# Computational Advances and Applications of Hidden (Semi-)Markov Models

Jan Bulla

▶ **To cite this version:**

Jan Bulla. Computational Advances and Applications of Hidden (Semi-)Markov Models. Computation [stat.CO]. Université de Caen, 2013. tel-00987183

**HAL Id: tel-00987183**

**https://theses.hal.science/tel-00987183**

Submitted on 5 May 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Computational Advances and Applications of Hidden (Semi-)Markov Models

Habilitation thesis

by

Jan Bulla

from

Hannover, Germany

Presented on November 29th 2013 at the Université de Caen Basse Normandie

**Reviewers**

- Prof. Dankmar Böning
- Prof. Peter Thomson
- Prof. Maurizio Vichi

**Members of the jury**

- Prof. Frédéric Abergel
- Prof. Marco Alfò
- Prof. Damien Davenne
- Prof. Dankmar Böning
- Prof. Mohamed Didi Biha
- Prof. Francesco Lagona
- Prof. Nikolaos Limnios

# Contents

**1 Introduction** **1**

**2 Computational aspects of hidden (semi-) Markov models** **3**

2.1 Efficient parameter and confidence interval estimation for HMMs . . 4

2.2 Algorithms for working with HSMMs . . . . . 6

2.3 Perspectives . . . . . 8

**3 Applications of hidden (semi-)Markov models** **8**

3.1 H(S)MMs in Finance . . . . . 8

3.2 HMMs in Environmental Modeling . . . . . 18

3.3 HMMs in Bioinformatics . . . . . 21

3.4 HMMs in Marketing . . . . . 22

3.5 Perspectives . . . . . 24

**4 Statistics in medicine and biology** **24**

4.1 Function of a neuroprotective system . . . . . 25

4.2 nNOS overexpression is cardioprotective . . . . . 25

4.3 Calibration and performance of a temperature sensor . . . . . 26

4.4 DNA methylation in Biomphalaria glabrata . . . . . 26

4.5 Perspectives . . . . . 27

**5 Miscellaneous** **28**

5.1 Estimation of the stationary distribution of a semi-Markov chain . . 28

5.2 On choosing a mixture model for clustering . . . . . 31

5.3 Perspectives . . . . . 34

## List of Figures

1 Simulated observations and states and inferred states . . . . . 9
2 Empirical and model ACF . . . . . 11
3 Mean and standard deviation of the sojourn time distributions . . . 12
4 International indices with smoothing probabilities, 1993-2007 . . . . 15
5 Components of the estimated 3-state model . . . . . 20
6 Proportion of customers classified at the aggregate level. . . . . 23
7 Histogram of CpGo/e ratio in B. glabrata transcripts . . . . . 27
8 Estimated values values of $N_i(M)/M$ and true values of $1/\mu_{ii}$ . . . . 31
9 Clustering of Old Faithful Geyser data . . . . . 34

## List of Tables

1 Out-of-sample performance of Markov-switching strategies . . . . . 17
2 Model selection by SAIC/SBIC . . . . . 33

## Acknowledgements

First of all, I would like to thank Dankmar Böning, Peter Thomson, and Maurizio Vichi for the time and effort they invested in reviewing my habilitation thesis and Frédéric Abergel, Marco Alfò, Dankmar Böning, Pierre Denise, Mohamed Didi Biha, Francesco Lagona, and Nikolaos Limnios for accepting to be members of the jury.
I would like to sincerely thank the members of the LMNO for their support in past years, in particular Francesco Amoroso, Mohamed Didi Biha, Bernard Leclerd, and André Sesboüé for their numerous helpful suggestions and continuous assistance. Special thanks go to Christophe Chesneau for hundreds of hours he invested in correcting my poor French in the many documents I had to prepare for teaching and, more importantly, all the administrative 'dossiers'.
Moreover, I would also like to render thanks to all members of COMETE for giving me insight into many medical research projects, specially Nicolas Bessot and Pierre Denise for their regular assistance and explanations. The Fédération Normandie-Mathématiques equally deserves my thanks for supporting my invited researchers over the past years, as well as Francesco Lagona, Antonello Maruotti, Marco Picone, Tanya Mark, John Sansom, and Peter Thomson for the great research stays I spent in Canada, Italy, and New Zealand. Additionally, I would like to thank all my other co-authors for the successful work and the administrative staff of the LMNO for their help. Last but not least, thanks to Walter Zucchini for introducing me to the world of Hidden Markov Models - no one could have predicted this outcome!
On a more personal note, I am grateful to Oana Serea and Ingo Bulla for the continual support provided over so many years. I would also like to thank my parents for supporting me many years and allowing me to achieve many goals I would not even have dreamt of some time ago. My heartfelt thanks also go to Heinrich Hering and Wolfgang Schwarzwäller for their valuable advice over the many years.

# 1 Introduction

In this document I describe my main research activities, covering several subjects of different nature, with applied statistics being the common theme. My main research field lies in the area of time series and panel data analysis. In time series analysis, I often employ hidden Markov models (HMMs) and hidden semi-Markov models (HSMMs). These models, regularly addressed by the more general terms 'latent class' or 'regime-switching' models as well, provide flexible, general-purpose models for univariate and multivariate time series, be it for single series or multiple sequences of observations. Both HMMs and HSMMs have been used for more than two decades in various fields, good overviews are provided, e.g., by Bartolucci et al. (2012), Zucchini & MacDonald (2009), Cappé et al. (2005), Yu (2010). In addition to my primary research field, I have also been contributing to research in other areas (e.g. medicine and biology), employing a broader range of methods. The following paragraphs provide a brief overview.

The growing popularity of HMMs in the past decades has led to numerous papers on applications to real-world problems, and to increased interest in computational aspects. In order to estimate the parameters of the model, most researchers employ maximum-likelihood (ML) parameter estimation, mostly by implementing either a numerical maximization of the log-likelihood function or alternatively by the so-called expectation-maximization (EM) algorithm. In Bulla & Berzel (2008), we compare these two approaches by means of several criteria and propose a minor modification of the EM algorithm allowing the estimation of stationary HMMs. In the case of HSMMs, the situation is, however, slightly different. These models are a generalization of the well-known HMM - the main difference being that they allow for a greater flexibility for the choice of the sojourn time distributions, which implicitly follow a geometric distribution in the case of a hidden Markov chain. Unfortunately, with this flexibility comes a much higher computational burden. In order to make this class of models accessible to a larger number of researchers, we introduced `hsmm` (Bulla et al. 2010), a software package for the statistical computing environment R (R Development Core Team 2013). This package provides the most important algorithms required for working with HSMMs.

Applications of HMMs to real data have grown since comparably high computational power has become widely available. Nevertheless, for a long time the lion's share of the investigated models used Markovian mixtures of Gaussian distributions as observational distribution. One of the main reasons for the preference of this model may be its frequent use in the (theoretical) literature. Moreover, the implementation of the EM algorithm often becomes much more challenging with increasing model complexity.
In the financial area, HMMs have regularly been employed in the context of daily returns modelling. A popular contribution to this subject was authored by Rydén et al. (1998), who showed that HMMs reproduce most of the stylized facts about daily series of returns established by Granger & Ding (1995*a*,*b*). In Bulla & Bulla (2006),

we show that certain stylized fact can be described better by means of HSMMs, which are moreover preferred by common model selection criteria. Another important aspect of modelling daily returns of financial time series is the occurrence of outliers, or extreme values. Motivated by this phenomenon, we present a further extension of the most common model to conditional $t$-distributions in Bulla (2011), including models with unequal distribution types in different states.
Anyhow, if a high number of models has to be fitted in the absence of big computational power, simple HMMs with conditional Gaussian distributions still constitute an alternative to be considered. In Bulla et al. (2011), we propose a simple Markov-switching asset allocation model, which reduces the market exposure to periods of high volatility. In an out-of-sample context, the strategy proves profitable after taking transaction costs into account, rendering it more attractive than a simple buy-and-hold strategy.
Another popular concept in finance constitutes the single factor capital asset pricing model (CAPM). In the context of the CAPM, the systematic risk Beta ($\beta$) has historically been assumed to be constant over time and was mostly estimated via ordinary least squares (OLS). In Mergner & Bulla (2008), we investigate the time-varying behaviour of systematic risk for 18 pan-European sectors. Using weekly data over the period 1987-2005, six different modelling techniques in addition to the standard constant coefficient model are employed, including two Markov-switching models with Gaussian error term. A comparison of ex-ante forecast performances of the different models indicates that the random walk process in connection with the Kalman Filter is the preferred model to describe and forecast the time-varying behaviour of sector betas in a European context.
Outside of the financial context, HMMs also find their application in Marine research. In Bulla et al. (2012), we propose a model allowing the identification of sea regimes from environmental multivariate time series. This task is complicated by the mixed linear-circular support of the data, the occurrence of missing values, the skewness of some variables, and the temporal autocorrelation of the measurements. The proposed procedure is illustrated for a multivariate marine time series, and identifies a number of wintertime regimes in the Adriatic Sea.
In biology, the application of HMMs has a long history in the context of sequence and genome analysis. Basic models have become known, in particular, by the work of Durbin et al. (1998). Since then, many more applications and more complex models have followed. In Unterthiner et al. (2011), we have developed the Unknown Subtype Finder (USF), an algorithm based on a probabilistic model, which automatically determines which parts of an HIV-1 Group M input sequence originate from a subtype yet unknown.
Marketing is a field relatively young to the application of HMMs. Although marketing data sets are often of longitudinal form and of high dimensionality, thus well-suited for the application of complex models, the popularity of HMMs in this field has been only recently increased, basically since the paper of Netzer et al. (2008). In Mark et al. (2013), we contribute to this growing literature by extending the hurdle model to capture customer dynamics using a hidden Markov chain.

Leaving H(S)MM-related methods, an important field of applied statistics is, for example, medical (and biological) research. The application of statistics in medicine has a long history, and due to the growing amount of data collected, the complexity of the models required has been continuously increasing. For example, simple linear models are regularly replaced by more advanced techniques for panel data and time series analysis. Naturally, medical researchers often lack either the necessary time or formation or both for selecting, applying, and developing complex statistical methods fitting their research goals. Therefore, a growing part of my current research work is dedicated to supporting and consulting researchers in this field. The tasks covered range from application of basic statistical techniques over study design to the application and development of complex models for time series or longitudinal data. In Unzicker et al. (2005), we studied the expression and function of a neuroprotective system, the cannabinoid CB1-receptors. The hypothesis that nNOS overexpression is cardioprotective after ischemia/ reperfusion because of inhibition of mitochondrial function and a reduction in reactive oxygen species generation is treated in Burkard et al. (2010). The study carried out by Chapon et al. (2012) aims to assess the relevance of 1-point calibration procedure, within the framework of the development of a new ingestible telemetric temperature sensor. Last but not least, in Fneich et al. (2013) we deal with DNA methylation in B. glabrata, which is a snail intermediate host of Schistosoma mansoni, a tropical flatworm responsible for parasitic infection of humans.

In addition to the previously described section, I have also been contributing to other topics in applied statistics. These do not yet fit into a larger framework, as my work in these fields is just at the beginning. In the article of Barbu et al. (2012), we deal with the estimation of the stationary distribution of a discrete-time semi-Markov process. Moreover, we treat a clustering algorithm and the corresponding model selection criteria in Ngatchou-Wandji & Bulla (2013).

The remainder of this document is structured as follows. Section 2 reviews contributions to computational methods for H(S)MMs, while Section 3 focuses on model development for and applications of H(S)MMs. The subsequent Section 4 summarizes the results achieved in medical and biological research, and Section 5 presents those subjects which cannot be thematically embedded into the previous three sections and further development is ongoing.

## 2 Computational aspects of hidden (semi-) Markov models

In the following, we provide a brief overview of the results obtained by Bulla & Berzel (2008) and Bulla et al. (2010), which are dealing with different computational aspects of HMMs and HSMMs, respectively.

## 2.1 Efficient parameter and confidence interval estimation for HMMs

The growing popularity of HMMs in the past decades led to a high number of research papers, published either on theoretical aspects of these models or on their application to real-world problems. In order to estimate the parameters of the model, most researchers use maximum-likelihood (ML) parameter estimation. To maximize the likelihood, one of the following two approaches is implemented in most cases: numerical maximization of the log-likelihood function or, more popularly, the expectation-maximization (EM) algorithm. Although neither algorithm is superior to the other in all respects, researchers and practitioners who work with HMMs tend to use either of them, and ignore the other. Direct numerical maximization (DNM) has appealing properties, especially concerning the treatment of missing observations, flexibility in fitting complex models and the speed of convergence in the neighbourhood of a maximum. The main disadvantage of this method is its relatively small circle of convergence.

The statistical software R (R Development Core Team 2013) provides, inter alia, the two functions `nlm()` and `optim()` to perform DNM of the log-likelihood. The function `nlm()` carries out minimization of a function using a Newton-type algorithm (Dennis & Moré 1977, Schnabel et al. 1985). On the other hand, `optim()` offers, among other things, the Nelder-Mead simplex algorithm (Nelder & Mead 1965), a popular adaptive downhill simplex method for multidimensional unconstrained minimization, which does not require the computation of derivatives. In general, the Nelder-Mead algorithm is more stable; however, it may also get stuck in local minima and is rather slow compared to Newton-type minimization. In Bulla & Berzel (2008) we compare the two algorithms by means of several criteria. Since both the functions `nlm()` and the Nelder-Mead algorithm can only perform unconstrained numerical minimization, the parameter constraints need to be taken into account by different transformation procedures. For the TPM, we apply the TR-transformation described in Zucchini & MacDonald (1998). In order to meet the non-negativity constraint of some of the parameters of the state-dependent distributions, we use different transformations and compare their performance. The general tendency, that the unordered log-parameterization provides the most stable results, was found to hold for all simulated series.
Moreover, we propose a minor modification of the EM algorithm. In its standard implementation, this algorithm is unsuitable for fitting stationary HMMs, which can be resolved by a modified E-step. After assigning initial values to the parameters, the EM algorithm is implemented by successively iterating the E-step and the M-step until convergence is achieved.

*E-step*: Compute the $Q$-function

$$Q(\theta, \theta^{(k)}) = \mathbf{E}_{C^{(T)}} \left[ \log P(X^{(T)}, C^{(T)} | \theta) \,|\, X^{(T)}, \theta^{(k)} \right],$$

where $\theta^{(k)}$ is the current estimate of the parameter vector $\theta$, $X^{(T)} := \{X_1, X_2,$

$\ldots, X_T\}$ the sequence of observations and $C^{(T)} := \{C_1, \ldots, C_T\}$ the states of the latent process generated by an $m$-state homogeneous and irreducible Markov chain.

