# MCMC methods for the estimation of MS-ARMA-GARCH Models

Jan S. Henneke [a,*], Svetlozar T. Rachev [b,1] Frank J. Fabozzi [c]

[a] *University of Karlsruhe*
*WestLB AG London*

[b] *University of Karlsruhe, Germany*
*University of California, Santa Barbera, USA*

[c] *Yale School of Management, USA*

## Abstract

Regime Switching models, especially Markov switching models, are regarded as a promising way to capture nonlinearities in time series. They can account for sudden changes in the structure of the mean or the variance of a process and give a straightforward interpretation of these shifts. Such shifts would cause regular ARMA-GARCH models to imply non-stationary processes. Combining the elements of Markov switching models with full ARMA-GARCH models poses severe difficulties when it comes to understanding their dynamic properties and for the computation of parameter estimators. Maximum Likelihood estimation can become completely unfeasible due to the full path dependence of such models. Estimation methods such as the EM algorithm can be used for the ML estimation of simple Markov switching AR-ARCH models but become unfeasible if MA or GARCH components are introduced. In this article we formulate a full Markov switching ARMA-GARCH model and estimate based on the Bayesian framework. This facilitates the use of Markov Chain Monte Carlo methods and allows us to develop an algorithm to compute the Bayes estimator of the regimes and parameters of our model. The approach is illustrated on simulated data and with returns from the New York Stock exchange. Our model is then compared to MS-AR-ARCH variants and proves clearly to be advantageous.

*Key words:* Regime Switching, Markov Switching, Markov Chain Monte Carlo Methods, MCMC, Bayesian Estimation
*JEL Classification*: C11, C13, C51, C52, C63

[*] Corresponding author.

[1] Prof. S.T. Rachev gratefully acknowledges research support by grants from the Division of Mathematical, Life and Physical Sciences, College of Letters and Science,

## 1 Introduction

A central property of economic time series, common to many financial time series, is that their volatility varies over time. Describing the volatility of an asset is a key issue for researchers in financial economics and analysts in financial markets. The prices of assets depend on the expected volatility (covariance structure) of the returns and some derivatives depend solely on the correlation structure of their underlyings. Second, statistical inference about the parameters of the conditional mean is dependent on the correct specification and estimation of the variance (and vice versa). Banks and other financial institutions apply value-at-risk models to assess risks in their marketable assets. Here a precise specification of a model is needed so that risks are neither over nor underestimated.

The most popular class of models for time varying volatility is represented by GARCH type models. Bollerslev et al. [1992] survey several hundred studies of financial markets with applications of various extensions of the GARCH models. The popularity of the GARCH models stem from their ability to parsimoniously capture complex patterns of dependency in the data.

In practical application the estimated GARCH models usually imply a very high level of persistence in the volatility. This lead to the IGARCH model of Bollerslev where the process for the volatility incorporates a unit root. But what if the data actually stems from stationary processes that differ in their parameters? This question stirred research about structural breaks in stochastic processes. It turned out that such changes in the parameters could account for a part of the high persistency and disentangle the persistency that stems from changes in the parameters and the one implied by the estimated GARCH model. Thus estimates on data from different stationary processes lead to an indication of nonstationarity which lead to methods that describe and test for structural breaks.

Maekawa et al. [2005] demonstrate that most of the Tokyo stock return data sets posses volatility persistence and in many cases it is a consequence of structural breaks in the GARCH process. Rapach and Strauss [2005] report significant evidence for structural breaks in the unconditional variance of exchange rate returns. Smith finds strong evidence of structural breaks in GARCH models of US index returns, foreign exchange rates, and individual stock returns. He concludes that standard diagnostic tests are no substitute for structural break tests and that the results suggest that more attention needs to be given to structural breaks when building and estimating GARCH models.

University of California, Santa Barbera, the German Research Foundation (DFG) and the German Academic Exchange Service (DAAD).

Another approach to this problem would be to describe changes in the parameters endogenously with a Markov Switching model. These models were introduced to the econometric mainstream by the seminal article of Hamilton [1989].
The difference is that the process can leave a state (parameter set) and returns with a positive probability. Let us assume, that a process has a ”normal” state and several other states with higher or lower volatilities. A structural break model will base its parameter estimates only on the data between changes in the structure and throws away the rest of the data. In such a scenario a Markov Switching model would retrieve much better estimates for the ”normal” state, because it operates on a much larger data set. In this case the Markov Switching model would yield a superior fit and more important, a better forecasting performance.

Markov Switching models are currently being considered for various markets. One example are the electricity markets, where the prices exhibit extreme jumps. These are due to generator outages, network problems or sudden increases in the demand. All of these represent exogenous events, which would represent the current regime of the price process, and suggest the use of Markov Switching models.
FX markets are subject to changes in the monetary policies of different countries. Changes in these policies that are triggered by external events can cause a change in the level of the variance, thus these economic series are also a candidate for the modelling with Markov Switching models.

In financial markets, spread products are becoming increasingly popular, such as a derivative paying X times the amount of the spread between the ten and two year swaprate: $X \times (s(t, t + 10y) - s(t, t + 2y))$. The prices of these derivatives are extremely sensitive to changes of the correlation between the two underlyings.
Under certain market conditions, that are again triggered by external events, seemingly uncorrelated processes suddenly become more correlated. This has strong implications for the pricing and hedging of derivatives on spreads. Here a Markov Switching framework for the level of the correlation will provide the means to capture these phenomena and conduct inferences about the current level.

While these effects are related to pricing and hedging in the derivatives markets, one naturally finds applications in the risk management of a portfolio. Modern portfolio theory is build on the concept of diversification, therefore a compelling reason for investing in certain funds, is the fact that their returns seem relatively uncorrelated with market indexes such as the S&P 500. In Chan et al. [2005] the authors describe how this diversification argument had to be reviewed for hedge funds by the lessons of the summer of 1998 when the default in Russian government debt triggered a global flight to quality. This changed many of the correlations overnight from 0 to 1.

For the successful application of Markov-Switching models to these problems, it is crucial to have reliable parameter estimators. In econometrics the usual route to derive parameter estimates is to choose the maximum likelihood approach. However it turned out that this approach becomes computationally infeasible for Markov-Switching ARMA-GARCH models and researchers such as Cai and Hamilton have dismissed these models as too untractable. Instead they use low order MS-AR-ARCH models for which they derived estimators. A considerable amount of research in the econometric society is currently being devoted to Markov-Switching models as these are perceived to be very promising by an ever growing amount of researchers and practitioners. On difficult data sets the simpler low order MS-AR-ARCH models do not yield the desired results and the need for advanced diagnostics and more sophisticated models rises.

In this paper we develop an algorithm for the estimation of the parameters of a full MS-ARMA-GARCH model. For this we chose the bayesian framework to formulate the estimator since this enables the application of Markov Chain Monte Carlo methods which are very powerful tools for the numerical computation of integrals. We proceed as follows: in section 2 we present several Markov Switching models and highlight their characteristics. In section 3 we briefly review the theory of bayesian estimation to prepare the ground for an introduction to Markov Chain Monte Carlo Methods in section 4. With those methods at hand we derive our estimation algorithm for the MS-ARMA-GARCH model in section 5. Thereafter we briefly present a diagnostic tool for the convergence of Markov Chain Monte Carlo method in section 6, before we evaluate our algorithm in section 7 on both simulated and empirical data. Section 8 concludes.

## 2 Markov Switching Models

All of the models exhibited in this section vary only slightly in their formulation, but as we will see in later sections, this will have a great impact on their analytical tractability and the derivation of estimators for their parameters. All of them specify a number of latent regimes or states, which control the parameters of the process. These states are themselves random and are assumed to follow a discrete $\mathcal{S}$ dimensional markov chain $\{S_t\}_{t\in\mathbb{N}}$ which is defined on the discrete state space $\{1, 2, \ldots, \mathcal{S}\}$ with the transition probability matrix

$$\Pi = \begin{pmatrix} \pi_{1,1} & \pi_{1,2} & \cdots & \pi_{1,\mathcal{S}} \\ \pi_{2,1} & \ddots & & \\ \vdots & & & \\ \pi_{\mathcal{S},1} & \cdots & \cdots & \pi_{\mathcal{S},\mathcal{S}} \end{pmatrix} \tag{1}$$

where $\pi_{i,j}$ is the probability that the Markov chain goes from state $i$ to $j$. This is common for all of the models considered in this article.

We will begin with the model proposed in Hamilton [1989]

**Model M1 (Hamilton '89)** *The state of the economy $S_t$ follows a two state Markov chain with a transition matrix $\Pi$ as defined in (1). The time series $\{y_t\}$ is modelled as a fourth-order autoregression around one of two constants, $\mu_1$ or $\mu_2$.*

$$(y_t - \mu_{S_t}) = \sum_{i=1}^{4} \phi_i (y_{t-i} - \mu_{S_{t-i}}) + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2)$$

Hamilton fits this to the the US GDP data and identifies recessions and recoveries of the business cycle.
A popular class of time series models for macroeconomic data such as the GDP is represented by ARMA processes. Macroeconomic data is usually modelled with an ARMA process, whereas time series of financial returns in the global markets exhibit strong signs of heteroskedasticity. By far the most popular way to model these returns is through a GARCH process. In order to transfer the idea of regime switching to financial markets, Hamilton extended his model to incorporate ARCH effects.
In Hamilton and Susmel [1994] the authors propose the following model to explain the weekly returns from the New York Stock Excange:

**Model M2 (Hamilton '94)** *The latent state governing the evolution of the model parameters is assumed to follow an $\mathcal{S}$ dimensional markov chain whose* *transition matrix is given through* (1).

$$\begin{aligned}
y_t &= \mu_{S_t} + \sum_{i=1}^{r} \phi_i y_{t-i} + \varepsilon_t \\
\varepsilon_t &= \sqrt{g_{S_t}} \cdot \tilde{u}_t \\
\tilde{u}_t &= \sqrt{h_t} \cdot u_t \quad u_t \sim t(v) \\
h_t &= a_0 + \sum_{i=1}^{q} a_i \tilde{u}_{t-i}^2 + d_{t-1} l_1 \cdot \tilde{u}_{t-1}^2 \\
d_{t-1} &= \mathbf{1}[\tilde{u}_{t-1} \leq 0]
\end{aligned}$$

*where the* $\phi_i$ *are the regression coefficients and* $a_i$, $g_i$, $l_1$ *are positive scalars.*

$g_{S_t}$ is a state dependent amplifier of the conditional variance. Since $h_t$ is defined on the pre-amplified residuals, the conditional variance of $\tilde{u}_t$ is thus modelled as a standard ARCH(q) process with a leverage effect. The idea is thus to model changes in regime of the conditional variance process as changes in the scale of the process. The dependency structure within each regime will remain unchanged since all values are amplified equally.
The specification of the leverage effect through the dummy regression variable $d_{t-1}$ is taken form Glosten et al. [1989]. This effect is often observed in financial data, where markets react more volatile to negative shocks.
The conditional mean is modelled as an AR(r) process with regime switching means $\mu_{S_t}$.
Low order GARCH specifications of the conditional variance offer a much more parsimonious representation than higher order ARCH models while they are able to capture an equally complex autocorrelation structure. That is the reason why they are usually preferred by practitioners. It therefore seems like a step back, to have only the ARCH component in a Markov switching model. But the GARCH component poses significant drawbacks in the estimation process when the maximum likelihood route is chosen. Hamilton dismissed MS-GARCH models as untractable and computationally to intensive, therefore he chose to model the conditional variance with higher order ARCH processes.

Through a different specification Gray [1996] is able to compute ML estimates of a Model which allows for ARMA and GARCH effects. But we can only speak of effects, since the model differs considerably from a classical ARMA-GARCH model, and properties such as stationarity cannot be transferred directly to the new specification. Furthermore he suggests to asses the goodness of fit through the Ljung-Box Q Test, applied to the estimated residuals. However, the residuals estimated trough his method are not i.i.d. and therefore the LBQ Test is likely to be spurious. Haas et al. [2004] also criticize this model specification since the conditional variance does not only depend on the past i.i.d. innovations but also on shocks caused by a change in regime.

In Haas et al. [2004] the authors propose a different formulation of the conditional variance process. This specifications allows the authors to derive analytical stationarity conditions and straightforward parameter estimators.

**Model M3 (Haas '04)** *Let the univariate time series $\varepsilon_t$ be given by*

$$\varepsilon_t = \sqrt{h_t(S_t)} \cdot u_t \qquad u_t \sim N(0,1)$$

*where the conditional variance $h_t$ is an $\mathcal{S} \times 1$ dimensional vector process given by*

$$h_t = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta h_{t-1}$$

*with*

$$\alpha_i = \begin{pmatrix} \alpha_{i,1} \\ \alpha_{i,2} \\ \vdots \\ \alpha_{i,\mathcal{S}} \end{pmatrix} \qquad \beta = diag(\beta_1, \beta_2, \ldots, \beta_{\mathcal{S}})$$

*where $h_t(S_t)$ denotes the element of $h_t$ at position $S_t$.*

The conditional variances of the regimes run in parallel and affect each other only through the realized values of the innovations. In contrast to model M2 it does not impose the same dependency structure for all regimes. It rather allows a clear-cut interpretation of the variance dynamics in each regime. The authors argue that it is the primary feature of a GARCH model, that shocks drive the volatility and that it can parsimoniously represent high-order ARCH models. A shift in the regime will therefore affect the conditional variance process only through the realized shock $\varepsilon_t$, whose variance is different. Therefore they speak of this as the "natural" generalization of the ARCH approach to a regime switching setting. Furthermore they also present closed form volatility forecasts.
The drawback of this model is, that it is only analytically tractable, as long as no process is specified for the conditional mean. For exchange rate dynamics, this is an appealing model, since the logarithmic percentage returns of the major exchange rates show no significant autocorrelations in the mean. This is different for returns on stocks or interest rates.

In Francq and Zakoian [2001b],Francq and Zakoian [2001a] and Francq and Zakoian [2002] the autors discuss the stationarity properties of markov switching processes, existence of moments and give estimators of the parameters based on the GMM technique. However, moment estimators do not give smoothed estimates of the states from the latent process. But it is one major aim of the Markov switching model, to provide both a model that captures characteristics of a time series and a "story". Their specification of a MS-ARMA-GARCH model is the same as that of M4, which is the straightforward extension of Hamiltons original regime switching model. This is the model specification for which we will derive our Bayesian MCMC estimator in the later sections.

**Model M4** *We assume that the state of the economy, $S_t$ follows a discrete $\mathcal{S}$ dimensional markov chain with transition probability matrix given by* (1)*. We will now consider an ARMA-GARCH model who's parameters are dependent on the state of this markov chain.*

$$y_t = c_{S_t} + \sum_{i=1}^{r} \phi_i(S_t) \cdot y_{t-i} + \varepsilon_t + \sum_{j=1}^{m} \psi_j(S_t) \cdot \varepsilon_{t-j} \tag{2a}$$

$$h_t = \omega_{S_t} + \sum_{i=1}^{q} \alpha_i(S_t) \cdot \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j(S_t) \cdot h_{t-j} \tag{2b}$$

$$\varepsilon_t = \sqrt{h_t} \cdot u_t \qquad u_t \sim N(0,1)$$

*We will then succeedingly extend the model. In a first step we will let the innovations $\varepsilon_t$ follow a student-t distribution and in a second step we include a leverage effect in the GARCH component. In the full model the conditional variance then becomes*

$$h_t = \omega_{S_t} + \sum_{i=1}^{q} \alpha_i(S_t) \cdot \varepsilon_{t-i}^2 + \sum_{k=1}^{q} d_k \cdot l_k(S_t)\varepsilon_{t-k}^2 + \sum_{j=1}^{p} \beta_j(S_t) \cdot h_{t-j}$$

*where $d_k$ is $\mathbf{1}[\varepsilon_{t-k} \leq 0]$.*

As we can see, there are many ways in which one can formulate a Markov-Switching Model, even within these categories one can choose between different specifications. We talk about different specifications when we distinguish two models from the same category, but with a different number of parameters. An ARMA(1,1) and an ARMA(2,1) are two different specifications of an ARMA(r,m) model.

## 3 Bayesian Estimation

Since bayesian estimation is not the mainstream inference technique, we give a short introduction to this topic.

