ARTICLE IN PRESS

Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Testing for the number of states in hidden Markov models

Hajo Holzmann *, Florian Schwaiger

Philipps-Universität Marburg, Fachbereich Mathematik und Informatik, Hans-Meerwein-Str., D-35032 Marburg, Germany

ARTICLE INFO

Article history:
Received 20 August 2013
Received in revised form 8 June 2014
Accepted 10 June 2014
Available online xxxx

Keywords:
Finite mixtures
Hidden Markov models
Hypothesis testing
Volatility states
Financial time series

ABSTRACT

Scale mixtures of normal distributions are frequently used to model the heavy tails of asset returns. A simple specification is a three component scale mixture, where the states correspond to high, intermediate and low volatility phases of the market. Tests for the number of states in hidden Markov models are proposed and used to assess whether in view of recent financial turbulences, three volatility states are still sufficient. The tests extend tests for independent finite mixtures by using a quasi-likelihood which neglects the dependence structure of the regime. The main theoretical insight is the surprising fact that the asymptotic distribution of the proposed tests for HMMs is the same as for independent mixtures with corresponding weights. As application the number of volatility states for logarithmic returns of the S&P 500 index in two HMMs is determined, one with state-dependent normal distributions and switching mean and scale, and the other with state-dependent skew-normal distributions with switching scale and structural mean and skewness parameters. It turns out that in both models, four states are indeed required, and a maximum-a-posteriori analysis shows that the highest volatility state mainly corresponds to the recent financial crisis.

1. Introduction

The mixture of distributions hypothesis for asset returns refers to specifications for which the marginal distribution of the returns is a scale-mixture of a standard, most often the normal distribution; see Shephard and Andersen (2009). A simple version is a finite scale-mixture of normal distributions, as proposed in Kon (1984), typically with three states corresponding to high, intermediate and low volatility. In order to induce volatility clustering one additionally requires positive serial correlation of the latent scale process, e.g. via a stationary finite-state Markov chain with high diagonal entries; see Rydén et al. (1998). For the resulting class of processes, called hidden Markov models (HMMs), we shall propose tests with a tractable asymptotic distribution for the number of states of the underlying unobserved regime. As application, we investigate whether in view of recent financial turbulences, three volatility states are still sufficient.

Formally, an HMM is a bivariate process $(S_t, X_t)_{t\geq 1}$, where $(S_t)_{t\geq 1}$ is an unobservable, finite-state Markov chain and $(X_t)_{t\geq 1}$ is the observable process with values in some Borel-subset of a Euclidean space, which are related as follows. Given $(S_t)_{t\geq 1}$, the $(X_t)_{t\geq 1}$ are conditionally independent, and for each $t \geq 1$, the conditional distribution of X_t depends on S_t only. The unobservable Markov chain is also called the regime or the latent process of the HMM. We shall assume that (S_t) is stationary and ergodic with state space $\mathcal{M} = \{1, \dots, k\}$, so that the stationary distribution $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)$ of the associated transition matrix $\gamma_{lm} = P(S_{t+1} = m | S_t = l)$, $l, m \in \mathcal{M}$, is uniquely determined.

The conditional distributions of X_t given $S_t = l$, $l = 1, \dots, k$, called the state-dependent distributions, are assumed to have densities $f(\cdot, \boldsymbol{\nu}, \boldsymbol{\vartheta}_l)$ from some parametric family w.r.t. some σ-finite measure. Thus, $\boldsymbol{\nu} \in \Theta_1 \subset \mathbb{R}^{d_1}$ is a structural parameter and $\boldsymbol{\vartheta}_l \in \Theta_2 \subset \mathbb{R}^{d_2}$ is state-dependent.

* Corresponding author.
E-mail addresses: holzmann@mathematik.uni-marburg.de (H. Holzmann), schwaige@mathematik.uni-marburg.de (F. Schwaiger).

http://dx.doi.org/10.1016/j.csda.2014.06.012

HMMs provide a flexible and very widely used class of models for dependent data, in particular in the presence of overdispersion (for series of count data) or unobserved heterogeneity; see the monographs by Zucchini and MacDonald (2009) for further examples of applications, and by Cappé et al. (2005) for a state-of-the-art overview of theoretical developments for HMMs. In order to model further stylized facts of asset-return series, various more general latent-state models were proposed, including hidden semi-Markov models (see Bulla and Bulla, 2006), switching GARCH models (see Augustyniak, 2014) as well as multivariate time-series models with switching dependence structure (see Stöber and Czado, 2014); see also Cappé et al. (2005) for an overview.