*M-step*: Compute $\theta^{(k+1)}$, the parameter values that maximize the function $Q$ w.r.t. $\theta$:

$$\theta^{(k+1)} = \underset{\theta}{\operatorname{argmax}}\ Q(\theta, \theta^{(k)}).$$

The EM algorithm for HMMs commonly presented in the literature works as follows. The $Q$-function of a HMM given by

$$Q(\theta, \theta^{(k)}) = \underbrace{\sum_{i=1}^{m} log\, \delta_i\, \psi_1(i)}_{A} + \underbrace{\sum_{i,j=1}^{m} \sum_{t=1}^{T-1} log\, \gamma_{ij}\, \xi_t(i,j)}_{B} + \underbrace{\sum_{i=1}^{m} \sum_{t=1}^{T} log\, p_i(s_t)\, \psi_t(i)}_{C},$$

$$\begin{aligned} \text{with } \psi_t(i) &:= P(\{C_t = i | S^{(T)} = s^{(T)}, \theta\}) \\ \text{and } \xi_t(i,j) &:= P(\{C_t = i, C_{t+1} = j | S^{(T)} = s^{(T)}, \theta\}, \end{aligned} \tag{1}$$

can be split of in three additive parts. In the M-step, the initial component $A$, the transition component $B$ and the observation component $C$ are maximized separately. In particular, the reestimation formulae for the first two components are

$$\delta_i^{(k+1)} = \psi_1(i) \text{ and } \gamma_{ij}^{(k+1)} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \psi_t(i)},\ i,j = 1, \ldots, m. \tag{2}$$

Clearly, this procedure fits a homogeneous, but non-stationary, HMM because the individual treatment of the initial and the transition component leads to an estimate $\hat{\boldsymbol{\delta}}$ which is not the stationary distribution of $\hat{\boldsymbol{\Gamma}}$. In order to fit a stationary Markov chain, we propose a modified M-step. The initial component $A$ and the transition component $B$ of Equation (1) have to be treated simultaneously with a stationarity constraint. I.e., the M-step for these two components becomes

$$\begin{aligned} &\max_{\gamma_{ij} \in \boldsymbol{\Gamma}} \left( \sum_{i=1}^{m} log\, \delta_i\, \psi_1(i) + \sum_{i,j=1}^{m} \sum_{t=1}^{T-1} log\, \gamma_{ij}\, \xi_t(i,j) \right) \\ &\text{with } \boldsymbol{\delta} \tilde{\boldsymbol{\Gamma}} = (0, 0, \ldots, 0, 1). \end{aligned} \tag{3}$$

The matrix $\tilde{\boldsymbol{\Gamma}}$ is obtained by replacing the last column of $\mathbf{1} - \boldsymbol{\Gamma}$ by the vector $(1, \ldots, 1)^T$ of length $m$. It results from the original form of the stationarity constraint for Markov chains $\boldsymbol{\delta}\boldsymbol{\Gamma} = \boldsymbol{\delta}$ with $\sum_{i=1}^{m} \delta_i = 1$ (see MacDonald & Zucchini 1997). The explicit calculation of a maximizing solution of the system of equations (3), which has to be carried out at each iteration, is more difficult than it appears at first glance. Even for the simplest non-trivial HMM with two states, the system becomes intractable. However, solving it with numerical methods is straightforward:

We implemented a Newton-type algorithm which takes the values of $\mathbf{\Gamma}$ resulting from the M-step of the preceding iteration as initial values for the M-step of the current iteration. Compared to the regular re-estimation given in Equation (2), the modified M-step does not slow down the estimation significantly.
In addition, we propose a hybrid algorithm that is designed to combine the advantageous features of the two algorithms and compare the performance of the three algorithms using simulated data from a designed experiment, and a real data set. For this hybrid algorithm, the estimation procedure starts with the EM algorithm and switches to a Newton-type algorithm when a certain stopping criterion is fulfilled. The hybrid algorithm would seem to provide an excellent compromise, because it is almost as stable as the EM-algorithm, but clearly faster.
Finally, we describe the results of a simulation experiment to assess the true coverage probability of bootstrap-based confidence intervals for the parameters. The mains finding is that the true coverage probability for bootstrap-based confidence intervals, obtained by parametric bootstrap, can be unreliable for models whose state-dependent parameters lie close to each other.

## 2.2 Algorithms for working with HSMMs

In the case of HSMMs, the situation is slightly different. These models, also referred to as explicit duration HMM or state duration HMM, is a generalization of the HMM that allows one to utilize more general sojourn time distributions. A HSMM consists of a pair of discrete-time stochastic processes $\{S_t\}$ and $\{X_t\}$. Similar to HMMs, the observed process $\{X_t\}$ is related to the unobserved semi-Markovian state process $\{S_t\}$ by the so-called conditional distributions.
Let $X_1^T := (X_1, \ldots, X_T)$ denote the observed sequence of length $T$. The same convention is used for the state sequence $S_t$. The set of parameters of the model is denoted by $\theta$. The state process is a finite-state semi-Markov chain, which is constructed as follows. A homogeneous Markov chain with $J$ states, labelled $1, \ldots, J$, models the transitions between different states. The stochastic process $\{S_t\}$ is specified by the initial probabilities $\pi_j := P(S_1 = j)$ with $\sum_j \pi_j = 1$, and the transition probabilities $p_{ij}$. For states $i, j \in \{1, \ldots, J\}$ with $j \neq i$, these are given by

$$p_{ij} := P(S_{t+1} = j \mid S_{t+1} \neq i, S_t = i)$$

satisfying $\sum_j p_{ij} = 1$, and $p_{ii} = 0$. The diagonal elements of the transition probability matrix (TPM) of a HSMM are required to be zero, since we separately model the runlength distribution. This distribution, also referred to as sojourn time distribution, is associated with each state. It models the duration the process $\{S_t\}$ remains in the state $j$ and is defined by

$$d_j(u) := P(S_{t+u+1} \neq j, S_{t+u} = j, \ldots, S_{t+2} = j \mid S_{t+1} = j, S_t \neq j).$$

The combination of a Markov chain, modelling state changes, and runlength distributions, determining the sojourn times in the states, define $\{S_t\}$ and illustrate

the main difference between the HMM and the HSMM. The semi-Markovian state process $\{S_t\}$ of a HSMM does not have the Markov property at each time $t$, but is Markovian at the times of state changes.
The observed process $\{X_t\}$ at time $t$ is related to the state process $\{S_t\}$ by the conditional distributions $b_j(x_t)$, which are either probability functions in the case of discrete conditional distributions or probability densities in the case of continuous conditional distributions:

$$b_j(x_t) = \begin{cases} P(X_t = x_t \mid S_t = j) & \text{for discrete } X_t \\ f(X_t = x_t \mid S_t = j) & \text{for continuous } X_t \end{cases}.$$

For the observation component, the so-called conditional independence property is fulfilled:

$$P(X_t = x_t \mid X_1^T = x_1^T, S_1^{t-1} = s_1^{t-1}, S_t = j, S_{t+1}^T = s_{t+1}^T) = P(X_t = x_t \mid S_t = j),$$

That is, the output process at time $t$ depends only on the value of $S_t$.

In his pioneering work, Ferguson (1980) introduced a particular HSMM with non-parametric sojourn time distributions in the field of speech recognition. Since then, the model was further investigated by various authors. Applications include, e.g., speech and pattern recognition (Levinson 1986, Sin & Kim 1995), the analysis of branching and flowering patterns, rainfall data, and user request patterns to a Web server (Guédon et al. 01, Sansom & Thomson 2001, Yu & Kobayashi 2003), gene finding (Burge & Karlin 1997, Lukashin & Borodovsky 1998), and protein secondary structure prediction (Schmidler et al. 2000). Unfortunately, the computational burden of estimating these models is much higher than in the case of HMMs (see Ferguson 1980, Levinson 1986, Guédon 2003). Therefore, despite the different fields of application, flexible software allowing the work with HSMMs has been available only to a limited extent. Before introducing `hsmm`, a software package for the statistical computing environment R (R Development Core Team 2013) in Bulla et al. (2010), the only existing implementation was included in the publicly available program `AMAPmod` (Godin & Guédon 2007). This program is tailored to specific problems and thus cannot be easily modified, in particular not without at least good knowledge of the `C++` programming language.
The `hsmm` package provides the most important algorithms required for working with HSMMs. The functions contained in the package address three important aspects of the HSMM. Firstly, the simulation of sequences of states and observations given the model specifications (sojourn time and conditional distributions) and parameters. Secondly, maximum likelihood estimation of the model parameters, given a sequence of observation and the model specifications. Thirdly, acquisition of information about the underlying state sequence via the Viterbi algorithm and the smoothing probabilities. The computationally intensive part, written in C, is independent of the selected distributional assumption. Extension to include other distributions requires minor modifications to the R code. Figure 1 shows an example of a simulated

sequence from a HSMM with logarithmic sojourn time distributions and Gaussian conditional distributions, together with the inferred latent states.

### 2.3 Perspectives

In its current form, both the package `hsmm` and a similar package presented by O'Connell & Højsgaard (2011) base on the algorithms of Guédon (2003). A flexible alternative is the approximation of HSMMs by specially structured HMMs as described, for example, by Durbin et al. (1998) or Langrock & Zucchini (2011). The advantage of the approximation technique is a high flexibility: in a current research project, covariate effects have been included in the parameters of the sojourn time distributions without major difficulties - this kind of extension would have required considerable work in the context of the original algorithms of Ferguson (1980) or those of Guédon (2003). Moreover, the computational complexity using the HMM approximation is low. Therefore, the next step might be an update of the package `hsmm` to take advantage of approximation techniques.

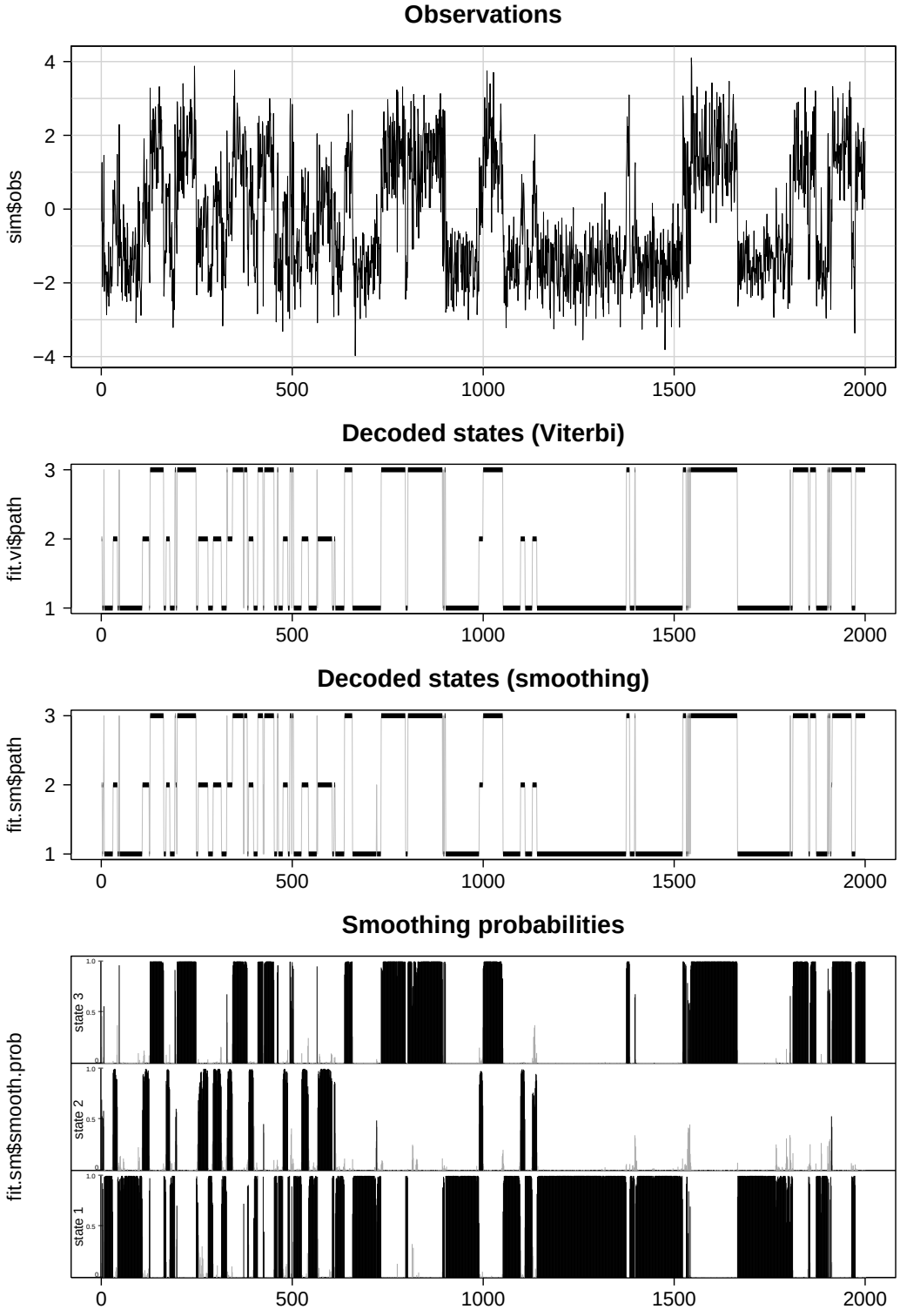
## 3 Applications of hidden (semi-)Markov models

In the past decades, rather simple as well as more complex H(S)MMs have regularly been applied to real data since computational power has become available at comparably low cost. Nevertheless, for a long time the lion's share of the investigated models concerned Markovian mixtures of Gaussian distributions. This may have been motivated, on the one hand, by the moderate computational effort due to the availability of closed formulae for the E-step of the popular EM algorithm. On the other hand, the Gaussian model has been very common in the theoretical literature as well. Nevertheless, as soon as this standard framework is left, often models which are more appropriate for a certain application can be developed - but the algorithmic challenges may increase substantially. We have been investigating the application of H(S)MMs in various fields, namely finance, environmental sciences, bioinformatics, and marketing.

### 3.1 H(S)MMs in Finance

In the financial area, HMMs are mostly termed 'regime-switching' models since the seminal paper of Hamilton (1989). Subsequently, these models have been used by an increasing number of researchers, in particular in the then growing field of daily returns modelling. A popular contribution to this subject was authored by Rydén et al. (1998), who showed that HMMs reproduce most of the stylized facts about daily series of returns established by Granger & Ding (1995*a*,*b*). A notable exception is the inability of these models to reproduce one ubiquitous feature of such time series, namely the slow decay in the autocorrelation function of absolute (or

Figure 1: Simulated observations and states and inferred states
The upper panel displays the sequence of observations. The second and third panel show the state sequence estimated by the Viterbi algorithm and the smoothing probabilities, respectively. The lower panel shows the smoothing probabilities (black if it is the maximum probability and gray otherwise).