In the parametric frequentist paradigm data stems from a statistical model with fixed parameters. These parameters are totally inaccessible. The bayesian paradigm is fundamentally different in this respect. The parameters of the model are also assumed to follow a distributional assumption, which is assumed by the statistician. This distribution which is referred to as the prior distribution summarizes the prior beliefs of the researcher. The only thing that is considered real is the observed data. The data itself is also assumed to follow a statistical law. The assumption about the type of distribution is also an expression of a belief. Inference about the distribution that generated the data is therefore the transition of a prior belief into a posterior belief. The bayes theorem formalizes this process of updating a prior belief through the observed data set into a posterior belief. Robert [1994] formulates the bayesian paradigm as a duality principle: compared with probabilistic modeling, the purpose of a statistical analysis is fundamentally an inversion purpose, since it aims at retrieving the causes (parameters of the probabilistic generating mechanism) from the effects (observations). A general description of the inversion of probabilities is given by Bayes theorem: if $A$ and $E$ are events such that $P(E) \neq 0$, $P(A|E)$ and $P(E|A)$ are related by

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)}$$

Bayes and Laplace considered that the uncertainty on the parameters $\theta$ could be modeled through a probability distribution $p(\theta)$ on the parameter space $\Theta$. The inversion is then described through an application of Bayes theorem to the full model, as a posterior distribution is computed conditional on the observed data $x$

$$p(\theta|x) = \frac{f(x|\theta)p(\theta)}{\int f(x|\theta)p(\theta)d\theta} \tag{3}$$

where $f$ is the likelihood of the assumed probabilistic model that generates the data. Since the integral in the denominator is only a constant and the nominator is enough to determine the distributional family of the posterior, it is common in the bayesian literature to use the notion of proportionality and write (3) as

$$p(\theta|x) \propto p(\theta)f(x|\theta).$$

A bayesian statistical model is therefore defined as follows ( Robert [1994]):

**Definition 3.1** *A Bayesian statistical model is made of a parametric statistical model, $f(x|\theta)$, and a prior distribution on the parameters, $p(\theta)$.*

The parameters of the prior distribution are usually called *hyperparameters*. The bayesian paradigm in statistical modelling is based on the same measure theoretic results as frequentist models, but it differs conceptually in the issue of inference.

### *3.1 Conjugate Priors*

When a statistical model is to be employed, one has to make a choice about the data generating distribution. In a bayesian setting one also has to choose a prior distribution. This prior distribution should hold the prior information about the parameters. For a given parameter it mainly reflects the beliefs where the parameter is expected and how sure one is about that. The rough position of the parameter is expressed through the mean, whereas the variance expresses the prior uncertainty. These are only the beliefs about the two first moments of a possible prior distribution, but one has to choose an entire distribution. For this choice of the class or family of the prior, several approaches have emerged in the bayesian literature. Seemingly the most prominent of these is the principle of *conjugate priors*.

Let $f$ be a likelihood function $f(\theta|x)$. A class $\mathsf{C}$ of prior distributions is said to form a *conjugate family* if the posterior density

$$p(\theta|x) \propto p(\theta)f(\theta|x)$$

is in the class $\mathsf{C}$ for all $x$ whenever the prior density is in $\mathsf{C}$.

To give an example we give the conjugate priors for the mean and the variance of a normal distribution: if

$$x \sim N(\theta, \sigma^2) \qquad \theta \sim N(\theta_0, \sigma_0^2)$$

then the likelihoods are

$$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma_0} exp\left\{-\frac{(\theta-\theta_0)^2}{\sigma_0^2}\right\} \tag{4}$$

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{(x-\theta)^2}{\sigma^2}\right\} \tag{5}$$

hence

$$p(\theta|x) \propto exp\left\{-\frac{\theta^2}{2}\left(\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2}\right)+\theta\left(\frac{\theta_0}{\sigma_0^2}+\frac{x}{\sigma^2}\right)\right\}$$

with

$$\sigma_1^2 = \frac{1}{\sigma_0^{-2}+\sigma^{-2}} \qquad \theta_1 = \sigma_1^2\left(\frac{\theta_0}{\sigma_0^2}+\frac{x}{\sigma^2}\right) \tag{6}$$

we can write

$$p(\theta|x) \propto exp\left\{-\frac{(\theta-\theta_1)^2}{2\sigma_1^2}\right\}$$

and since a density must integrate to one, it follows that

$$\theta|x \sim N(\theta_1, \sigma_1^2) \tag{7}$$

The use of a prior density that conjugates the likelihood allows for analytic expressions of the posterior density. Table 1 gives the conjugate priors for several common likelihood funcions.

| Likelihood | Conjugate prior |
|---|---|
| Binomial | Beta |
| Multinomial | Dirichlet |
| Poisson | Gamma |
| Normal (unknown mean) | Normal |
| Normal (unknown variance) | Inverse Chi-Square |

Table 1
Conjugate priors for common likelihood functions

### *3.2 The bayes estimator*

In this section we summarize the results on the optimal bayes estimators under quadratic loss.

The expression "bayes estimator" is used in the literature as a pseudonym for the optimal bayesian estimator under quadratic loss. In order to evaluate or find an estimator, an optimality property is required. One such optimality property is the so called average risk optimality. In this setup the average risk for some suitable non-negative weight function is minimized. In Lehman and Casella [1998] this is formalized as follows

$$\min \quad r(p,\delta) = \int R(\theta,\delta)dp(\theta) \tag{8}$$

where $\delta$ is the estimator, $R(.)$ is the risk function and $p$ is a probability distribution on $\Theta$ giving the weights. An estimator $\delta$ minimizing (8) is called a bayes estimator with respect to $p$: $\delta_p$(x)
When the risk $R(.)$ is given through the loss function for an estimator $d$

$$L(\theta,d) = [d - g(\theta)]^2$$

then the optimal estimator with respect to $p$ is

$$\delta_p(x) = E[g(\Theta)|x]$$

where $g(\theta)$ is usually taken to be $\theta$. The optimal bayes estimator under quadratic loss is therefore simply the *posterior mean*.

### *3.2.1 Analytical Bayes Estimators for selected Models*

In 3.1 we computed the posterior distribution of a bayesian model of a normal random variable with a conjugate prior. The posterior distribution was again normal and the posterior moments are given through (6). So that for a normal random variable with known variance, the bayes estimator for the mean is $\theta_1$ from (6). For a sample $x = (x_1, x_2, \dots, x_n)$, where $X_i \sim N(\theta, \sigma^2)$ with known variance and unknown mean, the posterior moments are found analogue to (6)

$$\sigma_1^2 = (\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2})^{-1} \tag{9}$$

$$\theta_1 = \sigma_1^2 \left( \frac{\theta_0}{\sigma_0^2} + \frac{n \cdot \bar{x}}{\sigma^2} \right) \tag{10}$$

*Normal variance*

Let $x = (x_1, x_2, \dots, x_n)$ be a sample from a normal distribution with known mean $\mu$ but unknown variance. Then the likelihood is

$$\begin{aligned} f(x|\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left\{ -\frac{(x_i-\mu)^2}{2\sigma^2} \right\} \\ &\propto \sigma^{-n} exp\left\{ -\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \right\} \end{aligned}$$

We write

$$S = \sum_{i=1}^{n} (x_i - \mu)^2$$

so the posterior is

$$p(\sigma^2|x) \propto \sigma^{2\cdot(-n/2)} exp\left\{ -\frac{S}{2\sigma^2} \right\}$$

A conjugate prior is then given by

$$p(\sigma^2) \propto \sigma^{2\cdot(-k/2)} exp\left\{ -\frac{S_0}{2\sigma^2} \right\}$$

If we substitute $k = v + 2$ we can write this as

$$p(\sigma^2|x) \propto (\sigma^2)^{-(v+n)/2-1} exp\left\{-\frac{S+S_0}{2\sigma^2}\right\}$$

which is proportional to an inverse chi-squared distribution ($\chi^{-2}$). For the posterior distribution of $\sigma^2$ one finds

$$\sigma^2 \sim (S_0 + S)\chi^{-2}_{v+n} \tag{11}$$

which is a *scaled inverse chi-square distribution* [see also Lee, 2003]. We will need these results again in section B.1.

### *3.2.2 Asymptotics of the bayes estimator*

This section states the main asymptotic results of bayesian estimation under quadratic loss. We will see that for large n, the estimator converges to the ML estimator.

In a bayesian model one imposes a distribution on the parameter space. This seems to contrast drastically with the frequentist approach, where the world of the parameters is completely determined a priori. However, the data is also assumed to be generated by a distribution whose parameters are fixed just as in the frequentist setting. The difference is only that the parameters in the frequentist world are inaccessible, whereas in the bayesian setting, a belief about the parameters is expressed through a prior distribution. But at the bottom line both paradigms try to estimate the true data generating parameters $\theta_0$.

ML estimators are asymptotically efficient and normally distributed. The same holds for the bayesian estimator under quadratic loss, independent of the prior distribution. The following theorem and its proof can be found in [Lehman and Casella, 1998, Theorem 8.3]:

**Theorem 3.1** *If the regularity conditions* (A1)-(A5) *hold, and if* $\tilde{\theta}_n$ *is the Bayes estimator when the prior density is p and the loss is squared error, then*

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} N(0, 1/I(\theta_0)),$$

*so that* $\tilde{\theta}_n$ *is consistent and asymptotically efficient.*

In order to compute the bayes estimator for the parameters of model M4, we need to specify the full bayesian statistical model. The model specification and the distributional assumption of the innovations yield an expression for the likelihood conditional on the observed data. It remains to specify the prior distributions for the individual parameters. Here we will choose conjugate priors wherever possible and normal priors with adequate hyperparameters in the other cases. Since we will work on large data sets, we do not consider the choice of the prior distribution as a critical issue and rely on the asymptotic results stated in section 3.2.2.

The parameter space $\Theta_{ARMA}$ of the ARMA component given in (2a) is the cartesian product $\{C \times \Phi \times \Psi\}$, with $C, \Phi$ and $\Psi$ given by $\mathbb{R}^{\mathcal{S}}$, $\mathbb{R}^{r \cdot \mathcal{S}}$ and $\mathbb{R}^{m \cdot \mathcal{S}}$ respectively. The parameter space of the transition probabilites $\mathbf{\Pi}$ is the unit hypercube in $\mathbb{R}_{+}^{\mathcal{S}-1}$. The complete parameter space is therefore given by:

$$\Theta = \{\mathbf{\Pi} \times \Theta_{ARMA} \times \Theta_{GARCH}\}$$

Note that this does include all nonstationary specifications of the parameters. We will impose stationarity through the prior distributions, which we will restrict to a subset on $\Theta$.

$$\begin{aligned}
\Theta_{ARMA_1} &\sim N(\mu_{ARMA_1}, \Sigma_{ARMA_1}) \cdot \mathbf{1}_S(\theta_{ARMA_1}) \\
\Theta_{GARCH_1} &\sim N(\mu_{GARCH_1}, \Sigma_{GARCH_!}) \cdot \mathbf{1}_S(\theta_{GARCH_!}) \\
&\vdots \\
\Theta_{ARMA_{\mathcal{S}}} &\sim N(\mu_{ARMA_{\mathcal{S}}}, \Sigma_{ARMA_{\mathcal{S}}}) \cdot \mathbf{1}_S(\theta_{ARMA_{\mathcal{S}}}) \\
\Theta_{GARCH_{\mathcal{S}}} &\sim N(\mu_{GARCH_{\mathcal{S}}}, \Sigma_{GARCH_{\mathcal{S}}}) \cdot \mathbf{1}_S(\theta_{GARCH_{\mathcal{S}}}) \\
\mathbf{\Pi} &\sim Dirichlet(\alpha_1, \ldots, \alpha_k)
\end{aligned}$$

where the indicator functions $\mathbf{1}_S(\theta) = 1$ for a parameter set which is stationary, 0 otherwise.
The complete prior distribution for $\theta$ is

$$[\theta] = Dirichlet(\alpha_1, \ldots, \alpha_k) \times \prod_{i=1}^{\mathcal{S}} N(\mu_{ARMA_i}, \Sigma_{ARMA_i}) \times N(\mu_{GARCH_i}, \Sigma_{GARCH_i}) \tag{12}$$

To compute the likelihood of the model, given a certain data set, one would have to integrate over all possible paths of the latent states. This is of course not feasible and can be circumvented in a bayesian context. This is called data augmentation; the parameter vector $\theta$ is augmented with the states $S_{[1,T]}$. We can then compute the posterior of all unobservable quantities and do not only get estimates for the model parameters $\hat{\theta}$, but also for the states $\hat{S}_{[1,T]}$.

$$\begin{aligned} p(\theta, S_{[1,T]}|y) &\propto f(y|\theta, S_{[1,T]})p(\theta, S_{[1,T]}) \\ &\propto f(y|\theta, S_{[1,T]})p(S_{[1,T]}|\theta)p(\theta) \end{aligned} \tag{13}$$

The full statistical model is then given through equations (12) and (13). We can see that computing the posterior mean is a difficult task.
In fact

*our major problem to be solved is to compute or approximate the posterior mean for the full statistical model.*

$$\{\hat{\theta}, \hat{S}_{[1,T]}\} = E[\theta, S_{[1,T]}|Y = y] \tag{14}$$

The posterior density is of very high dimension and only known up to a constant, since we cannot compute it analytically. To tackle the problem of high dimensionality and that of the unknown constants, our method of choice will be a Markov Chain Monte Carlo method, which are introduced in section 4.

# 4 MCMC methods in bayesian statistics

In this section we introduce the Markov chain Monte Carlo methods which are employed in bayesian estimation. This numerical procedure is definitely one reason of the increasing popularity of bayesian inference.
Markov Chain Monte Carlo methods can be employed to approximate the integral

$$\mathcal{I} = \int h(x) f(x) dx$$

with some distribution function $f$ and some function $h$. It might not be necessary to use an MCMC algorithm to compute this integral, since other ordinary Monte Carlo methods can also achieve this. But in some cases MCMC methods will yield superior results. While regular Monte Carlo methods and MCMC algorithms both satisfy the $\mathcal{O}(\frac{1}{\sqrt{n}})$ convergence requirement, there are many instances where a specific MCMC algorithm dominates the corresponding Monte Carlo approach in terms of its variance. Robert and Casella [1999] give several examples on this and compare the speed of convergence of different methods. More important for our task at hand is the fact that MCMC methods can efficiently sample from a distribution $f$, from which it is hard to obtain a sample in an ordinary way. As discussed in the previous section, bayesian estimators are based on the posterior distribution of the model parameters and it can be hard, if not impossible to determine this distribution analytically. Monte Carlo methods are a tool for numerical integration and are therefore a natural candidate for the construction of the posterior distributions which are hard to integrate analytically. In the particular case of bayesian posterior distributions MCMC methods are superior to normal Monte Carlo methods due to reasons that we will now explore in more detail.

The bayesian paradigm facilitates a straightforward way to compute the posterior density up to a constant. One hardly ever bothers to compute the full posterior density analytically, instead it is usually expressed as

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

MCMC methods can sample from the posterior distribution even if the posterior density is only known up to this constant. It should be pointed out that also other methods like *importance sampling* can be used, but we already mentioned above that MCMC methods can have superior characteristics variance wise. This is the case in our problem, where we have to sample from very complex high dimensional posterior distributions.
To render it more precise how these MCMC methods work, we will summarize the relevant results from Markov chain theory. We start with a definition from Robert and Casella [1999]:

**Definition 4.1** *A Markov Chain Monte Carlo (MCMC) method for the sim-*

*ulation of a distribution $f$ is any method producing an ergodic Markov chain $\{X^{(g)}\}_{g\in\mathbb{N}}$ whose stationary distribution is $f$.*

The ergodic theorem, often called central limit theorem for Markov chains ensures that a sequence $\{X^{(g)}\}$ produced by an MCMC algorithm can be employed just like an iid sample from the stationary distribution $f$. It guarantees that the empirical average

$$\hat{\mathcal{I}}_N = \frac{1}{N}\sum_{g=1}^{N} h(X^{(g)})$$

converges almost surely to $\mathbb{E}_f[h(X)]$. In our context the function of interest $h$ would simply be $\theta$, then we compute our bayesian estimate of the model parameters as the empirical average over the simulated Markov chain.
Our main objective is therefore:

*to construct an ergodic Markov chain on the parameter space $\Theta$, whose stationary distribution is the posterior distribution of our model.*

In the subsection 4.1 to 4.3 we begin with an informal presentation of some relevant material from Markov chain theory and then introduce two ways of constructing Markov chains that satisfy the properties in definition 4.1. We will use these methods in conjunction in section 5 to estimate the parameters of our model.

### *4.1 Markov Chains*

The results presented in this section can be found in more detail in Robert and Casella [1999]

A Markov chain is a sequence of random variables $X_0, X_1, \ldots, X_n$ on the state space $\Omega \subseteq \mathbb{R}^p$, if for any $t$, the conditional distribution of $X_t$ given $x_{t-1}, \ldots, x_0$ is the same as the distribution of $X_t$ given $x_{t-1}$; that is:

$$\begin{aligned} P(X_{t+1} \in A | x_0, \ldots, x_k) &= P(X_{t+1} \in A | x_t) \quad A \subseteq \Omega \\ &= \int_A P(x_t, dy) \end{aligned} \tag{15}$$

$P(x, A)$ is referred to as the transition kernel of the Markov chain. The transition kernel is the distribution of $X_{t+1}$ given $x_t$ and in the continuous case the kernel denotes the conditional density $P(x, dy)$ of the transition from $x$ to $dy$. The right hand side of (15) is the typical notation, but could also be expressed as:

$$P(X_{t+1} \in A|x_t) = \int_A p(x_t, y)dy$$

and we refer to $p(x_t, y)$ as the transition density.