In statistical applications of HMMs, the selection of the number of states k of the latent process is a task of major importance. To this end, in certain models for fixed $k_0 \in \mathbb{N}$ we shall propose tests for the hypothesis

$$H_0 : k = k_0 \quad \text{against} \quad H_1 : k > k_0.$$

Since Gassiat and Keribin (2000) show that the LRT statistic for testing $k = 1$ against $k \geq 2$ for an HMM diverges to ∞, we shall follow the quasi-likelihood-based approach in Lindgren (1978) and Dannemann and Holzmann (2008) and proceed via the marginal finite mixture.

Specifically, we use the testing approaches for the number of states in a finite mixture by Chen et al. (2012) for normal state-dependent distributions with switching means and scales, as well as that by Li and Chen (2010) for a univariate switching parameter, extended to allow for nuisance parameters. Our main theoretical insight is the surprising fact that the asymptotic distribution of these tests for HMMs is the same as for independent mixtures with corresponding weights. Thus, our results also state that existing tests for independent mixtures are indeed robust against Markov-dependence in the regime.

The structure of the paper is as follows. In Section 2 we develop the relevant testing methodology. Section 3 contains results of an extensive simulation study.

As application, in Section 4 we determine the number of volatility states for logarithmic returns of the S&P 500 index in two HMMs, one with state-dependent normal distributions and switching mean and scale, and the other with state-dependent skew-normal distributions with switching scale and structural mean and skewness parameters. It turns out that in both models, four states are indeed required, and a maximum-a-posteriori analysis shows that the highest volatility state mainly corresponds to the recent financial crisis.

Appendix A contains a proof of the main insight that the asymptotic distribution of the test by Li and Chen (2010) remains the same for HMMs as for independent finite mixtures. In the supplementary material in Holzmann and Schwaiger (2014), additional simulation results as well as technical details are provided (see Appendix B).

2. Quasi-likelihood-based estimation and testing

In this section we present the statistical methodology. Section 2.1 introduces quasi-likelihood estimation for HMMs. Sections 2.2 and 2.3 contain the test statistics for the number of states together with their asymptotic distributions for normal HMMs as well as HMMs with univariate switching parameter, respectively. Section 2.4 contains a discussion, where we compare the test statistics and give intuition for their quite distinct asymptotic distributions and relate the results to the literature.

2.1. Quasi-likelihood estimation

Following Lindgren (1978) and Dannemann and Holzmann (2008), we consider a quasi-log-likelihood which neglects the dependence in the regime. For a given number of states k, set $\boldsymbol{\theta} = (\boldsymbol{\nu}^T, \boldsymbol{\vartheta}_1^T, \dots, \boldsymbol{\vartheta}_k^T)^T \in \Theta_1 \times \Theta_2^k$,

$$f_{mix}(X_t; \boldsymbol{\theta}, \boldsymbol{\pi}) := \sum_{j=1}^k \pi_j f\left(X_t | S_t = j; \boldsymbol{\theta}\right) = \sum_{j=1}^k \pi_j f(X_t; \boldsymbol{\nu}, \boldsymbol{\vartheta}_j),$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)$ with $\pi_j \geq 0, \pi_1 + \dots + \pi_k = 1$, and

$$l_n(\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{t=1}^n \log\left(f_{mix}(X_t; \boldsymbol{\theta}, \boldsymbol{\pi})\right).$$

The quasi-maximum-likelihood estimator (QMLE) is then given by

$$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}}) := \arg\max \left\{ l_n(\boldsymbol{\theta}, \boldsymbol{\pi}) : \boldsymbol{\theta} \in \Theta_1 \times \Theta_2^k, \ \sum_{j=1}^k \pi_j = 1, \ \pi_j \geq 0 \right\}.$$

As above, we shall most often suppress the number of components k in the notation for simplicity and only indicate it if confusion might arise otherwise. We are mainly interested in two specific situations, for which we extend the testing methodology for mixtures to the case of HMMs.

Normal HMMs. One of the most important classes of HMMs are those with normal state-dependent distributions. If both mean μ and variance σ^2 are allowed to switch, we have that $f_{X_t|S_t=j}(x) = \phi(x; \mu_j, \sigma_j), j = 1, \dots, k$, where ϕ denotes