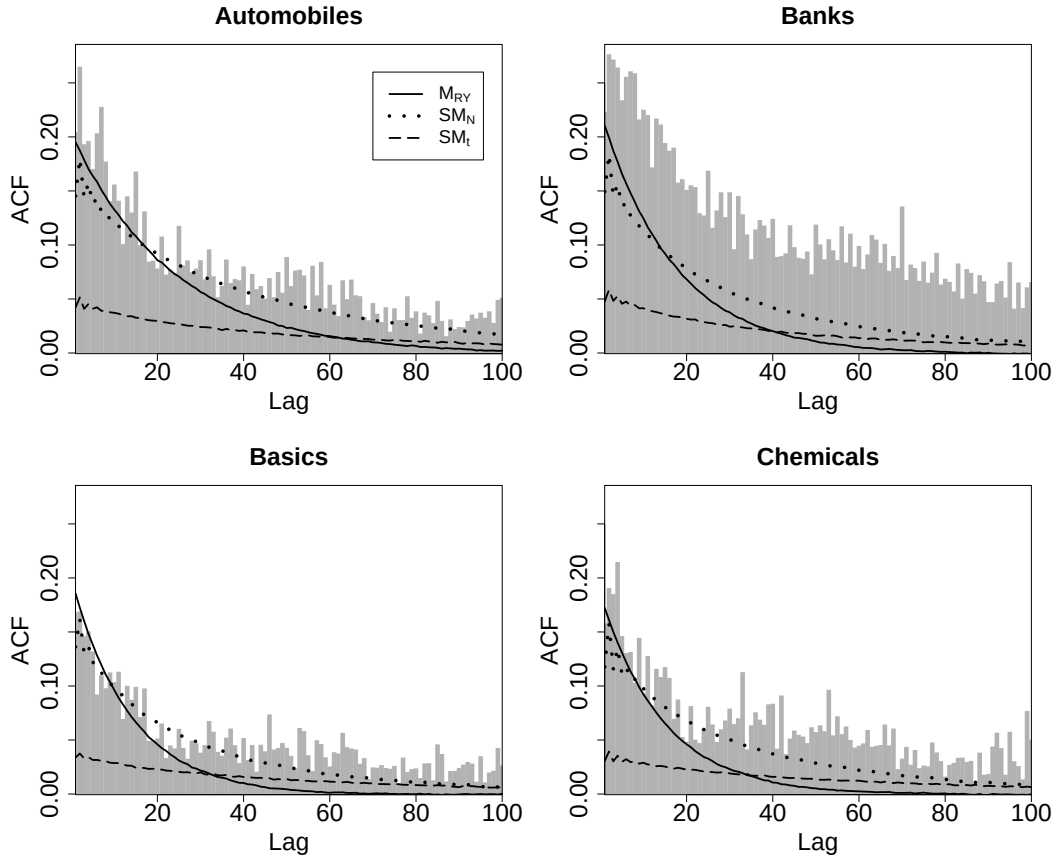
squared) returns. The lack of flexibility of a HMM to model the temporal higher order dependence can be explained by the implicit geometric distributed sojourn time in the hidden states.
In Bulla & Bulla (2006), we focus on modelling the distributional and temporal properties of daily return series by HSMMs. The two HSMMs explored are generalizations of the model presented by Rydén et al. (1998), termed 'RY' in the following. We show that slow decay in the autocorrelation function can be described much better by means of HSMMs, while all other stylized facts are equally well or better reproduced. Moreover, HSMMs are generally preferred by model selection criteria such as AIC or BIC. This is illustrated by examining the fit of two such models to 18 series of daily sector returns. It is remarkable that the estimated average sojourn times for the HSMMs are significantly lower than for HMMs, contradicting the generally assumed high persistence of volatility clusters.
More precisely, we generalize the model of RY by fitting a HSMM with negative binomial sojourn time distributions of the form

$$d_j(u) = \binom{u-2+r_j}{u-1} {p_j}^{r_j}(1-p_j)^{u-1},\ u = 1, 2, \ldots .$$

The number of parameters only increases by one per state, and for $r_j = 1,\ j \in 0, \ldots, J-1$ our model reduces to a HMM. While Granger & Ding (1995*a*,*b*) suggested a double exponential distribution to characterize daily returns, RY proposed mixtures of normal variables. We fit HSMMs with normal and $t$ distributed variables, respectively. In the following, the HMM of RY will be denoted by $M_{RY}$, the HSMM with normal conditional distributions by $SM_N$ and the HSMM with conditional $t$ distributions by $SM_t$. As to the number of states, all models investigated have two states, as RY noticed that the three-state models 'are less similar to each other' and that 'the estimation results seem heavily dependent on outlying observations' (Rydén et al. 1998). These findings were confirmed in our own preliminary analysis.
We treated two data sets: One containing the original returns, the second oultier-corrected returns following the approach of Granger & Ding (1995*a*). That is, setting values outside the interval $[\bar{r}_t - 4\hat{\sigma}, \bar{r}_t + 4\hat{\sigma}]$ equal to the value of the closest interval boundary to reduce extreme outliers, which may jeopardize the specification power of the ACF (Chan 1995). Fitting the models showed an average log-likelihood of $M_{RY}$ is 14200 and 14267 for the original and outlier-corrected data, respectively. It increases to 14236 (14299) and 14271 (14311) for $SM_N$ and $SM_t$, respectively. As the three models are hierarchically nested, a likelihood ratio test (LRT) may be applied with the null hypothesis of $r_1, r_2 = 1$ for the comparison $M_{RY}/SM_N$, and $\nu_1, \nu_2 = \infty$ (d.f. of $t$ distribution) for $SM_N/SM_t$. Using the original data, $SM_N$ is better than $M_{RY}$ at 0.1% level of significance for each of the 18 sectors. The same statement holds true for $SM_N/SM_t$, indicating that $SM_t$ provides the best fit to the data. The results for the outlier-corrected data are similar, with the limitation that level of significance is 1% for the comparison $SM_N/SM_t$. The only exception is the Utilities sector, where the test is not significant. The preference for the HSMMs

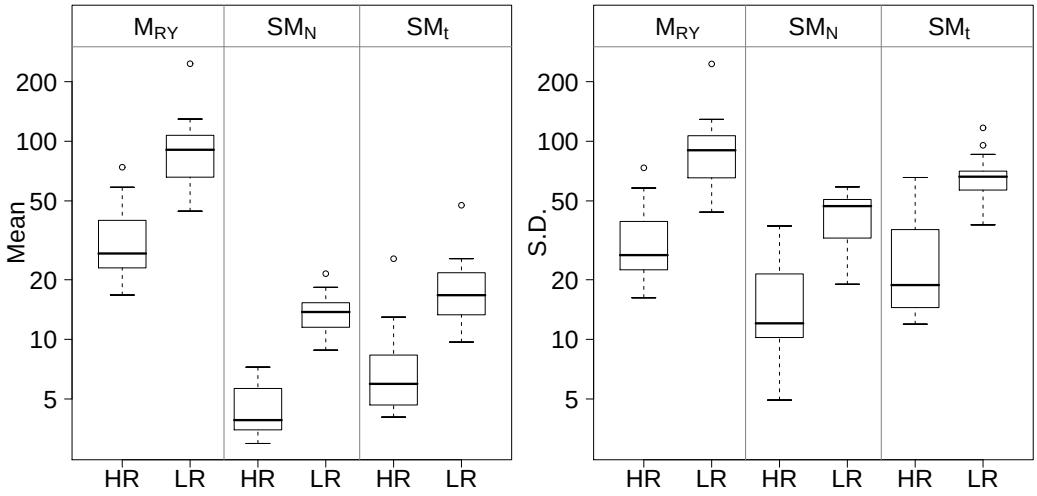
is supported by the Akaike information criterion which, on an average, decreases from -28388 (-28523) for the $M_{RY}$ to -28456 (28582) and -28522 (-28601) for $SM_N$ and $SM_t$, respectively.

Figure 2: Empirical and model ACF
The figure shows the empirical ACF (gray bars) and the ACF of the three models considered (solid, dotted, and dashed lines) at lag 1-100 of $r_t^2$ for the first four sectors.




For all details on the reproduction of stylized facts, we refer to Bulla & Bulla (2006), and present only results on the ACF of squared returns. Figure 2 displays the empirical ACF of squared returns as well as the ACF of the three models for the first four sectors. The other fourteen sectors and the outlier-corrected series show similar results and are therefore omitted. The solid line represents the ACF of $M_{RY}$, while the dotted and dashed lines represent $SM_N$ and $SM_t$, respectively.
The HMM shows the typical strong decay of the autocorrelations and is far from the gray empirical ACF, which confirms the results of RY. Both $SM_N$ and $SM_t$ reproduce this stylized fact much better than the HMM. However, the $SM_t$ looses some of its credibility due to the bad fit for the lags of lower order. Here, $SM_N$

Figure 3: Mean and standard deviation of the sojourn time distributions
The figure shows mean and standard deviation of the sojourn time distributions of all the 18 sectors, grouped by model and high-risk (HR)/low-risk (LR) states. The y-axis is logarithmic.




performs clearly better.
Finally, it may be noted that the main difference between HMMs and HSMMs is the sojourn time distribution. The results of the original and the outlier-corrected data do not differ substantially, and we therefore restrict our remarks to the analysis of the original data. Figure 3 shows the mean and standard deviation of the estimated sojourn time distributions by state. For every model, the expected sojourn time is higher in the 'low-risk state', where risk is measured in terms of variance of the respective conditional distribution.
It is remarkable that the average sojourn times for the HSMMs are significantly lower than for $M_{RY}$, i.e., the persistence of both the high- and the low-risk state is much lower.

Another important aspect of modelling daily returns of financial time series is the occurrence of outliers or extreme values. As mentioned before, these values may, if not excluded or capped, on the one hand have strong effects on the parameters of an estimated model. On the other hand, they may mask the empirical autocorrelation function (Chan 1995). Therefore, many analysis are preceded an outlier-correction, based on varying criteria. A different approach, based on the assumption that these rare values constitute integral part of the sample, is to keep them unchanged and adopt the distribution of the model considered to account for extreme values. In this case one may be obliged to depart from the Gaussian approach. In view of the application to return series, which are often heavy-tailed and leptokurtic (see, e.g., Gettinby et al. 2004, Harris & Küçüközmen 2001), a possible candidate for an extension of the Gaussian is the $t$-distribution.

In Bulla (2011), we present an extension of the model of Rydén et al. (1998) (RY) to conditional $t$-distributions, including models with unequal distribution types in different states. More precisely, we investigate two extensions: On the one hand, mixtures of Gaussians where the conditional means may take any value, allowing for skewed marginal distributions. This model is denoted by $M_N$ in the following. On the other hand, we introduce conditional $t$-distributions. The model denoted by $M_{Nt}$ is characterized by a $m-1$ Gaussian distributions and one $t$-distribution in $m^{\text{th}}$ state, i.e.

$$X_t = \mu_{s_t} + \epsilon_{s_t},\ \epsilon_{s_t} \sim \begin{cases} N(0, \sigma_i^2) & \text{for} \quad S_t \in \{1, \ldots, m-1\} \\ t(0, \sigma_m^2, \nu) & \text{for} \quad S_t = m \end{cases} .$$

Finally, $M_t$ denotes a model having exclusively conditions $t$-distributions. The choice of only one $t$-distribution is motivated by the application to daily returns: the $m^{\text{th}}$ state is supposed to represent that regime characterized by highest volatility and extreme observations. The last model is $M_t$ and has $m$ conditional $t$-distributions. In view of Robert & Titterington (1998) we require $\sigma_i < \sigma_{i+1}\, \forall\, i = 1, \ldots, m-1$ for all models considered to ensure their identifiability.

The main data analyzed in this paper are the daily returns calculated for the S&P500 index, covering the period from January 3$^{\text{rd}}$, 1928 to August 13$^{\text{th}}$, 2007. We segmented this long time series into periods of the length of eight calendar years, starting with 1928-1935 and ending with 2000-2007, which allows analyzing the performance of different models in many different time periods.

The main results are that a) the extended models reproduce various stylized facts of daily returns better than the common Gaussian model, and b) robustness to outliers and persistence of the visited states increases significantly. More precisely, the extensions to models with varying means and at least one conditional $t$-distribution seems to be reasonable. Often these models are more parsimonious than a Gaussian 3-state alternative, provide more stable parameter estimates, and are preferred by BIC and LRT. Moreover, $M_{Nt}$ and $M_t$ allow for skewed distributions, and reproduce the kurtosis as well as extreme observations (measured by outlier location tests) better than their competitors with Gaussian components. Generally, these two models show a superior performance when it comes to reproducing the stylized facts presented above in the context of HSMMs.

More important, in particular from a practical perspective, is the fact that that the introduction of conditional $t$-distributions often increases the state persistence significantly, resulting in longer and more stable volatility periods. This has considerable effects on the estimated state sequence, which is often utilized to link certain economic patterns to particular periods. We demonstrate this by means of an analysis of various international indices, Figure 4 displays the results. In this figure, we visualize the effect of extending to conditional $t$-distributions by plotting the smoothing probabilities and resulting state classifications. The top eight panels display the returns and smoothing probabilities for the S&P500 on the left and for the Nikkei on the right. For better identification, the background of the periods with

$P(S_t = 2) > P(S_t = 1)$ is shaded light gray. These two 2-state models visualize how large respectively small the effect of conditional $t$-distributions can be. The state classification of the S&P500 changes completely as the number of transitions (or state switches) reduces from 41 ($M_{RY}$) to 23 ($M_{Nt}$) and finally 5 ($M_t$). In case of the Nikkei, however, the (optical) difference between the models is much smaller. The lower eight panels show corresponding quantities resulting from the two 3-state models for CAC and DAX. The solid and dotted lines represent $P(S_t = 2)$ and $P(S_t = 3)$ respectively, and the background of the high-risk state is shaded dark grey. For all indices, the evolution of the estimated state sequence changes considerably.
The reason for the increased persistence of the models with conditional $t$-distribution(s) may most likely result from the excess kurtosis of the $t$-distributed component. Regarding the high-volatile state, the augmented probability mass around zero increases the persistence of this state in short periods of low volatility, while heavier tails still allow for catching extreme outliers. The argumentation for the low-risk state is similar: compared to the Gaussian distribution, heavier tails increase the state's persistence, because they allow for a higher robustness towards short periods of observations with comparably high volatility. Last but not least, on the one hand, the extended models with non-zero conditional mean confirm the link between periods of high volatility and falling stock prices. On the other hand, in contrast to other extensions of the commonly used Gaussian HMM, e.g. duration-dependent parameters (Maheu & McCurdy 2001, Peria 2002) or semi-Markovian models (Bulla & Bulla 2006), the estimation requires only a very moderate increase in computational complexity.

