WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Finite Mixture and Markov Switching Models

Sylvia Frühwirth-Schnatter

Western Swiss Doctoral School in Statistics and Probability, February 2020

Part I

Finite Mixture Models and Model-based Clustering

Part II

Hidden Markov and Markov Switching Models

Outline

Part II: Hidden Markov and Markov Switching Models

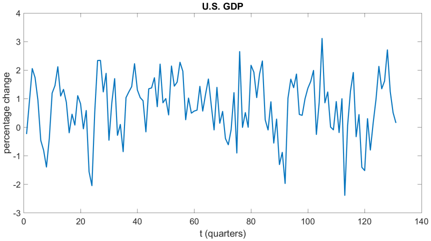
- Motivating Example
- Finite Markov mixture models
- Markov mixture modelling
- Bayesian Inference
- Applications
  - U.S. GDP Data – Markov switching autoregressive models
  - U.S./U.K. real exchange rate – Switching state space models
  - NYSE Data – Switching ARCH Model

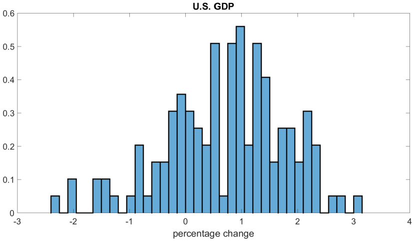
Motivating Example

- Consider the quarterly percentage growth rate

$$Y_t = 100(\log(\mathrm{GDP}_t) - \log(\mathrm{GDP}_{t-1}))$$

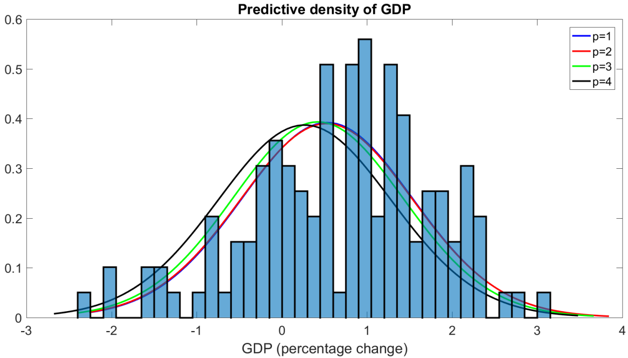
  of the U.S. real GDP series, for $t = 1, \ldots, T$.
- Quarterly data 1951.II to 1984.IV Time series plot of y_t (left) and empirical marginal distribution of y_t (right)





Fitting AR(p) models to the GDP data

- GDP data, modeled by an AR(p) model with $p = 1, \ldots, 4$



- Unimodal stationary distribution $p(Y_t|p)$ (full line) implied by each AR(p) model. .
- Surprisingly little difference in $p(Y_t|p)$ for the different model orders p.
- Striking difference to the bi-/multi-modality empirical histogram of Y_t.

Capturing multimodality through finite mixtures

- Introduce a hidden indicator S_t, where $\Pr(S_t = k) = \eta_k$, $k = 1, \ldots, K$.
- Define the conditional distribution of Y_t given S_t, e.g.

$$Y_t | S_t = k \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right).$$

- S_t models to which group (state) observation Y_t belongs.
- A finite mixture distribution results as marginal distribution:

$$p(y_t) = \eta_1 f_N(y_t; \mu_1, \sigma_1^2) + \cdots + \eta_K f_N(y_t; \mu_K, \sigma_K^2).$$

- See [Frühwirth-Schnatter, 2006] and [Kaufmann, 2019] for a review.

Finite mixture modeling of time series

- In a time series application, $\mathbf{S} = (S_1, \ldots, S_T)$ is a time series of discrete indicators.
- However, for standard finite mixture distributions, successive values are independent:

$$\xi_{jk} = \Pr(S_t = k | S_{t-1} = j) = \Pr(S_t = k) = \eta_k.$$

- The implied marginal distribution of Y_t could be multimodal, but marginally Y_t is a white noise process (uncorrelated over time).
- To capture both multimodality and autocorrelation for time series, Markov switching models have been developed.

Outline

Part II: Hidden Markov and Markov Switching Models

- Motivating Example
- Finite Markov mixture models
- Markov mixture modelling
- Bayesian Inference
- Applications
  - U.S. GDP Data – Markov switching autoregressive models
  - U.S./U.K. real exchange rate – Switching state space models
  - NYSE Data – Switching ARCH Model

Finite Markov mixture models

- The probability distribution of the stochastic process Y_t depends on the states of a hidden discrete stochastic process S_t.
- The stochastic process Y_t is directly observable.
- S_t is a latent random process that is observable only indirectly through the effect it has on the realizations of Y_t.
- This leads to a rich class of **nonlinear time series models.**
- The hidden process $\{S_t\}_{t=0}^{T}$ is an irreducible, aperiodic Markov chain of order one starting from its ergodic distribution $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$:

$$\Pr(S_0 = k|\boldsymbol{\xi}) = \eta_k.$$

The hidden Markov chain

- The properties of S_t are described by the $(K \times K)$ transition matrix $\boldsymbol{\xi}$:

$$\boldsymbol{\xi} = \begin{pmatrix} \xi_{11} & \cdots & \xi_{1K} \\ \vdots & \ddots & \vdots \\ \xi_{K1} & \cdots & \xi_{KK} \end{pmatrix}$$

- Each element ξ_{jk} is equal to the transition probability from state j to state k:

$$\xi_{jk} = \Pr(S_t = k | S_{t-1} = j), \qquad \forall j, k \in \{1, \ldots, K\}.$$

- The jth row of the transition matrix $\boldsymbol{\xi}$ defines the conditional distribution $S_t | S_{t-1} = j$ of S_t given the information that S_{t-1} is in state j (for all $t = 1, \ldots, T$).

The finite mixture model as a special case

- A random variable Y_t drawn from a standard finite mixture with weight distribution $\boldsymbol{\eta}$ is observationally equivalent with a process Y_t generated by a finite Markov mixture distribution where all rows of the transition matrix of S_t are identical to $\boldsymbol{\eta}$:

$$\boldsymbol{\xi} = \begin{pmatrix} \xi_{11} & \cdots & \xi_{1K} \\ \vdots & \ddots & \vdots \\ \xi_{K1} & \cdots & \xi_{KK} \end{pmatrix} = \begin{pmatrix} \xi_{1} & \cdots & \xi_{K} \\ \vdots & \ddots & \vdots \\ \xi_{1} & \cdots & \xi_{K} \end{pmatrix}$$

The invariant distribution of a K-state Markov chain

- Any probability distribution $\boldsymbol{\eta} = (\eta_1, \dots, \eta_K)$ that fulfills the **invariance property**

$$\boldsymbol{\xi}^{'}\boldsymbol{\eta} = \boldsymbol{\eta}, \tag{8}$$

- is called an invariant distribution of the Markov chain S_t .
- If the states of S_{t-1} are drawn from an invariant distribution of $\boldsymbol{\xi}$, then

$$\Pr(S_t = k|\boldsymbol{\xi}) = \sum_{j=1}^{K} \Pr(S_t = k|S_{t-1} = j, \boldsymbol{\xi})\Pr(S_{t-1} = j|\boldsymbol{\xi}) = \sum_{j=1}^{K} \xi_{jk}\eta_j = \eta_k,$$

and $\Pr(S_t = k|\boldsymbol{\xi})$ is again equal to $\boldsymbol{\eta}$.

The invariant distribution of a K-state Markov chain

- It is possible to show that such an invariant distribution *exists* for any finite Markov chain.
- For $K > 2$, numerical methods have to be used for solving (8) in $\boldsymbol{\eta}$.
- The invariant distribution is not unique for arbitrary transition matrices.
- E.g., for $\boldsymbol{\xi} = \mathbf{I}_K$ any arbitrary probability distribution $\boldsymbol{\eta}$ is invariant.

The long run behaviour of a Markov chain

- Consider the hth power of $\boldsymbol{\xi}$:

$$\boldsymbol{\xi}^h = \underbrace{\boldsymbol{\xi} \cdots \boldsymbol{\xi}}_{h \quad \text{times}} .$$

- Interpretation of the element (k, ℓ) of $\boldsymbol{\xi}^h$:

$$(\boldsymbol{\xi}^h)_{k\ell} = \Pr(S_{t+h} = \ell | S_t = k, \boldsymbol{\xi}),$$

  i.e. probability to end up in ℓ after h periods, given a start in k (what happens in between does not matter).
- The kth row of $\boldsymbol{\xi}^t$ is the distribution $\Pr(S_t | \boldsymbol{\xi}, S_0 = k)$
- $\boldsymbol{\xi}^h$ determines the long-run behavior of the Markov chain.

Irreducibility of a Markov chain

- Uniqueness of the invariant distribution follows for any transition matrix that leads to an *irreducible* Markov chain.
- Irreducibility means that starting S_t from an arbitrary state $k \in \{1, \ldots, K\}$
- any state $\ell \in \{1, \ldots, K\}$ must be reachable in finite time:

$$\forall (k, \ell) \in \{1, \ldots, K\} \quad \Rightarrow \quad \exists h(k, \ell) : (\boldsymbol{\xi}^{h(k,\ell)})_{k\ell} > 0.$$

- Sufficient condition for irreducibility: $(\boldsymbol{\xi}^h)_{k\ell} > 0$ for some $h \geq 1$ independent of k, ℓ.
- E.g. all elements $\xi_{k\ell}$ of $\boldsymbol{\xi}$ are positive.