The distribution for $X_n$ given $x_t$ is the $m$-step ahead distribution ($n > t$, $m = n - t$). It is obtained from the $m$-step ahead kernel given by

$$P^{(m)}(x_t, A) = \int_\Omega P(x_t, dy)P^{(m-1)}(y, A)$$

**Invariance** It can be shown that, under certain conditions, the $m^{th}$ iterate of the transition kernel converges to the *invariant* distribution $\pi^*$ as $m$ goes to infinity ( also called stationary distribution). It is given by

$$\pi^*(dy) = \int_\Omega P(x, dy)\pi^*(dx) \tag{16}$$

This condition states that if $X_i$ is distributed according to $\pi^*$, then so are all subsequent elements of the chain.

**Reversibility** A Markov chain with transition kernel $P(y, x)$ is said to satisfy the *detailed balance condition* if there exists a function $\pi$ satisfying

$$P(y, x)\pi(y) = P(x, y)\pi(x) \tag{17}$$

for everery $(x, y)$. Such a chain is also said to be *reversible* and has $\pi^*$ as an invariant distribution ($\pi^*(dy) = \pi(y)dy$) [see Robert and Casella, 1999, Theorem 6.2.2].

**Irreducibility** Another important notion is that of *irreducibility*. A Markov chain is said to be $\pi^*$-irreducible if for every $x \in \Omega$, $\pi^*(A) > 0 \Rightarrow P(X_i \in A|x_0) > 0$ for some $i \geq 1$. This condition states that all sets with positive probability under $\pi^*$ can be reached from any starting point in $\Omega$. It is a first measure of the sensitivity of the markov chain to the initial conditions. This is crucial in the setup of MCMC algorithms, because it leads to a guarantee of convergence.

**Aperiodicity** The next important property is *aperiodicity*, which ensures that the chain visits all regions and not only a finite number of sets. A Markov chain is aperiodic if there exists no partition of $\Omega = (D_0, D_1 \ldots, D_{p-1})$ for some $p > 2$ such that $P(X_i \in D_{imod(p)}|X_0 \in D_0) = 1$ for all $i$.

With these definitions we can state the following important result for MCMC methods [see Tierny, 1994]:

**Proposition 4.1** *If $P(x, y)$ is $\pi^*$-irreducible and has an invariant distribution $\pi^*$, then $\pi^*$ is the unique invariant distribution of $P(x, y)$. If $P(x, y)$ is also aperiodic, then for $\pi^*$-almost every $x \in \Omega$, and all sets $A$*

*(1)* $|P^m(x, A) - \pi^*(A)| \overset{m\to\infty}{\longrightarrow} 0$

*(2) for all $\pi^*$-integrable real valued functions $h$,*

$$\frac{1}{m}\sum_{i=1}^{m} h(X_i) \to \int h(x)\pi^*(x) \quad \textit{as } m \to \infty, a.s$$

The first part of this proposition states that the probability density of the $m^{th}$ iterate of the Markov chain is very close to its unique, invariant density (for large m). This means that for drawings made from $P^m(x, dy)$, the probability distribution of the drawings is the invariant distribution. The second part states that the ergodic averages converge to their expected value under the invariant density.

The first algorithm that we introduce, was proposed by Metropolis et al. [1953] and later generalized by Hastings [1970] and is now known as the Metropolis-Hastings algorithm. This is in fact the most general MCMC algorithm which offers (almost) universal applicability and can be used as a black-box method to obtain bayesian parameter estimates.

### *4.2 The Metropolis Hastings Algorithm*

Many articles and books have been published that address the Metropolis Hastings algorithm. We would like to point out two particular sources. A very detailed discussion of the algorithm can be found in the book by Robert and Casella [1999], where the algorithm is embedded in the larger context of MCMC methods and an article by Chib and Greenberg [1995], which provides an excellent intuitive exposition of the algorithm.

The MH algorithm constructs a Markov chain that fulfills the requirements of proposition 4.1 and is easily implemented. The objective of the algorithm is to generate samples from an absolutely continuous *target density* $f(x) = \frac{p(x)}{C}$, where $p(x)$ is an unnormalized density on $\mathbb{R}^n$ with the possibly unknown normalizing constant $C$. This directly shows its possible importance for bayesian inference, where posterior distribution of the parameters are easily computed up to a constant.

The MH algorithm can be viewed as an improvement on the Acceptance-Rejection sampling, where a complicated target distribution is sampled via an *instrumental* or *proposal* distribution that can be sampled from by some known method. The MH algorithm is then defined as:

**Algorithm A1 (Metropolis-Hastings-Algorithm)**
*Let $f(y)$ be the target density and $g(y|x)$ the proposal density, then given an arbitrary starting value $x^{(0)}$ repeat for $g = 1, 2, \ldots, N$:*

*(1) Draw a sample $Y_g \sim g(y|x^{(g)})$,*

*(2) Set*

$$X^{(g+1)} = \begin{cases} Y_g \ \textit{with probability} \quad \alpha(x^{(g)}, Y_g) \\ x^{(g)} \ \textit{with probability} \ 1 - \alpha(x^{(g)}, Y_g) \end{cases} \tag{18}$$

*(3) whith*

$$\alpha(x, y) = \begin{cases} min\left(\frac{f(y)}{f(x)} \cdot \frac{g(x|y)}{g(y|x)}, 1\right) \ \textit{if} \ f(x)q(x, y) > 0 \\ \qquad 1 \qquad\qquad\qquad \textit{otherwise} \end{cases} \tag{19}$$

Note that this algorithm depends only on the ratios

$$\frac{f(y)}{f(x)} \qquad \frac{g(x|y)}{g(y|x)}$$

and is therefore independent of any normalizing constants in the definition of $f$ and $g$. We note that the proposal density does not have to be a conditional density on $x$.

In particular, the algorithm A1 defines a Markov chain that is governed by the transition kernel

$$P(x, dy) = g(y|x)\alpha(x, y)dy + r(x)\delta_x(dy) \tag{20}$$

where $p_{MH}(x, y) = g(y|x)\alpha(x, y)$ may be referred to as the transition density with the properties that

$$\textstyle\int g(y|x)dy = 1 \ , \quad \delta_x(dy) = \begin{cases} 1 \qquad \text{if} \qquad x \in dy \\ 0 \ \text{otherwise} \end{cases}$$

$$r(x) = 1 - \textstyle\int_\Omega g(y|x)\alpha(x, y)dy.$$

That is, transitions from $x$ to $y$ occur according to $p(x, y)$ and the probability that $x$ remains unchanged occurs with probability $r(x)$. It is straightforward to verify the two equalities

$$\begin{aligned} g(y|x)\alpha(x,y)f(x) &= g(x|y)\alpha(y,x)f(y) \\ r(x)\delta_x(y)f(x) &= r(y)\delta_y(x)f(y) \end{aligned}$$

which together establish the detailed balance for the Metropolis-Hastings chain. The stationarity of f is therefore established for almost any conditional distribution $g$. In fact $\alpha(x, y)$ is constructed so that the algorithm fulfills this property [see Chib and Greenberg, 1995].

A condition for algorithm A1 to satisfy the requirements of Proposition 4.1 can be based on a result by Roberts and Tweedie [1996]

**Proposition 4.2** *Assume $f$ is bounded and positive on every compact set of its support $\mathcal{E}$. If there exist positive numbers $\varepsilon$ and $\delta$ such that*

$$g(y|x) > \varepsilon \quad \textit{if} \quad |x - y| < \delta$$

*then the Metropolis-Hastings Markov chain $(X^{(g)})$ is f-irreducible and aperiodic, and the conditions of Propostion 4.1 are fulfilled.*

This result basically requires that the support of the proposal density encompasses that of the target density. If moreover both distributions are continuous, the central limit theorem for markov chains applies.

### *4.2.1 Block-at-a-Time Algorithm*

In the previous section we could see how the MH algorithms overcomes the problem of finding the unknown normalizing constant in bayesian models. The "Block at a Time" algorithm, as discussed in [Hastings, 1970, sec.2.4] often simplifies the search for an adequate proposal distribution and will help us to treat the curse of high dimensionality in our model.
We will illustrate this idea with a case where the random vector $x$ can be split into two blocks $x = (x_1, x_2)$ with $x_i \in \mathbb{R}^{d_i}$ [see Chib and Greenberg, 1995]:

Let the transition kernel $P_1(x_1, dy_1|x_2)$ have an invariant distribution with density $f_{1|2}$:

$$f_{1|2}(dy_1|x_2) = \int P_1(x_1, dy_1|x_2) f_{1|2}(x_1|x_2) dx_1 \tag{21}$$

and let the $2^{nd}$ transition kernel $P_2(x_2, dy_2|x_1)$ have $f_{2|1}$ as its invariant distribution analogous to (21). The kernel $P_1$ could for example be generated by a Metropolis-Hastings chain applied to the block $x_1$, with $x_2$ fixed for all iterations. It turns out that the product of the transition kernels has $f^*(x_1, x_2)$ as its invariant density. So that instead of having to run each kernel to convergence for every value of the conditioning variable, we can simply draw the individual variables in succesion. This is called the *product of kernels* principle.

$$f^*(dy_1, dy_2) = \int \int P_1(x_1, dy_1|x_2) P_2(x_2, dy_2|y_1) f(x_1, x_2) dx_1 dx_2 \tag{22}$$

The product of kernels in (22) arises, when the "scan" through the elements of $x$ is systematic:

(1) $y_1$ is produced by $P_1(.|x_2)$ conditional on the realization $x_2$ from the last draw.

(2) $y_2$ is then produced by $P_2(.|y_1)$ conditional on the realization of $y_2$ from the current iteration.

It should be noted that several other possibilities for the "scanning" method can be used, for example a random scan, as we will discuss in the section on the Gibbs sampler.

The results on the product of kernels principle are directly related to the Gibbs sampler, which turns out to be a special case of the Metropolis-Hastings algorithm. It also gives rise to interesting hybrid methods, which are, in their essence, based on the product of kernels.

In the light of these results, the MH algorithm is fascinating in its universality. It provides a sample from an arbitrary distribution $f$ with support $\mathcal{E}$ given another arbitrary distribution $g$ on the same support. However, this universality may only be a formality. Though we have stated the convergence theorems, we have not said anything about the speed of convergence. If the proposal distribution $g$ only rarely simulates points in the domain of the support of $f$ where the most mass is located, the convergence rate will be extremely poor. This leads to the problem of a good choice for $g$.

### *4.2.2 Choice of the proposal distribution*

Typically, the proposal density is selected from a family of distributions that require the specification of parameters such as the location and scale. Intuitively one would like to choose $g(x, y)$ so that the location varies along the chain. This would reduce the possibility of undersampling certain regions. There are several possible strategies for this:

(1) The *random walk* chain [see Robert and Casella, 1999],
(2) the *autoregressive* chain proposed by Tierny [1994],
(3) proposal distributions that exploit the knowledge about $f$ [see Chib and Greenberg, 1994],

(4) and there also exist some fully automated algorithms such as the ARMS (*Adaptive Rejection Metropolis Sampling*) [see Robert and Casella, 1999],

just to mention a few.

For an such a complex estimation problem as ours, black box methods are likely to be very inefficient. We will therefore choose a strategy from category 3. in the later sections to solve our estimation problem. We will see that our strategy implies certain location and scale parameters of the proposal distribution, but we will use the freedom we have to fine tune these. Especially the question of tuning the spread, or scale is critical. This has strong implications for the efficiency of the algorithm. The scale of the proposal density affects the behavior of the chain in two ways: one is the *acceptance rate*, and the other is the region of the sample space that is traversed by the chain.

In the next section we introduce the Gibbs Sampler which is another algorithm for the construction of a Markov Chain. It is closely related to the MH algorithm but in some ways a more intuitive approach.

## *4.3 Gibbs sampling*

The Gibbs sampling algorithm dates back at least to Suomela [1976] in a Ph.D. thesis on Markov random fields. It was formulated independently by Creutz [1979] in statistical Physics, Ripley [1979] in spatial statistics and by Grenander [1983] and Geman and Geman [1984]. The name Gibbs Sampler is due to the simulation of the *Gibbs* distribution in statistical physics, which corresponds to Markov random fields in spatial statistics. The equivalence was established through the Clifford-Hammersley theorem Besag [2001].

For a good introduction to the algorithm we refer to an article by Casella and George [1992]. The authors give an intuitive and a simple explanation of how and why the Gibbs samlper works. For a more thorough discussion we refer to the works of Robert [1994], where the Gibbs sampler is introduced in the context of bayesian estimation techniques and to Robert and Casella [1999], where a detailed discussion of the algorithm is given in the context of MCMC methods.

As with the Metropolis-Hastings algorithm we will only state the algorithm and then recall the main theorems and regularity conditions which ensure the convergence of the resulting Markov chain to its stationary distribution. Then the connection to the MH algorithm in A1 is shown and different variants of the algorithm are discussed. By the end of this section we will have summarized the relevant MCMC techniques and will use the results of the article by Chib and Greenberg [1994] as an example, to see how these methods are used to compute the bayes estimates for the parameters of classical ARMA models.

Let the random vector $X \in \mathcal{X}$, $X' = (X_1, \ldots, X_k)$ have the joint density $f(x_1, \ldots, x_k)$. Where the individual $x_i$ are either uni- or multidimensional. Suppose that we can simulate from the corresponding conditional densities $f_1, \ldots, f_k$. The associated Gibbs Sampler is given by this algorithm:

**Algorithm A2 (The Gibbs Sampling Algorithm)**
*Given $x^{(g)} = (x_1^{(g)}, \ldots, x_k^{(g)})$, generate*

```
(1) $X_1^{(g+1)} \sim f_1(x_1|x_2^{(g)}, \ldots, x_k^{(g)})$
(2) $X_2^{(g+1)} \sim f_2(x_2|x_1^{(g+1)}, x_3^{(g)}, \ldots, x_k^{(g)})$
    $\vdots$
(k) $X_k^{(g+1)} \sim f_k(x_1|x_1^{(g+1)}, \ldots, x_{k-1}^{(g+1)})$
```

These steps generate a markov chain $\{X^{(g)}\}_{g\in\mathbb{N}}$ which will converge to the joint posterior distribution $f$ on $\mathcal{X}$. The Clifford-Hammersley theorem establishes the result, that the full conditional distributions fully characterize the joint distribution. The theorem holds if the following *positivity condition* (taken from  Robert and Casella [1999]) is satisfied:

**Definition 4.2** *Let $(X_1, X_2, \ldots, X_k) \sim f(x_1, \ldots, x_k)$, where $f^{(i)}$ denotes the marginal distribution of $X_i$. If $f^{(i)}(x_i) > 0$ for every $i = 1, \ldots, k$, implies that $f(x_1, \ldots, x_k) > 0$ then $f$ satisfies the positivity condition.*

Note that only the full conditional densities are used for the simulation. If these are only univariate, a high dimensional problem is reduced to sampling from one-dimensional distributions, which can greatly reduce the complexity.

The algorithm A2 constructs a Markov chain where the transition from $x = x^{(g)}$ to $y = x^{(g+1)}$ takes place according to the transition density

$$p_G(x, y) = \prod_{i=1}^{k} f(y_i | y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_k) \tag{23}$$

It can then be shown that this transition density has the joint density $f$ as its invariant distribution.

$$f(dy) = \int p_G(x, dy) f(y) dy$$

The Gibbs sampler can be interpreted as a componentwise MH algorithm in which proposals are made from the full conditional distributions. Since transitions to the same point occur with probability zero, $r(x)$ in (20) equals 0 and the acceptance probability is equal to one. We can see how this construction of the Markov Chain is different from the MH algorithm in the aspect, that it is not necessarily reversible. Thus other conditions than those of Proposition 4.2 have to be met. There exist several sets of conditions which ensure that the Gibbs sampler satisfies the conditions of Proposition 4.1. A convenient set is due to Roberts and Smith [1994]:

**Proposition 4.3** *Suppose that*

*(i)* *$f(x) > 0$ implies there exists an open neighborhood $N_x$ containing $x$ and $\varepsilon > 0$ such that for all $y \in N_x$, $f(y) \geq \varepsilon > 0$;*
*(ii)* *$\int f(x) dx_i$ is bounded for all $k$ and all $y$ in an open neighborhood of $x$; and*
*(iii)* *the support of $x$ is arc connected.*

*then $p_G(x, y)$ satisfies the conditions of Proposition 4.1.*

If some of the full conditional densities are difficult to sample by traditional means, that density can be sampled by the MH algorithm [Mller, 1991]. This variant of the MH algorithm basically has some normal MH components, and some where the next element is drawn from the full conditional distribution and accepted with probability one. This has become known as the Metropolized Gibbs sampler. In Robert and Casella [1999] it is shown that the needed regularity conditions for the resulting markov chain are satisfied.

The Gibbs sampler from A2 has also been called the Gibbs sampler with systematic scan or sweep, as the path of iteration is to proceed systematically in one direction. Such a sampler results in a non-reversible Markov chain. Liu [1995] has proposed an alternative which is called Gibbs sampling with random scan. The simulation is done in a random order $\sigma$, which is drawn as a permutation on $1, \ldots, k$. (This Gibbs sampler produces a reversible chain $\{X^{(g)}\}$).