The single factor capital asset pricing model (CAPM) constitutes a very popular concept in finance. In the context of the CAPM, beta has historically been assumed to be constant over time, that is, market risk is treated as being constant. This assumption yield a benchmark for time-varying betas, an the excess-return market model with constant coefficients where an asset's unconditional beta can be estimated via OLS:

$$R_{it} = \alpha_i + \beta_i R_{0t} + \epsilon_{it}, \quad \epsilon_{it} \sim (0,\, \sigma_i^2), \tag{4}$$

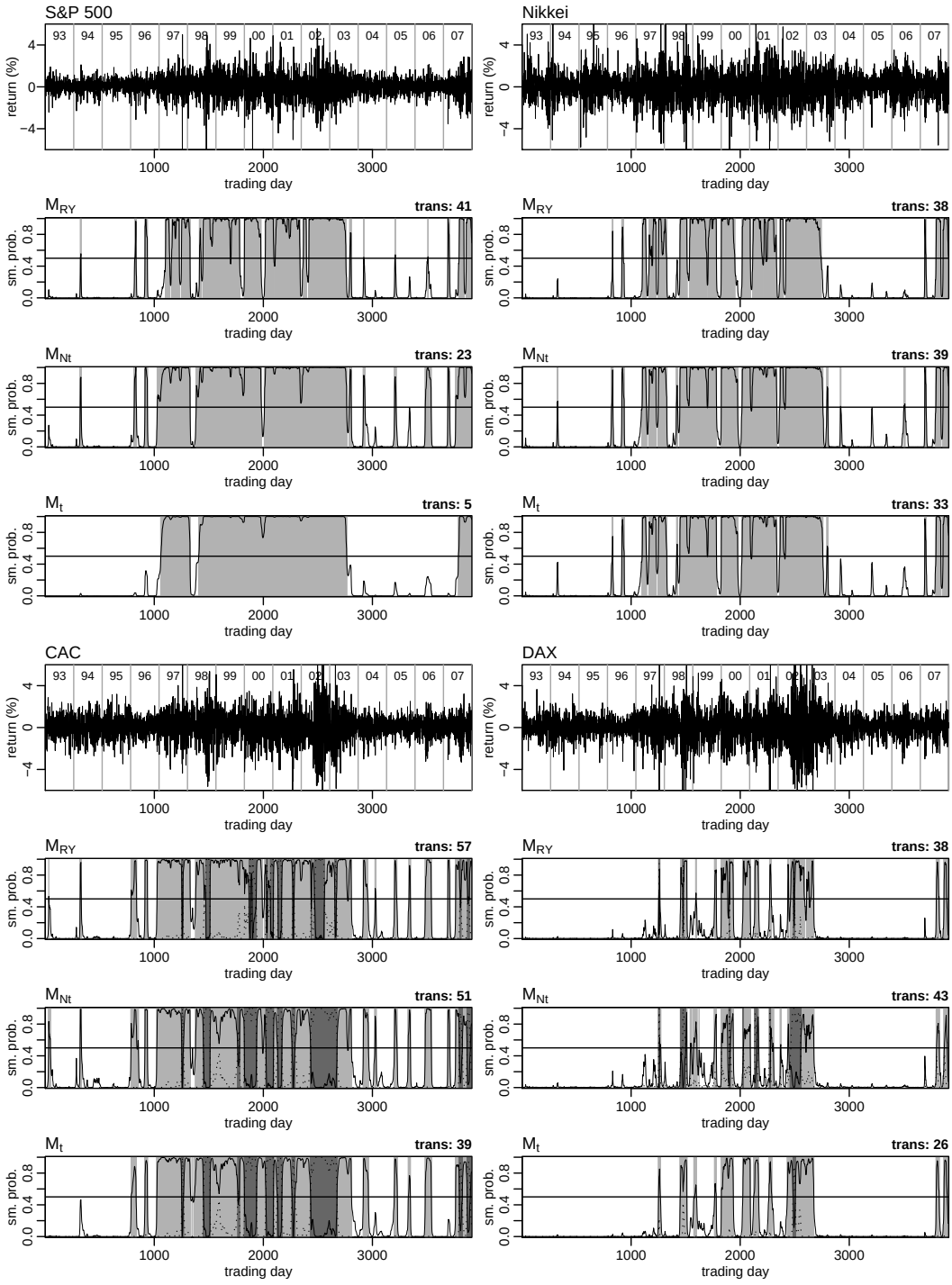
with

$$\hat{\beta}_i = \frac{Cov(R_0, R_i)}{Var(R_0)}, \tag{5}$$

where $R_{0t}$ denotes the excess return of the market portfolio and $R_{it}$ denotes the excess return to sector $i$ for $i = 1, \ldots, I$, each for period $t = 1, \ldots, T$. The error terms $\epsilon_{it}$ are assumed to have zero mean, constant variance $\sigma_i^2$ and to be independently and identically distributed (IID). Following the Sharpe (1964) and Lintner (1965) version of the CAPM, where investors can borrow and lend at a risk-free rate, all returns are in excess over a risk-free interest rate and $\alpha_i$ is expected to be zero; see Campbell et al. (1997, Ch. 5) for a review of the CAPM.
However, inspired by theoretical arguments that the systematic risk of an asset depends on microeconomic as well as macroeconomic factors, various studies over the

Figure 4: International indices with smoothing probabilities, 1993-2007

The figure shows percentage returns of the S&P 500, Nikkei, CAC, and DAX from 1993 to 2007. Below each return series, three panels display the corresponding the smoothing probabilities $P(S_t = i \,|\, X_1^T)$ for $M_{RY}$, $M_{Nt}$, and $M_t$, respectively. The background of periods with $\hat{s}_t = 2$ is shaded light gray. For the two 3-state models (DAX and CAC), the background of periods with $\hat{s}_t = 3$ is shaded dark gray. The smoothing lines themselves are solid and dotted for state 2 and 3, respectively. We omit $M_N$, because there is almost no visual difference to $M_{RY}$.



last three decades have rejected the assumption of beta stability (see, for example, Fabozzi & Francis 1978, Sunder 1980, Bos & Newbold 1984, Collins et al. 1987). In Mergner & Bulla (2008), we investigate the time-varying behaviour of systematic risk for 18 pan-European sectors. Using weekly data over the period 1987-2005, six different modelling techniques in addition to the standard constant coefficient model are employed: a bivariate t-GARCH(1,1) model, two Kalman Filter (KF)-based approaches, a bivariate stochastic volatility model estimated via the efficient Monte Carlo likelihood technique as well as two Markov switching models.
The two Markov switching approaches extend the relatively thin literature dealing with time-varying betas. Before our work, only two authors contributed to this subject. Fridman (1994) considers monthly data from 1980 to 1991 to analyze the excess returns of three oil corporation securities by fitting a two-state regression model, which leads to an improved assessment of systematic risk associated with each security. He also notes two effects. Firstly, beta increases whenever the process is in the more volatile state and, secondly, the higher volatility state tends to be less persistent than the lower volatility state. Huang (2000) also considers a Markov switching model with one high-risk and one low-risk state. Using monthly return data from April 1986 to December 1993, he performs several test to check the consistency of different states with the CAPM and rejected the hypothesis that the data were from the same state.
In Mergner & Bulla (2008), the data used are weekly excess returns calculated from the total return indices for eighteen pan-European industry portfolios, covering the period from 2 December 1987 to 2 February 2005. The DJ STOXX$^{SM}$ 600 index serves as a proxy for the overall market, and 3-month Frankfurt Interbank Offered Rate (FIBOR) as risk-free interest rate for calculating weekly excess returns.
Comparing the in- and out-of-sample forecast performances of the various techniques, the results of this study indicate that time-varying sector betas are best described by a random walk process, estimated by the use of the Kalman filter. While the in-sample results overwhelmingly support the KF approach, its superiority is only partly maintained out-of-sample where the advantage over its competitors is less pronounced. It is noteworthy that the out-of-sample forecast performance of the two proposed Markov switching models is inferior to that of any time-varying alternative and also to OLS. This suggests that HMMs may better serve for the ex-post identification of periods with different structure in-sample in this context, while preference may be given to other methods for forecasting tasks.

In situation where a high number of models has to be fitted having access only to limited computational power, simple Gaussian HMMs constitute an alternative to be considered. In Bulla et al. (2011), we propose a straightforward Markov-switching asset allocation model, which reduces the market exposure to periods of high volatility. The model employed is a simple two-state HMM, which nevertheless forms the basis of a profitable investment strategy. The idea behind this strategy is rather simple:

1. For time $t$, predict the hidden state $\hat{s}_t$.

Table 1: Out-of-sample performance of Markov-switching strategies

This table displays annualized mean returns (in %), standard deviations (in %) for the five indices and the Markov-switching strategies based on $\hat{s}_{t+1}^{f}$. Sharpe ratios are only reported for positive mean returns. "$Str.^{Vit.}$" denotes the Viterbi-based strategy. Every index is followed by the statistics of the respective strategy in the subsequent row.

| Name | Mean | S.D. | Sharpe ratio | # Forecasts | # Transitions |
|---|---|---|---|---|---|
| DAX | 7.24 | 22.0 | 0.292 | 5,796 | - |
| $Str.^{Vit.}$ | 7.76 | 13.0 | 0.437 | - | 84 |
| DJIA | 8.93 | 16.8 | 0.417 | 5,823 | - |
| $Str.^{Vit.}$ | 9.82 | 11.2 | 0.646 | - | 60 |
| NASDAQ | 6.82 | 32.0 | 0.272 | 3,383 | - |
| $Str.^{Vit.}$ | 8.63 | 14.0 | 0.464 | - | 31 |
| Nikkei | -4.30 | 22.6 | – | 3,925 | - |
| $Str.^{Vit.}$ | -2.28 | 13.9 | – | - | 89 |
| S&P 500 | 8.37 | 16.5 | 0.390 | 5,834 | - |
| $Str.^{Vit.}$ | 8.56 | 10.3 | 0.577 | - | 46 |

2. Determine the weights of the portfolio at time $t$. If $\hat{s}_t = 1$, invest 100% in the index $X_t$, else 100% in the risk-free asset (Cash),

where, without loss of generality, we assume $\sigma_1 < \sigma_2$. The intention is to reduce overall portfolio risk during volatile market periods by shifting from equities into the risk-free asset class. The data analyzed are daily returns for five major equity indices, each covering over 20 years: DAX, DJIA, NASDAQ 100, Nikkei 225 and S&P 500. The data for the DAX, DJIA and S&P 500 start in January 1976, whereas the records of the NASDAQ and Nikkei begin in October 1985 and January 1983 respectively. Following Ang & Bekaert (2002), we fix the return of our risk-free asset to an annual rate of 3%.

We examine the performance of a regime-based asset allocation strategy under realistic assumptions, and compare the results to a buy-and-hold strategy. As Table 1 shows, the strategy proves profitable In an out-of-sample context after taking transaction costs into account (fixed at 10 basis points (0.10%) for a one-way trade).

For all indices, the exposure to highly volatile periods is reduced. Investors following the strategy significantly reduce their risk in terms of the annualized standard deviation, on average by 41% (p-value of a paired t-test: 0.013). The highest degree of risk reduction is observed for the NASDAQ where the standard deviation of the strategy (14%) is not even half the risk of the index (32%). Moreover, all strategies outperform the respective index in terms of annual returns. The highest average annual excess return is realized for the Nikkei (201.6 bp), the lowest difference occurs for the S&P 500 (18.5 bp). This is naturally much lower than in-sample, however a not undesirable side-effect of avoiding volatile periods (p-value of a paired t-test: 0.0385). Consequently, the strategies exhibit much better Sharpe ratios than the respective indices (p-value of a paired t-test: 0.00163).

## 3.2 HMMs in Environmental Modeling

Leaving the financial context, H(S)MMs have been applied for some time in environmental sciences. Part of this area is marine research, where the development of models that help scientists to understand how air-sea interactions influence the sea surface constitutes a major goal. In Bulla et al. (2012), we propose a model allowing the identification of sea regimes from environmental multivariate time series. This task is complicated by the mixed linear-circular support of the data, by the occurrence of missing values, by the skewness of some variables, and by the temporal autocorrelation of the measurements. We address these issues simultaneously by an HMM-based approach, and segment the data into pairs of toroidal and skew-elliptical clusters by means of the inferred sequence of latent states.
More precisely, toroidal clusters are defined by a class of bivariate von Mises densities for modelling wind and wave direction simultaneously. The bivariate von Mises density in the form introduced by Singh et al. (2002) is a parametric distribution on the torus, which naturally embeds the bivariate normal distribution when the range of observations is small. Its density is given by

$$f(\boldsymbol{x};\boldsymbol{\beta}) = \frac{\exp\left(\beta_{11}\cos(x_1-\beta_1)+\beta_{22}\cos(x_2-\beta_2)+\beta_{12}\sin(x_1-\beta_1)\sin(x_2-\beta_2)\right)}{C(\boldsymbol{\beta})},$$

with normalizing constant

$$C(\boldsymbol{\beta}) = 4\pi^2 \sum_{m=0}^{\infty} \binom{2m}{m} \left(\frac{\beta_{12}^2}{4\beta_{11}\beta_{22}}\right)^m I_m(\beta_{11}) I_m(\beta_{22}),$$

where

$$I_m(x) = \frac{1}{\pi}\int_0^{\pi} e^{x\cos t}\cos(mt)dt$$

is the modified Bessel function of order $m$. This density can be viewed as a bivariate generalization of the von Mises distribution, where $\beta_{12}$ accounts for the statistical dependence between $x_1$ and $x_2$.
Moreover, a bivariate skew normal distribution is employed to define skew-elliptical clusters of wind speeds and wave heights. Following Lin (2009), we specify a bivariate skew normal density as a linear mixed model with positive random effects. More precisely, let $\varphi(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})$ be the bivariate normal density function with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We first introduce a bivariate random effect $\boldsymbol{v} = (v_1, v_2)$ with independent components, distributed according to two standard normal distributions truncated at 0, say

$$\boldsymbol{v} \sim f(\boldsymbol{v}) = \frac{\varphi(\boldsymbol{v};\boldsymbol{0},\boldsymbol{I})}{\int_{(0,+\infty)^2}\varphi(\boldsymbol{u};\boldsymbol{0},\boldsymbol{I})d\boldsymbol{u}} = \frac{2}{\pi}\exp\left(-\frac{1}{2}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{v}\right) \qquad \boldsymbol{v}\in[0,+\infty)^2.$$

Second, we assume that $\boldsymbol{y}$ follows a bivariate normal distribution conditionally on $\boldsymbol{v}$

$$f(\boldsymbol{y}|\boldsymbol{v};\boldsymbol{\gamma}) = \varphi(\boldsymbol{y};\boldsymbol{\mu}(\boldsymbol{v};\boldsymbol{\gamma}),\boldsymbol{\Sigma}(\boldsymbol{\gamma}))$$

with mean

$$\mu(\boldsymbol{v};\boldsymbol{\gamma}) = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} + \begin{pmatrix} \gamma_1' & 0 \\ 0 & \gamma_2' \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

and covariance matrix

$$\boldsymbol{\Sigma}(\boldsymbol{\gamma}) = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{12} & \gamma_{22} \end{pmatrix}.$$

In this setting, a bivariate skew normal distribution is obtained as

$$f(\boldsymbol{y};\boldsymbol{\gamma}) = \int_{(0,+\infty)^2} \varphi(\boldsymbol{y}|\boldsymbol{v};\boldsymbol{\gamma}) f(\boldsymbol{v}) d\boldsymbol{v}$$

and reduces to a bivariate normal distribution when the skewness parameters $\gamma_1' = \gamma_2' = 0$.