Reducible Markov chains

- If any element $(\boldsymbol{\xi}^h)_{k\ell} \equiv 0$ for all $h \geq 1$, then the Markov chain is reducible.
- E.g., transition matrix of a change point model:

$$\boldsymbol{\xi} = \begin{pmatrix} \xi_{11} & \xi_{12} & \xi_{13} & \xi_{14} \\ 0 & \xi_{22} & \xi_{23} & \xi_{24} \\ 0 & 0 & \xi_{33} & \xi_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

- $(\boldsymbol{\xi}^h)_{k\ell} \equiv 0$ for all $\ell < k$ for all $h \geq 1$.
- e.g. for $K = 2$

$$\boldsymbol{\xi} = \begin{pmatrix} \xi_{11} & 1 - \xi_{11} \\ 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\xi}^h = \begin{pmatrix} \xi_{11}^h & 1 - \xi_{11}^h \\ 0 & 1 \end{pmatrix}.$$

Aperiodic Markov chains

- Consider, for each state k, *all h for which $(\xi^h)_{kk} > 0$*.
- The period of state k is the greatest common divisor (GCD) of all h.
- A Markov chain is aperiodic, if the period of each state is equal to one:

$$\mathrm{GCD}\{h \geq 1 : (\xi^h)_{kk} > 0\} = 1, \qquad \forall k \in \{1, \ldots, K\}.$$

- Less formally, aperiodicity is defined as the absence of periodicity.
- Sufficient condition: a Markov chain is aperiodic, if all diagonal elements of ξ are positive.

An example of a periodic Markov chain

- Consider following irreducible transition matrix ξ:

$$\xi = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad \xi^3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$

- The invariant distribution is unique (due to irreducibility) and equal to the uniform distribution.
- The period of each state is equal to 3, e.g. if $S_0 = 1$:

$$\begin{aligned} \Pr(S_t = 1 | S_0 = 1, \xi) = 1, & \qquad \text{iff } t = 3, 6, 9, \ldots \\ \Pr(S_t = 2 | S_0 = 1, \xi) = 1, & \qquad \text{iff } t = 1, 4, 7, \ldots \\ \Pr(S_t = 3 | S_0 = 1, \xi) = 1, & \qquad \text{iff } t = 2, 5, 8, \ldots \end{aligned}$$

- The distribution $\Pr(S_t | S_0, \xi)$ does not converge to the invariant distribution.

Irreducible Aperiodic Markov Chains

- Ergodicity:

Ergodicity of a Markov chain

For an *ergodic* Markov chain,

- the invariant distribution $\boldsymbol{\eta}$ is unique (called ergodic distribution);
- the distribution $\mathrm{Pr}(S_t|\boldsymbol{\xi}, S_0 = k)$ converges to the invariant distribution, regardless of the state k the initial value S_0.

- A Markov chain is ergodic, if the transition matrix $\boldsymbol{\xi}$ is *irreducible* and *aperiodic*.

Two-state Markov chains

- Consider, for illustration, a two-state Markov chain with transition matrix

$$\boldsymbol{\xi} = \begin{pmatrix} \xi_{11} & 1-\xi_{11} \\ 1-\xi_{22} & \xi_{22} \end{pmatrix}.$$

- The invariant probability distribution $\boldsymbol{\eta} = (\eta_1, \eta_2)$ given by:

$$\eta_1 = \frac{\xi_{21}}{\xi_{12}+\xi_{21}}, \qquad \eta_2 = \frac{\xi_{12}}{\xi_{12}+\xi_{21}}.$$

  - For a „symmetric" Markov chain with $\xi_{11} = \xi_{22}$, the invariant probability distribution is uniform: $\eta_1 = \eta_2 = 0.5$;
  - For an „asymmetric" Markov chain $\xi_{11} > \xi_{22}$ favors state 1: $\eta_1 > \eta_2$, whereas $\xi_{11} < \xi_{22}$ favors state 2: $\eta_1 < \eta_2$.
- A two-state Markov chain is ergodic, if $0 < \xi_{11} + \xi_{22} < 2$.
- In the long-run, an ergodic Markov chain converges from any initial state S_0 to $\boldsymbol{\eta}$.

Two-state Markov chain

- State persistence depends on the eigenvalues of $\boldsymbol{\xi}$, obtained from

$$\begin{vmatrix} \xi_{11}-\lambda & 1-\xi_{11} \\ 1-\xi_{22} & \xi_{22}-\lambda \end{vmatrix} = (\lambda-1)(\lambda-(\xi_{11}+\xi_{22}-1)) = 0.$$

- One eigenvalue is equal to 1, the second eigenvalue is equal to:

$$\lambda = \xi_{11} - (1-\xi_{22}) = \xi_{11} - \xi_{21}.$$

- Representation of $\boldsymbol{\xi}^h$ in terms of the invariant probability distribution is possible:

$$\boldsymbol{\xi}^h = \begin{pmatrix} \eta_1 & \eta_2 \\ \eta_1 & \eta_2 \end{pmatrix} + \lambda^h \begin{pmatrix} \eta_2 & -\eta_2 \\ -\eta_1 & \eta_1 \end{pmatrix},$$

with λ being the second eigenvalue of $\boldsymbol{\xi}$.
- Persistence of S_t is higher, the closer λ is to 1.

Outline

Part II: Hidden Markov and Markov Switching Models

- Motivating Example
- Finite Markov mixture models
- Markov mixture modelling
- Bayesian Inference
- Applications
  - U.S. GDP Data – Markov switching autoregressive models
  - U.S./U.K. real exchange rate – Switching state space models
  - NYSE Data – Switching ARCH Model

The Basic Markov mixture model

- Conditional on knowing $\mathbf{S} = (S_0, \ldots, S_T)$, the random variables $Y_1, \ldots, Y_T$ are stochastically independent.
- The distribution of Y_t arises from one out of K distributions with density $p(y_t|\boldsymbol{\theta}_1), \ldots, p(y_t|\boldsymbol{\theta}_K)$, depending on the state of S_t:

$$Y_t|S_t = k \sim p(y_t|\boldsymbol{\theta}_k).$$

- The unconditional distribution of Y_t is a finite mixture distribution with the ergodic distribution $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ acting as weight distribution:

$$p(y_t|\boldsymbol{\vartheta}) = \sum_{k=1}^{K} \eta_k p(y_t|\boldsymbol{\theta}_k),$$

Markov mixture of two normal distributions

- Markov mixture models are able to generate time series data with asymmetry and fat tails in the marginal distribution [Timmermann, 2000].
- Consider a Markov mixture of two normal distributions:

$$Y_t = \begin{cases} \mu_1 + \varepsilon_t, & \varepsilon_t \sim \mathcal{N}\left(0, \sigma_1^2\right), & S_t = 1, \\ \mu_2 + \varepsilon_t, & \varepsilon_t \sim \mathcal{N}\left(0, \sigma_2^2\right), & S_t = 2. \end{cases}$$

- Multimodality of the marginal distribution is possible for appropriate choices of $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \xi_{11}, \xi_{21})$ [Ray and Lindsay, 2005].

Capturing Asymmetry

- Coefficient of skewness in the marginal distribution of Y_t, with $\mu = \mathrm{E}(Y_t|\boldsymbol{\vartheta})$ and $\sigma^2 = \mathrm{Var}(Y_t|\boldsymbol{\vartheta})$:

$$\frac{\mathrm{E}((Y_t-\mu)^3|\boldsymbol{\vartheta})}{\mathrm{E}((Y_t-\mu)^2|\boldsymbol{\vartheta})^{3/2}} = \eta_1\eta_2(\mu_1-\mu_2)\frac{3(\sigma_2^2-\sigma_1^2)^2+(\eta_2-\eta_1)(\mu_2-\mu_1)^2}{\sigma^3},$$

- No skewness is present, if the means are the same ($\mu_1 = \mu_2$).
- Skewness is present whenever both the means and the variances are different.
- If the means are different ($\mu_1 \neq \mu_2$), but the variances the same ($\sigma_1 = \sigma_2$), asymmetry is introduced only through asymmetry in the persistence probabilities, because $\eta_1 \neq \eta_2$ iff $\xi_{11} \neq \xi_{22}$.

Excess kurtosis

- Excess kurtosis is given by

$$\frac{\mathrm{E}((Y_t-\mu)^4|\boldsymbol{\vartheta})}{\mathrm{E}((Y_t-\mu)^2|\boldsymbol{\vartheta})^2}-3=\eta_1\eta_2\frac{3(\sigma_2^2-\sigma_1^2)^2+c(\mu_1,\mu_2)}{\sigma^4},$$

where $c(\mu_1,\mu_2)=6(\eta_1-\eta_2)(\sigma_2^2-\sigma_1^2)(\mu_2-\mu_1)^2+(\mu_2-\mu_1)^4(1-6\eta_1\eta_2)$.

- If $\mu_1=\mu_2$, then the marginal distribution has fatter tails than a normal distribution as long as $\sigma_1^2\neq\sigma_2^2$.

Capturing autocorrelation

- A finite Markov mixture model might generate an autocorrelated process Y_t, even if the process Y_t is uncorrelated conditional on knowing S_t.
- Autocorrelation in the marginal process Y_t, where S_t is unknown, enters through persistence in S_t.
- Note that Y_t, in contrast to S_t, is no longer a Markov process of first order.

