Austrian Research Institue for Artificial Intelligence
Österreichisches Forschungsinstitut für Artificial Intelligence
http://www.oefai.at/oefai/oefai.html

Freyung 6/6, Vienna
A-1010, Austria
(+43-1) 5336112-0

ÖFAI Neural Computation Group Technical Report

October 16, 2003

OEFAI–TR–2003–30

Variational Bayesian Autoregressive Conditional Heteroskedastic Models

Brian Sallans
ÖFAI Neural Computation Group

Abstract

A variational Bayesian autoregressive conditional heteroskedastic (VB-ARCH) model is presented. The ARCH class of models is one of the most popular for economic time series modeling. It assumes that the variance of the time series is an autoregressive process. The variational Bayesian approach results in an approximation to the full posterior distribution over ARCH model parameters, and provides a method for model selection. A novel application of Monte Carlo sampling is presented, wherein sampling is used to evaluate difficult terms in the variational free energy. A description of the variational approximation is followed by encouraging experimental results on model selection and volatility prediction on synthetic and historical financial data.

Variational Bayesian Autoregressive Conditional Heteroskedastic Models

Brian Sallans
ÖFAI Neural Computation Group

1 Introduction

Financial time series have unique statistical properties. Stock returns, exchange rates, stock indices, and many other time series share traits such as: Zero or near zero mean; low or no autocorrelations; significant autocorrelations of the absolute or squared values (called "volatility clustering"); and high kurtosis in the marginal empirical distribution [Cont 2001]. Econometric models have been developed to model these properties. One of the most widely used is the class of autoregressive conditional Heteroskedastic (ARCH) models [Engle 1982].

The ARCH family was introduced to model the high kurtosis and volatility clustering properties of financial time series. They model the mean and, more importantly, the variance of time series. Accurately modeling variance (or volatility), is an extremely important problem in econometrics. Important applications include value at risk estimation, valuation of options, and modeling of exchange rate volatility.

ARCH models has been studied both theoretically and empirically [Engle 1982; Bollerslev 1986; Bera and Higgins 1993; Kaufmann and Frühwirth-Schnatter 2002]. Empirically, ARCH models have been treated under a maximum likelihood framework [Engle 1982], and more recently under a Bayesian framework [Kaufmann and Frühwirth-Schnatter 2002], with the help of Markov chain Monte Carlo (MCMC) techniques. Under the Bayesian framework, uncertainty in model parameters is taken into account by finding a posterior distribution over parameters, given the prior assumptions and evidence.

The variational Bayesian framework is an alternative to MCMC methods. Under the variational Bayesian (VB) framework, it is assumed that the posterior distribution over model parameters and latent variables takes on a particular restricted form. In some cases, the optimal functional form can then been found directly by functional maximization [Attias 2000; Ghahramani and Beal 2000; Beal and Ghahramani 2002; Penny and Roberts 2002]. In others, a particular form must be chosen a priori [Attias 1999]. Given the form of the approximate posterior, a lower bound on the log evidence of the data can be computed. The lower bound is maximized with respect to the parameters of the variational approximation. Given the maximizing parameters, approximate predictive distributions can be calculated. Model selection is also possible, by selecting the model which maximizes the lower bound.

Markov chain Monte Carlo techniques have the advantage that they will eventually draw a sample from the true posterior distribution over models. They are also very generally applicable. New lower bounds need not be derived for each model class, although proposal distributions and other aspects of the sampler must be carefully considered.

In contrast, variational approximations produce biased estimates, because of the assumption of a restricted class of posteriors. They can be technically complex to use, because new approximations are needed for each new class of models. However, they can be fit quickly. This is particularly advantageous in the econometrics domain, where many new time series models must be re-fit frequently. In addition, variational approximations are deterministic, and convergence of the variational parameters can be assessed by monitoring the lower bound on the log-evidence. The variational approximation provides a very compact representation of the posterior, in the form of the variational parameters. This is useful when each sample is a large, complex model. Finally, the variational approximation can be used as a principled proposal distribution for importance sampling, which can quickly generate an unbiased estimate of the true predictive density; the Kullback-Leibler divergence between the approximate and true posterior; and the true log-evidence [Ghahramani and Beal 2000].

In this paper we present a variational Bayesian treatment of two ARCH models, normal-ARCH and Student t-ARCH. ARCH models present particular difficulties under the VB framework, because the autoregressive component appears in the noise model. In order to circumvent these difficulties, we present a novel application of Monte Carlo sampling to compute expectations under the variational approximation. We show that the variational Bayesian approximation allows us to select the true model order, and make good predictions of future volatility. We show results on a synthetic test problem, and on a "value-at-risk" problem using historical financial time series data.