The core of the classification procedure is an EM algorithm accounting for missing measurements, unknown cluster membership, and random effects as different sources of incomplete information. The proposed procedure is illustrated for a multivariate marine time series, and identifies a number of wintertime regimes in the Adriatic Sea. The data that analyzed are time series of semi-hourly wave and wind directions, as well as wind speeds and wave heights, recorded in the period 12/12/2009 - 12/1/2010 by the buoy of Ancona, located in the Adriatic Sea at about 30 km from the coast.

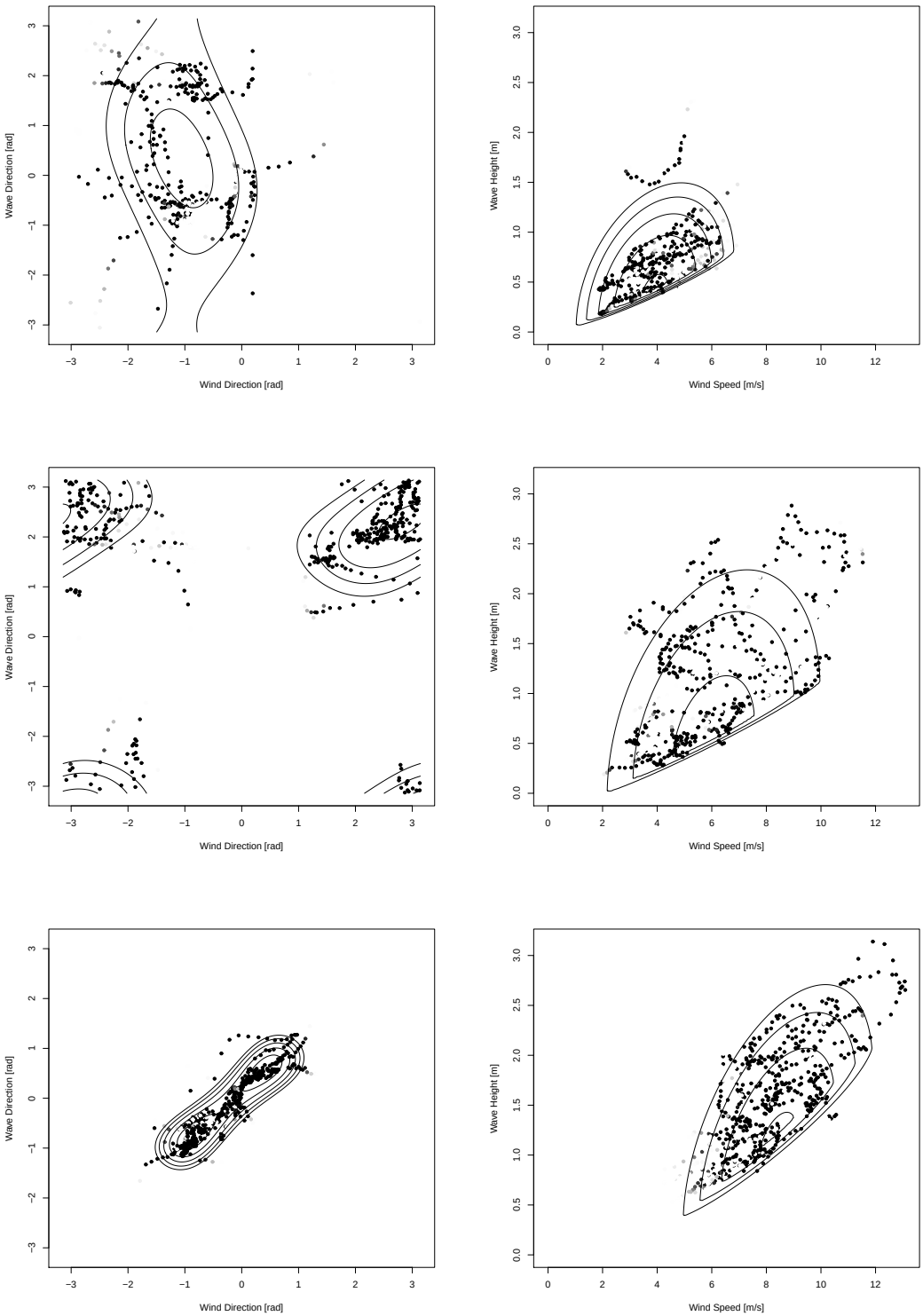
Figure 5 displays the components of the estimated 3-state model (selected by ICL) as log-densities through contour lines. Each scatter plot in Figure 5 includes the data points, filled with grey levels according to the posterior membership probabilities $\hat{p}_{tk}$ (black indicates $\hat{p}_{tk} = 1$). The first component of the model is associated with periods of calm sea: weak winds ($\gamma_2 = 3.665$) generate small waves ($\gamma_1 = 0.400$). In this regime, the shape of the joint distribution of wave and wind directions is essentially spherical ($\beta_{12}$ is barely significant) and centred at the average wind direction $\hat{\beta}_2 = -1.053$, corresponding to northwesterly Mistral episodes. As expected, wind and wave directions are poorly synchronized under this regime, because wave direction is more influenced by marine currents than by wind direction during weak wind episodes.

The second component is associated with Sirocco episodes ($\beta_2 = 2.840$). Compared to the first regime, wind and wave directions appear more synchronized ($\beta_{12} = 1.758$) and characterized by winds of higher speed ($\gamma_2 = 5.740$) and higher waves ($\gamma_1 = 0.514$). In this second regime, waves travel southeasterly along the major axis of the basin ($\beta_1 = 2.305$), driven by winds that blow from a similar directional angle ($\beta_2 = 2.840$). As there are neither coastlines nor mountains, there is little dispersion of energy in the interaction between wind and wave and, as a result, waves can reach significant heights. In studies of the Adriatic Sea, detection of Sirocco regimes is very important because it exposes Venice to the famous flooding tides when occurring in combination with luni-solar astronomical forces.

A similar phenomenon, although in the opposite direction, is captured by the third component of the model. In this regime, northern Bora jets ($\beta_2 = -0.210$) generate

Figure 5: Components of the estimated 3-state model

Log-densities of the circular (left) and linear (right) component of a three-states hidden Markov models. Contour lines are computed at the levels -0.5, -1.25, -2, -2.75, -3.5, and points are filled on a grey level scale according to their posterior probability of class membership, where black is associated with probability 1.



high waves ($\gamma_2 = 1.119$) that travel along the major axis of the basin ($\beta_1 = -0.081$). Compared to the other two regimes, waves and winds are much more synchronized ($\beta_{12} = 18.840$) and highly concentrated around one modal direction. Most of the wind energy is transferred to the sea surface and, as a result, the correlation between wind speed and wave height is larger than that observed under Sirocco or Mistral episodes. As expected, most of the profiles with the highest waves in the sample are clustered in this last regime.
The model describes the plasticity of the wind-wave interaction in the Adriatic Sea, indicating that the joint distribution of wind and wave data changes under different environmental regimes. Regime-switching does not only change directional and linear averages but also, and more interestingly, the correlation structure of the data. As a result, on the one side the weak (marginal) correlation between wind and wave observations is explained by the presence of a Mistral-specific regime of good weather conditions. On the other side, the model indicates that wind is an accurate predictor of wave-metric processes during a Bora episode, but that the level of accuracy decreases under Sirocco and almost vanishes under Mistral episodes. In summary, weather conditions should not be used to predict wave direction and height, without accounting for the latent, environmental heterogeneity of the data under study.

## 3.3 HMMs in Bioinformatics

In bioinformatics, the application of HMMs has a long history in the area of sequence and genome analysis. Basic as well as more advanced models have become known to a wider audience in the past two decades. Probably the most widely known application consists in profile HMMs (pHMM), presented in the well known book from Durbin et al. (1998). pHMMs are generative probabilistic models for families of nucleotide or protein sequences (our application deals with nucleotide sequences). Based on a multiple sequence alignment (MSA) of the family of nucleotide sequences to be modelled, a special HMM is constructed whose topology is determined by the MSA. If sampled from (which is the application the most easy to grasp and the least relevant in practice), this HMM yields nucleotide sequences, whereby the probability of yielding a particular sequence reflects how plausible it is that this sequence is a member of the given sequence family. When used for decoding, a pHMM provides the likelihood that a given query sequence is a member for a particular nucleotide family. Like this, one can determine to which family a query sequence probably belongs to when given multiple families.
One particular task in sequence analysis of HIV consists in determining whether or not a given HIV-1 Group M sequence stems - completely or in part - from some unknown HIV-1 Group M subtype (for HIV the term 'subtype' has been established instead of 'family'). This is important for phylogenetic inference as well as epidemiological monitoring. Nevertheless, a single algorithm only, the Branching Index (BI), has been developed for this task so far. Moving along the genome of a query sequence in a sliding window, the BI computes a ratio quantifying how closely the

query sequence clusters with a subtype clade. In its current version, however, the BI does not provide predicted boundaries of unknown fragments.
In Unterthiner et al. (2011), we have developed the Unknown Subtype Finder (USF), an algorithm based on a probabilistic model, which automatically determines which parts of an input sequence originate from a subtype yet unknown. The underlying model is based on a simple pHMM for each known subtype and an additional pHMM for an unknown subtype. The emission probabilities of the latter are estimated using the emission frequencies of the known subtypes by means of a (position-wise) probabilistic model for the emergence of new subtypes. We have applied USF to SIV and HIV-1 sequences formerly classified as having emerged from an unknown subtype. Moreover, we have evaluated its performance on semi-artificial HIV-1 recombinants and non-recombinant HIV-1 sequences. The results have been compared with the corresponding results of the BI.

## 3.4 HMMs in Marketing

Our last area of interest is marketing, a field in which the application of HMMs has only been recently pursued. Although marketing data sets are often of longitudinal form and of high dimensionality, thus well-suited for the application of complex models, one may note that only recently the popularity of HMMs has grown, in principal since the paper of Netzer et al. (2008). Few studies have examined the influence of marketing activities while accounting for customer dynamics over time. In Mark et al. (2013), we contribute to this growing literature by extending the hurdle model (Mullahy 1986) to capture customer dynamics using a hidden Markov chain. The resulting model can be interpreted as random coefficients hurdle model, extending the work of Alfò & Maruotti (2010).
In detail, the first part of the model, often called 'decision' or 'participation' component, which treats the probability of a non-zero observation is represented by a conditional Bernoulli distributions with parameter $\tau_{itj}$, i.e.,

$$\begin{aligned} P(y_{it} = 0 \mid \mathbf{x}_{it}^{(1)}, s_{it} = j) &= \tau_{itj} \quad \text{with} \\ \operatorname{logit}(\tau_{itj}) &= \alpha_{0j} + \alpha_{1j} x_{it1}^{(1)} + \cdots + \alpha_{lj} x_{itl}^{(1)}. \end{aligned}$$

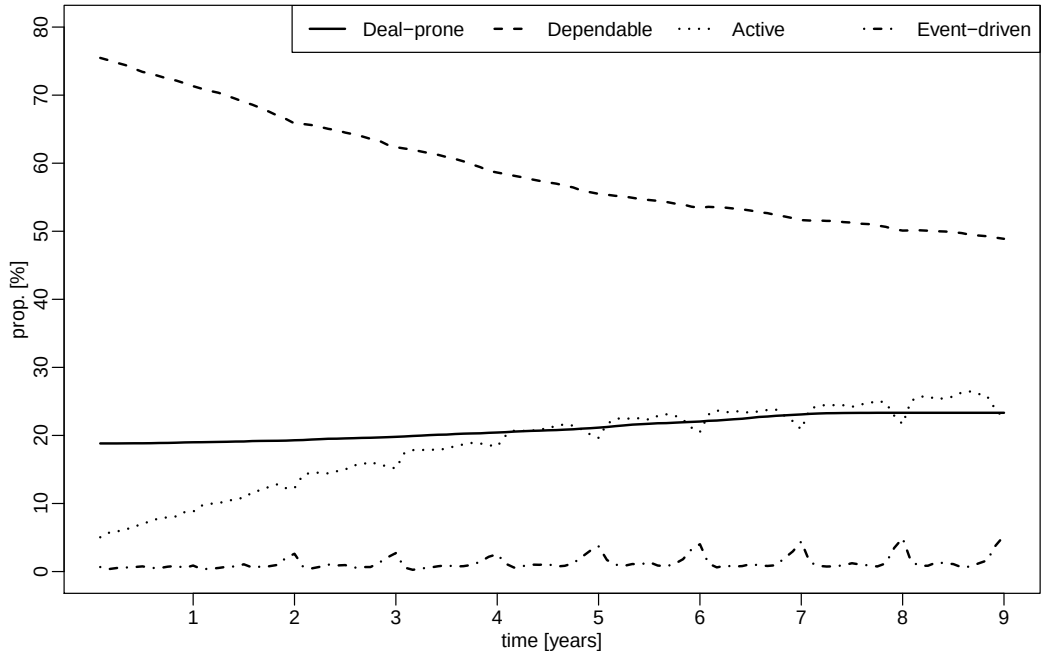
The second part of the model, which is often termed the 'utilization' component, consists of a conditional truncated (at 0) Poisson distribution with parameter $\lambda_{itj}$, i.e.,

$$\begin{aligned} P(y_{it} = k \mid \mathbf{x}_{it}^{(2)}, s_{it} = j) &= \frac{\frac{\lambda_{itj}^k}{k!} e^{-\lambda_{itj}}}{1 - e^{-\lambda_{itj}}} \quad \text{for } k = 1, 2, \ldots \text{ and} \\ \log(\lambda_{itj}) &= \beta_{0j} + \beta_{1j} x_{it1}^{(2)} + \cdots + \beta_{mj} x_{itm}^{(2)} \end{aligned}$$

Here, $y_{it}$ represents the observation recorded for individual $i$ at time $t$. The not necessarily identical sets of covariates in the two steps are $x_{it1}^{(1)}, \ldots, x_{itl}^{(1)}$ and $x_{it1}^{(2)}, \ldots, x_{itm}^{(2)}$,

respectively. Finally, $\alpha_{0j}, ..., \alpha_{lj}$ and $\beta_{0j}, ..., \beta_{mj}$ represent random coefficients driven by latent Markov chains $\{s_i\}_t$ with $j = 1, \ldots, J$ states.
We find our dynamic model performing better than static and latent class models. Our results suggest the customer base can be segmented into four segments: Deal-prone, Dependable, Active, and Event-driven. Each segment reacts differentially to marketing activities. Catalogues influence both purchase incidence and the number of orders, and this marketing activity has the largest impact on purchase incidence across all four segments. In contrast, retail promotions are more likely to influence the number of orders a customer will make for all of the segments except for the Deal-prone segment.