Example: Autocorrelation for a Two-State Model

- Autocorrelation function of a two-state Markov mixture model:

$$\rho_{Y_t}(h|\boldsymbol{\vartheta}) = \frac{\mathrm{E}(Y_t Y_{t+h}|\boldsymbol{\vartheta}) - \mu^2}{\sigma^2} = \frac{\eta_1\eta_2(\mu_1 - \mu_2)^2}{\sigma^2}\lambda^h, \tag{9}$$

  with $\lambda = \xi_{11} + \xi_{22} - 1$ being the second largest eigenvalue of $\boldsymbol{\xi}$.
- No autocorrelation in Y_t is present if $\mu_1 = \mu_2$.
- Autocorrelation of Y_t is introduced through the hidden Markov chain S_t, whenever $\xi_{11} + \xi_{22} \neq 1$.
- The process Y_t exhibits positive autocorrelation provided that $\xi_{11} + \xi_{22} > 1$.

Autocorrelation Function of the Squared Process

- Finite Markov mixture models might generates processes with Y_t^2 being autocorrelated.
- E.g., for a Markov mixture of two normal distributions:

$$\rho_{Y_t^2}(h|\boldsymbol{\vartheta}) = \frac{\eta_1\eta_2(\mu_1^2 - \mu_2^2 + \sigma_1^2 - \sigma_2^2)^2}{\mathrm{E}(Y_t^4|\boldsymbol{\vartheta}) - \mathrm{E}(Y_t^2|\boldsymbol{\vartheta})^2}\lambda^h. \tag{10}$$

- Y_t^2 exhibits positive autocorrelation provided that $\xi_{11} + \xi_{22} > 1$.
- Interestingly, state dependent variances are neither necessary nor sufficient for autocorrelation in the squared process.
- Even if $\sigma_1^2 = \sigma_2^2$, Y_t shows conditional heteroscedasticity, as long as S_t does not degenerate to an $\mathrm{i.i.d.}$ process.

Relation to ARMA Models

- There exists a close relationship between Markov mixture models and non-normal ARMA models.
- For a two-state Markov mixture model, for instance, the autocorrelation function of Y_t given in (9) fulfills, for $h > 1$, the following recursion,

$$\rho_{Y_t}(h|\boldsymbol{\vartheta}) = \lambda \rho_{Y_t}(h-1|\boldsymbol{\vartheta}),$$

- This corresponds to the autocorrelation function of an ARMA$(1,1)$ process, whereas the nonnormality of the unconditional distribution of Y_t is preserved through the mixture distribution.
- In general, [Poskitt and Chung, 1996] proved for a univariate K-state hidden Markov chain $Y_t = \mu_{S_t} + u_t$ the existence of an ARMA$(K-1, K-1)$ representation with a homogeneous zero-mean white noise process.

Outline

Part II: Hidden Markov and Markov Switching Models

- Motivating Example
- Finite Markov mixture models
- Markov mixture modelling
- Bayesian Inference
- Applications
  - U.S. GDP Data – Markov switching autoregressive models
  - U.S./U.K. real exchange rate – Switching state space models
  - NYSE Data – Switching ARCH Model

Difficulties with ML estimation for mixture models

- The sample size T has to be very large, before asymptotic theory of ML applies.
- Regularity conditions are often violated (overfitting mixtures with too many states, zero transitions between certain states).
- The provision of standard errors is not straightforward in particular when using the EM algorithm (singularity of the matrix of second partial derivatives of the log likelihood function)
- Mixtures of normal distributions with switching variances:
  - the mixture likelihood is unbounded,
  - the ML estimator as a global maximizer of the likelihood function does not exist,
  - it usually exists as a local maximizer
  - difficult to find this local maximum and to avoid spurious modes in the course of maximizing the log likelihood function.
- see also [McLachlan and Peel, 2000].

Prior distributions

- Invariant, possibly hierarchical priors for $\boldsymbol{\theta}_k$:

$$p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K | K) = p(\boldsymbol{\delta}) \prod_{k=1}^{K} p(\boldsymbol{\theta}_k | \boldsymbol{\delta}).$$

- For random hyperparameters $\boldsymbol{\delta}$, a hierarchical prior $p(\boldsymbol{\delta})$ is employed
- Yields a joint marginal prior $p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K | K)$
- Each row of the transition matrix follows a Dirichlet distribution:

$$\boldsymbol{\xi}_{k,\cdot} \sim \mathcal{D}\left(e_{k1}^0, \ldots, e_{kK}^0\right),$$

  where $e_{kk}^0 \equiv e_p > 0$ for all k and $e_{kj}^0 \equiv e_t > 0$ for all $k \neq j$ to ensure invariance with respect to relabelling the states of S_i.
- Prior distribution of the initial value S_0:
  - equal to the ergodic distribution $\boldsymbol{\eta}_{\xi}$ corresponding to the transition matrix $\boldsymbol{\xi}$.
  - or assumed to be uniform.

MCMC Estimation

Gibbs sampling for Markov mixture models [Frühwirth-Schnatter, 2006]

Choose path $\mathbf{S}^{(0)}$ and repeat for $m = 1, \ldots, M_0, \ldots, M + M_0$:

(a) Parameter estimation conditional on the classification $\mathbf{S}^{(m-1)}$:

  (a1) Sample the model parameter $\boldsymbol{\theta}_1^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)}$ from the complete-data posterior $p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K | \mathbf{y}, \mathbf{S}^{(m-1)})$.

  (a2) Sample the transition matrix $\boldsymbol{\xi}^{(m)}$ from the complete-data posterior distribution $p(\boldsymbol{\xi} | \mathbf{S}^{(m-1)})$.

(b) State simulation conditional on knowing $\boldsymbol{\vartheta}^{(m)}$ by sampling a path $\mathbf{S}^{(m)}$ of the hidden Markov chain from the conditional posterior $p(\mathbf{S} | \boldsymbol{\vartheta}^{(m)}, \mathbf{y})$:

- Forward filtering
- Backwards sampling

After the burn-in period M_0, the sampled values of $(\boldsymbol{\vartheta}^{(m)}, \mathbf{S}^{(m)})$ are draws from the joint posterior $p(\boldsymbol{\vartheta}, \mathbf{S} | \mathbf{y})$.

MCMC Estimation

- Step (a1) is exactly the same sampling step as for a standard finite mixture distributions, because for state parameter estimation only the number of observations in state k are relevant but not the number of transitions.
- The number of transitions are relevant only for sampling of the transition matrix ξ in step (a2) (product of K Dirichlet distribution)
- Sampling $\mathbf{S}$ is much more involved for a Markov mixture than is the corresponding step for a standard finite mixture model:
  - for a finite mixture model the indicators are independent conditional on $\mathbf{y}$ and ϑ.
  - $\mathbf{S}$ is a path of a stochastic process with dependence among successive values of S_t, even if the parameters are known.
  - Efficient methods for sampling a path of $\mathbf{S}$ are based on forward-filtering-backward-sampling.

Sampling the hidden Marcov Chain

Forward Filtering:

Filter at $t-1$:	$\Pr(S_{t-1} = l \vert \mathbf{y}^{t-1}, \boldsymbol{\vartheta})$	
	$\Downarrow$	
Prediction for t:	$\Pr(S_t = l \vert \mathbf{y}^{t-1}, \boldsymbol{\vartheta})$	Data at t: y_t
	$\Downarrow$	$\Downarrow$
Filter at t:	$\Pr(S_t = l \vert \mathbf{y}^{t}, \boldsymbol{\vartheta})$	

Forward Filtering

Forward Filtering

1. One-step ahead prediction of S_t for $t = 1, \ldots, T$:

$$\Pr(S_t = l|\mathbf{y}^{t-1}, \boldsymbol{\vartheta}) = \sum_{k=1}^{K} \Pr(S_{t-1} = k|\mathbf{y}^{t-1}, \boldsymbol{\vartheta})\xi_{kl},$$

   for $l = 1, \ldots, K$, where ξ_{kl} are the transition probabilities.

2. Filtering for S_t, $t = 1, \ldots, T$:

$$\Pr(S_t = l|\mathbf{y}^{t}, \boldsymbol{\vartheta}) = \frac{p(y_t|S_t = l, \mathbf{y}^{t-1}, \boldsymbol{\vartheta})\Pr(S_t = l|\mathbf{y}^{t-1}, \boldsymbol{\vartheta})}{p(y_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta})},$$

$$p(y_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta}) = \sum_{k=1}^{K} p(y_t|S_t = k, \mathbf{y}^{t-1}, \boldsymbol{\vartheta})\Pr(S_t = k|\mathbf{y}^{t-1}, \boldsymbol{\vartheta}).$$

Sampling the hidden Markov Chain

Backward sampling:

(a) Sample $S_T^{(m)}$ from the filtered state probability distribution $\Pr(S_T = j|\mathbf{y}^T, \boldsymbol{\vartheta})$.

(b) For $t = T-1, T-2, \ldots, 0$ sample $S_t^{(m)}$ from the conditional distribution $\Pr(S_t = j|S_{t+1}^{(m)} = l, \mathbf{y}^t, \boldsymbol{\vartheta})$ given by

$$\Pr(S_t = j|S_{t+1}^{(m)} = l, \mathbf{y}^t, \boldsymbol{\vartheta}) = \frac{\xi_{jl}\Pr(S_t = j|\mathbf{y}^t, \boldsymbol{\vartheta})}{\sum_{k=1}^{K} \xi_{kl}\Pr(S_t = k|\mathbf{y}^t, \boldsymbol{\vartheta})}.$$

Outline

Part II: Hidden Markov and Markov Switching Models

- Motivating Example
- Finite Markov mixture models
- Markov mixture modelling
- Bayesian Inference
- Applications
  - U.S. GDP Data – Markov switching autoregressive models
  - U.S./U.K. real exchange rate – Switching state space models
  - NYSE Data – Switching ARCH Model

Markov switching autoregressive models, 1

- Markov switching autoregressive (MSAR) models introduce a hidden Markov chain $S_0, S_1, \ldots, S_T$ into an AR(p)-model.
- Allow for a random shift in the mean level μ of an AR(p)-process through a hidden Markov chain:

$$Y_t - \mu_{S_t} = \phi_1(Y_{t-1} - \mu_{S_{t-1}}) + \cdots + \phi_p(Y_{t-p} - \mu_{S_{t-p}}) + \varepsilon_t. \tag{11}$$

- Suggested independently by [Neftçi, 1984] and [Sclove, 1983], became popular in econometrics for analyzing economic time series such as the GDP data through the work of [Hamilton, 1989].