2 Autoregressive Conditional Heteroskedastic Models

Consider a time series $y_t, t \in \{1, ..., T\}$ (denoted hereafter $\{y_t\}_{t=1}^T$). Under an ARCH model, the time series is assumed to be generated according to the following dynamics:[1]

$$\begin{aligned} h_t &= a_0 + \sum_{i=1}^{M} \frac{a_i}{1 + y_{t-i}^2} \\ y_t &= c_0 + \sum_{i=1}^{K} c_i y_{t-i} + \frac{\epsilon_t}{\sqrt{h_t}} \end{aligned} \tag{1}$$

For a normal-ARCH model, the noise term ϵ_t is drawn from a zero-mean, unit variance Normal distribution. For a t-ARCH model, the noise term is drawn from a zero-mean, unit variance Student-t distribution with ν degrees of freedom. We will restrict ourselves to zero-mean ARCH models.

3 Variational Bayesian Modeling

Consider a class of models $\mathcal{M}(\theta)$ parameterized by θ. Under a Bayesian framework, the posterior distribution over models is found, given a prior distribution and the evidence:

$$P(\theta|\{y_t\}_{t=1}^T) = \frac{P(\theta; \mathcal{M}) P(\{y_t\}_{t=1}^T|\theta; \mathcal{M})}{\int_{\widehat{\theta}} \partial\widehat{\theta} P(\widehat{\theta}; \mathcal{M}) P(\{y_t\}_{t=1}^T|\widehat{\theta}; \mathcal{M})} \tag{2}$$

We will drop the reference to the model class for brevity. With the posterior distribution in hand, we can compute quantities of interest, such as the predictive distribution:

$$P(y_{T+1}|\{y_t\}_{t=1}^T) = \int_\theta \partial\theta P(\theta|\{y_t\}_{t=1}^T) P(y_{T+1}|\theta) \tag{3}$$

as well as uncertainty on model parameters. Models fit under a Bayesian framework are inherently robust against overfitting and tend to make more accurate predictions. Unfortunately, computation of the posterior is intractable for many models of interest.

In order to overcome this limitation, we can assume that the posterior takes on some restricted form $Q(\theta; \phi)$ parameterized by ϕ. The approximation should be chosen so that expectations and other quantities of interest can be computed. Given this assumption, the log-evidence is lower-bounded using

[1] We differ from the standard formulation of ARCH models in that h_t is the precision rather than the variance of the noise model. This is for mathematical convenience in the formulation of the variational approximation. The addition of 1 in the denominator is to avoid numerical instabilities for near-zero data.

Jensen's inequality:

$$\begin{aligned}
\log P(\{y_t\}_{t=1}^T) &= \log\left(\int_\theta \partial\theta P(\{y_t\}_{t=1}^T|\theta)P(\theta)\right) && (4)\\
&= \log\left(\int_\theta \partial\theta \frac{P(\{y_t\}_{t=1}^T|\theta)P(\theta)Q(\theta;\phi)}{Q(\theta;\phi)}\right) && (5)\\
&\geq \int_\theta \partial\theta Q(\theta;\phi)\log\left(\frac{P(\{y_t\}_{t=1}^T|\theta)P(\theta)}{Q(\theta;\phi)}\right) && (6)\\
&= \log P(\{y_t\}_{t=1}^T) - \mathrm{KL}(Q(\theta;\phi)\|P(\theta|\{y_t\}_{t=1}^T)) && (7)
\end{aligned}$$

Equation (6) is called the negative variational free energy, by analogy to the free energy of statistical physics. The last line shows that the negative free energy lower bounds the log-evidence, with the difference being the Kullback-Leibler divergence between the variational approximation and the true posterior. To find the best approximation, the lower bound is maximized with respect to the variational parameters. Given the variational approximation, we can compute approximate predictive distributions and uncertainty on model parameters. By minimizing the free energy, we can also perform model selection.

4 Variational Bayesian ARCH Models

Consider a zero-mean ARCH model (Eq.(1)). We want to integrate over the parameters a_i. The priors over parameters will be independent Gamma distributions: $a_i \sim Ga(\alpha_i, \beta), i \in \{0, ..., M\}$, where $Ga(\alpha, \beta)$ denotes a Gamma distribution with shape parameter $\alpha > 0$ and inverse scale $\beta > 0$:

$$Ga(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{8}$$

In the above, $\Gamma(\cdot)$ denotes the gamma function.

The parameters are assumed independent in the approximate posterior. The approximate posterior distribution over parameter i is assumed to be Gamma with shape parameter $\widehat{\alpha}_i$ and inverse scale parameter $\widehat{\beta}$.

$$P(\{a_i\}_{i=1}^M|\mathbf{y}) \approx Q(\{a_i\}_{i=1}^M) = \Pi_i Ga(\widehat{\alpha}_i, \widehat{\beta}) \tag{9}$$

Note that all of the prior distributions share a single inverse scale parameter β, and all of the approximate posterior distributions share a single inverse scale parameter $\widehat{\beta}$. We do this for simplicity, but it is not strictly necessary. Gamma distributions were chosen throughout, because they are appropriate for the positive precisions.