Figure 6: Proportion of customers classified at the aggregate level.
The figure shows the proportion of customers classified in the four states at the aggregate level over the observation period of 9 years.




Furthermore, empirical inquiry into the estimated state sequences also provides insight into the evolution of the relationships between the customers and retailer. Figure 6 displays the proportion of customers classified in the four states at the aggregate level. We find that the estimated proportion of Deal-prone customers increases relatively slowly over the observation period, from 18.8% to 23.3%. The estimated transitions (via a maximum a posteriori analysis of the posterior probabilities) underline that once customers enter this state, they basically remain in it.

As for the Dependable state, the estimated proportion of customers in this state diminishes significantly over the observation period. The initial state probabilities attribute 75.5% of our sample to this state; however, state membership decreases to 48.9% by year nine. In contrast, the Active state gains the largest number of customers over time. Initially, this state has a smaller proportion of customers, namely 5.0%, and grows to 22.5% of the customer base by the ninth year. Finally, we find that the trajectory of the Event-driven costumers is highly seasonal with peaks mostly during the holiday seasons. Finally, our empirical findings suggest that when customers make a transition, they are more likely to transition to more valuable states.
Summarizing, our results suggest that retailers would benefit from a segmentation model that incorporates customer dynamics. The model proposed and tested here will enable marketers to better understand the impact of marketing variables on buying behaviour.

### 3.5 Perspectives

For the future, many extensions and generalizations of the works above can be thought of. For example, currently the option of including seasonal dynamics into the model presented in Bulla et al. (2012) is examined. Moreover, in a different project, the marketing model is improved by including channel purchase behaviour, on the one hand by mixed HMMs, on the other hand by including a multinomial step in the hurdle model presented in Mark et al. (2013). In the context of a different research project, the application of mixed HSMMs is examined for modelling heart rates over a 24 hour period. Finally, the possibility of modelling spatial rainfall patterns by H(S)MMs is an ongoing project with researchers from Wellington (NZ). Altogether, the increased amount of panel data available in good quality may very probably require the development of many more sophisticated models, and (mixed) H(S)MMs constitute a promising approach for capturing the various dynamics potentially present in the data.

## 4 Statistics in medicine and biology

The application of statistics in medicine (and biology, and other fields) has a long history, as the large majority of papers published in medical journals contains at least basic statistical techniques. Unfortunately, most medical researchers lack the necessary formation for selecting and applying the correct methods, ranging from research design over statistical analysis to the presentation of the (statistical) results. This has been outlined in a large number of papers, e.g., Ioannidis (2005) or the regular contributions of Altman (1982, 1991, 1994, 2000) provide a good entrance to the subject.
The fact that the volume of recorded data has been continuously increasing in the

past years, in medicine not less than in other fields. Consequently, medical researches find themselves in a rather difficult situation: high amounts of complex data sets often require relatively sophisticated statistical techniques, often from the time series modelling or panel data analysis framework. Consequently, as medical researchers rarely possess the necessary time for working with complex statistical methods, a growing part of my current research work is dedicated to supporting and consulting researchers in this field. In the following, I provide a brief overview of these activities.

## 4.1 Function of a neuroprotective system

In Unzicker et al. (2005), we studied the expression and function of a neuroprotective system, the cannabinoid CB1-receptors, in an Endothelin (ETB)-deficient hippocampus. We show that CB1 expression in the hippocampus increases postnatally in all rats, but that the increase in CB1-receptor expression is significantly higher in ETB-deficient compared to wildtype littermates. Neuronal apoptosis decreases during brain maturation but remains on a significantly higher level in the ETB-deficient, compared to wildtype dentate. When investigating survival of hippocampal neurons in culture, we found significant protection against hypoxia-induced cell death with CB1-analogs (noladin, 9-tetrahydrocannabinol) only in ETB-deficient neurons. We suggest that CB1-receptor upregulation in the ETB-mutant hippocampus reflects an attempt to compensate for the lack of ETB-receptors.
The statistical methods used in this article are rather basic non-parametric statistics, such as the Wilcoxon rank-sum test (also called Mann-Whitney-U test) and the Spearman rank correlation test. In order to account for multiple comparisons, a Bonferroni correction has been applied.

## 4.2 nNOS overexpression is cardioprotective

The hypothesis that nNOS overexpression is cardioprotective after ischemia/ reperfusion because of inhibition of mitochondrial function and a reduction in reactive oxygen species generation is treated in Burkard et al. (2010). We succeed to demonstrate that conditional transgenic overexpression of nNOS resulted in myocardial protection after ischemia/reperfusion injury. More precisely, the corresponding ischemia/reperfusion experiments in isolated hearts showed a cardioprotective effect of nNOS overexpression. Infarct size in vivo was also significantly reduced. Besides a reduction in reactive oxygen species generation, this might be caused by nitrite-mediated inhibition of mitochondrial function, which reduced myocardial oxygen consumption already under baseline conditions.
As before, robust non-parametric methods such as Wilcoxon rank-sum test and Kruskal Wallis test were performed. For posthoc pairwise comparisons, the p-values were subject to Bonferroni correction. For analysing several longitudinal data sets, I employed generalized estimated equation (GEE) techniques. The target function was non-linear, a model specification which is not directly available in the `geepack`

package (Halekoh et al. 2006). Therefore, I determined the model parameters by a numerical optimization of the mean squared error of the model. To account for intra-individual dependencies, I utilized and unstructured working correlation matrix, and standard errors were determined by a Jackknife approach.

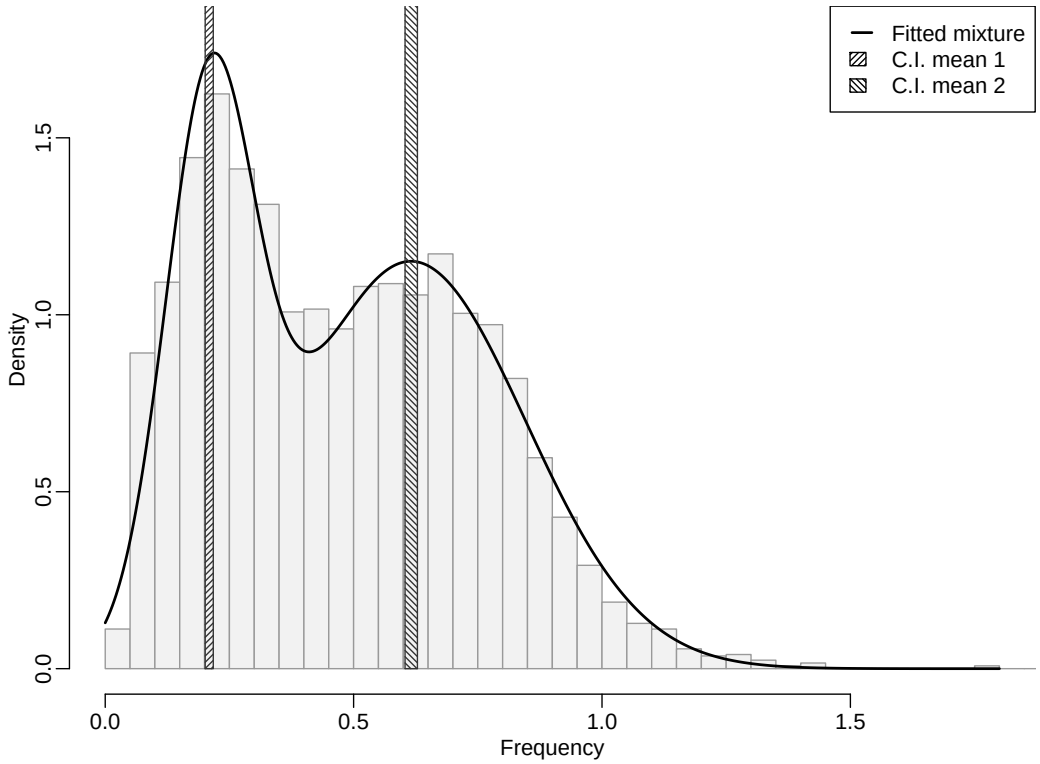
## 4.3 Calibration and performance of a temperature sensor

The study carried out by Chapon et al. (2012) aims to assess the relevance of 1-point calibration procedure, within the framework of the development of a new ingestible telemetric temperature sensor. The criteria used for performance assessment were the level of accuracy, and the time of inertia of the temperature sensor prototype (TSP) tested. First, the stability of the calibration bath was assessed. Then, the accuracy of 16 prototypes was evaluated for different target temperatures (ranging from 29℃ to 45℃). Finally, the inertia of TSP response was evaluated while increasing and decreasing the bath temperature. The results show that the difference between prototype and target temperature increases as bath temperature moves away from 37℃, however, the accuracy of the sensor conforms to applicable standards. Most TSP remain in the range of ±0.2℃ for each temperature level tested, but a linear, decreasing slope is observed. Data from time of inertia assessment show that probes were within the range of ±0.2℃ from the target temperature with a maximal delay of 150 seconds, which satisfies standard norms. However, our results indicate that a 1-point calibration procedure of the sensors appears non optimal, a 2-point calibration procedure should be performed to avoid the observed temperature data slope.
Most of the statistical methods utilized for this study fall into the class of linear models, such as uni- and multivariate regression, analysis of variance, and analysis of covariance. Only the non-linear model developed for the inertia analysis required a combination of generalized least squares and numerical optimization of likelihood for parameter estimation. In order to account for heteroscedasticity, the residual variance was assumed to follow an exponential structure of the variance function, and the correlation structure of the error was accounted for by imposing an AR(1) pattern.

## 4.4 DNA methylation in Biomphalaria glabrata

In the paper of Fneich et al. (2013), we report that DNA methylation, which is one of the carriers of epigenetic information, occurs in B. glabrata; approximately 2% of cytosine nucleotides are methylated. We describe the methylation machinery of B. glabrata. Methylation occurs predominantly at CpG sites, present at high ratios in coding regions of genes associated with housekeeping functions. We also demonstrate by bisulfite treatment that methylation occurs in multiple copies of Nimbus, a transposable element.
Apart from basic correlation analyses, the question whether observed/expected CpG

Figure 7: Histogram of CpGo/e ratio in B. glabrata transcripts
CpGo/e ratio was measured as a proxy to estimate the CpG methylation in transcripts from RNA-seq libraries from B. glabrata guadeloupian strain (Bg Gua). X axis: CpGo/e ratio, Y-axis density of transcripts. The figure displays a histogram of Bg Gua CpGo/e ratios with a fitted mixture distribution. The grey shaded bars represent 95% confidence intervals for the two mean values of the components. The estimated mean values of the two components are 0.209 and 0.616.




ratios are subject to a bimodal distribution with distinct modes required more advanced methods. To analyze this, I fitted simple Gaussian and a mixtures of Gaussians to the data and determined confidence bands for the two means, Figure 7 shows the results. Not surprisingly, the model selection criteria AIC and BIC also both indicate a clear preference for the mixture model in comparison to a single Gaussian distribution.

## 4.5 Perspectives

The work started on telemetric sensors commenced in Chapon et al. (2012) is currently being continued. Two paths are pursued: First, we examine the effect of sensor location in rats, and secondly we examine the performance of a more advanced device, the development of which took the findings of the prototype stage into ac-

count. Furthermore, I intend to improve the model for analyzing the methylation in transcripts from RNA-seq libraries, as mixtures of Gaussians are not very adequate due to the bounded support and skewness of the data. Moreover, I am working on measuring driving performance, because the models which are currently considered 'state of the art' allow many improvements. In particular, none of the current approaches takes any mean reversion (to the centre of the road) into account, which may lead to major improvements in describing driver's behaviour. Finally, I intend to further encourage the utilization of advanced statistical models for the projects I provide statistical council on, for example (non-) linear (generalized) mixed effects models and survival analysis for treating longitudinal data in medical studies. This is motivated by the currently frequently occurring loss of significant information when the data are pre-transformed to allow for the application of simpler, but less appropriate methods.

# 5 Miscellaneous

In this section, contributions to topics in applied statistics which do not fit into any of the previous sections are described. These may, in future, lead to larger or more numerous research projects in the same directions. However, due to the rather low amount of time spent on these projects, it seems too early for embedding them into a larger framework.

## 5.1 Estimation of the stationary distribution of a semi-Markov chain

In the article of Barbu et al. (2012), we deal with the estimation of the stationary (or limit) distribution of a discrete time, irreducible, and aperiodic semi-Markov process $Z = (Z_k)_{k\in\mathbb{N}}$ with finite mean sojourn times. The limit distribution of a semi-Markov chain (SMC) is given by

$$\pi_j = \frac{1}{\mu_{jj}} m_j = \frac{\nu(j)m_j}{\sum_{i\in E}\nu(i)m_i} = \frac{\nu(j)m_j}{\overline{m}}, \quad j \in E,$$

where the row vector $\nu = (\nu(1), \ldots, \nu(s))$ is the stationary distribution of the embedded Markov chain (EMC) $(J_n)_{n\in\mathbb{N}}$ and $E = \{1, ..., s\}$ the finite state space. Moreover, we denote by $\overline{m} := \sum_{i\in E}\nu(i)m_i$ the mean sojourn time of the SMC, by $\mu_{jj}$ the mean recurrence time of state $j$ for the SMC, and by $m_j$ the mean sojourn time in state $j$.
In the following, let us denote by $S = (S_n)_{n\in\mathbb{N}}$ the successive time points when state changes in $(Z_n)_{n\in\mathbb{N}}$ occur and by $J = (J_n)_{n\in\mathbb{N}}$ the successively visited states at these time points. Set also $X = (X_n)_{n\in\mathbb{N}^*}$ for the successive sojourn times in the visited states. Then, let us assume that we have an observed sequence of a SMC, censored at

fixed arbitrary time $M \in \mathbb{N}^*$, $(Z_0, \ldots, Z_M)$, or, equivalently, an observation of the associated Markov renewal chain $(J_n, S_n)_{n\in\mathbb{N}}$, $(J_0, X_1, \ldots, J_{N(M)-1}, X_{N(M)}, J_{N(M)}, u_M)$, where $u_M := M - S_{N(M)}$ is the censored sojourn time in the last visited state $J_{N(M)}$. All quantities required for an estimator of the stationary distribution of a SMC base on simple counts. For all states $i, j \in E$, let us introduce the two quantities:

- $N_i(M) := \sum_{n=0}^{N(M)-1} \mathbf{1}_{\{J_n=i\}} = \sum_{n=0}^{M} \mathbf{1}_{\{J_n=i, S_{n+1}\leq M\}}$ the number of visits to state $i$ of the EMC $(J_n)_{n\in\mathbb{N}}$, up to time $M$;
- $N_{ij}(M) := \sum_{n=1}^{N(M)} \mathbf{1}_{\{J_{n-1}=i, J_n=j\}} = \sum_{n=1}^{M} \mathbf{1}_{\{J_{n-1}=i, J_n=j, S_n\leq M\}}$ the number of transitions of the EMC $(J_n)_{n\in\mathbb{N}}$ from $i$ to $j$, up to time $M$.

