

# Testing for two states in a hidden Markov model

Hajo Holzmann

Institut für Stochastik  
Universität Karlsruhe (TH)

`www.mathematik.uni-karlsruhe.de/stoch/`

Zürich, September 2007

Institut *für* Stochastik Karlsruhe

# Overview

1. Hidden Markov Models (HMMs)
2. Likelihood inference for HMMs
3. Testing w.r.t. the marginal distribution
4. Testing for two states in an HMM

# Overview

1. Hidden Markov Models (HMMs)
2. Likelihood inference for HMMs
3. Testing w.r.t. the marginal distribution
4. Testing for two states in an HMM

# Hidden Markov Models

- $(Y_n)_{n \geq 0}$  (observed): Sequence of  $\mathbb{R}$ -valued random variables.
- $(X_n)_{n \geq 0}$  (unobserved): Stationary Markov chain with finite state space, transition matrix  $P$  and stationary distribution  $\pi$ .

If

- Conditional on  $X_0, \dots, X_n$ , the  $Y_0, \dots, Y_n$  are independent.
- The conditional distribution of  $Y_n$  given  $(X_j)_{j \geq 0}$  depends on  $X_n$  alone.
- Given  $X_n = i$ , the  $Y_n$  have density  $f_{\theta_i}$ ,  $\theta_i \in \Theta$ .

Then  $(X_n, Y_n)_{n \geq 0}$  is called a *hidden Markov model* (HMM).

# Relation to finite mixtures

- $(X_n, Y_n)_{n \geq 0}$  : HMM
- Then  $(Y_n)_{n \geq 0}$  is stationary, and the marginal distribution of the  $Y_i$  is the **finite mixture**

$$\sum_{i=1}^d \pi_i f_{\theta_i}.$$

Parameters of the HMM:

- Entries  $(\alpha_{i,j})_{i,j=1,\dots,d}$  of  $P$ ,
- Parameters  $\theta_i$  of the **state-dependent distributions**  $f_{\theta_i}$ .

# Example: Epileptic seizure counts

- **Data:** Number of epileptic seizures each day of a patient at the British Columbia's Children's Hospital over 204 days.
- Count Data  $\rightarrow$  Poisson distribution. **But:**

$$\text{Mean} = 0.75, \quad \text{Variance} = 1.1$$

$\rightarrow$  **Overdispersion.**

- Standard model: Finite mixture of Poisson distributions.
- Due to time-series structure (one-step correlation = 0.25):  
 $\rightarrow$  HMM with Poisson-distributed state-dependent distributions.

# Overview

1. Hidden Markov Models (HMMs)
2. Likelihood inference for HMMs
3. Testing w.r.t. the marginal distribution
4. Testing for two states in an HMM

# Likelihood in HMMs

Parameter vector:  $\omega = (\alpha_{11}, \dots, \alpha_{1,d-1}, \alpha_{2,1}, \dots, \alpha_{d,d-1}, \theta_1, \dots, \theta_d)$ .

Log likelihood function:

$$L_n(\omega) = \log \left( \sum_{x_1=1}^d \cdots \sum_{x_n=1}^d \pi_{x_1} f(Y_1; \theta_{x_1}) \prod_{j=2}^n \alpha_{x_{j-1}, x_j} f(Y_j; \theta_{x_j}) \right).$$

Score Vector:  $D_\omega L_n(\omega)$ .

Maximum Likelihood Estimator:  $\hat{\omega}$ : Argmax of  $L_n(\omega)$ .

# Likelihood Inference I

- Consistency of the MLE:  $\hat{\omega} \rightarrow \omega_0$  almost surely (Leroux 1992).
- Law of large numbers for the observed Fisher Information

$$\frac{1}{n} D_{\omega} D_{\omega}^T L_n(\tilde{\omega}) \rightarrow -\mathcal{J},$$

$\mathcal{J}$  : Fisher Information

- Asymptotic normality of the score:

$$\frac{1}{\sqrt{n}} D_{\omega} L_n(\omega) \xrightarrow{\mathcal{L}} N(0, \mathcal{J}).$$

→ Bickel, Ritov und Ryden (1999).