**Algorithm A3 (Random Sweep Gibbs Sampler)**

```
Given $x^{(g)} = (x_1^{(g)}, \ldots, x_k^{(g)})$,

  (1)  Generate a permutation $\sigma$ of $\{1, \ldots, k\}$
  (2)  Simulate $X_{\sigma_1}^{(g+1)} \sim f_1(x_1 | x_2^{(g)}, \ldots, x_k^{(g)})$
         ⋮
(k+1)  Simulate $X_{\sigma_k}^{(g+1)} \sim f_k(x_k | x_1^{(g)}, \ldots, x_{k-1}^{(g)})$
```

The Gibbs sampler is more intuitive in its construction and appears to be preferable over the MH algorithm since it uses the true distribution to derive the conditional distributions. A Metropolis-Hastings method can have a "bad" proposal distribution which will lead to many useless simulations (rejections). However, a Gibbs sampler faces different problems, which we can compare to those of an MLE algorithm. An MLE algorithm maximizes a high dimensional function one component at a time. It is well known how nonlinear optimization algorithm get stuck in local maxima. This strong attraction to the closest local mode is similar in the Gibbs sampler, where one component at a time is simulated. This restricts the possible excursions of the chain $\{X^{(g)}\}$ and explains why the gibbs sampling methods are usually slow to converge.

Now we have the methods at hand to construct a Markov chain whose invariant distribution is the joint posterior distribution of our model parameters given in (13). With these methods we can solve the problem of the unknown constants and that of high dimensionality. We will construct a transition kernel from different Gibbs and MH steps. Whenever we can compute an analytical full conditional posterior density, we will use this as the transition density of the kernel. If this is not possible, we will have to decide on an adequate proposal distribution.

In the next section we give a qualitative overview about how MCMC methods have been used to estimate ARMA and GARCH models before we derive our estimation procedure in section 5

### *4.4 MCMC methods for ARMA models*

In this section we give the results from Chib [1993], Chib [1996] and Chib and Greenberg [1994] who make use of the above methods to estimate ARMA models. Then we discuss how Nakatsuma [1998] and Yoo [2004] extend the approach of Chib and Greenberg to estimate basic markov switching models in a bayesian context. Then we show why this approach does not work, when not only the constant terms $c$ and $w$ are state dependent, but also the ARMA and GARCH components depend on the state of the latent markov chain. Then we have all the necessary background to explain our estimation technique.

In Chib and Greenberg [1994] the authors provide methods to estimate the parameters of ARMA(p,q) regression error models in a bayesian framework using the Gibbs sampler and Metropolis-Hastings algorithms. They consider the following Gaussian model in which the observation at time t is generated by

$$\tilde{y}_t = x_t'\beta + y_t \tag{24}$$

where $x_t$ is a $k \times 1$ vector of covariates, $\beta$ is the vector of regression parameters, and $y_t$ is a random error which follows an ARMA(p,q) process:

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \psi_j \varepsilon_{t-j} \qquad \varepsilon_t \sim N(0, \sigma^2) \tag{25}$$

which is expressed in terms of a polynomial in the backshift operator $L$ as

$$\phi(L)y_t = \psi(L)\varepsilon_t$$

They make the following assumptions:

(1) Stationarity: All roots of $\phi(L)$ lie outside the unit circle
(2) Invertibility: All roots of $\psi(L)$ lie outside the unit circle.
(3) Prior distributions:

$$[\beta, \phi, \psi, \sigma^2] = N_k(\beta_0, B_0) \times N_p(\phi_0, \Phi_0)\mathbf{1}_{S_\phi} \times N_q(\psi_0, \Psi_0)\mathbf{1}_{S_\psi} \times \mathcal{IG}(v_0/2, \delta_0/2)$$

where $\mathbf{1}_{S_\cdot}$ are indicator functions, securing the stationarity and invertibility. The parameters $\beta$ and $\sigma^2$ do not pose any difficulties and one can use standard results from the bayesian literature to arrive at analytical posterior distributions. These are the Normal and Inverse Gamma respectively, since the priors are conjugate [2].

This is different for $\phi$ and $\psi$. The posterior density can only be derived up to a constant so that these vectors have to be drawn in a MH-step.
A novelty in their article is the way in which they recursively transform the model equations to diagonalize the covariance matrix of the errors $y_t$. This results in a very "compact" expression for the unnormalized posterior density. This expression is not only compact, it yields a formulation of the covariance matrix for the parameter vectors $\phi$ and $\psi$, which are $p$ and $q$ dimensional. They can thus be drawn and tried in a single Metropolis-Hastings step instead of drawing each $\phi_i$ and $\psi_j$ individually. This can be a desirable feature, since the computation of the acceptance probability in the MH algorithm is based on the likelihood and can become very intensive. If the parameters are drawn in individual steps, the product of the MH-kernels would account for the correlation between them, but the likelihood would have to be evaluated $p+q$ times.

Then they construct a Markov chain from what some call a hybrid Gibbs-MH procedure. $\beta$ and $\sigma^2$ are drawn from their analytical posterior distribution whereas $\phi$ and $\psi$ are generated by an MH step. For the MH step they exploit the knowledge about the posterior distribution, expressed through the unnormalized density (as we pointed out in subsection 4.2.2).
They also prove that the conditions of proposition 4.3 are satisfied.

There does not seem to be much clarity in the literature, when an algorithm that constructs a Markov chain based on the product of kernels principle is to be called a Gibbs-MH algorithm, a hybrid Gibbs-MH procedure or a Metropolized Gibbs sampler. In Chib and Greenberg [1995] the authors criticize the expression of Metropolized-Gibbs sampler, since Hastings introduced the idea of drawing "blocks at-a-time" much earlier. When we compared the Gibbs sampler to the MH algorithm, we could see that the Gibbs sampler is indeed a special case of the "block-at-a-time" algorithm where the proposal distribution is the posterior distribution and the difference is that the acceptance probability is set to one. Since the MH chain was, by construction, reversible, we can see how this reversibility might be lost through a Gibbs kernel.

We will refer to a *Gibbs step*, when we draw from an analytically derived posterior distribution and always accept the value. We refer to an *MH step*, when we draw from a proposal distribution and accept the value according to

[2] These results correspond to table 1; note that the Inverse Gamma distribution is the general form of a scaled Inverse Chi Square distribution. They choose to formulate it in terms of the Inverse Gamma whereas we will prefer the form of a scaled inverse Chi Square distribution in later sections

(18)-(19). Also if the proposal distribution is an analytically derived posterior distribution!
We will not speak of a Metropolized Gibbs sampler etc, but of a Markov chain (MC) algorithm whose kernel consists of the product of Gibbs and MH kernels, or of a MC algorithm with Gibbs and MH steps.

In the next subsection we will briefly review how Nakatsuma [1998] uses these results to estimate the parameters of a GARCH model and how Yoo [2004] has "tweaked" the algorithm from Nakatsuma to estimate the parameters of a basic Markov switching GARCH model.

### *4.5 MCMC methods for GARCH models*

Nakatsuma extended the methods of Chib and Greenberg to estimate the parameters of a GARCH process. This is shown in more detail in section 5.2.2. Yoo uses the framework from (24) and uses the vector of covariates $x_t$ as the unobserved states of the economy. The regression coefficients in the parameter vector $\beta$ can now be thought of as the means of the different regimes. All other parameters are independent of the regime and remain as in equation (25). Thus, this framework does not allow a switching in the ARMA or GARCH parameters of the conditional mean or variance.
It remains the problem that the states are latent. This can be solved in a bayesian context through the method of *data augmentation*. This technique has been developed to deal with missing values in the data and applies equivalently to a problem where certain quantities are not observable. This has the convenient side-effect that one retrieves posterior estimates of the latent states at the same time.

## 5 Estimating the Model Parameters

This section introduces our estimation technique which is inspired by the results of the authors mentioned above.
This is done very thorough, recapitualting the steps from the sections, where the bayes estimator and the MCMC methods were introduced.

To estimate the model parameters with the bayesian technique we need to compute the posterior mean

$$\hat{\theta} = E[\theta | Y = y] = \int \theta \; p(\theta|y) \; d\theta$$

Therefore, we need to compute the posterior density of our model parameters. The posterior density is determined by the prior density and the likelihood as previously described in the section on bayesian estimation. We had

$$p(\theta|y) = \frac{f(y|\theta) \; p(\theta)}{\int f(y|\theta) \; p(\theta) \; d\theta}$$

following the common notation in bayesian statistics, we write

$$p(\theta|y) \propto f(y|\theta) \; p(\theta). \tag{26}$$

Since we do not only want to obtain the estimates of the model parameters, but also an estimate of the state $\hat{S}_{[1,T]}$, we need to use the posterior of all unobservable quantities

$$\begin{aligned} p(\theta, \; S_{[1,T]}| \; y) &\propto f(y| \; \theta, S_{[1,T]}) \; p(\theta, S_{[1,T]}) && (27) \\ &\propto f(y| \; \theta, S_{[1,T]}) \; p(S_{[1,T]}| \; \theta) \; p(\theta) && (28) \end{aligned}$$

By the Clifford-Hammersley theorem the joint posterior distribution is fully characterized by the complete conditional distributions

$$p(\theta \mid S_{[1,T]}, y) \qquad p(S_{[1,T]} \mid \theta, y)$$

and the markov chain

$$\{\theta^{(g)}, S^{(g)}\}, \; g \in \mathbb{N}$$

converges to the joint posterior distribution $p(\theta, S_{[1,T]}|y)$

The parameter space $\Theta$ is the cartesian product of the individual parameter spaces:

$$\Theta = \{\Pi \times C \times \Phi \times \Psi \times \Omega \times \alpha \times \beta\}$$

The joint posterior is $p(\theta, S_{[1,T]}|y)$, which is characterized by

$$p(p_{01}|\ \{\theta \backslash p_{01}\}, S_{[1,T]}, y),\ \ldots,\ p(\beta_2|\ \{\theta \backslash \beta_2\}, S_{1,T]}, y)$$

We will now construct an MCMC algorithm that produces a series of samples

$$\{\theta_1^{(g)}, \ldots, \theta_m^{(g)}, S_{[1,T]}^{(g)}\} \qquad g \in \mathbb{N}, \quad m = dim(\Theta)$$

which will converge to the joint posterior distribution. To obtain the bayes estimators $\hat{\theta}_i$ we compute the mean from the sample of the stationary distribution of the simulated $\theta_i$.

### 5.1 Implementing the MCMC Algorithm

To sample from the individual full conditional posterior distributions, we need to choose adequate prior distributions for the parameters. We use the priors as proposed in section 3.3. If we can obtain an analytic expression for the full conditional posterior density, then we use a Gibbs step to obtain the sample since an MH step is computationally more intensive. Otherwise we can just use a rather diffuse normal prior because its influence will vanish on samples of the size that we consider. Therefore we use normal priors for all ARMA and GARCH coefficients.

It is not always the case that an MH step is computationally more intensive than a Gibbs step. If the posterior distribution of the Gibbs sampler is very complicated and the likelihood of the MH step is not, then an MH algorithm can even be faster than a Gibbs algorithm. But in our case the computation of the likelihood is time consuming due to the path dependency of the model.

The steps in the MCMC algorithm are as follows:

- Draw the parameters of the transition probability matrix of the regime generating markov chain from a Dirichlet distribution
- Draw the States $S_t$ from $p(S_t|\{S_{[1,T]} \backslash S_t\}, \Theta, y)$ by the "Single Move" procedure.
- Draw the parameter of the t-distributed innovations
- Draw the ARMA-GARCH parameters

#### 5.1.1 Sampling the transition probabilities

The posterior distribution of $\pi_{i,j}$ is given by

$$p(\pi_{1,1}|y, S, \Theta \backslash \pi_{1,1}) \propto p(\pi_{1,1}) p(S, y|\Theta)$$

Since $S_t$ is independent of $y$, this is

$$p(\pi_{1,1}|y, S, \Theta\backslash\pi_{1,1}) \propto p(\pi_{1,1})p(S|\Theta)$$

Let $\eta_{i,j}$ be the cumulated number of transitions from state $i$ to state $j$ in the current sample $S^{(g)}_{[1,T]}$. Then we can write this as:

$$\begin{aligned} p(S|\Theta) &= \prod_{t=1}^{T} p(S_{t+1}|S_t, \Theta) & (29) \\ &= (\pi_{1,1})^{\eta_{1,1}}(\pi_{1,2})^{\eta_{1,2}}(\pi_{2,2})^{\eta_{2,2}}(\pi_{2,1})^{\eta_{2,1}} & (30) \\ &= (\pi_{1,1})^{\eta_{1,1}}(1-\pi_{1,1})^{\eta_{1,2}}(\pi_{2,2})^{\eta_{2,2}}(1-\pi_{2,2})^{\eta_{2,1}} \end{aligned}$$

This has the form of a beta density. The conjugate prior is therefore a beta distribution with the hyperparameters $h_{1,1}$, $h_{1,2}$, $h_{2,2}$ and $h_{2,1}$. The posterior distribution becomes:

$$\begin{aligned} p(\pi_{1,1}|y, S, \Theta\backslash\pi_{1,1}) &\propto p(\pi_{1,1})p(S|\Theta) \\ &\propto (\pi_{1,1})^{h_{1,1}-1}(1-\pi_{1,1})^{h_{1,2}-1}(\pi_{1,1})^{\eta_{1,1}}(1-\pi_{1,1})^{\eta_{1,2}} \\ &\propto (\pi_{1,1})^{\eta_{1,1}+h_{1,1}-1}(1-\pi_{1,1})^{\eta_{1,2}+h_{1,2}-1} \end{aligned}$$

Up to a constant this is the Beta density function. Therefore we sample $\pi_{i,j}$ in a Gibbs sampling step from the following Beta distribution:

$$\begin{aligned} \pi_{1,1}|S_{[1,T]} &\sim Beta(h_{1,1}+\eta_{1,1}, h_{1,2}+\eta_{1,2}) \\ \pi_{2,2}|S_{[1,T]} &\sim Beta(h_{2,2}+\eta_{2,2}, h_{2,1}+\eta_{2,1}) \end{aligned}$$

Higher dimensions of the chain:
(29) would become

$$p(S|\Theta) = (\pi_{1,1})\eta^{1,1}(\pi_{1,2})^{\eta_{1,2}} \dots (\pi_{1,\mathcal{S}})^{1,\mathcal{S}} \cdot (\pi_{2,1})^{\eta_{2,1}} \dots$$

for each row of $\Pi$, $\pi_s = (\pi_{s,1}, \dots, \pi_{s,\mathcal{S}})$, this is proportional to the density from a Dirichlet distribution. A conjugate prior would thus be a Dirichlet distribution with the hyperparameters $\alpha_s = (\alpha_{s,1}, \dots, \alpha_{s,\mathcal{S}})'$:

$$f(x|\alpha) = \frac{1}{B(\alpha)}\prod_{i=1}^{k} x_i^{\alpha_i-1} \quad B(\alpha) = \frac{\prod_{i=1}^{k}\Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k}\alpha_i)}$$

The posterior is then again a Dirichlet distribution with the parameters $\alpha+\eta$. We therefore obtain a sample of the transition probabilities from state $s$ to all others by generating a draw from

$$P(\Pi_s|\alpha_s) = Dirichlet(\alpha_{s,1}+\eta_{s,1}, \dots, \alpha_{s,\mathcal{S}}+\eta_{s,\mathcal{S}})$$

In the next step, we need to obtain a sample of the states. We will follow the single move scheme suggested by  Carlin et al. [1992].

### *5.1.2 Sampling $S_{[1,T]}$*

In this step of the MCMC algorithm we obtain a sample from the distribution of the entire markov chain $S_{[1,T]}$. One possibility is to compute the measure $p(S_t|y,\Theta)$, but because of the path dependency of the likelihood $p(y|S_{[1,T]},\Theta)$ the time complexity is in $\mathcal{O}(\mathcal{S}^T)$ and therefore computationally not feasible. The single move procedure breaks this step down into a Gibbs cycle of $T$ consecutive draws from the conditional distribution of the state at a single point in time, conditional on all other states. This is done as follows:

We compute the measure $p(S_t|\{S_{[1,T]}\backslash S_t\},\Theta,y)$. We write $\{S_{[1,T]}\backslash S_t\}$ as $S_{\neq t}$, $S_{[1,T]}$ as $S$ and omit to explicitly condition on $\Theta$.

$$\begin{aligned}
p(S_t|S_{\neq t},y) &= \frac{p(S_t,S_{\neq t},y)}{p(y,S_{\neq t})} \\
&= \frac{p(y|S)\cdot p(S)}{p(y|S_{\neq t})\cdot p(S_{\neq t})} \\
&= \frac{p(y|S)\cdot p(S|S_{\neq t})}{p(y|S_{\neq t})}
\end{aligned}$$

$p(y|S)$ is computed easily. With a given sample of $S$ this is simply the likelihood of the model. $p(S_t|S_{\neq t})$ is only dependent on $S_{t-1}$ and $S_{t+1}$ due to the markov property of the chain.

$$\begin{aligned}
p(S_t=i|y,S_{\neq t}) &= p(S_t=i|S_{t-1},S_{t+1}) \\
&= \frac{\pi_{l,i}\cdot\pi_{i,k}}{\sum_{i=1}^{\mathcal{S}}\pi_{l,i}\cdot\pi_{i,k}}
\end{aligned}$$

with $S_{t-1}=l$ , $S_{t+1}=k$ and $\pi_{i,j}$ the respective transition probabilities from $\Pi$.
Since for all $S_t=i$, $i\in\{1,\ldots,\mathcal{S}\}$, $p(y|S_{\neq t})$ is constant, we write

$$p(S_t=i|y,S_{\neq t})\propto p(y|S_{t=i},S_{\neq t})\cdot p(S_t=i|S_{\neq t})$$

Because $p(S_t|y,S_{\neq t})$ is a probability measure, we can now compute it as

$$p(S_t=i|y,S_{\neq t})=\frac{p(y|S_t=i,S_{\neq t})\cdot p(S_t=i|S_{\neq t})}{\sum_{i=1}^{\mathcal{S}}p(y|S_t=i,S_{\neq t})\cdot p(S_t=i|S_{\neq t})}$$

One sample of $S_{[1,T]}$ is thus obtained by cycling through these steps:
For each $t \in \{1, \dots, T\}$, starting with $t = 1$:

- Compute the distribution $p(S_t = i | S_{\neq t}, y)$ on $\{1, \dots, \mathcal{S}\}$.
- Draw $S_t$ from this distribution.
- Update $S_{[1,T]}$ with this value.