Subsequently, we consider the empirical estimator of the stationary distribution of the EMC $(J_n)_{n\in\mathbb{N}}$ defined by:

$$\widehat{\nu}(i, M) = \frac{N_i(M)}{N(M)}, i \in E.$$

and the estimator for $m_i$,

$$\widehat{m}_i(M) = \frac{1}{N_i(M)} \sum_{k=1}^{N_i(M)} X_{ik}.$$

Consequently, an estimator of the mean sojourn time of the SMC, $\overline{m}$, is

$$\widehat{\overline{m}}(M) = \frac{1}{N(M)} \sum_{j\in E} \sum_{k=1}^{N_j(M)} X_{jk} = \frac{1}{N(M)} \sum_{k=1}^{N(M)} X_k,$$

and we obtain the following estimator of the stationary distribution of the SMC

$$\widehat{\pi}_i(M) = \frac{1}{\widehat{\overline{m}}(M) N(M)} \sum_{k=1}^{N_i(M)} X_{ik}, i \in E.$$

Our first results concern the following asymptotic properties:

$$\begin{aligned} N_i(M)/N(M) &\xrightarrow[M\to\infty]{a.s.} \nu(i), \\ N_{ij}(M)/N(M) &\xrightarrow[M\to\infty]{a.s.} \nu(i)p_{ij}, \\ N_i(M)/M &\xrightarrow[M\to\infty]{a.s.} 1/\mu_{ii}. \end{aligned}$$

Moreover, the estimators $\widehat{\nu}(i, M)$, $\widehat{m}_i(M)$, $\widehat{\overline{m}}(M)$, and $\widehat{\pi}_i(M)$ for the stationary distribution of the EMC, mean sojourn time in state $i$, mean sojourn time of the SMC, and stationary distribution of the SMC, respectively, are strongly consistent,

as $M$ tends to infinity for any state $i \in E$ of the SMC. Furthermore, for any fixed arbitrary state $i \in E$, we have

$$\sqrt{M}[\widehat{\pi}_i(M) - \pi_i] \xrightarrow[M\to\infty]{\mathcal{D}} \mathcal{N}(0, \sigma^2_{\pi_i}),$$

with asymptotic variance

$$\sigma^2_{\pi_i} = \frac{1}{\mu_{ii}} \frac{\frac{\sigma_i^2}{m_i^2} + \frac{\rho_{ii}^2 - \sigma_i^2}{(\mu_{ii}-m_i)^2}}{\left(\frac{1}{m_i} + \frac{1}{\mu_{ii}-m_i}\right)^2},$$

where $\rho_{ii}^2$ is the variance of the recurrence time of state $i$ and $\sigma_i^2$ is the variance of the sojourn time in state $i$.
We demonstrate the theoretical findings presented above by means of a short simulation study in Barbu et al. (2012). More precisely, we chose a 3-state SMC with shifted Poisson sojourn time distributions, that is,

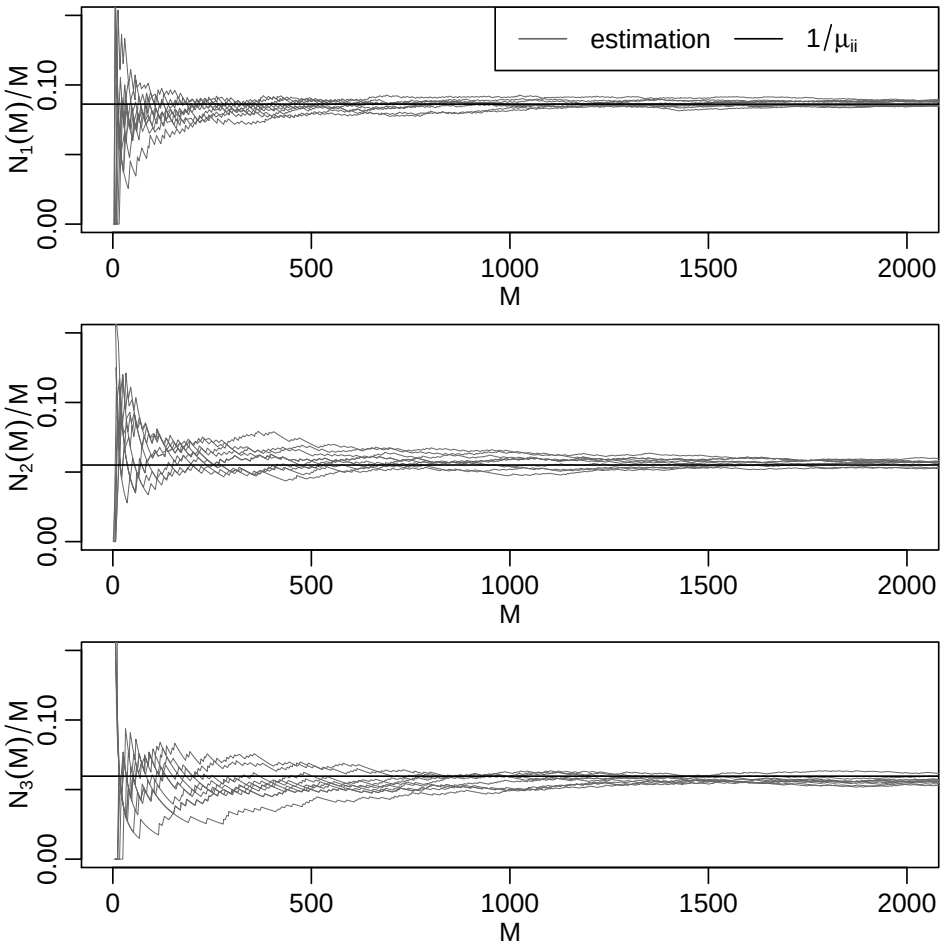
$$h_i(k) = \frac{\lambda_i^{k-1}}{(k-1)!} e^{-\lambda_i}.$$

The true parameter values of the model equal

$$\mathbf{p} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.7 & 0 & 0.3 \\ 0.8 & 0.2 & 0 \end{pmatrix} \text{ and } \lambda = \begin{pmatrix} 4 & 5 & 3 \end{pmatrix}.$$

Additionally, a uniform distribution is assumed for the initial distribution $\alpha$. From this parameterization directly follows $\nu = \begin{pmatrix} 0.429 & 0.274 & 0.297 \end{pmatrix}$, $\mu = \begin{pmatrix} 11.6 & 18.2 & 16.8 \end{pmatrix}$, $\pi = \begin{pmatrix} 0.431 & 0.330 & 0.239 \end{pmatrix}$, as well as $m = \begin{pmatrix} 5 & 6 & 4 \end{pmatrix}$ and $\sigma^2 = \begin{pmatrix} 4 & 5 & 3 \end{pmatrix}$. Thus, the true values of $\mu$ and $\pi$ are available for checking the consistency of the estimators. Therefore, we simulate 200 sequences with $N(M) = 500$ each, which is equivalent to values of $M$ moderately superior to 2000. As small example, Figure 8 provides a visual impression of convergence towards the true parameter values by means of 20 randomly selected sample paths. While the black horizontal lines represent to the true values of $1/\mu$ the gray lines result from the corresponding estimators.

Figure 8: Estimated values values of $N_i(M)/M$ and true values of $1/\mu_{ii}$
The figure shows the estimated values values of $N_i(M)/M$ (gray lines) from simulated series together with the true value of $1/\mu_{ii}$ (black lines) for states $i = 1, 2$, and 3.




## 5.2 On choosing a mixture model for clustering

In Ngatchou-Wandji & Bulla (2013), we treat a new clustering algorithm and the corresponding, newly introduced model selection criteria SAIC and SBIC. More precisely, we propose a method for clustering data and choosing a mixture model. A $d$-variate finite mixture model assumes that the data $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{R}^{dn}$ are a sample from a probability distribution with density of the form

$$f(\mathbf{u}|K, \theta) = \sum_{k=1}^{K} p_k \phi_k(\mathbf{u}|\mathbf{a}_k), \quad \mathbf{u} \in \mathbb{R}^d, \tag{6}$$

where $K$ is the number of components of the mixture, the $p_k$'s represent the mixing proportions, and the components $\phi_k(\cdot|\mathbf{a}_k)$'s are density functions, each with a

known form and depending on the parameter vector $\mathbf{a}_k$. Finally, $\theta := (\theta_1, \theta_2) := ((p_1, \ldots, p_K), (\mathbf{a}_1, \ldots, \mathbf{a}_K))$ represents the full parameter vector of the mixture $(m, K)$ at hand. The most popular mixture is the Gaussian mixture model, where $\phi_k(\cdot|\cdot)$ are Gaussian densities with mean $\mu_k$ and covariance matrix $\Sigma_k$. That is, $\phi_k(\cdot|\mathbf{a}_k) = \phi(\cdot|\mathbf{a}_k)$ is a $d$-variate Gaussian density with $\mathbf{a}_k = (\mu_k, \Sigma_k)$ for $k = 1, \ldots, K$.
The first main result is the derivation of a classification algorithm based on the so-called classification likelihood. Then, the likelihood conditioned on these clusters is written as the product of likelihoods of each cluster, and AIC- and BIC-type approximations, respectively, are applied. The resulting criteria, termed SAIC and SBIC, turn out to be the sum of the classical AIC or BIC, respectively, relative to each cluster plus an entropy term. More precisely, they are given by

$$\mathrm{SAIC}(K|\hat{\theta}_1, \mathbf{z}) = \sum_{k=1}^{K} \left( \sum_{\mathbf{x}_j \in C_k} \log \phi_k(\mathbf{x}_j|\hat{\mathbf{a}}_k) + n_k \log \hat{p}_k - d_{\mathbf{a}_k} \right) \text{ and}$$

$$\mathrm{SBIC}(K|\hat{\theta}_1, \mathbf{z}) = \sum_{k=1}^{K} \left[ \sum_{\mathbf{x}_j \in C_k} \log \phi_k(\mathbf{x}_j|\hat{\mathbf{a}}_k) + n_k \log \hat{p}_k - \frac{d_{\mathbf{a}_k}}{2} \log(n_k) \right],$$

where $C_1, C_2, \ldots, C_K$ correspond to the $K$ clusters, $n_k$ to the number of observations per component, and $d_{\mathbf{a}_k}$ the number of parameters in each component $k$.
The performance of our methods is evaluated by Monte-Carlo methods and on a real data set, showing in particular that the iterative estimation algorithm converges quickly in general, and thus the computational load is rather low. In the following, we present an extract of these results, which concern data for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. The observations are waiting times between eruptions and the durations of the eruption. This data set with 272 observations is included in the `datasets` package of `R`.
In order to initialize our clustering algorithm, called mb1 in the following, we follow two approaches. On the one hand, we use the k-means algorithm (function `kmeans` in `R`) to estimate an initial trajectory of $\mathbf{z} = (\mathbf{z}_1, ..., \mathbf{z}_n)$, which represent the missing data via sets of binary variables indicating whether $\mathbf{x}_i$ arises from the component $k$. The k-means algorithm itself is started by 100 different random sets, and we estimate models with two, three, and four components. On the other hand, we generate 1000 random paths for $\mathbf{z}$ (identical sampling probability for each component).
The initialization by random paths requires higher computational effort, however, also attains higher likelihoods. Therefore, this method is preferred for this example with relatively small sample size, and we do not further comment results from the k-means initialization. Fitting the 2-component model, the algorithm estimates clusters containing less than 5% of the sample for only 5% of the initial paths. However, this figure rises to ∼30% for the models with three/four components. These models have been removed, as they do not really utilize three respectively four components. Table 2 presents the results, showing an almost constant SAIC. Thus, according to this criterion, the parsimonious 3-component model should be

Table 2: Model selection by SAIC/SBIC

This table displays log-likelihood, SAIC, and SBIC of the estimated models with 2, 3, and 4 components, initialized by k-means or random paths.

| no. comp. | 2 | 3 | 4 |
|---|---|---|---|
| logL | -1131 | -1125 | -1120 |
| SAIC | -1141 | -1140 | -1140 |
| SBIC | -1155 | -1157 | -1158 |

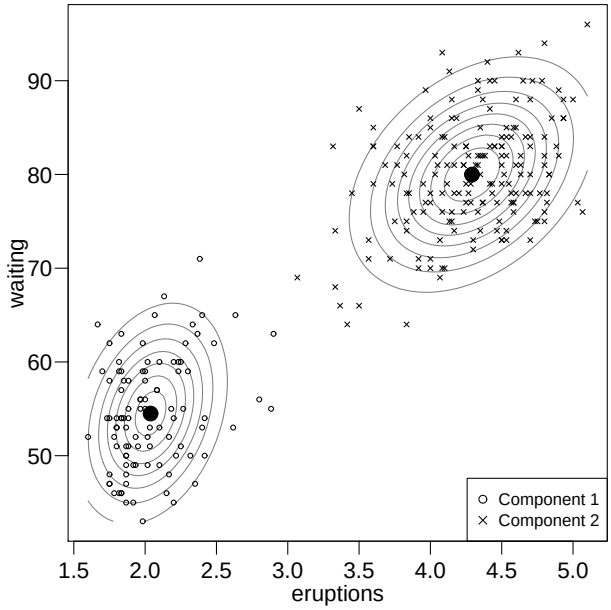
selected. The SBIC attains the highest value for two components, therefore the model with two components is chosen. Here, we set $K_{max} = 4$ because both SBIC and SAIC do not increase anymore when increasing the number of states from three to four.
Figure 9 displays the data, the estimated densities of the two components and the mapping of the observations to the components. The estimated parameters are

$$\mu_1 = \begin{pmatrix} 2.04 \\ 54.5 \end{pmatrix}, \mu_2 = \begin{pmatrix} 4.29 \\ 80.0 \end{pmatrix},$$

$$\Sigma_1 = \begin{pmatrix} 0.0712 & 0.452 \\ 0.452 & 34.1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.169 & 0.918 \\ 0.918 & 35.9 \end{pmatrix}.$$

The estimated values of $\mathbf{z}$ indicate that 35.7% and 64.3% of the observations belong to the respective components. Finally, the speed of convergence of the algorithm and its stability towards the initialization is of interest. The number of iterations required by the algorithm is rather manageable in the majority of cases. Considering the random initializations, the third quartile of the number of iterations lies at 14, 16, and 15 for models with 2, 3, and 4 components, respectively. The corresponding figures for the k-means initialization are 3, 9, and 13, confirming a low computational load. Concerning the stability of the algorithm towards initialization, it should be noted that mb1 failed to converge in 12% of the cases in the 2-component case. This may be attributed to a very poor initialization of the components. Convergence problems mainly occur because less than three observations belong to one of the components, such that the variance-covariance matrix cannot be estimated anymore. This phenomenon is mostly present at the initialization stage, but also happens rarely during the iteration steps. Moreover, the algorithm converged to the maximum likelihood of -1131 in 69.6% of the cases, which corresponds to the maximum attained by the k-means initialization.
For three and four components, respectively, the results are less satisfactory: First, almost all estimated models are (slightly) different to each other. Moreover, in 48%/76% of the samples the algorithm does not converge properly, determines components with very few observation ($< 10$), or estimates two or more components with (almost) identical parameters. Keeping in mind 'Garbage in, garbage out', this behaviour may however be viewed as the initialization paths are purely random and

Figure 9: Clustering of Old Faithful Geyser data
The figure shows bivariate data from the Old Faithful Geyser, clustered by mb1. The preferred model has two components, the centers of which are marked by filled circles. Contours result from the two estimated Gaussian densities.




may also underline the preference for the model with two components. Summarizing, random path initialization does not seem to provide better results than the k-means initialization, but rather entails convergence problems.