Markov switching autoregressive models, 2

- Alternatively, [McCulloch and Tsay, 1994] introduced the hidden Markov chain into an AR(p) model by assuming that the intercept is driven by the hidden Markov chain rather than the mean level:

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \zeta_{S_t} + \varepsilon_t. \tag{12}$$

- Although the two parameterizations are equivalent for the standard AR model, a model with a Markov switching intercept turns out to be different from a model with a Markov switching mean level.
- Model (12) is more convenient numerically, because $p(y_t|S_t, \vartheta)$ depends only on the current value of S_t. For model (11), $p(y_t|S_t, \ldots, S_{t-p}, \vartheta)$ depends also on past value of S_t and recursive filtering is much more involved.

Markov switching autoregressive models, 2

- In a more general form the MSAR model allows that the autoregressive coefficients are also affected by S_t [McCulloch and Tsay, 1994]:

$$Y_t = \phi_{S_t,1} Y_{t-1} + \cdots + \phi_{S_t,p} Y_{t-p} + \zeta_{S_t} + \varepsilon_t. \tag{13}$$

- The MSAR model can be extended to deal with the presence of exogenous variables $\boldsymbol{z}_t$ [McCulloch and Tsay, 1994, Albert and Chib, 1993].
- In a Markov switching dynamic regression models all parameters, including the regression coefficient $\boldsymbol{\beta}$, are affected by endogenous regime shifts following a hidden Markov chain:

$$Y_t = \phi_{S_t,1} Y_{t-1} + \cdots + \phi_{S_t,p} Y_{t-p} + \boldsymbol{z}_t \boldsymbol{\beta}_{S_t} + \zeta_{S_t} + \varepsilon_t.$$

- In any of these models the variance may be assumed to be constant, irrespective of the state of S_t, or it is possible to assume a shift in the variance, $\varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_{\varepsilon,S_t}\right)$.

Markov switching autoregressive models,3

- Autocorrelation introduced through the hidden Markov chain as well as through the observation equation, leading to rather flexible autocorrelation structures [Timmermann, 2000].
- For an MSAR-model with $K = 2$, $p = 1$, switching mean, fixed variance, and fixed AR coefficient ϕ_1, for instance, the autocorrelation function of Y_t reads:

$$\rho_{Y_t}(h|\boldsymbol{\vartheta}) = \frac{1}{\mathrm{Var}(Y_t|\boldsymbol{\vartheta})} \left(\lambda^h (\mu_1 - \mu_2)^2 \eta_1 \eta_2 + {\phi_1}^h \frac{\sigma_\varepsilon^2}{1 - \phi_1^2} \right), \tag{14}$$

with $\lambda = \xi_{11} - \xi_{21}$ being the second eigenvalue of the transition matrix $\boldsymbol{\xi}$.
- The autocorrelation function fulfills, for $h > 2$, the following recursion,

$$\rho_{Y_t}(h|\boldsymbol{\vartheta}) = (\phi_1 + \lambda)\rho_{Y_t}(h-1|\boldsymbol{\vartheta}) - \phi_1 \lambda \rho_{Y_t}(h-2|\boldsymbol{\vartheta}),$$

and corresponds to the autocorrelation function of an ARMA$(2, 1)$ model, but has a nonnormal unconditional distribution.

Features of the MSAR model

- The assumption that the autoregressive parameters switch between the two states implies different dynamic patterns in the various states, and introduces asymmetry over time.
- Asymmetry over time between the states is introduced also through the hidden Markov chain as different persistence probabilities imply different state durations:

$$\mathrm{E}(D_j) = \frac{1}{1 - \xi_{jj}}.$$

- This combined asymmetry leads to a rather flexible model that is able to capture asymmetric patterns observed in economics time series, such as the fast rise and the slow decay in the U.S. quarterly unemployment rate.

Dealing with Spurious Unit Roots

- Consider a two-state Markov mixture of normal distributions with $\mu_2 \neq \mu_1$, no autocorrelation within the two regimes ($\phi_1 = 0$) and a highly persistent transition matrix where ξ_{11} and ξ_{22} are close to one (i.e. $\lambda = \xi_{11} - \xi_{21}$ close to 1).
- As evident from (14), high autocorrelation in the marginal process Y_t is present although there exists no autocorrelation within the two regimes.
- A unit root test applied to Y_t is biased toward nonrejection of the unit root hypothesis under a sudden change in the mean (spurious unit root) with increasing rate of non rejection as $|\mu_2 - \mu_1|$ increases.
- Markov switching models are to a certain degree able to deal with spurious unit roots caused by structural breaks.
- [Garcia and Perron, 1996]:
  - model interest rates by a three-state MSAR model with state-invariant autocorrelation and heteroscedastic variances
  - show that the autocorrelation within in the various regimes actually nearly disappears

Model selection for the U.S. quarterly GDP series

- [Frühwirth-Schnatter, 2004] compared 25 different models
- Standard AR(p)-models for $p = 1, \dots, 5$ ($\mathcal{M}_1$)
- K-state MSAR model with switching intercept, but state-independent AR(p) parameters and state-independent variances [Chib, 1996] for $K = 2, 3$ and $p = 1, \dots, 5$ ($\mathcal{M}_2$);
- K-state MSAR model with switching intercept, switching AR parameters, and switching error variance ("totally switching") [McCulloch and Tsay, 1994] for $K = 2, 3$ and $p = 1, \dots, 5$ ($\mathcal{M}_3$).
- The priors are selected to be rather vague and state-independent (intercept $\sim \mathcal{N}(0, 1)$, AR parameters $\sim \mathcal{N}(0, 0.25)$; variances $\sim \mathcal{G}^{-1}(2, 0.5)$).

Model selection for the U.S. quarterly GDP series

- Log marginal likelihoods $\log p(\mathbf{y}|\mathcal{M}_j, K, p)$ computed using bridge sampling [Frühwirth-Schnatter, 2004]:

	$\mathcal{M}_1$	$\mathcal{M}_2$		$\mathcal{M}_3$	
p	$K=1$	$K=2$	$K=3$	$K=2$	$K=3$
0	–199.71	–193.54	–192.25	–194.25	–193.10
1	**–194.22**	–192.54	–192.75	–193.58	–194.71
2	–196.30	–194.15	–194.38	**–191.62**	–194.33
3	–197.26	–194.59	–194.74	–193.67	–196.78
4	–199.18	–195.70	–195.72	**–195.34**	–199.88

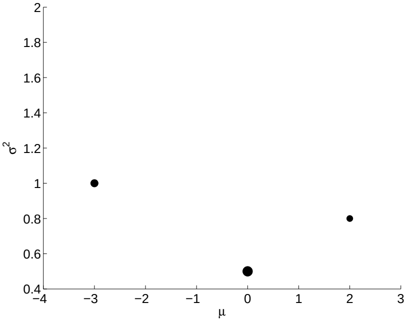
- A two-state totally switching MSAR model of order $p = 2$ has the highest marginal likelihood – confirms empirical results of [McCulloch and Tsay, 1994].
- See [Frühwirth-Schnatter, 2019] for a recent improvement to compute marginal likelihoods.

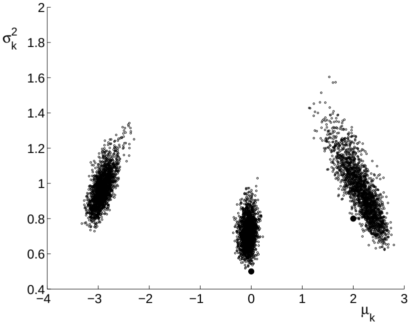
Model selection for the U.S. quarterly GDP series

- Results indicate the importance of simultaneously testing for Markov switching heterogeneity and selecting the appropriate model order
- Compare a two-state totally switching model of order four [McCulloch and Tsay, 1994] with an AR(1) model (optimal among all AR(p) models) $\Rightarrow$ evidence in favor of no Markov switching heterogeneity.
- Comparing a two-state totally switching MSAR model with the optimal model order $p = 2$ with the AR(1) model $\Rightarrow$ evidence in favor of Markov switching heterogeneity.
- Explains why several studies have produced somewhat conflicting evidence concerning the presence or absence of Markov switching heterogeneity in this time series.