- Asymptotic normality of the MLE:

$$\sqrt{n}(\hat{\omega} - \omega_0) \xrightarrow{\mathcal{L}} N(0, \mathcal{J}^{-1}).$$

- LRT for regular hypotheses (Guidici, Ryden und Vandekerckhove 2000)

# MLE: Epileptic seizures

- Parameters of the **state-dependent distributions**:

$$\lambda_1 = 0.262 \text{ and } \lambda_2 = 1.167$$

- **Transition matrix**:

$$P = \begin{pmatrix} 0.973 & 0.027 \\ 0.035 & 0.965 \end{pmatrix}$$

→ **Stationary distribution**:

$$\hat{\pi} = (0.567, 0.433).$$

# Likelihood inference II

Asymptotics for the LRT for hypotheses with parameters on the boundary

→ Dannemann and Holzmann (2007a).

Examples:

- Testing  $H : \alpha_{i,j} = 0$ .
- Testing  $H : \alpha_{i,j} \geq \alpha_{k,l}$ .
- Testing for boundary values of  $\theta$ .
- Testing bimodality for  $d = 2$  for a normal HMM.

# Testing: Epileptic seizure counts

**Question:** Is there only a single state in which epileptic seizures occur?

Have  $\lambda_1 = 0.262$  and  $\lambda_2 = 1.167$ , thus test

$$H : \lambda_1 = 0.$$

LRT statistic:  $T_n = 10.25 \rightarrow$  p-value = 0.

**Question:** Is the state with low seizure probability significantly more often visited than the other state?

Have  $\hat{\pi} = (0.567, 0.433)$ , thus test

$$H : \pi_1 \leq \pi_2 \quad \text{against} \quad K : \pi_1 > \pi_2.$$

LRT Statistic:  $T_n = 0.111 \rightarrow$  p-value 0.369. Thus **not** significant.

# Overview

1. Hidden Markov Models (HMMs)
2. Likelihood inference for HMMs
3. Testing w.r.t. the marginal distribution
4. Testing for two states in an HMM

# Testing w.r.t. the marginal distribution

Marginal density:

$$g(x; \pi_1, \dots, \pi_{d-1}, \theta_1, \dots, \theta_d) = \sum_{i=1}^d \pi_i f(x; \theta_i),$$

where  $\pi P = \pi$ .

Parameters:  $\omega = (\pi_1, \dots, \pi_{d-1}, \theta_1, \dots, \theta_d)$ .

- Estimate  $\omega$  and test hypotheses for  $\omega$ .
- Test for the number of states, in particular  $H : d = 2$  against  $K : d \geq 3$ .

# Likelihood under independence assumption

$$L_n^{ind}(\omega) = \sum_{k=1}^n \log(g(Y_k; \omega)).$$

For the **observed Fisher Information**:

$$\frac{1}{n} D_\omega^T D_\omega L_n^{ind}(\omega_0) \rightarrow \Sigma = E D_\omega^T D_\omega (g(Y_k; \omega_0)),$$

For the **score vector**:

$$\frac{1}{\sqrt{n}} D_\omega L_n^{ind}(\omega_0) \xrightarrow{\mathcal{L}} N(0, \text{Cov}_0),$$