This is also a Gibbs sampler with systematic scan.

### *5.2 Sampling the ARMA-GARCH Parameters*

We will use a Metropolis Hastings step to obtain samples from the full conditional posterior distributions of these parameters. The so sampled posterior distribution will converge to the true posterior distribution for (almost) any proposal distribution. However, for the speed of the convergence it is crucial to select an adequate proposal distribution. We will therefore exploit all the knowledge we have about the full conditional posterior.

#### *5.2.1 Sampling the parameters of the conditional mean*

At first we demonstrate our procedure for a simple ARMA model without regime switching in the parameter $c$. In such a model the ARMA coefficients are generated as follows:

*Sampling c*

$$\begin{aligned} p(c|\Theta\backslash c, S, y) &\propto f(y|\Theta, S)p(c) \\ &\propto \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi h_t}} exp\left\{ -\frac{(y_t - c - \phi_{S_t} y_{t-1} - \psi_{S_t} \varepsilon_{t-1})^2}{2h_t} \right\} p(c) \end{aligned}$$

The last expression is only a function of $c$ and we can treat all other parameters as constants and with

$$\mathcal{C}_t = y_t - \phi_{S_t} y_{t-1} - \psi_{S_t} \varepsilon_{t-1}$$

we rewrite it as:

$$\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi h_t}}exp\left\{-\frac{(\mathcal{C}_t-c)^2}{2h_t}\right\}p(c)=\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi h_t}}exp\left\{-\frac{\mathcal{C}_t^2-2\mathcal{C}_t\cdot c+c^2}{2h_t}\right\}p(c)$$

$$\propto exp\left\{\frac{\sum_{t=1}^{T}\mathcal{C}_t^2}{2h_t}\right\}\cdot exp\left\{\sum_{t=1}^{T}\frac{-2\mathcal{C}_t\cdot c+c^2}{2h_t}\right\}p(c)$$

$$\propto exp\left\{\frac{\sum_{t=1}^{T}\mathcal{C}_t^2}{2h_t}\right\}\cdot exp\left\{c\cdot\sum_{t=1}^{T}\frac{-\mathcal{C}_t}{h_t}+c^2\sum_{t=1}^{T}\frac{1}{2h_t}\right\}p(c)$$

This has the form of a normal density with

$$\sigma^{-2}=\sum_{t=1}^{T}\frac{1}{h_t}\qquad \mu=\sum_{t=1}^{T}\frac{\mathcal{C}_t}{h_t}\cdot\left(\sum_{t=1}^{T}\frac{1}{h_t}\right)^{-1}$$

As the proposal distribution we choose $N(\mu_c,\sigma_c^2)$ with

$$\mu_c=\sum_{t=1}^{T}\frac{y_t-\phi_{S_t}y_{t-1}-\psi_{S_t}\varepsilon_{t-1}}{h_t}\cdot\left(\sum_{t=1}^{T}\frac{1}{h_t}\right)^{-1}$$

$$\sigma_c^{-2}=\sum_{t=1}^{T}\frac{1}{h_t}$$

For the other parameters we can proceed in a similar fashion. Next we introduce regime switching only into the mean and then outline the complete algorithm for the full model.
The model is

$$y_t=c_1\cdot\mathbf{1}_{[S_t=1]}+c_2\cdot\mathbf{1}_{[S_t=2]}+\phi_{S_t}y_{t-1}+\psi_{S_t}\varepsilon_{t-1}+\varepsilon_t$$

The posterior becomes

$$p(c_1|\Theta\backslash c_1,S,y)\propto\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi h_t}}exp\left\{-\frac{(y_t-c_1\cdot\mathbf{1}_{[S_t=1]}-c_2\cdot\mathbf{1}_{[S_t=2]}-\phi_{S_t}y_{t-1}-\psi_{S_t}\varepsilon_{t-1})^2}{2h_t}\right\}p(c)$$

$$\mathcal{C}_t=y_t-\mathbf{1}_{[S_t=2]}c_2-\phi_{S_t}y_{t-1}-\psi_{S_t}\varepsilon_{t-1}$$

Even though we cannot observe the state $S_t$ in reality, the MCMC algorithm provides us with a sample $S_{[1,T]}$ which we simply plug into the above formula. And following the previous steps we arrive at:

$$p(c_1|\Theta\backslash c_1,S,y)\propto exp\left\{c_1\sum_{t=1}^{T}\frac{-\mathcal{C}_t\mathbf{1}_{[S_t=1]}}{h_t}+c_1^2\sum_{t=1}^{T}\frac{\mathbf{1}_{[S_t=1]}}{2h_t}\right\}$$

The proposal Distribution therefore is $N(\mu_{c_1},\sigma_{c_1}^2)$ with

$$\mu_{c_1}=\sum_{t=1}^{T}\frac{\mathbf{1}_{[S_t=1]}y_t-\mathbf{1}_{[S_t=1]}\phi_1y_{t-1}-\mathbf{1}_{[S_t=1]}\psi_1\varepsilon_{t-1}}{h_t}\cdot\left(\sum_{t=1}^{T}\frac{\mathbf{1}_{[S_t=1]}}{h_t}\right)^{-1}$$

$$\sigma_{c_1}^{-2} = \sum_{t=1}^{T} \frac{\mathbf{1}_{[S_t=1]}}{h_t}$$

*Sampling $\phi$*

$$p(\phi_1|\Theta\phi_1, S, y) \propto \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi h_t}} exp\left\{-\frac{(y_t - c_{S_t} - \phi_{S_t} y_{t-1} - \psi_{S_t}\varepsilon_{t-1})^2}{2h_t}\right\} p(\phi_1)$$

Now treat only $\phi_1$ as variable and with

$$\mathcal{C}_t = y_t - c_{S_t} - \mathbf{1}_{[S_t=2]}\phi_2 y_{t-1} - \psi_{S_t}\varepsilon_{t-1}$$

we rewrite the above as

$$\begin{aligned} p(\phi_1|\Theta\backslash\phi_1, S, y) &\propto \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi h_t}} exp\left\{-\frac{(\mathcal{C}_t - \mathbf{1}_{[S_t=1]}\phi_1 y_{t-1})^2}{2h_t}\right\} p(\phi_1) \\ &\propto exp\left\{\phi_1 \sum_{t=1}^{T} \frac{-\mathbf{1}_{[S_t=1]}\mathcal{C}_t y_{t-1}}{h_t} + \phi_1^2 \sum_{t=1}^{T} \frac{\mathbf{1}_{[S_t=1]} y_{t-1}^2}{2h_t}\right\} p(\phi_1) \end{aligned}$$

Analogue to the above results we get

$$\mu_{\phi_1} = \sum_{t=1}^{T} \frac{-\mathbf{1}_{[S_t=1]}\mathcal{C}_t y_{t-1}}{h_t} \cdot \left(\sum_{t=1}^{T} \frac{\mathbf{1}_{[S_t=1]} y_{t-1}^2}{h_t}\right)^{-1}$$

$$\sigma_{\phi_1}^{-2} = \sum_{t=1}^{T} \frac{\mathbf{1}_{[S_t=1]} y_{t-1}^2}{h_t}$$

*Sampling $\psi$*

$$p(\psi_1|\Theta\backslash\psi_1, S, y) \propto \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi h_t}} exp\left\{-\frac{(y_t - c_{S_t} - \phi_{S_t} y_{t-1} - \psi_{S_t}\varepsilon_{t-1})^2}{2h_t}\right\} p(\psi_1)$$

Treat only $\psi_1$ as variable

$$\mathcal{C}_t = y_t - c_{S_t} - \phi_{S_t} y_{t-1} - \mathbf{1}_{[S_t=2]}\psi_2\varepsilon_{t-1}$$

and the conditional posterior distribution of $\psi_1$ is

$$p(\psi_1|\Theta\backslash\psi_1, S, y) \propto exp\left\{\psi_1 \sum_{t=1}^{T} \frac{-\mathbf{1}_{[S_t=1]}\mathcal{C}_t\varepsilon_{t-1}}{h_t} + \psi_1^2 \sum_{t=1}^{T} \frac{\mathbf{1}_{[S_t=1]}\varepsilon_{t-1}^2}{2h_t}\right\} p(\psi_1)$$

Therefore we choose $N(\mu_{\psi_1}, \sigma^2_{\psi_1})$ as the proposal distribution with

$$\mu_{\psi_1} = \sum_{t=1}^{T} \frac{-\mathbf{1}_{[S_t=1]}\mathcal{C}_t \varepsilon_{t-1}}{h_t} \cdot \left( \sum_{t=1}^{T} \frac{\mathbf{1}_{[S_t=1]}\varepsilon^2_{t-1}}{h_t} \right)^{-1}$$

$$\sigma^{-2}_{\psi_1} = \sum_{t=1}^{T} \frac{\mathbf{1}_{[S_t=1]}\varepsilon^2_{t-1}}{h_t}$$

### *5.2.2 Sampling the GARCH Coefficients*

As shown by Bollerslev [1986] a GARCH(p,q) process is expressed as an ARMA(l,s) process of:

$$\varepsilon_t^2 = \omega + \sum_{j=1}^{l} (\alpha_j + \beta_j)\varepsilon^2_{t-j} + \tilde{w}_t - \sum_{j=1}^{s} \beta_j \tilde{w}_{t-j}$$

with: $\alpha_j = 0$ for $j > p$, $\beta_j = 0$ for $j > q$, $l = min(p,q)$, $s = q$
and

$$\begin{aligned} \tilde{w}_t &:= \varepsilon_t^2 - \sigma_t^2 \\ &= \left( \frac{\varepsilon_t^2}{\sigma_t^2} - 1 \right) \sigma_t^2 \\ &= (\chi^2(1) - 1)\sigma_t^2 \end{aligned}$$

The conditional mean of $\tilde{w}_t$ is $E[\tilde{w}_t|\mathcal{F}_{t-1}) = 0$, and the conditional variance is $Var(\tilde{w}_t|\mathcal{F}_{t-1}) = 2\sigma_t^4$ . Nakatsuma [1998] suggests to replace this $\tilde{w}_t$ with $w^* \sim N(0, 2\sigma_t^4)$. Then we have an auxiliary ARMA model for the squared errors $\varepsilon_t^2$:

$$\varepsilon_t^2 = \omega + \sum_{j=1}^{l} (\alpha_j + \beta_j)\varepsilon^2_{t-j} + w_t - \sum_{j=1}^{s} \beta_j w_{t-j} \qquad w_t \sim N(0, 2\sigma_t^4) \tag{31}$$

Rewriting this expression and factoring out $\beta_j$, we get

$$\varepsilon_t^2 = \omega + \sum_{j=1}^{l} \alpha_j \varepsilon^2_{t-j} + w_t + \sum_{j=1}^{s} \beta_j (\varepsilon^2_{t-j} - w_{t-j})$$

$$w_t = \varepsilon_t^2 - \omega - \sum_{j=1}^{l} \alpha_j \varepsilon^2_{t-j} - \sum_{j=1}^{s} \beta_j \underbrace{(\varepsilon^2_{t-j} - w_{t-j})}_{=:v_t} \sim N(0, 2\sigma_t^4)$$

In our GARCH(1,1) setting $l = 1$ and $s = 1$ we arrive at a posterior distribution for $\omega$ as follows:

$$p(\omega_1|\Theta\backslash\omega_1, S, y) \propto \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi 2h_t^2}} exp\left\{-\frac{(\varepsilon_t^2 - \omega_{S_t} - \alpha_{S_t}\varepsilon_{t-1}^2 - \beta_{S_t}v_{t-1})^2}{4h_t^2}\right\} p(\omega_1)$$

As before we write

$$\mathcal{C}_t = \varepsilon_t^2 - \mathbf{1}_{[S_t=1]}\omega_2 - \alpha_{S_t}\varepsilon_{t-1}^2 - \beta_{S_t}v_{t-1}$$

and obtain

$$p(\omega_1|\Theta\backslash\omega_1, S, y) \propto exp\left\{\omega_1 \sum_{t=1}^{T} \frac{-\mathbf{1}_{[S_t=1]}\mathcal{C}_t}{2h_t^2} + \omega_1^2 \sum_{t=1}^{T} \frac{\mathbf{1}_{[S_t=1]}}{4h_t^2}\right\} p(\omega_1)$$

Our proposal distribution for $\omega_1$ therefore is $N(\mu_{\omega_1}, \sigma_{\omega_1}^2)$ with

$$\mu_{\omega_1} = \sum_{t=1}^{T} \frac{-\mathbf{1}_{[S_t=1]}\mathcal{C}_t}{2h_t^2} \cdot \frac{1}{\sum_{t=1}^{T} \frac{\mathbf{1}_{[S_t=1]}}{4h_t^2}}$$

$$\sigma_{\omega_1}^{-2} = \sum_{t=1}^{T} \frac{\mathbf{1}_{[S_t=1]}}{2h_t^2}$$

$\alpha$ and $\beta$ are obtained analogue to these results

## 5.3 Estimating the parameters of the t-distributed innovations

In the model where the innovations are student t-distributed, the algorithm changes slightly. The MH-steps essentially stay the same, but we will adjust the proposal distribution. We also have to estimate the degree of freedom parameter $v_{S_t}$.
First let us consider the posterior distribution of the degree of freedom parameter $v$. Again the model can be specified in several ways, one possibility is to model the innovations independent of the states, that is to say, there is only one $v$ for all $t$. Or this parameter could also be chosen to be state dependent. To start with we will consider the former case of a regime independent degree of freedom parameter. It will then be straight forward to extend the approach.

### 5.3.1 Sampling $v$

We follow Jacquier et al. [2003] and choose a discrete flat prior for $v$, whereas Geweke [1993] uses a continous prior to estimate the degree of freedom in student-t linear models. The posterior distribution of $v$ is proportional to the product of $t$ distribution ordinates:

$$\begin{aligned} p(v|\Theta\backslash v, S, y) &\propto p(v)p(y|\Theta, S) \\ &\propto \prod_{t=1}^{T} \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\,\Gamma(\frac{v}{2})}(1+\frac{\hat{e}_t^2}{h_t v})^{-(\frac{v+1}{2})} \end{aligned}$$

where

$$\hat{e}_t = y_t - \sum_{i=1}^{r}\phi_i(S_t)y_{t-i} - \sum_{i=1}^{m}\psi_i(S_t)\varepsilon_{t-i}$$

Theoretically $v$ is in $\mathbb{N}^+$, but we choose a flat prior on $\{3,\ldots,40\}$. The posterior distribution can then be calculated analytically as follows:
Let

$$\tilde{p}(v|\Theta\backslash v, S, y) = \prod_{t=1}^{T} \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\,\Gamma(\frac{v}{2})}(1+\frac{\hat{e}_t^2}{h_t v})^{-(\frac{v+1}{2})} \tag{32}$$

Then for $p(v|\Theta\backslash v, S, y)$ with a flat prior on $\{3,\ldots,40\}$, $p(v)=\frac{1}{37}$, we get:

$$p(v|\Theta\backslash v, S, y) = \frac{\tilde{p}(v|\Theta\backslash v, S, y)}{\sum_{i=3}^{40}\tilde{p}(i|\Theta\backslash i, S, y)}\cdot\frac{1}{37}$$

Truncating the interval of $v$ on $\{3,\ldots,40\}$ will not result in inaccuracies as long as the sampled values $v^{(g)}$ do not touch the boundaries. We will only choose to model data with a t-distribution if the degree of freedom is significantly smaller than 30. Above 30 it is common in statistical literature to approximate the t-distribution with the normal since they are very close.

### *5.3.2 Proposal Distributions*

With normally distributed innovations we were able to find a good proposal distribution analytically. It is also common in the literature on MCMC algorithms to choose a proposal distribution who's moments are obtain by a Maximum Likelihood estimate. This is legitimate even though it appears as one is mixing up the different inferential approaches. The ML estimates merely serve as proxies for the means and the variances of the proposal distributions. In fact the parameters $\mu$ and $\sigma$ as they were computed in section (5.2) correspond exactly to the ML estimates of the mean and variance of the conditional distribution.
The scheme we used to obtain the paramters for the normal do not work for t-distributed innovations, where we end up with a nonlinear expression. The ML estimates are also obtained from a nonlinear equation and in practice they are calculated numerically.