Exploring the point process representation

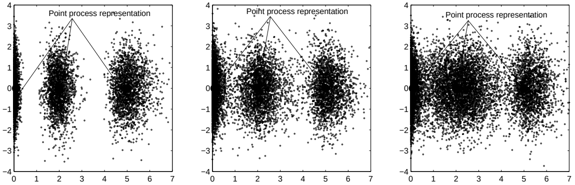
The MCMC draws scatter around the points corresponding to the “true” point process representation e.g. $K = 3$, $\mu_1 = -3$, $\mu_2 = 0$, $\mu_3 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 0.5$, $\sigma_3^2 = 0.8$





Exploring the point process representation

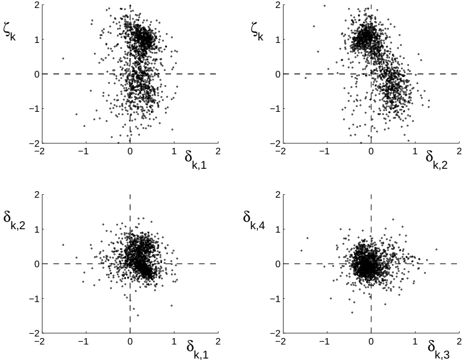
The point process representation of the MCMC draws will cluster around the point process representation of the true model even if the mixture is overfitting, although the spread of these simulation clusters increases with K. Asymptotically, the number of simulation clusters in these figures indicate the true number of components.



Point process representation of the posterior density $p(\mu_k|\mathbf{y}, \mathcal{M}_K)$ for $K = 3$ (left-hand side), $K = 4$ (middle), and $K = 5$ (right-hand side)

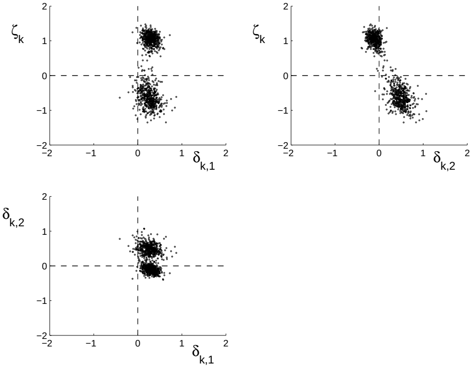
Exploratory Bayesian Analysis for an Overfitting Model

GDP data, totally Markov switching model with $K = 3$ and $p = 4$ (overfitting); explore point process representation of the MCMC output



Model selection for the U.S. quarterly GDP series

GDP data, totally Markov switching model with $K = 2$ and $p = 2$ (selected model); explore point process representation of the MCMC output



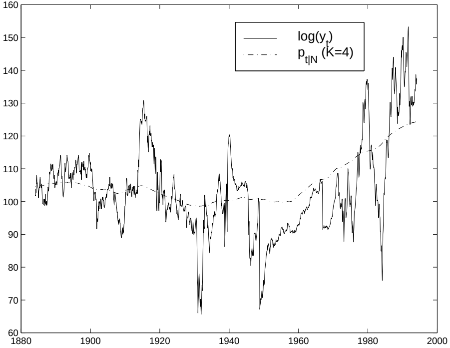
Parameter Estimation for the selected model

Parameter	Contraction ($k = 1$)	Expansion ($k = 2$)
$\phi_{k,1}$	0.249 (0.164)	0.295 (0.116)
$\phi_{k,2}$	0.462 (0.164)	–0.114 (0.098)
ζ_k	–0.557 (0.322)	1.060 (0.175)
$\sigma_{\varepsilon,k}$	0.768 (0.161)	0.692 (0.115)
$\xi_{kk'}$	0.489 (0.165)	0.337 (0.145)

- Positive growth in expansion is followed by negative growth in contraction.
- The dynamic behavior of the U.S. GDP growth rate is different between contraction and expansion with reaction to a percentage change of the GDP growth being faster in expansion than in contraction.
- The expected duration of expansion is longer than that of contraction.

Switching State Space Model

Log of the U.S./U.K. real exchange rate from January 1885 to November 1995 [Grilli and Kaminsky, 1991]



State Space Modeling

- [Engel and Kim, 1999] suggested decomposing the log of the real exchange rate Y_t into a permanent component μ_t and a transitory component c_t:

$$\log Y_t = \mu_t + c_t,$$

where c_t is assumed to follow an AR(p) process:

$$c_t = \phi_1 c_{t-1} + \cdots + \phi_p c_{t-p} + w_{t,1},$$

and μ_t follows a random walk process:

$$\mu_t = \mu_{t-1} + w_{t,2}.$$

Switching variances

- The conditional variance of c_t is assumed to switch between K_1 values according to a hidden Markov chain S_t^1 with transition matrix ξ^1,

$$w_{t,1} \sim \mathcal{N}\left(0, \sigma^2_{1,S_t^1}\right),$$

- The conditional variance of the permanent component μ_t is assumed to switch between K_2 values according to a hidden Markov chain S_t^2 with transition matrix ξ^2:

$$w_{t,2} \sim \mathcal{N}\left(0, \sigma^2_{2,S_t^2}\right).$$

State Space Modeling

The model can be put into state space form with the following state vector $\mathbf{x}_t$ and matrix $\mathbf{F}$,

$$\mathbf{x}_t = \begin{pmatrix} \mu_t \\ c_t \\ \vdots \\ c_{t-p+1} \end{pmatrix}, \qquad \mathbf{F} = \begin{pmatrix} 1 & \mathbf{0}_{1\times p} \\ \mathbf{0}_{p\times 1} & \mathbf{F}(\phi) \end{pmatrix},$$

$$\mathbf{F}(\phi) = \begin{pmatrix} \phi_1 \ldots \phi_{p-1} & \phi_p \\ \mathbf{I}_{p-1} & \mathbf{0}_{(p-1)\times 1} \end{pmatrix}. \tag{15}$$

Model selection problems

:

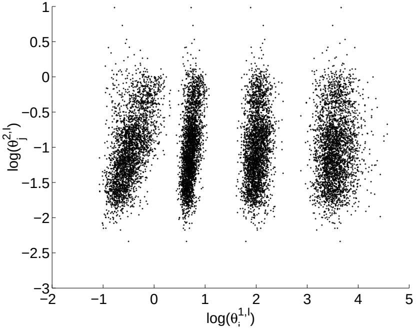
- How many states K_1 for the variance of the transitory component?
- Testing $K_1 = 1$ versus $K_1 > 1$ is a nonregular testing problem.
- How many states K_2 for the variance of the permanent component?
- Testing $K_2 = 1$ versus $K_2 > 1$ nonregular testing problem.
- Order selection p for the AR-model.
- Compare various models differing in K_1, K_2 and p using a Bayesian approach.

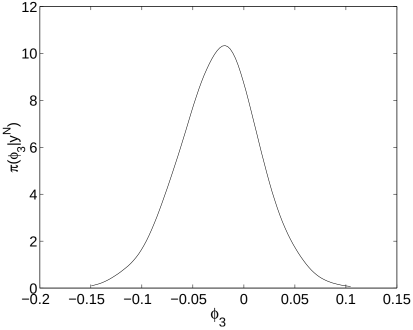
MCMC Estimation for Fixed Model Structure

- MCMC estimation of switching linear Gaussian state space model – data augmentation and Gibbs sampling [Frühwirth-Schnatter, 2001a]
- Sample the state processes $\mu_t, t = 1, \ldots, T$ and $c_t, t = 0, \ldots, T$ (FFBS, e.g. [Frühwirth-Schnatter, 1994]);
- Sample the hidden Markov processes S_t^1 and S_t^2 for $t = 0, \ldots, T$ (discrete FFBS);
- Sample the switching variances $\sigma^2_{1,k}, k = 1, \ldots, K_1$ and $\sigma^2_{2,k}, k = 1, \ldots, K_2$ (inverted Gamma densities) and the transition matrices $\boldsymbol{\xi}^1$ and $\boldsymbol{\xi}^2$ of the hidden Markov chains (Dirichlet densities);
- Sample the AR parameters $\phi_1, \ldots, \phi_p$ (normal likelihood with nonconjugate prior, if stationarity is assumed).

Assume $K_1 = 4$, $K_2 = 2$, and $p = 3$

left-hand side: $\log(\sigma^2_{1,k})$ versus $\log(\sigma^2_{2,k})$ for all possible k; right-hand side: posterior of ϕ_3





Exploratory Bayesian Analysis

- Estimation based on $K_1 = 4$, $K_2 = 2$, and $p = 3$
- For S_t^1 we have allowed for four states and there are actually four simulation clusters;
- for S_t^2, we have allowed for two states, however, there is just one simulation cluster.
- this provides empirical evidence in favor of a homogeneous rather than a switching variance of the permanent component.
- This hypothesis is further supported by the point process representation of $(\sigma^2_{1,k})^{(m)}$ versus $(\sigma^2_{2,k})^{(m)}$.
- The mode of the posterior of the AR parameter ϕ_3 is close to 0 providing evidence for the hypothesis that ϕ_3 is equal to zero.
- Exploratory analysis provides evidence in favor of a model with $K_1 = 4$, $K_2 = 1$, and $p = 2$.