where in general

$$\Sigma \neq -\text{Cov}_0.$$

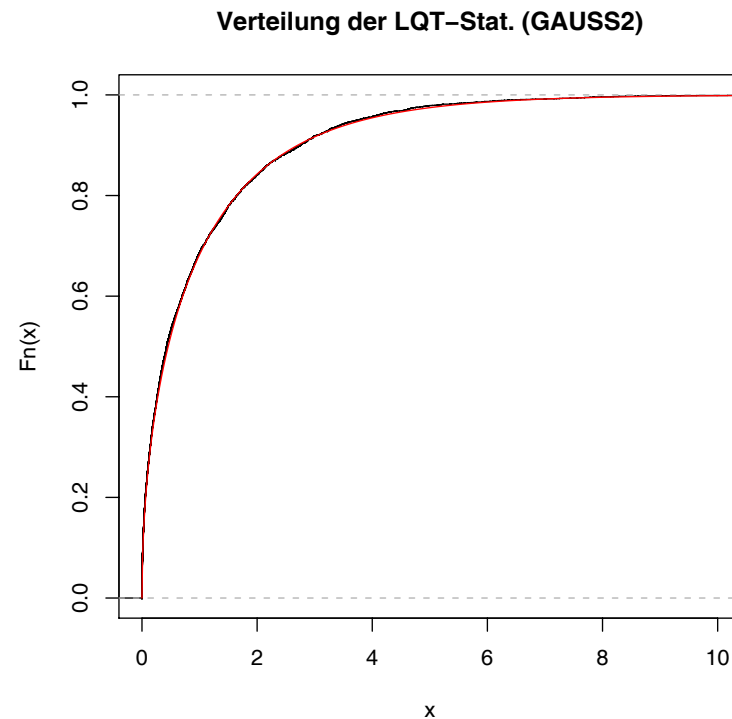
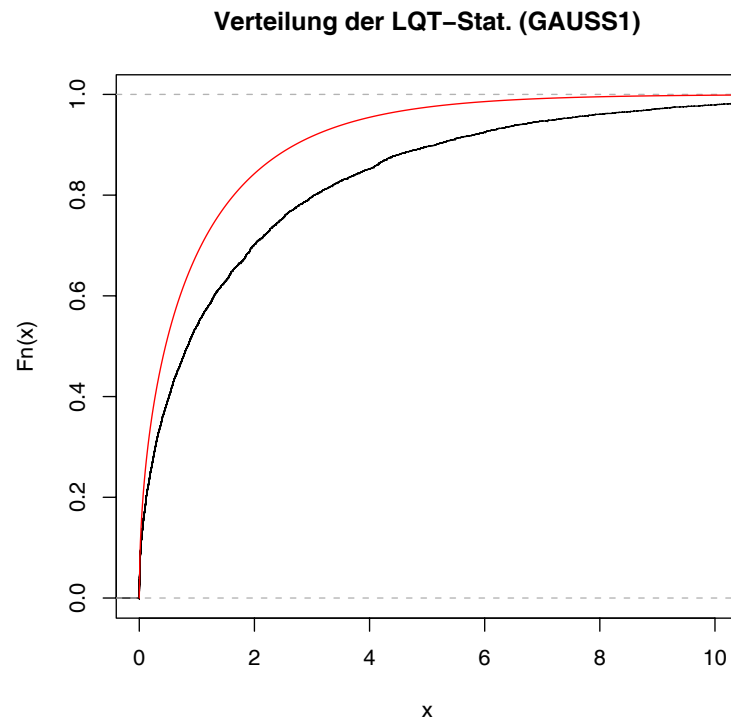
# LRT under independence

Since  $\Sigma \neq \text{Cov}_0 \rightarrow$  LRT not  $\chi^2$ -distributed.

Example:  $\omega = (\pi_1, \mu_1, \mu_2)$ ,  $\sigma_1, \sigma_2$  fixed,  $H : \pi_1 = 1/2$ ,  $n = 10^6$ .

GAUSS1:  $\alpha_{12} = 0.3$ ,  $\alpha_{21} = 0.3$ ,  $\mu_1 = 0$ ,  $\mu_2 = 3$ ;  $\sigma_1 = \sigma_2 = 1$

GAUSS2:  $\alpha_{12} = 0.5$ ,  $\alpha_{21} = 0.5$ ,  $\mu_1 = 0$ ,  $\mu_2 = 3$ ;  $\sigma_1 = \sigma_2 = 1$



# Overview

1. Hidden Markov Models (HMMs)
2. Likelihood inference for HMMs
3. Testing w.r.t. the marginal distribution
4. Testing for two states in an HMM

# Testing for homogeneity in finite mixtures

$f(\cdot; \theta)$  : One-parameter family of distributions on  $(\mathbb{R}, \mathcal{B})$ .