In our bayesian context we can circumvent this inconvenience due to the fact that for Student's t-distribution, there exists a hidden mixture representation through the normal distribution. Since $p(x|\theta)$ is the mixture of a normal distribution and an inverse gamma distribution:

$$
\begin{aligned}
&x|z \sim N(\theta, z\sigma^2),\\
&z^{-1} \sim Gamma(\frac{v}{2}, \frac{v}{2}).
\end{aligned}
$$

This is a well known trick in the bayesian literature, see for example Robert [1994].
We can now rewrite our model as

$$
\begin{aligned}
y_t &= c_{S_t} + \sum_{i=1}^{r} \phi_i(S_t) \cdot y_{t-i} + \eta_t + \sum_{j=1}^{m} \psi_j(S_t) \cdot \eta_{t-j} && \text{(33a)}\\
h_t &= \omega_{S_t} + \sum_{i=1}^{p} \alpha_i(S_t) \cdot \eta_{t-i}^2 + \sum_{j=1}^{q} \beta_j(S_t) \cdot h_{t-j} && \text{(33b)}\\
\eta_t &= \sqrt{h_t} \cdot \tilde{\eta}_t && \text{(33c)}\\
\tilde{\eta}_t &= \sqrt{\lambda_t} \cdot u_t \quad \sim \quad t(v) && \text{(33d)}\\
u_t &\sim N(0,1) && \text{(33e)}\\
\lambda_t &\sim \mathcal{IG}(\frac{v}{2}, \frac{v}{2}) && \text{(33f)}
\end{aligned}
$$

This helps us to find the proposal distributions in the following way: Conditional on a sample $\lambda^{(g)}_{[1,T]}$ the normalized residuals are again normal:

$$\varepsilon_t = \left( y_t - c_{S_t} - \sum_{i=1}^{r} \phi_i(S_t) \cdot y_{t-i} - \sum_{j=1}^{m} \psi_j(S_t) \cdot \eta_{t-j} \right) \frac{1}{\sqrt{h_t \lambda_t}} \tag{34}$$

The sample $\lambda^{(g)}_{[1,T]}$ is obtained in an individual Gibbs step.

### *5.3.3 Sampling $\lambda_t$*

The posterior distribution of $\lambda_t$ is given by

$$p(\lambda_t | \eta_t, h_t, v) \propto \lambda_t^{-\frac{(v+3)}{2}} exp \left\{ -(\frac{\eta_t^2}{h_t} + v) \frac{1}{2\lambda_t} \right\}$$

This is proportional to the density of a $\chi^2$ random variable and we have

$$(\frac{\eta_t^2}{h_t} + v) \frac{1}{\lambda_t} \sim \chi^2(v+1)$$

To obtain the sample $\lambda_t$, we draw from $x \sim \chi^2(v+1)$ and compute

$$\lambda_t = (\frac{\eta_t^2}{h_t} + v) \frac{1}{x}$$

In the model specified in equations (33a) through (33f) we set

$$\tilde{h}_t = h_t \cdot \lambda_t$$

Replacing $h_t$ in section 5.2 with $\tilde{h}_t$ we obtain the proposal distribution for the parameters of the conditional mean in the case of t-distributed innovations.

For the proposal distributions of the parameters in the conditional variance of the model, we set

$$\tilde{\varepsilon}_t = \frac{\eta_t}{\sqrt{\lambda_t}}$$

and use it to replace $\varepsilon_t$. The generalized Gibbs sampling algorithm ensures that the markov chain converges to the true posterior distribution of hierarchical representation of the parameters of the t-distribution.

We have now developed all the necessary steps of the algorithm which we will summarize in the next section.

### *5.4 The Complete Algorithm*

At the beginning of this chapter we outlined the MCMC algorithm that samples a markov chain $\{\theta^{(g)}, S^{(g)}\}$ which converges to the joint posterior distribution $p(\Theta, S_{[1,T]}|y)$. Now we can present a precise descritption of the algorithm:

$\theta^{(g)}$ denotes the parameter set obtained in the $g^{th}$ step. The sample value $\theta^{(g+1)}$ is obtained by iterating through the following steps:

(1) Sample $\Pi^{(g+1)}$: draw $\pi_{i,j}$ from $p(\pi_{i,j}|y, S^{(g)}, \theta^{(g)}\backslash\pi_{i,j})$

$$\begin{aligned}\pi_{1,1}|S_{[1,T]} &\sim Beta(h_{1,1} + \eta_{1,1}, h_{1,2} + \eta_{1,2})\\ \pi_{2,2}|S_{[1,T]} &\sim Beta(h_{2,2} + \eta_{2,2}, h_{2,1} + \eta_{2,1})\end{aligned}$$

(2) Sample $S^{(g+1)}$ by the single move procedure from

$$p(S_t = i|y, S_{\neq t}) = \frac{p(y|S_{[1,T]}) \cdot p(S_t = i|S_{\neq t})}{\sum_{i=1}^{S} p(y|S_{[1,T]}) \cdot p(S_t = i|S_{\neq t})}$$

(3) Sample $\lambda_{[1,T]}^{(g+1)}$: compute

$$\begin{aligned}\hat{\eta}_t &= y_t - c_{S_t} - \sum_{i=1}^{r} \phi_i(S_t) \cdot y_{t-i} - \sum_{j=1}^{m} \psi_j(S_t) \cdot \hat{\eta}_{t-j}\\ \hat{h}_t &= \omega_{S_t} + \sum_{i=1}^{p} \alpha_i(S_t) \cdot \hat{\eta}_{t-i}^2 + \sum_{j=1}^{q} \beta_j(S_t) \cdot \hat{h}_{t-j}\end{aligned}$$

with the parameters taken from the current sample $\theta^{(g)}$ and $S_t$ from $S^{(g+1)}$. Then draw a sample $x_{[1,T]}$ from $\chi^2(v + 1)$ with $v$ from $\theta^{(g)}$ and compute

$$\lambda_t = (\frac{\eta_t^2}{h_t} + v)\frac{1}{x_t}$$

(4) Sample $v^{(g+1)}$: use the same $\hat{\eta}_t$ and $\hat{h}_t$ from the previous step to sample from

$$p(v|\theta^*\backslash v, S, y) = \frac{\tilde{p}(v|\theta^*\backslash v, S^{(g+1)}, y)}{\sum_{i=3}^{40} \tilde{p}(i|\theta^*\backslash i, S^{(g+1)}, y)} \cdot \frac{1}{37}$$

with $\tilde{p}(i|\theta^*\backslash i, S^{(g+1)}, y)$ from equation (32). $\theta^*$ is the current parameter set $\theta^{(g)}$, updated with the parameters sampled in steps 1 to 3. Using the same residuals and conditional variances from step 3 helps to avoid problems documented by Eraker et al. [1998], in which $v$ can get absorbed into the lower bound.

(5) Cycle through the ARMA-GARCH parameters. For each parameter draw from the respective proposal distribution by the scheme outlined above. The candidate is then tried in a Metropolis Hastings step according to section 4.2 and either accepted or rejected.
Let $\hat{\vartheta}$ denote the proposed parameter value. We update $\theta^*$ with this value and refer to this parameter set as $\hat{\theta}$. We accept $\hat{\theta}$ as the new $\theta^*$ with probability

$$\alpha_{MH}(\theta^*, \hat{\theta}) = min\left\{\frac{p(\hat{\theta}|\, y, S^{(g+1)})}{p(\theta^*|\, y, S^{(g+1)})}\Big/\frac{g(\hat{\theta})}{g(\theta^*)}, 1\right\}$$

where $p(\hat{\theta}|\, y, S^{(g+1)})$ is the likelihood of the data $y$ and the current sample $S^{(g+1)}$, calculated with the parameter set $\hat{\theta}$. $g(.)$ is the respective proposal density. This is done for all parameters of the ARMA-GARCH specification. Finally we have obtained the new sample $\{\theta^{(g+1)}, S^{(g+1)}\}$.

### *5.4.1 Variations of the Algorithm*

As discussed in section 4.3 there are several variations of the MH- and Gibbs Sampling algorithm. These differ in their performance especially when it comes to sample the ARMA-GARCH parameters. For the different regimes these parameters are highly correlated and their generation is computationally intensive. Chib and Greenberg report that they could increase the performance of MH algorithms when highly correlated parameters were sampled in a block. We can confirm this observation. Especially for the conditional mean we observed that the chain converged much faster to its stationary distribution. Even though more proposals were rejected, the chain moved around more freely in the parameter space and thereby explored the shape of the posterior faster because it did not get stuck in certain regions for too long. We also found the algorithm proposed by Liu [1995] to be beneficial in this respect.

Most numerical implementations of parameter estimation algorithms perform quite well, if the data is well behaved and fits well. The more interesting cases are the ones, were the parameters of a model are at the border of the respective parameter space. For ARMA-GARCH processes such a 'stress' situation arises, when the process is close to being integrated. We simulated a stationary MS-ARMA-GARCH model with the sum of the GARCH parameters being close to unity. The acceptance rates of some parameters dropped to about ten percent, resulting in a slow convergence of the Markov chain. Here we found that a specification of the algorithm presented in 5.4, were each parameter was tried separately in an MH step could increase the acceptance rate. Such a low acceptance rate can create problems with the convergence of the estimates of the regime process. Since the Gibbs step for the states is the most time consuming, we only do this every 10 iterations of the algorithm, allowing the

ARMA-GARCH parameters to explore their changed posterior distributions. We argued above, that a drop in the acceptance rate does not necessarily have a negative effect as long as the chain explores the whole parameter space better. But if the ARMA-GARCH parameters get only updated once in between to samples of the states, this can slow down the convergence considerably.

Another problem, that arises in the case, when either the conditional mean or variance is close to being integrated, is the choice of the proposal distribution. Say for example that $\beta_1(1) = 0.95$, then the normal approximation for $\tilde{w}_t$ is very questionable and the proposed mean $\mu_{\beta_1}$ will often be greater than one. This leads to an extremely high rejection of the proposed value. Since we want to impose stationarity, we do not accept $\beta_i$'s that are equal or greater than one. If $\mu_{beta_i}$ should be greater than one, we will set $\mu_{beta_i} = \beta^{(g)}$. This has proven to be very useful when we estimated simulated data with one $\beta_i$ being close to one.

### *5.5 Estimating Hamilton's '94 model with this algorithm*

For the purpose of comparison, we will also estimate a variant of the model proposed in Hamilton and Susmel [1994]. With the method developed in this section, it is not a big step to compute the estimators for M2. We will deviate slightly from Hamilton's proposal and formulate the model as

**Model M5 (Modified Hamilton '94)** *The latent state governing the evolution of the model parameters is again assumed to follow an $\mathcal{S}$ dimensional time discrete Markov chain whose transition matrix is given through* (1). *The conditional mean of the time series $\{y_t\}$ is as in* (2a)

$$\begin{aligned}
y_t &= c_{S_t} + \sum_{i=1}^{r} \phi_i(S_t) \cdot y_{t-i} + \varepsilon_t + \sum_{j=1}^{m} \psi_j(S_t) \cdot \varepsilon_{t-j} \\
h_t &= w + \sum_{i=1}^{q} \alpha_i \cdot \varepsilon_{t-i} + \sum_{j=1}^{p} \beta_j \cdot h_{t-j} \\
\varepsilon_t &= \sqrt{g_{S_t}} \cdot \sqrt{h_t} \cdot u_t \\
u_t &\sim N(0,1)
\end{aligned}$$

The mean is as in model M4 and the conditional variance is an *amplified* GARCH process. The main difference between model M2 and M4 is the specification of the variance. The other parameters of the model can be estimated through a straightforward application of the methods shown in the previous sections. Therefore we will now discuss how the amplifying parameter $g_{S_t}$ can be estimated within our MCMC setting.

The likelihood $f(\theta|y, S_{[1,T]})$ is given through

$$f(\theta|y, S_{[1,T]}) = \prod_{t_1 \in \mathcal{I}_1} \frac{1}{\sqrt{2\pi g_1 h_t}} exp\left\{-\frac{\varepsilon_{t_1}^2}{2g_1 h_t t_1}\right\} \times \ldots \times \prod_{t_S \in \mathcal{I}_S} \frac{1}{\sqrt{2\pi g_S h_t}} exp\left\{-\frac{\varepsilon_{t_s}^2}{2g_S h_{t_s}}\right\}$$

where $\mathcal{I}_s$ is the index set containing all points in time when $S_t = s$. It is easy to see that for a certain $g_s$ the likelihood of $g_s$ is proportional to

$$\begin{aligned} f(g_s|\{\theta \backslash g_s\}, y, S_{[1,T]}) &\propto \prod_{t_s \in \mathcal{I}_s} \frac{1}{\sqrt{2\pi h_t}} g^{-\frac{1}{2}} exp\left\{-\frac{\varepsilon_t^2/(2h_t)}{g_s}\right\} \\ &\propto g_s^{-N_s/2} exp\left\{-\frac{1}{g_s} \sum_{t \in \mathcal{I}_s} \frac{\varepsilon_t^2}{2h_t}\right\} \end{aligned}$$

where $N_s = card(\mathcal{I}_s)$. This has the form of an *Inverse Gamma* density, which is

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} exp\left\{\frac{-\beta}{x}\right\}$$

with shape parameter $\alpha$ and scale parameter $\beta$. Therefore a conjugate prior $p(g_s)$ is an inverse gamma distribution with the hyperparameters $\alpha_0, \beta_0$. The posterior distribution for $g_s$ would then be

$$f(g_s|\{\theta \backslash g_s\}, y, S_{[1,T]}) \propto g^{-N_s/2} exp\left\{-\frac{1}{g_s} \sum_{t \in \mathcal{I}_s} \frac{\varepsilon_t^2}{2h_t}\right\} \cdot p(g_s|\alpha_0, \beta_0)$$

and hence

$$g_s \sim \mathcal{IG}(\frac{1}{2}N_s + \alpha_0, \sum_{t \in \mathcal{I}_s} \frac{\varepsilon_t^2}{2h_t} + \beta_0)$$

We specified the innovations $u_t$ as coming from a normal distribution. We could also let them be student t-distributed, in the estimation procedure we would then simply have to apply the result from subsection 5.3.2.

## 6 Diagnosing Convergence

This section provides some words about how to check wether the markov chain of the MCMC algorithm has 'converged'. This can of course never be said for sure, but there are some visual indicators and some statistics that can be calculated.

Robert and Casella [1999] differentiate between three types of convergence criteria. These differ in increasingly stringent conditions:

(1) Convergence to the Invariant Distribution
This criterion considers the convergence of the chain $X^{(g)}$ to the invariant distribution $f$. However, $f$ is only the limiting distribution of $X^{(g)}$ and invariance is only an asymptotical property.

(2) Convergence of Averages
This type considers the convergence of the empirical average

$$\frac{1}{N}\sum_{g=1}^{N} h(X^{(g)})$$

to $\mathbb{E}_f[h(X)]$.

(3) Convergence to iid Sampling
This convergence criterion measures how close a sample is to being iid and looks at the independence requirements for the simulated values.

We will be most concerned with the convergence of averages. Even if the initial values, $x^{(0)}$ are distributed according to $f$, the exploration of the support of $f$ by the chain can be poor. This will depend heavily on the transition kernel, chosen in the MCMC algorithm. The purpose of the convergence assessement is therefore to determine whether the chain has scanned $f$ to a satisfactory degree. While the ergodic theorem guarantees the convergence theoretically, it is most relevant in a practical implementation to have a stopping rule. This rule would yield a minimal $N$ which justifies the approximation of $\mathbb{E}_f[h(X)]$.

Robert and Casella provide a graphical method which is due to Yu and Mykland [1998]. It is based on cumulative sums (CUSUM), where they plot the differences

$$D_N^i = \sum_{g=1}^{i}[h(X^{(g)}) - \hat{h}_N], \qquad i = 1, 2, \ldots, N \tag{35}$$

where

$$\hat{h}_N = \frac{1}{N}\sum_{g=1}^{N} h(X^{(g)}).$$

They derive a qualitative evaluation of the mixing speed of the chain and the correlation between the $X^{(g)}$'s. When the mixing of the chain is high, the graph of $D_N^i$ is highly irregular and concentrated around zero. Slowly mixing chains produce regular graphs with long excursions away from 0. The difficulty with this method is, that it does not give any indications for unexplored regions of the chain. This is because the diagnose is only based on a single chain and one gets no idea of the regions that have not been seen.

## 7 Estimation Results

### *7.1 Performance of the Estimator on Simulated Data*

We simulate different parameterizations of the model and take a closer look at the posterior statistics. These results can then be compared to the ones retrieved on real data and help to judge the performance of the estimator and its convergence.
We commence by assessing the quality of the estimates on single regime processes and compare them to the respective ML estimates. Then we compare the results that we achieved with our estimation methodology to results previously reported in the literature. Bauwens and Lubrano [1998] derive a Griddy Gibbs sampler to estimate the parameters of simple GARCH processes. Their main focus is directed to the performance of their method on small samples. This is the area where the ML methodology differs mainly from the bayesian approach. On samples of about 100-150 data points the influence brought about by the choice of the prior distributional assumption clearly shows.
Chib and Greenberg [1994] develop a hybrid mcmc method in which they use both MH and Gibbs sampling steps to estimate processes with ARMA(p,q) errors. They too, consider small samples when they compare their estimators to the ML alternative.