Model Choice Using Marginal Likelihoods

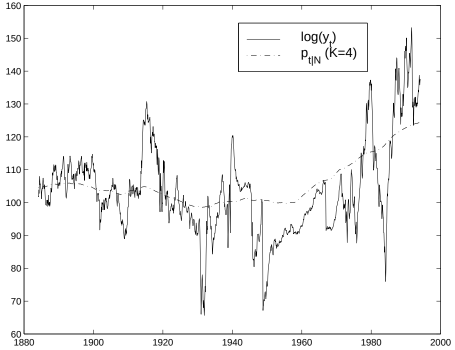
- Formal model selection using marginal likelihoods

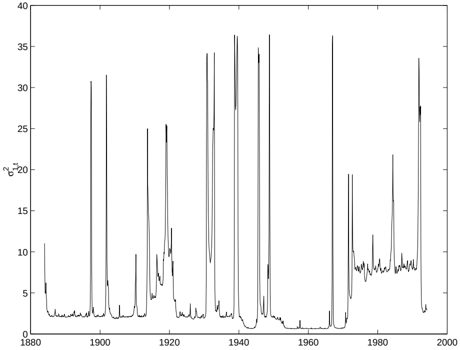
Model	$\log p(\mathbf{y}\vert\mathrm{Model})$
$K_1 = 4,\ K_2 = 2,\ p = 3$	–2562.4
$K_1 = 4,\ K_2 = 1,\ p = 2$	**–2515.5**
$K_1 = 4,\ K_2 = 1,\ p = 1$	–2612.5
$K_1 = 3,\ K_2 = 1,\ p = 2$	–2605.9
$K_1 = 5,\ K_2 = 1,\ p = 2$	–2880.2
No switching, $p = 2$	–2914.4

- Marginal likelihoods are computed using bridge sampling (hidden Markov processes S_t^1 and S_t^2 are integrated out)

Inference for the Selected Model

- Four-state model ($K_1 = 4,\ K_2 = 1,\ p = 2$)





- left-hand side: smoothed real exchange rate $\hat{p}_{t|T}$;
- right-hand side: estimated time-varying variance $\hat{\sigma}^2_{1,t}$

$$\hat{\sigma}^2_{1,t} = \frac{1}{M}\sum_{m=1}^{M}\left(\sigma^2_{1,s}\right)^{(m)}, \qquad s = (S^1_t)^{(m)}.$$

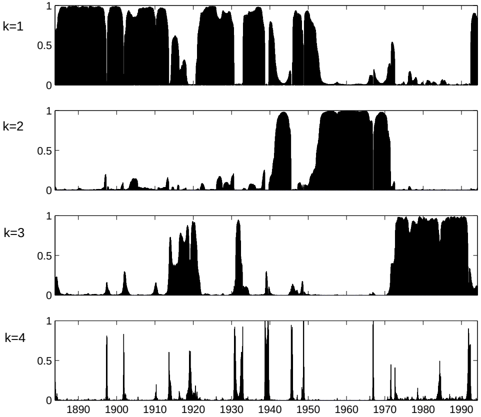
Inference for the Selected Model

Impose identifiability constraint $\sigma^2_{1,1} < \sigma^2_{1,2} < \sigma^2_{1,3} < \sigma^2_{1,4}$ on MCMC draws

Parameter	Mean	Std.Dev.	95%-H.P.D. Regions	
$\sigma^2_{1,1}$	0.634	0.151	0.371	0.93
$\sigma^2_{1,2}$	2.05	0.196	1.67	2.42
$\sigma^2_{1,3}$	7.63	1.07	5.9	9.88
$\sigma^2_{1,4}$	36.4	9.13	20.7	53.9
σ^2_2	0.366	0.132	0.121	0.608
ϕ_1	1.06	0.0474	0.967	1.14
ϕ_2	–0.0729	0.046	–0.158	0.0139
ξ_{11}	0.968	0.0132	0.943	0.991
ξ_{22}	0.973	0.00853	0.957	0.988
ξ_{33}	0.956	0.0222	0.916	0.992
ξ_{44}	0.691	0.116	0.438	0.865

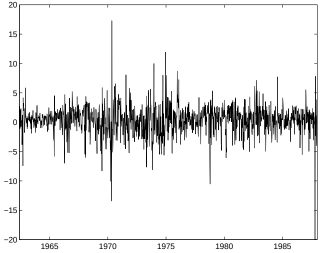
Inference for the Selected Model

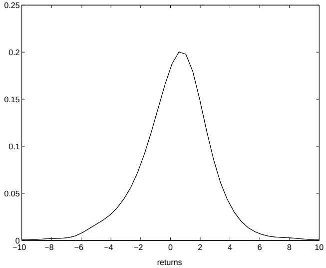
Smoothed state probabilities for S_t^1 for a switching state space model with $K_1 = 4$, $K_2 = 1$, and $p = 2$



Switching ARCH Model

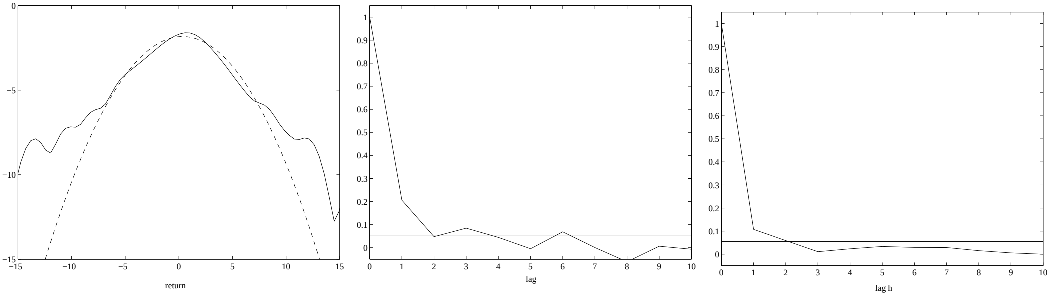
- Markov switching models are often used by researchers to account for specific features of financial time series such as asymmetries, fat tails, and volatility clusters.
- New York Stock Exchange Data, weekly observations from July 3, 1962 to December 29, 1987 (1330 observations); left: time series plot; right: smoothed histogram of the marginal distribution





NYSE Data

New York Stock Exchange Data, left: log of the smoothed histogram (solid line) in comparison to the log of a normal distribution with same mean and variance (dashed line); middle: empirical autocorrelogram of the returns; right: empirical autocorrelogram of the squared



ARCH and GARCH Models

- Volatility clustering implies persistence of states of high volatility and leads to the rejection of standard time series models in favor of models that allow the conditional variance $\mathrm{Var}(Y_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta})$ to depend on the history $y_{t-1}, y_{t-2}, \ldots$ of the observed process.
- Well-known models:
  - ARCH models [Engle, 1982]:

$$\mathrm{Var}(Y_t|\mathbf{y}^{t-1}, \boldsymbol{\vartheta}) = \gamma_t + \alpha_1 y_{t-1}^2 + \cdots + \alpha_m y_{t-m}^2,$$

  - GARCH models [Bollerslev, 1986]

.

Markov Mixture Models for Financial Time Series

- *Finite mixtures of normal distributions* to deal with skewness and excess kurtosis in the unconditional distribution of daily stock returns [Fama, 1965, Granger and Orr, 1972, Kon, 1984, Tucker, 1992] (which implies zero autocorrelation in Y_t and Y_t^2)
- *Markov mixture model* where the variance of a location-scale family is driven by a hidden Markov capture simultaneously autocorrelation in the processes Y_t and Y_t^2 [Engel and Hamilton, 1990, McQueen and Thorely, 1991, Rydén et al., 1998].
- More general (though) limited autocorrelation functions of Y_t^2 are possible if Y_t is generated by an *MSAR model* with or without switching AR coefficients [Hamilton, 1988, Turner et al., 1989, Cecchetti et al., 1990, Engel, 1994, Gray, 1996, Ang and Bekaert, 2002].

Markov Mixture Models for Financial Time Series

To obtain even more flexibility in the autocorrelation of Y_t^2, for a given marginal distribution of Y_t,

- [Hamilton and Susmel, 1994], [Cai, 1994], and [Gray, 1996] proposed to combine ARCH and Markov switching effects to formulate the *switching ARCH model*.
- [Francq et al., 2001] considered a switching GARCH model.
- [So et al., 1998] considered a stochastic volatility model with Markov switching.

Spurious Persistency in Squared Returns

- A common finding when fitting GARCH models to high-frequency financial data is the somewhat unexpected persistence of shocks to the variance implied by the estimated coefficients.
- [Lamoureux and Lastrapes, 1990] show that a deterministic structural shift in the unconditional variance, caused by exogenous shocks such as changes in the monetary policy, will increase persistency of squared residuals, however, when the structural break is accounted for, persistency often decreases dramatically.
- Introducing a hidden Markov chain into a variance model helps to explain spurious persistence in squared returns.

Spurious Persistency in Squared Returns

- Consider, for illustration, a simple Markov mixture of two normal distributions with $\mu_1 = \mu_2$ and $\sigma_1^2 \neq \sigma_2^2$ driven by a highly persistent transition matrix $\boldsymbol{\xi}$ with $\lambda = \xi_{11} - \xi_{21}$ being close to 1.
- Although the process Y_t^2 is not autocorrelated within each regime, marginally the persistence in Y_t^2 decays slowly, in particular if $\sigma_2^2 - \sigma_1^2$ is large:

$$\rho_{Y_t^2}(h|\boldsymbol{\vartheta}) = \frac{\eta_1\eta_2(\sigma_1^2 - \sigma_2^2)^2}{\mathrm{E}(Y_t^4|\boldsymbol{\vartheta}) - \mathrm{E}(Y_t^2|\boldsymbol{\vartheta})^2}\lambda^h,$$

- Also for the more general switching ARCH model, [Hamilton and Susmel, 1994] attribute part of the high marginal persistence in Y_t^2, which is typically much larger than autocorrelation of Y_t^2 in the various regimes, to this effect.