- Consider: Two-component mixture

$$g(x; p, \theta_1, \theta_2) = pf(x; \theta_1) + (1 - p)f(x; \theta_2),$$

where  $0 < p \leq 1/2$ ,  $\theta_1, \theta_2 \in \Theta \subset \mathbb{R}$ .

- Testing for **homogeneity**

$$H : p = 0 \text{ oder } \theta_1 = \theta_2$$

with the LRT for an independent two-component mixture.

- Asymptotic  $\chi^2$ -approximation **not valid!**

# Modified LRT for homogeneity

Chen, Chen and Kalbfleisch (2001) proposed:

Consider **modified** (log)-likelihood function

$$L_n^{mod}(p, \theta_1, \theta_2) = \sum_{k=1}^n \log(g(Y_k; p, \theta_1, \theta_2)) + C \log(4p(1-p)),$$

where  $(Y_k)$  from independent mixture.

Modified LRT statistic:

$$T_n^{mod} = 2(L_n^{mod}(\hat{p}, \hat{\theta}_1, \hat{\theta}_2) - L_n^{mod}(1/2, \hat{\theta}, \hat{\theta})).$$

Under regularity conditions under  $H$ :

$$T_n^{mod} \xrightarrow{\mathcal{L}} \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2.$$

# LRT for two components

- Chen, Chen and Kalbfleisch (2004): modified LRT for two components in finite mixture
- For HMMs: Use this test for the marginal distribution.
- Equivalent for testing  $d = 2$  in an HMM, if all  $\theta_i$  are distinct.

Dannemann and Holzmann (2007b): There is **no** correction needed for the dependence structure in the HMM.

# Example: Foetal Movements of a lamb

Data: Number of movements of a foetal lamb in 240 consecutive intervals of 5 seconds.

Count Data  $\rightarrow$  Poisson distribution. **But:**

$$\text{Mean} = 0.36, \quad \text{Variance} = 0.66$$

$\rightarrow$  **Overdispersion.**

Model with Poisson HMM.

- BIC : two states
- AIC : three states
- modified LRT for two components  $T_n^{mod} = 2.55$   
 $\rightarrow$  p-value 0.085.

# Summary

- HMMs: Time series models, generalize independent finite mixtures
- in particular useful for series of counts
- Likelihood theory, LRT under boundary conditions
- Likelihood theory w.r.t. the marginal distribution  
→ Testing for two states

# Outview

- Testing for switching regression / autoregression
- Testing for switching GARCH

# References

- Albert, P. S. (1991) A two-state Markov mixture model for time series of epileptic seizure counts. *Biometrics*, **47**, 1371–1381.
- Bickel, P. J., Ritov, Y. and Rydén, T. (1998) Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models, *Ann. Statist.*, **26**, 1614–1635.
- Chen, H., Chen, J. and Kalbfleisch, J. D. (2001) A modified likelihood ratio test for homogeneity in finite mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63**, 19–29.
- Chen, H., Chen, J. and Kalbfleisch, J. D. (2004) Testing for a finite mixture model with two components. *J. R. Stat. Soc. Ser. B*, **66**, 95–115.
- Leroux, B. G. and Puterman, M. L. (1992) Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. *Biometrics*, **48**, 545–558.

# Paper

Dannemann, J. und Holzmann, H. (2007a) Likelihood ratio testing for hidden Markov models under nonstandard conditions. Submitted.

Dannemann, J. und Holzmann, H. (2007b) Testing for two components in an HMM. Submitted.

Holzmann, H., Munk, A., Suster, M. and Zucchini, W. (2006) Hidden Markov models for circular and linear-circular time series.

*Environ. Ecol. Stat.* **13**(3), 325–347.