For the results in table 2 we used a constant scaling factor for the variance of the proposal distribution of 1.5. The proposal distribution was a Student-t distribution with degree of freedom $v^{(g)}$. They use a flat prior for the degree of freedom parameter, which seems to result in a bias on such small samples. This influence seems to vanish on larger samples as we will see later. For the IS (Importance Sampling) algorithm they base their importance function on a Student-t density and fix the location parameter at the posterior mode. The scaling parameters were tuned experimentally. For the MH estimator (see MH(1) in table 2) they use the same Student-t distribution as a proposal distribution. The results that we achieve with our algorithm (MH(2)) compare

| Parameter | ML | Griddy Gibbs | IS | MH (1) | MH (2) |
|---|---|---|---|---|---|
| $w$ (0.1) | 0.19 [0.10] | 0.26 [0.10] | 0.26 [0.10] | 0.21 [0.08] | 0.17 [0.07] |
| $\alpha$ (0.4) | 0.24 [0.12] | 0.30 [0.13] | 0.31 [0.13] | 0.26 [0.10] | 0.43 [0.10] |
| $\beta$ (0.4) | 0.38 [0.25] | 0.31 [0.17] | 0.31 [0.17] | 0.37 [0.17] | 0.41 [0.08] |
| $v$ (5.0) | 6.09 [2.49] | 9.7 [5.34] | 9.6 [5.02] | 9.90 [5.48] | 10.5 [7.85] |

Table 2
Posterior results by different methods for a Student-t-GARCH model on simulated data.

well to those of the other methods, if not better. This is however hard to determine on such a small sample.
Even though we are less concerned with smaller samples, using their results as a benchmark for our estimation procedure underpins the correctness of our approach.

In table 2 we can see how different the results can be. Whereas Bauwens and Lubrano's MH algorithm seems less capable on the small sample, our specification seems to produce the best result. This is due to the issues that we already addressed in section 4.2.2. The method of determining the location and scale parameters of the proposal distribution is critical for the MH algorithm to explore the domain of the posterior distribution. Our algorithm provides an automatic choice for the location parameter of the proposal distribution which varies along the chain. It does also produce a value for the scale parameter (variance). Even though these values are produced using our knowledge about the posterior distribution, this does not necessarily make them the best choice. To see this we compared several choices of the variance scaling parameter.

The results from table 2 were found on a small sample of 150 data points. Our focus is on larger samples, where the influence of the information implied by the prior distribution vanishes as we discussed in section 3.2.2. We will now study the behavior of our estimator on larger samples and under more difficult parameterizations.

Table 3 reports the results for the estimator on a simulated GARCH process with 1500 data points. The estimates for $w$ and $\beta$ seem "a bit off". Running the algorithm on the same data set, with the proposal variance scaled by factor 1.5, the results reported in table y are now very close to the true values. The parameters of the GARCH process were chosen to be rather moderate. That is, that they were not close to being integrated. As with a maximization algorithm, the performance of the MCMC methods could change, when the algorithm operates close to the border of a constrained domain. To see how our algorithm reacts in such cases, we studied the results on parameterizations where the sum $\alpha + \beta$ was in between 0.9 and 1. Especially for high $\beta$'s the acceptance rate drops very fast, which makes it necessary to increase the number of iterations of the algorithm. Nevertheless, our algorithm did not get "stuck" at any point and we conclude that it is reliable.

Our algorithm produces the correct results on samples with normally distributed innovations. Now we can examine its performance for student-t-distributed innovations. Here we found that with several parameterizations the degree of freedom parameter can be biased towards higher values. This

| Variance scaling |  | 1 |  |  | 1.5 |  |
|---|---|---|---|---|---|---|
| Parameter | $w$ | $\alpha$ | $\beta$ | $w$ | $\alpha$ | $\beta$ |
| True value | 2.3 | 0.2 | 0.6 | 2.3 | 0.2 | 0.6 |
| $\hat{\theta}$ | 2.78 | 0.195 | 0.542 | 2.269 | 0.203 | 0.600 |
| $\hat{\sigma}$ | 0.678 | 0.042 | 0.087 | 0.454 | 0.033 | 0.056 |
| Median | 2.693 | 0.194 | 0.5499 | 2.228 | 0.201 | 0.603 |
| Lower 5% limit | 1.843 | 0.1281 | 0.3863 | 1.594 | 0.148 | 0.500 |
| Upper 95% limit | 4.025 | 0.2671 | 0.6717 | 3.129 | 0.259 | 0.689 |
| Acceptance rate | 0.44 | 0.43 | 0.44 | 0.33 | 0.34 | 0.34 |

Table 3
This table compares the performance of the algorithm with two different choices of the variance scaling parameter. The GARCH process that is estimated was simulated with normal innovations and had a length of 1500 data points. The MCMC chain was run 50000 iterations of which the first 10000 were discarded (burn in time). The variance scaling parameter was set to 1 and 1.5 respectively.

bias for $v$ seems to be a problem in those cases where $\sum_{i=1}^{q} \alpha + \sum_{j=1}^{p} \beta$ is about 0.9 or higher. For other parameterizations the estimates where again very close to the true values, and exhibited small posterior standard deviations. The results presented in table 4 illustrate our findings for a specific simulation. Here we can see that for $\alpha = 0.3$ and $\beta = 0.6$ the true $v$ is even below the estimated 5% quantile. In the second panel we estimated a process with the same GARCH parametrization but normal innovations and see that the true parameters are recovered very precisely.

We recall that the degree of freedom parameter is drawn from a closed form (full conditional) posterior distribution in a Gibbs step. In Robert and Casella [1999] the authors report, that a chain from a Gibbs sampler can be very slow to converge. A bias such as the one we observed can occur when the chain does not explore all modes of the parameter space and gets trapped in a certain region. The speed of convergence can then be ameliorated by including an additional MH step in the algorithm. This MH step helps to avoid getting trapped in certain regions and allows the chain to move around more freely in the parameter space.

Having verified the performance of the algorithm on single regime processes, we can now examine the performance on simulated data from multi-regime Markov switching processes. In table 5 we see that the algorithm performs well, estimating the parameters of a Markov switching process. Again, the sum of the GARCH parameters is considerably smaller than 0.9. Although

| Parameter | True value | $\hat{\theta}$ | $\hat{\sigma}$ | Median | 5% quant. | 95% quant. | AR |
|---|---|---|---|---|---|---|---|
| $w$ | 0.1 | 0.132 | 0.024 | 0.131 | 0.093 | 0.173 | 0.321 |
| $\alpha$ | 0.3 | 0.374 | 0.041 | 0.374 | 0.308 | 0.443 | 0.322 |
| $\beta$ | 0.6 | 0.535 | 0.037 | 0.535 | 0.474 | 0.597 | 0.327 |
| $v$ | 7 | 14.07 | 5.13 | 13 | 8 | 24 | |
| $w$ | 0.1 | 0.098 | 0.017 | 0.097 | 0.071 | 0.126 | 0.259 |
| $\alpha$ | 0.3 | 0.318 | 0.033 | 0.316 | 0.265 | 0.376 | 0.255 |
| $\beta$ | 0.6 | 0.616 | 0.030 | 0.617 | 0.560 | 0.665 | 0.253 |
| $v$ | 0 | | | | | | |

Table 4
The simulated GARCH process spanned 1500 data points. The MCMC chain was run 50000 iterations of which the first 10000 were discarded (burn in time). The variance scaling parameter was set to 1.5 .

not depicted here, the estimates for the latent states reliably reproduced the true values.

Studying the performance on simulated data, we could verify the reliability of our procedure and can now move on to real data. In the next section we examined data from the New York Stock exchange. The estimated GARCH parameters were in those ranges where we did not experience these problems for the degree of freedom. Therefore we did not implement a version of the algorithm where the $v$ was sampled in an MH-step.

## *7.2 Empirical Results*

In this section we provide the estimation results of our model specification and make an effort to compare them to other results reported in the literature.

In Hamilton and Susmel [1994], estimates of the parameters for model M2 are reported and several different parameterizations are compared to each other. The authors conclude that the t-SWARCH-L(3,2) specifications is the most appropriate model for that data. They assessed the suitability of their model mainly through the forecasting properties. They observed that the single regime GARCH models implied a very high persistence, but did a worse

| Parameter | $w_1$ | $\alpha_1$ | $\beta_1$ | $w_2$ | $\alpha_2$ | $\beta_2$ | $v$ | $\pi_{1,1}$ | $\pi_{2,2}$ |
|---|---|---|---|---|---|---|---|---|---|
| True value | 3.3 | 0.1 | 0.4 | 0.6 | 0.2 | 0.08 | 8 | 0.998 | 0.997 |
| $\hat{\theta}$ | 3.94 | 0.145 | 0.376 | 0.557 | 0.191 | 0.095 | 7.19 | 0.9962 | 0.9974 |
| $\hat{\sigma}$ | 1.294 | 0.065 | 0.163 | 0.081 | 0.052 | 0.015 | 1.64 | 0.0031 | 0.0022 |
| 5%'tile | 1.855 | 0.046 | 0.111 | 0.428 | 0.111 | 0.0073 | 5 | 0.9901 | 0.9934 |
| 95%'tile | 6.135 | 0.264 | 0.646 | 0.691 | 0.281 | 0.241 | 10 | 0.9986 | 0.9997 |

Table 5
The simulated GARCH process spanned 1500 data points. The MCMC chain was run 50000 iterations of which the first 10000 were discarded (burn in time). The variance scaling parameter was set to 1.5 .

job forecasting the volatility than a simple constant variance model measured by mean squared errors. This is counterintuitive, since if the variance was persistent, that would imply better forecasts with a model that accounts for this persistence.

GARCH(1,1) models are generally outperforming low order ARCH(q) models when it comes to forecasting the conditional variance. It is a more parsimonious approach to capture the autocorrelation structure of the process and is therefore often preferred over an ARCH(q) specification. For this reason we believe that regime switching processes should be modelled with switching GARCH processes rather than low order ARCH processes. Also the conditional mean should be modelled as either a one state ARMA(r,m) process or a switching ARMA(r,m) process. If the autocorrelation structure of the mean provides a better fit to the data, this will have positive repercussions on the estimates of the conditional variance as well.

We use the same data set to see if the our specification improves the model fit. We looked at the sample autocorrelation function (sample ACF) for the weekly returns and found that a simple AR(1) specification of the conditional mean seemed to be an unlikely candidate to capture this rather complex structure. We decided to start with the standard ARMA(1,1)-GARCH(1,1) model with gaussian innovations and fitted it to the data. In a first step we assessed the goodness of fit through an examination of the sample ACF for the estimated residuals of the one day ahead forecasts. These exhibited significant autocorrelations in the mean at lag 1,3,6 and 8 at the 5 percent level. The sum of the estimates for the GARCH coefficients is close to unity, indicating that the process of the conditional variance could be integrated, hence the estimates would be of no great value.

We therefore estimated the parameters of an ARMA(1,1)-GARCH(1,1) model with t-distributed innovations, which seemed to provide an adequate description of the dependency structure, measured by the sample ACF of the estimated standardized residuals $\hat{u}_t$. When we examined the distributional shape of these residuals we could see in a qqplot, that they did not seem to be as heavy tailed as implied by a student t-distribution as estimated. Except for the one outlier produced by the crash in '87.

The observed values of the sample ACF of the residuals were extremely small. This is caused by the extreme innovation of the crash in '87. In this model such an event would have to be regarded as an outlier. If we would not regard this value as an outlier, this indicates that the data actually stems from a distribution that has much heavier tails. In fact such a model would not even have a variance nor a finite covariance. The autocorrelation as a measure for dependance is then not very meaningful. Never the less, we could truncate the data set to exclude this extreme event, and then try to describe the dependency with the sample ACF. This gives us a proxy by which we can diagnose the model fit of the rest of the data.

The sample ACF of the residuals of the truncated data set showed no significant autocorrelations. This indicates that the single regime ARMA might be enough to describe the mean of the weekly return series. In fact, as we fitted a two regime model of the mean, the results of the algorithm were very unstable in the estimated states, also an indication of no significant switching in the mean.

We then fitted an MS(1,2)-ARMA(1,1)-GARCH(1,1) specification of M4 to the data. Compared to the single regime model the estimates for the GARCH parameters were significantly smaller which shows that some of the persistence in the conditional variance is now attributed to a change in the regime.
We could observe a similar effect when we estimated the parameters of a two regime specification of the modified Hamilton'94 model (M5). Though only the amplification parameter varies with the regime, the persistence implied by the GARCH component decreased.
The fit of both of these models as measured by the sample ACF of the estimated residuals was still unsatisfactory.

Therefore we chose to model the conditional variance as a 3 regime process and estimated such a specification of model M5 and model M4 on the weekly NYSE returns. The resulting estimators and posterior density statistics are reported in table B.1 and B.2 respectively. From figure 1 we see how the states as estimated with our modified Hamilton'94 model are not as clearly identified as in the original model Hamilton and Susmel [1994].
In both models the third regime could be termed the high volatility regime and we can regard the second regime as the normal state. The first regime

only appears once. Therefore we cannot be sure wether this is a structural break in the process or a state that can govern the parameters again.

In our estimates we can observe that the $\beta_3$ is the highest value amongst the $\beta$'s in our model, at the same time it is smaller than the $\beta$ from the modified Hamilton'94 model. This shows that the high level of persistence in the variance process is attributed more to Markov switching than to a GARCH effect.

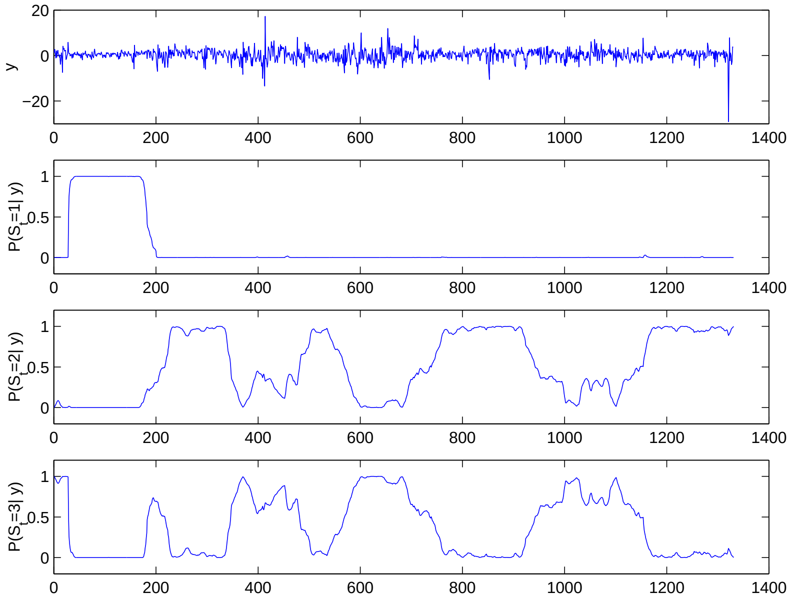

Fig. 1. Posterior probabilities of the different regimes for model M5 on the weekly NYSE returns. The corresponding parameter estimates are found in table B.1.

We compare the estimated posterior probabilities for both models which are depicted in figure 2 and 1. We see how our model is able to distinguish the different states much sharper, which naturally is a very desirable feature. Next we compare the sample ACF of the two models. Since the ARMA parameters are estimated to be about the same in both models, we only show the sample ACF for the squared residuals. In figure 4 and 3 we can see that our model captures the autocorrelation structure of the data much better. This is a bit at the cost of higher standard deviations of the posterior distribution as we can see from tables B.1 and B.2, but we still conclude that model M4 clearly outperformes M5.

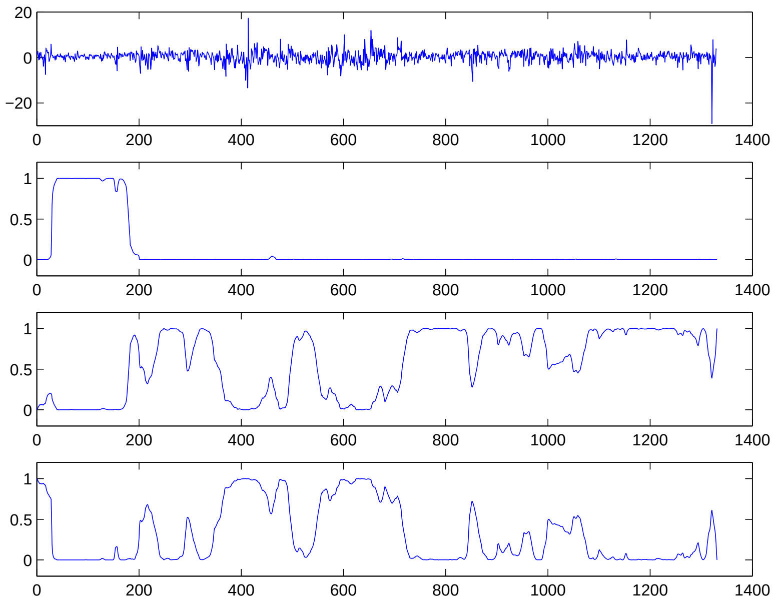

Fig. 2. Posterior probabilities of the regime 1 through 3 for the model M4. The corresponding parameter estimates are found in table B.2.