Example: switching AR(1)-ARCH model

- To account for the autocorrelation found in y_t and y_t^2, as well as for the fat tails and the asymmetry observed in the marginal distribution,
- fit a switching AR(1)-ARCH model which includes a leverage term [Hamilton and Susmel, 1994, Kaufmann and Frühwirth-Schnatter, 2002]:

$$
\begin{aligned}
y_t &= \zeta + \phi_1 y_{t-1} + u_t, \\
u_t &= \sigma_t \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1), \\
\sigma_t^2 &= \gamma_{S_t} + \alpha_1 u_{t-1}^2 + \cdots + \alpha_m u_{t-m}^2 + \varrho d_{t-1} y_{t-1}^2.
\end{aligned}
$$

- S_t is a hidden Markov chain with K states

Application to the NYSE Data

NEW YORK STOCK EXCHANGE DATA, modeled by a switching AR-ARCH model with leverage with different numbers of states K and different model orders m; log of the marginal likelihoods computed under different priors on the switching ARCH intercept using bridge sampling

		$\log p(\mathbf{y} \mid K, m)$	
K	m	(prior 1)	(prior 2)
3	2	–2858.5	–2858.0
3	3	–2858.2	–2857.7
3	4	**–2857.1**	–2856.4
4	2	–2861.0	–2859.7
4	3	–2860.7	–2859.4
4	4	–2859.1	**–2855.9**

Summary

- The introduction of a hidden Markov chain generates time series models which combine autocorrelation in Y_t and Y_t^2 with non-normal (skewed, fat tails) marginal distributions.
- Conditional on knowing the hidden Markov chain, standard time series models like AR, ARCH or non-Gaussian distributions are assumed
- This simplifies the analysis of the theoretical properties of the models
- This enables straightforward Bayesian inference using data augmentation and MCMC

Aitkin, M. (1996).

A general maximum likelihood analysis of overdispersion in generalized linear models.

Statistics and Computing, 6:251–262.

Albert, J. H. and Chib, S. (1993).

Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts.

Journal of Business & Economic Statistics, 11:1–15.

Alspach, D. L. and Sorenson, H. W. (1972).

Nonlinear Bayesian estimation using Gaussian sum approximations.

IEEE Transactions on Automatic Control, 17:439–448.

Ang, A. and Bekaert, G. (2002).

Regime switches in interest rates.

Journal of Business & Economic Statistics, 20:163–182.

Banfield, J. D. and Raftery, A. E. (1993).

Model-based Gaussian and non-Gaussian clustering.

Biometrics, 49:803–821.

Baudry, J., Raftery, A. E., Celeux, G., Lo, K., and Gottardo, R. (2010).
Combining mixture components for clustering.
Journal of Computational and Graphical Statistics, 19:332–353.

Bennett, D. A., Schneider, J. A., Buchman, A. S., de Leon, C. M., Bienias, J. L., and Wilson, R. S. (2005).
The Rush Memory and Aging Project: Study Design and Baseline Characteristics of the Study Cohort.
Neuroepidemiology, 25:163–175.

Bensmail, H., Celeux, G., Raftery, A. E., and Robert, C. P. (1997).
Inference in model-based cluster analysis.
Statistics and Computing, 7:1–10.

Biernacki, C., Celeux, G., and Govaert, G. (2000).
Assessing a mixture model for clustering with the integrated completed likelihood.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 22:719–725.

Binder, D. A. (1978).
Bayesian cluster analysis.
Biometrika, 65:31–38.

Bollerslev, T. (1986).
Generalized autoregressive conditional heteroskedasticity.
Journal of Econometrics, 31:307–327.

Cai, J. (1994).
A Markov model of switching-regime ARCH.
Journal of Business & Economic Statistics, 12:309–316.

Carter, C. K. and Kohn, R. (1997).
Semiparametric Bayesian inference for time series with mixed spectra.
Journal of the Royal Statistical Society, Ser. B, 59:255–268.

Cecchetti, S. G., Lam, P., and Mark, N. C. (1990).
Mean reversion in equilibrium asset prices.
The American Economic Review, 80:398–418.

Celeux, G. (1998).
Bayesian inference for mixture: The label switching problem.
In Green, P. J. and Rayne, R., editors, *COMPSTAT 98*, pages 227–232. Physica, Heidelberg.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006).

Deviance information criteria for missing data models.
Bayesian Analysis, 1:651–674.

Celeux, G., Frühwirth-Schnatter, S., and Robert, C. P. (2019).
Model selection for mixture models – perspectives and strategies.
In Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors, *Handbook of Mixture Analysis*, chapter 7, pages 117–154. CRC Press, Boca Raton, FL.

Celeux, G., Hurn, M., and Robert, C. P. (2000).
Computational and inferential difficulties with mixture posterior distributions.
Journal of the American Statistical Association, 95:957–970.

Chib, S. (1996).
Calculating posterior distributions and modal estimates in Markov mixture models.
Journal of Econometrics, 75:79–97.

Chib, S., Nardari, F., and Shephard, N. (2002).
Markov chain Monte Carlo methods for stochastic volatility models.
Journal of Econometrics, 108:281–316.

Dasgupta, A. and Raftery, A. E. (1998).
Detecting features in spatial point processes with clutter via model-based clustering.

Journal of the American Statistical Association, 93:294–302.

Diebolt, J. and Robert, C. P. (1994).
Estimation of finite mixture distributions through Bayesian sampling.
Journal of the Royal Statistical Society, Ser. B, 56:363–375.

Engel, C. (1994).
Can the Markov switching model forecast exchange rates?
Journal of International Economics, 36:151–165.

Engel, C. and Hamilton, J. D. (1990).
Long swings in the Dollar: Are they in the data and do markets know it?
The American Economic Review, 80:689–713.

Engel, C. and Kim, C.-J. (1999).
The long-run U.S./U.K. real exchange rate.
Journal of Money, Credit, and Banking, 31:335–356.

Engle, R. F. (1982).
Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation.
Econometrica, 50:987–1007.

Escobar, M. D. and West, M. (1998).
Computing nonparametric hierarchical models.
In Dey, D., Müller, P., and Sinha, D., editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, number 133 in Lecture Notes in Statistics, pages 1–22. Springer, Berlin.

Everitt, B. S. (1979).
Unresolved problems in cluster analysis.
Biometrics, 35:169–181.

Fama, E. (1965).
The behavior of stock market prices.
Journal of Business, 38:34–105.

Fraley, C. and Raftery, A. E. (2002).
Model-based clustering, discriminant analysis, and density estimation.
Journal of the American Statistical Association, 97:611–631.

Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012).
mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.
Technical Report 597, Department of Statistics, University of Washington.

Francq, C., Roussignol, M., and Zakoian, J. (2001).
Conditional heteroscedasticity driven by hidden Markov chains.
Journal of Time Series Analysis, 22:197–220.

Frühwirth-Schnatter, S. (1994).
Data augmentation and dynamic linear models.
Journal of Time Series Analysis, 15:183–202.

Frühwirth-Schnatter, S. (2001a).
Fully Bayesian analysis of switching Gaussian state space models.
Annals of the Institute of Statistical Mathematics, 53:31–49.

Frühwirth-Schnatter, S. (2001b).
Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models.
Journal of the American Statistical Association, 96:194–209.

Frühwirth-Schnatter, S. (2004).
Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques.
Econometrics Journal, 7:143–167.

Frühwirth-Schnatter, S. (2006).
Finite Mixture and Markov Switching Models.
Springer, New York.

Frühwirth-Schnatter, S. (2011a).
Dealing with label switching under model uncertainty.
In Mengersen, K., Robert, C. P., and Titterington, D., editors, *Mixture estimation and applications*, chapter 10, pages 213–239. Wiley, Chichester.

Frühwirth-Schnatter, S. (2011b).
Panel data analysis - A survey on model-based clustering of time series.
Advances in Data Analysis and Classification, 5:251–280.

Frühwirth-Schnatter, S. (2019).
Keeping the balance – Bridge sampling for marginal likelihood estimation in finite mixture, mixture of experts and markov mixture models.
Brazilian Journal of Probability and Statistics, 33:706–733.

Frühwirth-Schnatter, S. and Frühwirth, R. (2007).
Auxiliary mixture sampling with applications to logistic models.
Computational Statistics and Data Analysis, 51:3509–3528.

Frühwirth-Schnatter, S. and Frühwirth, R. (2010).
Data augmentation and MCMC for binary and multinomial logit models.
In Kneib, T. and Tutz, G., editors, *Statistical Modelling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, pages 111–132. Physica-Verlag, Heidelberg.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., and Rue, H. (2009).
Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data.
Statistics and Computing, 19:479–492.

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019).
From here to infinity: Sparse finite versus Dirichlet process mixtures in model-based clustering.
Advances in Data Analysis and Classification, 13:33–64.

Frühwirth-Schnatter, S., Pamminger, C., Weber, A., and Winter-Ebmer, R. (2012).
Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering.
Journal of Applied Econometrics, 27:1116–1137.

Frühwirth-Schnatter, S., Pittner, S., Weber, A., and Winter-Ebmer, R. (2018).
Analysing plant closure effects using time-varying mixture-of-experts Markov chain clustering.
Annals of Applied Statistics, 12:1786–1830.

Frühwirth-Schnatter, S. and Pyne, S. (2010).
Bayesian inference for finite mixtures of univariate and multivariate skew normal and skew-t distributions.
Biostatistics, 11:317 – 336.

Frühwirth-Schnatter, S., Tüchler, R., and Otter, T. (2004).
Bayesian analysis of the heterogeneity model.
Journal of Business & Economic Statistics, 22:2–15.

Frühwirth-Schnatter, S. and Wagner, H. (2006).
Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling.
Biometrika, 93:827–841.

Garcia, R. and Perron, P. (1996).
An analysis of real interest rate under regime shift.
The Review of Economics and Statistics, 78:111–125.

