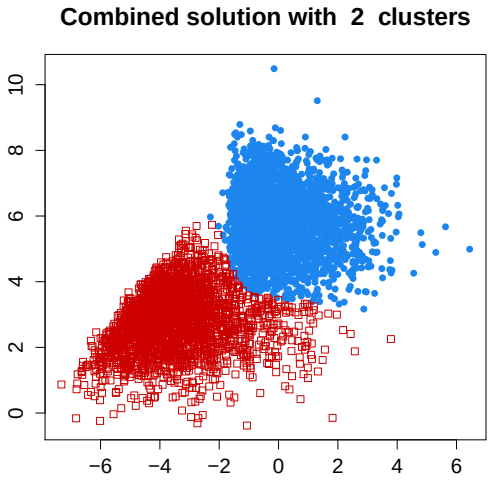
(a) visualization of the mixture model

(b) BIC values

Figure 2: Results of applying model selection of the R package `mclust`.



(a) cluster selection



(b) clustering result

Figure 3: Results of applying the function `clustCombi` of the R package `mclust`.

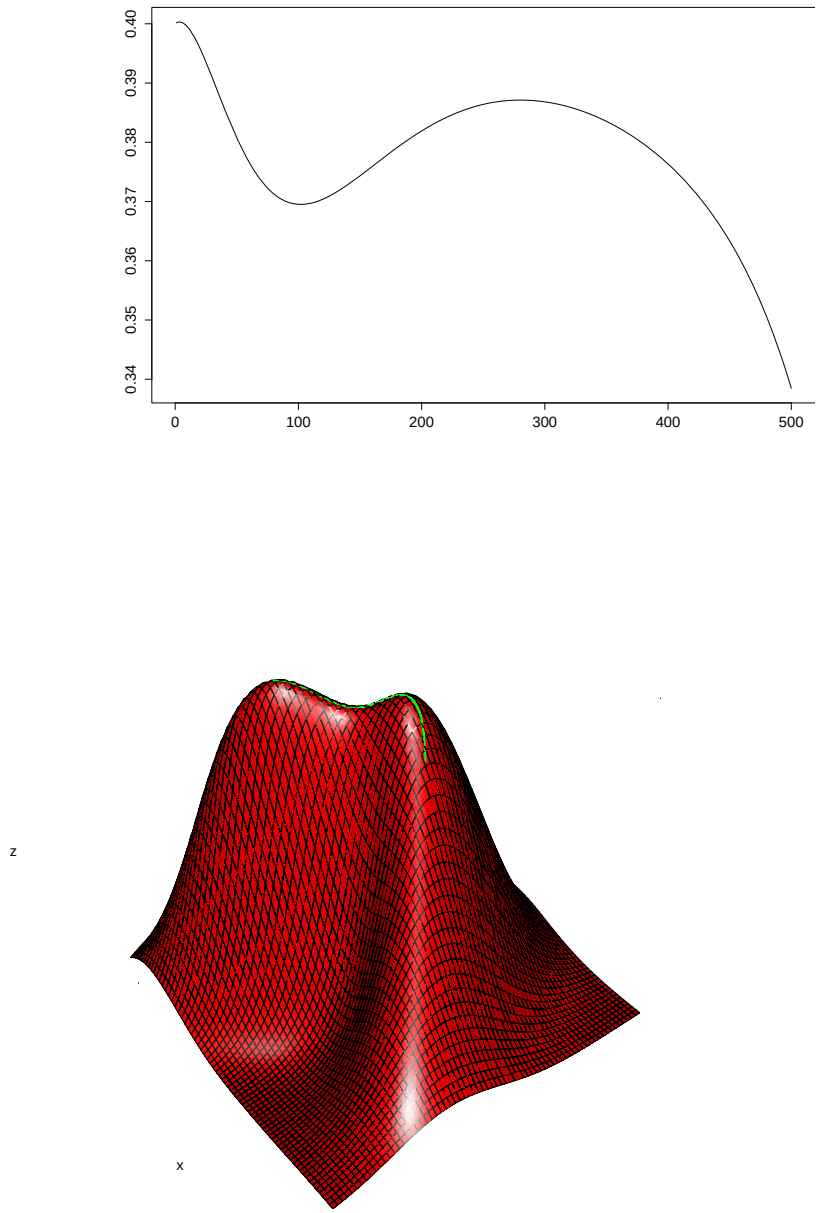


Figure 4: Ridgeline plots.