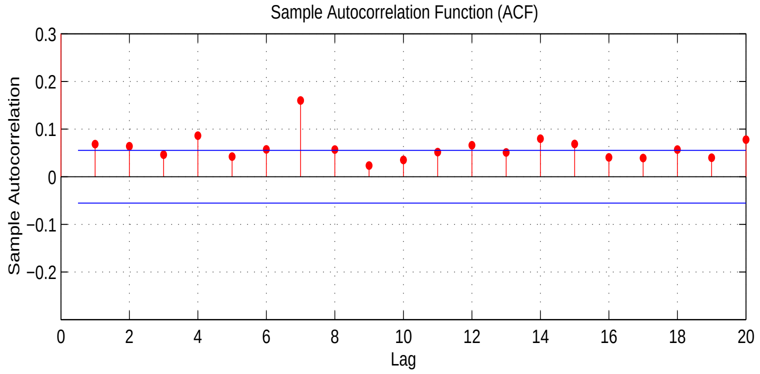

Fig. 3. Autocorrelation of the squared residuals computed without the crash in'87 for the modified version of Hamiltons model (M5) parameterized with the estimates from table B.1.

## 8 Conclusion

We developed a Markov Chain Monte Carlo method to compute the parameter estimates for a full MS-ARMA-GARCH model. These models are regarded as a promising class to describe certain phenomena in econometric time series observed on different markets. They seem appealing since they can tell a "story" that is readily interpreted. But due to their severe intractability their power was hard to assess in practice. In fact the only thing that is straightforward about Markov switching models is their specification. Due to their full path dependence, models with MA or GARCH components could not be estimated. But this is does not allow to unleash the full power of ARMA-GARCH

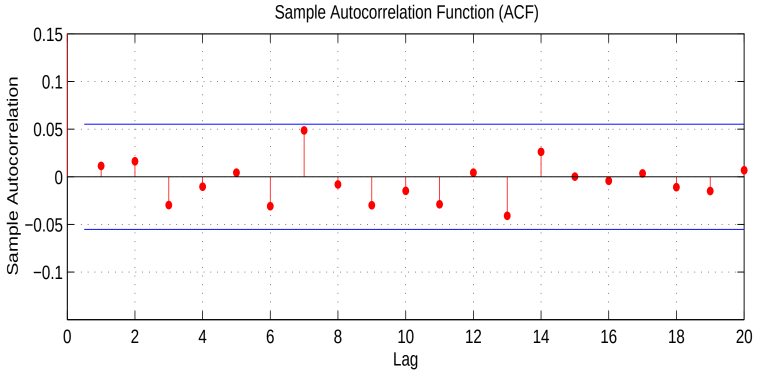

Fig. 4. Autocorrelations of the squared residuals computed without the crash in '87 for our model (M4) parameterized with the estimates from table B.2.

models. The algorithm that we presented overcomes this problem and is an important step for the further study of Markov switching models. It can easily be extended and employed for all specifications of Markov switching models. The results obtained on the NYSE weekly returns demonstrate the advantage of specifying an ARMA process for the mean and a switching GARCH process for the variance

If the data shows only weak signs of autocorrelation in the mean, the model M3 of Haas et al. seems to be a good alternative to our specification, since it is analytically more tractable and less intensive to estimate.

A lot of attention is directed towards these models at the moment and other research, as that of Francq and Zakoian [2001b], has established important results regarding the conditions for stationarity or the tail behavior of these models. Though Markov switching models are hard to handle, a lot of progress has been made and we are confident that with increasing computational power these models are a valuable tool for financial modelers.

## References

Luc Bauwens and Michel Lubrano. Bayesian inference on garch models using the gibbs sampler. *Econometrics Journal*, 1(0):C23–C46, 1998.

Julian Besag. Markov chain monte carlo for statistical inference. Technical report, University of Washington, 2001.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31:307–327, 1986.

Tim Bollerslev, Ray Y. Chou, and Kenneth F. Kroner. Arch modeling in finance: A review of the theory and empirical evidence. *Journal of Econometrics*, 52:5–59, 1992.

N.G. Carlin, B.P. Polson, and D.S. Stoffer. A monte carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association*, 87, 1992.

George Casella and Edward I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

Nicholas Chan, Mila Getmansky, Shane M. Haas, and Andrew W. Lo. Systemic risk and hedge funds. 2005.

Siddhartha Chib. Bayes regression with autoregressive errors, a gibbs sampling approach. *Journal of Econometrics*, (58):275–294, 1993.

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

Siddhartha Chib and Edward Greenberg. Bayes inference in regression models with arma(p,q) errors. *Journal of Econometrics*, 64:183–206, 1994.

Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

M. Creutz. Confinement and the critical dimensionality of space-time. *Physics Review Letters*, 43:553–556, 1979.

B. Eraker, Jacquier, and E. Polson. Pitfalls in mcmc algorithms. Technical report, Department of Econometrics and Statistics. Graduate School of Business, University of Chicago, April 1998.

Christian Francq and J.M. Zakoian. Conditional heteroskedasticity driven by hidden markov chains. *Journal of Time Series Analysis*, 22:197–220, 2001a.

Christian Francq and J.M. Zakoian. Stationarity of multivariate markov-switching arma models. *Journal of Econometrics*, (102):339–364, 2001b.

Christian Francq and J.M. Zakoian. $l^2$- structures of standard and switching-regime garch models and their implications for statistical inference. Working paper, Univ. du Littoral-Cote d'Opale , Universite de Lille 3 and CREST, 2002.

S. Geman and D. Geman. Stochastic relaxation, gibbs distribution and the bayesian restoration of images. *Institute of Electrical Engineers, Transitions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

J Geweke. Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics*, 8(0):S19–40, 1993. available at http://ideas.repec.org/a/jae/japmet/v8y1993isps19-40.html.

Lawrence R. Glosten, Jagannathan Ravi, and David Runkle. Relationship between the expected value and the volatility of the nominal excess return on stocks. Technical report, Northwestern University, Evanston IL, 1989.

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

U. Grenander. Tutorial in pattern theory. Technical report, Division of Applied Mathematics, Brown University, 1983.

M. Haas, S. Mittnik, and M. S. Paolella. A new approach to markov-switching garch models. *Journal of Financial Econometrics*, 2(4):493–530, 2004.

James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–84, 1989. available at http://ideas.repec.org/a/ecm/emetrp/v57y1989i2p357-84.html.

James D. Hamilton. Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45(1-2):39–70, 1990. available at http://ideas.repec.org/a/eee/econom/v45y1990i1-2p39-70.html.

James D. Hamilton and Raul Susmel. Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, 64(1-2):307–333, 1994. available at http://ideas.repec.org/a/eee/econom/v64y1994i1-2p307-333.html.

W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

Eric Jacquier, Nicholas G. Polson, and P.E. Rossi. Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics*, 122:185–212, 2003.

Peter M. Lee. *Bayesian Statistics*. Arnold, 2003.

E.L Lehman and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics. Springer, second edition edition, 1998.

J. Liu. Metropolized gibbs sampler: An improvement. Technical report, Dept. of Statistics, Stanford University, CA., 1995.

Koichi Maekawa, Sangyeol Lee, and Yasuyoshi Tokutsu. A note on volatility persistence and structural changes in garch models. Technical report, University of Hiroshima, Faculty of Economics, 2005.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953.

P. Mller. A generic approach to posterior integration and gibbs sampling. Technical report, Department of Statistics, Purdue University, 1991.

Teruo Nakatsuma. A markov-chain sampling algorithm for garch models. *Studies in Nonlinear Dynamics & Econometrics*, 3(2), 1998.

David E. Rapach and Jack K. Strauss. Structural breaks and garch models of exchange rate volatility. Technical report, Saint Louis University, 2005.

B.D. Ripley. Algorithm as 137: simulating spatial patterns: dependent samples from a multivariate density. *Applied Statistics*, 28:109–112, 1979.

Christian P. Robert. *The Bayesian Choice*. Springer-Verlag New York, 1994.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, 1999.

G.O. Roberts and A.F.M. Smith. Some simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic Processes and Applications*, 49:207–216, 1994.

G.O. Roberts and R.L. Tweedie. Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika*, 83:95–110, 1996.

P. Suomela. *Unpublished PhD thesis*. PhD thesis, University of Jyvskyl, 1976.

L. Tierny. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22:1701–1762, 1994.

Byoung Hark Yoo. A bayesian analysis of markov switching models with arma-garch error. Technical report, Rutgers University, 2004.

Bin Yu and Per Mykland. Looking at Markov samplers through cusum path plots: A simple diagnostic idea. *Statistics and Computing*, 8(3):275–286, 1998.

## A Bayesian Estimation

**Assumption A1** *The log likelihood function $l(\theta)$ satisfies the usual set regularity conditions as they are imposed in the context of asymptotic efficiency. These can for example be found in Lehman and Casella [see 1998, Theorem 6.2.6]*

- *The parameter space $\Omega$ is an open interval.*
- *The distributions $P_\theta$ of the $X_i$ have common support, so that the set $A = x : f(x|\theta) > 0$ is independent of $\theta$.*
- *For every $x \in A$, the density $f(x|\theta)$ is twice differentiable with respect to $\theta$, and the second derivative is continuous in $\theta$.*
- *The integral $\int f(x|\theta)d\mu(x)$ is twice differentiable under the integral sign (where $\mu$ is a $\sigma$-finite measure).*
- *The Fisher information $I(\theta)$ satisfies $0 < I(\theta) < \infty$.*
- *For any given $\theta_0 \in \Omega$, there exists a $c > 0$ and a function $M(x)$ such that $|\partial^2 log f(x|\theta)/\partial\theta^2 \leq M(x)|$ for all $x \in A$, $\theta_0 - c < \theta < \theta_0 + c$ and $E_{\theta_0}[M(X)] < \infty$.*

**Assumption A2** *Given any $\varepsilon > 0$, there exists $\delta > 0$ such that in the expansion*

$$l(\theta) = l(\theta_0) + (\theta - \theta_0)l'(\theta_0) - \frac{1}{2}(\theta - \theta_0)^2[nI(\theta_0) + R_n(\theta)]$$

*where*

$$\frac{1}{n}R_n(\theta) \xrightarrow{\mathcal{P}} 0 \quad as \quad n \to \infty$$

*the probability of the event*

$$sup\left\{\left|\frac{1}{n}R_n(\theta)\right| : |\theta - \theta_0| \leq \delta\right\} \geq \varepsilon$$

*tends to zero as $n \to \infty$.*

**Assumption A3** *For any $\delta > 0$, there exists $\varepsilon > 0$ such that the probability of the event*

$$sup\left\{\frac{1}{n}[l(\theta) - l(\theta_0)] : |\theta - \theta_0| \geq \delta\right\} \leq -\varepsilon$$

*tends to $1$ as $n \to \infty$.*

**Assumption A4** *The prior density $p$ of $\theta$ is continuous and positive for all $\theta \in \Theta$.*

**Assumption A5** *The expectation of $\theta$ under $p$ exists*

# B Empirical Results

| Parameter | $\hat{\theta}$ | $\hat{\sigma}$ | Median | 5% Quantile | 95% Quantile | Acceptance Rate |
|---|---|---|---|---|---|---|
| $c$ | 0.3307 | 0.0847 | 0.3278 | 0.2010 | 0.4883 | 0.3824 |
| $\phi_1$ | 0.3388 | 0.1429 | 0.3481 | 0.0917 | 0.5756 | 0.3824 |
| $\psi_1$ | -0.0930 | 0.1512 | -0.0970 | -0.3374 | 0.1776 | 0.3824 |
| | | | | | | |
| $w$ | 2.1884 | 1.0259 | 2.0585 | 0.7517 | 4.2162 | 0.4107 |
| $\alpha$ | 0.1785 | 0.0347 | 0.1766 | 0.1253 | 0.2421 | 0.3992 |
| $\beta$ | 0.4119 | 0.0988 | 0.4165 | 0.2384 | 0.5745 | 0.3918 |
| | | | | | | |
| $g_1$ | 0.1477 | 0.0877 | 0.1231 | 0.0620 | 0.3506 | - |
| $g_2$ | 0.5672 | 0.3466 | 0.4678 | 0.2372 | 1.4084 | - |
| $g_3$ | 1.1706 | 0.7623 | 0.9489 | 0.4375 | 2.9625 | - |
| | | | | | | |
| $v$ | 6.3998 | 1.0703 | 6.0000 | 5.0000 | 8.0000 | - |
| | | | | | | |
| $\pi_{1,1}$ | 0.9850 | 0.0108 | 0.9893 | 0.9689 | 0.9981 | - |
| $\pi_{1,2}$ | 0.0013 | 0.0012 | 0.0005 | 0.0000 | 0.0037 | - |
| $\pi_{1,3}$ | 0.0036 | 0.0034 | 0.0019 | 0.0000 | 0.0105 | - |
| | | | | | | |
| $\pi_{2,1}$ | 0.0319 | 0.0178 | 0.0301 | 0.0015 | 0.0654 | - |
| $\pi_{2,2}$ | 0.9861 | 0.0094 | 0.9888 | 0.9669 | 0.9967 | - |
| $\pi_{2,3}$ | 0.0090 | 0.0082 | 0.0068 | 0.0002 | 0.0243 | - |
| | | | | | | |
| $\pi_{3,1}$ | 0.0040 | 0.0062 | 0.0013 | 0.0000 | 0.0167 | - |
| $\pi_{3,2}$ | 0.0172 | 0.0106 | 0.0155 | 0.0020 | 0.0368 | - |
| $\pi_{3,3}$ | 0.9713 | 0.0207 | 0.9774 | 0.9288 | 0.9932 | - |

Table B.1
Estimated parameter values and posterior statistics for model M5 on the weekly NYSE returns

| Parameter | $\hat{\theta}$ | $\hat{\sigma}$ | Median | 5% Quantile | 95% Quantile | Acceptance Rate |
|---|---|---|---|---|---|---|
| $c$ | 0.3168 | 0.0788 | 0.3115 | 0.1977 | 0.4531 | 0.3211 |
| $\phi_1$ | 0.3642 | 0.1345 | 0.3727 | 0.1254 | 0.5699 | 0.3211 |
| $\psi_1$ | -0.1208 | 0.1438 | -0.1270 | -0.3519 | 0.1266 | 0.3211 |
| $w_1$ | 0.3298 | 0.0694 | 0.3216 | 0.1765 | 0.5116 | 0.5374 |
| $\alpha_1$ | 0.3517 | 0.1369 | 0.3456 | 0.1455 | 0.5858 | 0.5417 |
| $\beta_1$ | 0.1780 | 0.1182 | 0.1576 | 0.0225 | 0.3979 | 0.5345 |
| $w_2$ | 1.1423 | 0.3944 | 1.1304 | 0.5217 | 1.7951 | 0.4243 |
| $\alpha_2$ | 01082 | 0.1333 | 0.1029 | 0.0314 | 0.2032 | 0.4201 |
| $\beta_2$ | 0.3794 | 0.1840 | 0.3759 | 0.0846 | 0.6731 | 0.4192 |
| $w_3$ | 3.9438 | 1.1507 | 3.9218 | 2.0629 | 5.9045 | 0.4932 |
| $\alpha_3$ | 0.1804 | 0.0521 | 0.1730 | 0.0759 | 0.2978 | 0.4888 |
| $\beta_3$ | 0.2658 | 0.1467 | 0.2535 | 0.0466 | 0.5288 | 0.4923 |
| $v$ | 7.5089 | 1.5672 | 7.0000 | 5.0000 | 10.0000 | - |
| $\pi_{1,1}$ | 0.9925 | 0.0098 | 0.9913 | 0.9689 | 0.9981 | - |
| $\pi_{1,2}$ | 0.0001 | 0.0015 | 0.0005 | 0.0000 | 0.0037 | - |
| $\pi_{1,3}$ | 0.0074 | 0.0038 | 0.0019 | 0.0000 | 0.0105 | - |
| $\pi_{2,1}$ | 0.0017 | 0.0178 | 0.0018 | 0.0015 | 0.0154 | - |
| $\pi_{2,2}$ | 0.9811 | 0.0111 | 0.9835 | 0.9616 | 0.9975 | - |
| $\pi_{2,3}$ | 0.0172 | 0.0082 | 0.0168 | 0.0002 | 0.0243 | - |
| $\pi_{3,1}$ | 0.0005 | 0.0062 | 0.0003 | 0.0000 | 0.0067 | - |
| $\pi_{3,2}$ | 0.0368 | 0.0106 | 0.0355 | 0.0020 | 0.0468 | - |
| $\pi_{3,3}$ | 0.9627 | 0.0175 | 0.9668 | 0.9318 | 0.9980 | - |

Table B.2
Estimated parameters and posterior statistics of model M4 on the weekly returns from the New York Stock Exchange