Gormley, I. C. and Frühwirth-Schnatter, S. (2019).
Mixture of experts models.
In Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors, *Handbook of Mixture Analysis*, chapter 12, pages 271–307. CRC Press, Boca Raton, FL.

Gormley, I. C. and Murphy, T. B. (2008).
Exploring voting blocs within the Irish electorate: A mixture modeling approach.
Journal of the American Statistical Association, 103:1014–1027.

Granger, C. W. J. and Orr, D. (1972).
Infinite variance and research strategy in time series analysis.
Journal of the American Statistical Association, 67:275–285.

Gray, S. F. (1996).
Modeling the conditional distribution of interest rates as a regime switching process.
Journal of Financial Economics, 42:27–62.

Grilli, V. and Kaminsky, G. (1991).
Nominal exchange rate regimes and the real exchange rate: Evidence from the United States and Great Britain, 1885-1986.
Journal of Monetary Economics, 27:191–212.

Grün, B. (2019).
Model-based clustering.
In Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors, *Handbook of Mixture Analysis*, chapter 8, pages 157–192. CRC Press, Boca Raton, FL.

Hamilton, J. D. (1988).
Rational expectations econometric analysis of changes in regime: An investigation on the term structure of interest rates.
Journal of Economic Dynamics and Control, 12:385–423.

Hamilton, J. D. (1989).
A new approach to the economic analysis of nonstationary time series and the business cycle.
Econometrica, 57:357–384.

Hamilton, J. D. and Susmel, R. (1994).
Autoregressive conditional heteroskedasticity and changes in regime.
Journal of Econometrics, 64:307–333.

Hennig, C. (2010).
Methods for merging Gaussian mixture components.
Advances in Data Analysis and Classification, 4:3–34.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005).
Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling.
Statistical Science, 20:50–67.

Kaufmann, S. (2019).

Hidden Markov models in time series, with applications in economics.

In Frühwirth-Schnatter, S., Celeux, G., and Robert, C. P., editors, *Handbook of Mixture Analysis*, chapter 13, pages 308–341. CRC Press, Boca Raton, FL.

Kaufmann, S. and Frühwirth-Schnatter, S. (2002).

Bayesian analysis of switching ARCH models.

Journal of Time Series Analysis, 23:425–458.

Keribin, C. (2000).

Consistent estimation of the order of mixture models.

Sankhyā A, 62:49–66.

Kiefer, N. M. and Wolfowitz, J. (1956).

Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters.

Annals of Mathematical Statistics, 27:887–906.

Kim, S., Shephard, N., and Chib, S. (1998).

Stochastic volatility: Likelihood inference and comparison with ARCH models.

Review of Economic Studies, 65:361–393.

Kon, S. J. (1984).
Models of stock returns – A comparison.
The Journal of Finance, 39:147–165.

Lamoureux, C. G. and Lastrapes, W. D. (1990).
Persistence in variance, structural change and the GARCH model.
Journal of Business & Economic Statistics, 8:225–234.

Lee, S. X. and McLachlan, G. J. (2013).
EMMIXuskew: An R package for fitting mixtures of multivariate skew t-distributions via the EM algorithm.
Journal of Statistical Software, 55(12):1–22.

Leisch, F. (2004).
Exploring the structure of mixture model components.
In Antoch, J., editor, *COMPSTAT 2004. Proceedings in Computational Statistics*, pages 1405–1412. Physica-Verlag/Springer, Heidelberg.

Li, J. (2005).
Clustering based on a multi-layer mixture model.
Journal of Computational and Graphical Statistics, 14:547–568.

MacQueen, J. (1967).
Some methods for classification and analysis of multivariate observations.
In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume I, pages 281–297.

Malsiner Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016).
Model-based clustering based on sparse finite Gaussian mixtures.
Statistics and Computing, 26:303–324.

Malsiner Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2017).
Identifying mixtures of mixtures using Bayesian estimation.
Journal of Computational and Graphical Statistics, 26:285–295.

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005).
Bayesian modelling and inference on mixtures of distributions.
In Dey, D. and Rao, C., editors, *Bayesian Thinking, Modelling and Computation*, volume 25 of *Handbook of Statistics*, chapter 16, page ADD. North-Holland, Amsterdam.

McCulloch, R. E. and Tsay, R. S. (1994).
Statistical analysis of economic time series via Markov switching models.
Journal of Time Series Analysis, 15:523–539.

McLachlan, G. J. and Peel, D. (2000).
Finite Mixture Models.
Wiley Series in Probability and Statistics. Wiley, New York.

McQueen, G. and Thorely, S. (1991).
Are stock returns predictable? A test using Markov chains.
The Journal of Finance, 46:239–263.

Melnykov, V. (2016).
Merging mixture components for clustering through pairwise overlap.
Journal of Computational and Graphical Statistics, 25:66–90.

Meng, X.-L. and Schilling, S. (1996).
Fitting full-information item factor models and an empirical investigation of bridge sampling.
Journal of the American Statistical Association, 91:1254–1267.

Meng, X.-L. and Wong, W. H. (1996).
Simulating ratios of normalizing constants via a simple identity: A theoretical exploration.
Statistica Sinica, 6:831–860.

Miller, J. W. and Harrison, M. T. (2018).

Mixture models with a prior on the number of components.
Journal of the American Statistical Association, 113:340–356.

Neftçi, S. N. (1984).
Are economic time series asymmetric over the business cycle?
Journal of Political Economy, 92:307–328.

Nobile, A. (2004).
On the posterior distribution of the number of components in a finite mixture.
The Annals of Statistics, 32:2044–2073.

Nobile, A. and Fearnside, A. (2007).
Bayesian finite mixtures with an unknown number of components: The allocation sampler.
Statistics and Computing, 17:147–162.

Omori, Y., Chib, S., Shephard, N., and Nakajima, J. (2007).
Stochastic volatility with leverage: Fast and efficient likelihood inference.
Journal of Econometrics, 140:425–449.

Peng, F., Jacobs, R. A., and Tanner, M. A. (1996).
Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition.

Journal of the American Statistical Association, 91:953–960.

Polson, N. G., Scott, J. G., and Windle, J. (2013).
Bayesian inference for logistic models using Pólya-Gamma latent variables.
Journal of the American Statistical Association, 108:1339–49.

Poskitt, D. S. and Chung, S.-H. (1996).
Markov chain models, time series analysis and extreme value theory.
Advances in Applied Probability, 28:405–425.

Ray, S. and Lindsay, B. (2005).
The topography of multivariate normal mixtures.
The Annals of Statistics, 33:2042–2065.

Richardson, S. and Green, P. J. (1997).
On Bayesian analysis of mixtures with an unknown number of components.
Journal of the Royal Statistical Society, Ser. B, 59:731–792.

Rousseau, J. and Mengersen, K. (2011).
Asymptotic behaviour of the posterior distribution in overfitted mixture models.
Journal of the Royal Statistical Society, Ser. B, 73:689–710.

Rydén, T., Teräsvirta, T., and Åsbrink, S. (1998).
Stylized facts of daily return series and the hidden Markov model.
Journal of Applied Econometrics, 13:217–244.

Sclove, S. L. (1983).
Time series segmentation: A model and a method.
Information Science, 29:7–25.

Scott, A. J. and Symons, M. (1971).
Clustering methods based on likelihood ratio criteria.
Biometrics, 27:387–397.

Shephard, N. (1994).
Partial non-Gaussian state space.
Biometrika, 81:115–131.

So, M. K. P., Lam, K., and Li, W. K. (1998).
A stochastic volatility model with Markov switching.
Journal of Business & Economic Statistics, 16:244–253.

Sorenson, H. W. and Alspach, D. L. (1971).

Recursive Bayesian estimation using Gaussian sums.
Automatica, 7:465–479.

Sperrin, M., Jaki, T., and Wit, E. (2010).
Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models.
Statistics and Computing, 20:357–366.

Spezia, L. (2009).
Reversible jump and the label switching problem in hidden Markov models.
Journal of Statistical Planning and Inference, 139:2305–2315.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002).
Bayesian measures of model complexity and fit.
Journal of the Royal Statistical Society, Ser. B, 64:583–639.

Stephens, M. (2000a).
Bayesian analysis of mixture models with an unknown number of components – An alternative to reversible jump methods.
The Annals of Statistics, 28:40–74.

Stephens, M. (2000b).

Dealing with label switching in mixture models.
Journal of the Royal Statistical Society, Ser. B, 62:795–809.

Symons, M. J. (1981).
Clustering criteria and multivariate normal mixtures.
Biometrics, 37:35–43.

Timmermann, A. (2000).
Moments of Markov switching models.
Journal of Econometrics, 96:75–111.

Tucker, A. (1992).
A reexamination of finite- and infinite-variance distributions as models of daily stock returns.
Journal of Business & Economic Statistics, 10:73–81.

Turner, C. M., Startz, R., and Nelson, C. R. (1989).
A Markov model of heteroscedasticity, risk, and learning in the stock market.
Journal of Financial Economics, 25:3–22.
CHECK.

Wilson, R., Bienias, J., Evans, D., and Bennett, D. (2004).

The Religious Orders Study: Overview and Change in Cognitive and Motor Speed.
Aging, Neuropsychol, Cogn., 11:280–303.

Wolfe, J. H. (1970).
Pattern clustering by multivariate mixture analysis.
Multivariate Behavioral Research, 5:329–350.
CHECK.

Yerebakan, H. Z., Rajwa, B., and Dundar, M. (2014).
The infinite mixture of infinite Gaussian mixtures.
In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27 of *Proceedings from the Neural Information Processing Systems Conference*.