## 5.3 Perspectives

The two aforementioned works on semi-Markov chains and clustering algorithms do not fall into my main research interests. Therefore, it is hard to say when the next research project will fall into one of these two categories. However, together with some colleagues I am currently working on integer-valued data in order to obtain some experience in this new subject. More precisely, we are working on bivariate integer-valued autoregressive processes, and a bivariate Skellam distribution which allows for modelling bivariate correlated integer-valued data.

# References

Alfò, M. & Maruotti, A. (2010), 'Two-part regression models for longitudinal zero-inflated count data', *Can. J. Stat.* **38**(2), 197–216.

Altman, D. G. (1982), 'Statistics in medical journals', *Stat. Med.* **1**(1), 59–71.

Altman, D. G. (1991), 'Statistics in medical journals: Developments in the 1980s', *Stat. Med.* **10**(12), 1897–1913.

Altman, D. G. (1994), 'The scandal of poor medical research', *Brit. Med. J.* **308**(6924), 283–284.

Altman, D. G. (2000), 'Statistics in medical journals: some recent trends', *Stat. Med.* **19**(23), 3275–3289.

Ang, A. & Bekaert, G. (2002), 'International asset allocation with regime shifts', *Rev. Finan. Stud.* **15**(4), 1137–1187.

Barbu, V., Bulla, J. & Maruotti, A. (2012), 'Estimation of the stationary distribution of a semi-markov chain', *J. Reliab. Stat. Stud.* **Issue Special**, 15–26.

Bartolucci, F., Farcomeni, A. & Pennoni, F. (2012), *Latent Markov Models for Longitudinal Data*, Statistics in the Social and Behavioral Sciences, Chapman and Hall/CRC, London.

Bos, T. & Newbold, P. (1984), 'An empirical investigation of the possibility of stochastic systematic risk in the market model', *J. Bus.* **57**(1), 35–41.

Bulla, I., Bulla, J. & Nenadić, O. (2010), '`hsmm` - an **R** package for analyzing hidden semi-Markov models', *Comput. Statist. Data Anal.* **54**(3), 611–619.

Bulla, J. (2011), 'Hidden Markov models with $t$ components. Increased persistence and other aspects', *Quant. Financ.* **11**(3), 459–475.

Bulla, J. & Berzel, A. (2008), 'Computational issues in parameter estimation for stationary hidden Markov models', *Computation. Stat.* **23**(1), 1–18.

Bulla, J. & Bulla, I. (2006), 'Stylized facts of financial time series and hidden semi-Markov models', *Comput. Statist. Data Anal.* **51**(4), 2192–2209.

Bulla, J., Lagona, F., Maruotti, A. & Picone, M. (2012), 'A multivariate hidden markov model for the identification of sea regimes from incomplete skewed and circular time series', *J. Agr. Biol. Envir. St.* **17**(4), 544–567.

Bulla, J., Mergner, S., Bulla, I., Sesboüé, A. & Chesneau, C. (2011), 'Markov-switching asset allocation: Do profitable strategies exist?', *J. Ass. Man.* **12**(5), 310–321.

Burge, C. & Karlin, S. (1997), 'Prediction of complete gene structures in human genomic DNA', *J. Mol. Biol.* **268**(1), 78–94.

Burkard, N., Williams, T., Czolbe, M., Blömer, N., Panther, F., Link, M., Fraccarollo, D., Widder, J. D., Hu, K., Han, H., Hofmann, U., Frantz, S., Nordbeck, P., Bulla, J., Schuh, K. & Ritter, O. (2010), 'Conditional overexpression of neuronal nitric oxide synthase is cardioprotective in ischemia/reperfusion', *Circulation* **122**(16), 1588–1603.

Campbell, J. Y., Lo, W. & Craig, M. A. (1997), *The Econometrics of Financial Markets*, Princeton University Press, Princeton, NJ, USA.

Cappé, O., Moulines, E. & Ryden, T. (2005), *Inference in Hidden Markov Models*, Springer Series in Statistics, Springer-Verlag, New York - Heidelberg - Berlin.

Chan, W.-s. (1995), 'Time series outliers and spurious autocorrelations', *J. Appl. Stat. Sci.* **2**(2), 153–162.

Chapon, P. A., Gauthier, A., Bulla, J. & Moussay, S. (2012), 'Calibration and performance assessment of a temperature sensor prototype using a 1-point calibration procedure', *Rev. Sci. Instrum.* **83**(11), 114907.

Collins, D. W., Ledolter, J. & Rayburn, J. D. (1987), 'Some further evidence on the stochastic properties of systematic risk', *Journal of Business* **60**(3), 425–48.

Dennis, Jr., J. E. & Moré, J. J. (1977), 'Quasi-Newton methods, motivation and theory', *SIAM Rev.* **19**(1), 46–89.

Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. (1998), *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*, Cambridge University Press, Cambridge, UK.

Fabozzi, F. J. & Francis, J. C. (1978), 'Beta as a random coefficient', *J. Financ. Quant. Anal.* **13**(1), 101–116.

Ferguson, J. D. (1980), Variable duration models for speech, *in* 'Proceedings of the Symposium on the Applications of Hidden Markov Models to Text and Speech', Princeton, New Jersey, pp. 143–179.

Fneich, S., Dheilly, N., Adema, C., Rognon, A., Reichelt, M., Bulla, J., Grunau, C. & Cosseau, C. (2013), '5-methyl-cytosine and 5-hydroxy-methyl-cytosine in the genome of biomphalaria glabrata, a snail intermediate host of schistosoma mansoni', *Parasite Vector* **6**(167), 1–11.

Fridman, M. (1994), A two state capital asset pricing model, Ima preprint series, Institute of Mathematics and its Applications, University of Minnesota.

Gettinby, G. D., Sinclair, C. D., Power, D. M. & Brown, R. A. (2004), 'An analysis of the distribution of extreme share returns in the uk from 1975 to 2000', *J. Bus. Fin. Account.* **31**(5), 607–646.

Godin, C. & Guédon, Y. (2007), *AMAPmod and reference manual*, UMR CIRAD/INRA AMAP, Montpellier, France.
**URL:** *http://amap.cirad.fr/amapmod/refermanual15/accueil.html*

Granger, C. W. J. & Ding, Z. (1995*a*), 'Some properties of absolute return: An alternative measure of risk', *Ann. Economie Stat.* **40**, 67–91.

Granger, C. W. J. & Ding, Z. (1995*b*), Stylized facts on the temporal and distributional properties of daily data from speculative markets. Department of Economics, University of California, San Diego, unpublished paper.

Guédon, Y. (2003), 'Estimating hidden semi-Markov chains from discrete sequences', *J. Comput. Graph. Statist.* **12**(3), 604–639.

Guédon, Y., Barthélémy, D., Caraglio, Y. & Costes, E. (01), 'Pattern analysis in branching and axillary flowering sequences', *J. Theor. Biol.* **212**(4), 481–520.

Halekoh, U., Højsgaard, S. & Yan, J. (2006), 'The R package geepack for generalized estimating equations', *J. Stat. Softw.* **15**(2), 1–11.

Hamilton, J. D. (1989), 'A new approach to the economic analysis of nonstationary time series and the business cycle', *Econometrica* **57**(2), 357–384.

Harris, R. D. & Küçüközmen, C. C. (2001), 'The empirical distribution of uk and us stock returns', *J. Bus. Fin. Account.* **28**(5-6), 715–740.

Huang, H.-C. (2000), 'Tests of regimes-switching CAPM', *Appl. Financ. Econ.* **10**, 573–578.

Ioannidis, J. P. A. (2005), 'Why most published research findings are false', *PLoS Med.* **2**(8), e124.

Langrock, R. & Zucchini, W. (2011), 'Hidden markov models with arbitrary state dwell-time distributions', *Comput. Statist. Data Anal.* **55**(1), 715–724.

Levinson, S. E. (1986), 'Continuously variable duration hidden Markov models for automatic speech recognition', *Comput. Speech Lang.* **1**, 29–45.

Lin, T. I. (2009), 'Maximum likelihood estimation for multivariate skew normal mixture models', *J. Multivar. Anal.* **100**(2), 257â€“265.

Lintner, J. (1965), 'The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets', *Rev. Econ. Stat.* **47**, 13–37.

Lukashin, A. V. & Borodovsky, M. (1998), 'Genemark.hmm: new solutions for gene finding', *Nucleic Acids Res.* **26**(4), 1107–1115.

MacDonald, I. L. & Zucchini, W. (1997), *Hidden Markov and other models for discrete-valued time series*, Vol. 70 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

Maheu, J. M. & McCurdy, T. H. (2001), 'Identifying bull and bear markets in stock returns', *J. Bus. Econ. Statist.* **18**(1), 100–112.

Mark, T., Lemon, K. N., M., V., Bulla, J. & Maruotti, A. (2013), 'Capturing the evolution of customerâ€"firm relationships: How customers become more (or less) valuable over time', *J. Retailing* **89**(3), 231–245.

Mergner, S. & Bulla, J. (2008), 'Time-varying beta risk of pan-european industry portfolios: A comparison of alternative modeling techniques', *Europ. J. Finance* **14**(8), 771–802.

Mullahy, J. (1986), 'Specification and testing of some modified count data models', *J. Econometrics* **33**(3), 341–365.

Nelder, J. A. & Mead, R. (1965), 'A simplex method for function minimization', *Computer J.* **7**, 308–313.

Netzer, O., Lattin, M. & V., S. (2008), 'A hidden markov model of consumer relationship dynamic', *Market. Sci.* **27**(2), 185–204.

Ngatchou-Wandji, J. & Bulla, J. (2013), 'On choosing a mixture model for clustering', *J. Data Sci.* **11**(1), 157–179.

O'Connell, J. & Højsgaard, S. (2011), 'Hidden semi markov models for multiple observation sequences: The `mhsmm` package for R', *J. Stat. Softw.* **39**(4), 1–22.

Peria, M. S. M. (2002), 'A regime-switching approach to the study of speculative attacks: A focus on ems crises', *Empirical Econ.* **27**(2), 299–334.

R Development Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
**URL:** *http://www.R-project.org*

Robert, C. P. & Titterington, D. M. (1998), 'Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation', *Stat. Comput.* **8**, 145–158.

Rydén, T., Terasvirta, T. & Asbrink, S. (1998), 'Stylized facts of daily return series and the hidden Markov model', *J. Appl. Econom.* **13**(3), 217–244.

Sansom, J. & Thomson, P. J. (2001), 'Fitting hidden semi-Markov models to breakpoint rainfall data', *J. Appl. Probab.* **38A**, 142–157.

Schmidler, S. C., Liu, J. S. & Brutlag, D. L. (2000), 'Bayesian segmentation of protein secondary structure', *J. Comput. Biol.* **7**(1/20), 233–248.

Schnabel, R. B., Koontz, J. E. & Weiss, B. E. (1985), 'A modular system of algorithms for unconstrained minimization', *ACM Trans. Math. Software* **11**(4), 419–440.

Sharpe, W. F. (1964), 'Capital asset prices: A theory of market equilibrium under conditions of risk', *J. Financ.* **19**, 425–442.

Sin, B. & Kim, J. H. (1995), 'Nonstationary hidden Markov model', *Signal Process.* **46**(1), 31–46.

Singh, H., Hnizdo, V. & Demchuk, E. (2002), 'Probabilistic model for two dependent circular variables', *Biometrika* **89**(3), 719–723.

Sunder, S. (1980), 'Stationarity of market risk: Random coefficients tests for individual stocks', *J. Financ.* **35**(4), 883–896.

Unterthiner, T., Schultz, A.-K., Bulla, J. ., Morgenstern, B., Stanke, M. & Bulla, I. (2011), 'Detection of viral sequence fragments of HIV-1 subfamilies yet unknown', *BMC Bioinformatics* **12:93**.

Unzicker, C., Erberich, H., Moldrich, G., Woldt, H., Bulla, J., Mechoulam, R., Ehrenreich, H. & Sirén, A.-L. (2005), 'Hippocampal cannabinoid-1 receptor upregulation upon endothelin-B receptor deficiency: A neuroprotective substitution effect?', *Neurochem. Res.* **30**(10), 1305–1309.

Yu, S.-Z. (2010), 'Hidden semi-Markov models', *Artif. Intell.* **174**, 215–243.

Yu, S.-Z. & Kobayashi, H. (2003), 'An efficient forward-backward algorithm for an explicit-duration hidden Markov model', *Institute of Electrical and Electronics Engineers. Signal Processing Letters* **10**(1), 11–14.

Zucchini, W. & MacDonald, I. L. (1998), 'Hidden Markov time series models: Some computational issues', *Comput. Sc. Stat.* **30**, 157–163. Interface Foundation of North America.

Zucchini, W. & MacDonald, I. L. (2009), *Hidden Markov for Time Series: An Introduction Using R*, CRC Monographs on Statistics and Applied Probability, Chapman & Hall, London.