For the normal-ARCH model, the variational free energy (ignoring constants) is given by:

$$\begin{aligned}
\mathcal{F} &= -\sum_{i=0}^M \left[\log\frac{\beta^{\alpha_i}}{\Gamma(\alpha_i)} + (\alpha_i - 1)\left\langle \log a_i\right\rangle_Q - \beta\left\langle a_i\right\rangle_Q + H(Q(a_i;\widehat{\alpha}_i,\widehat{\beta}))\right] \\
&\quad -\sum_{t=1}^T \frac{1}{2}\left\langle \log h_t\right\rangle_Q + \left\langle \frac{y_t^2 h_t}{2}\right\rangle_Q
\end{aligned} \tag{10}$$

where $\langle\cdot\rangle_Q$ denotes expectations with respect to the approximate posterior, and $H(Q(a_i;\widehat{\alpha}_i,\widehat{\beta}))$ denotes the entropy of parameter i:

$$H(Q(a_i;\widehat{\alpha}_i,\widehat{\beta})) = \log\Gamma(\widehat{\alpha}_i) - \log\widehat{\beta} + (1-\widehat{\alpha}_i)\psi(\widehat{\alpha}_i) + \widehat{\alpha}_i \tag{11}$$

In the above, $\psi(\cdot)$ denotes the digamma function $\psi(x) = \frac{\partial}{\partial x}\log\Gamma(x)$.

In order to evaluate the variational free energy, we need the following expectations [Beal 2002]:

$$\begin{aligned} \langle a_i \rangle_Q &= \widehat{\alpha}_i / \widehat{\beta} \\ \langle \log a_i \rangle_Q &= \psi(\widehat{\alpha}_i) - \log \widehat{\beta} \\ \left\langle \frac{1}{2} y_t^2 \left(a_0 + \sum_i a_i / (1 + y_{t-i}^2) \right) \right\rangle_Q &= \frac{1}{2} y_t^2 \left(\widehat{\alpha_0} / \widehat{\beta} + \sum_i \widehat{\alpha}_i / [\widehat{\beta}(1 + y_{t-i}^2)] \right) \end{aligned}$$

We must still deal with the final expectation, $\left\langle \log \left(a_0 + \sum_i a_i / (1 + y_{t-i}^2) \right) \right\rangle_Q$. Under the approximate posterior, each term is Gamma distributed: $a_0 \sim Ga(\widehat{\alpha_0}, \widehat{\beta})$, $a_i \sim Ga(\widehat{\alpha}_i, \widehat{\beta}(1 + y_{t-i}^2))$. The latter is problematic, because the sum of Gamma variables remains Gamma only when they share the same (inverse) scale parameter. In our case, the scales change because of the multiplier $(1 + y_{t-i}^2)$. The resulting distribution does not have a simple closed form.

The t-ARCH model is similar, but with a likelihood from a zero-mean Student-t distribution with ν degrees of freedom:

$$t_\nu(y_t; h_t) = \frac{\Gamma((\nu+1)/2)\sqrt{h_t}}{\Gamma(\nu/2)\sqrt{\nu\pi}} \left(1 + \frac{y_t^2 h_t}{\nu} \right)^{-(\nu+1)/2} \tag{12}$$

where h_t is the precision.

The variational free energy is given by:

$$\begin{aligned} \mathcal{F} &= -\sum_{i=0}^{M} \left[\log \frac{\beta_i^\alpha}{\Gamma(\alpha_i)} + (\alpha_i - 1) \langle \log a_i \rangle_Q - \beta \langle a_i \rangle_Q + H(Q(a_i; \widehat{\alpha}_i, \widehat{\beta})) \right] \\ &\quad - \sum_{t=1}^{T} \left[\frac{1}{2} \langle \log h_t \rangle_Q - \left\langle \frac{\nu+1}{2} \log \left(1 + \frac{1}{\nu} y_t^2 h_t \right) \right\rangle_Q \right] \end{aligned} \tag{13}$$

Again, there are problematic terms in the likelihood: $\left\langle \log \left(a_0 + \sum_i a_i / (1 + y_{t-i}^2) \right) \right\rangle_Q$ and $\left\langle \log \left(1 + (y_t^2/\nu) \cdot \left(a_0 + \sum_i a_i / (1 + y_{t-i}^2) \right) \right) \right\rangle_Q$. It may be possible to find analytic lower bounds for these terms. Instead we will evaluate these terms numerically using simple Monte Carlo sampling.

It is not new to combine variational methods with sampling. Variational approximations have been used as proposal distributions for importance sampling [Ghahramani and Beal 2000], and for sophisticated Markov chain Monte Carlo techniques [de Freitas, Højen-Sørensen, Jordan, and Russell 2001]. Here we do the opposite, using simple Monte Carlo approximations to estimate the expectations of difficult terms in the variational free energy. This is very straightforward, because we assume that the posterior is the product of independent, standard distributions (in this case Gamma distributions).

Consider an expectation $\langle f(a_0, ..., a_M) \rangle_Q$ which cannot be evaluated analytically. Generating a set of R samples $\{\widehat{a}_0^j, ..., \widehat{a}_M^j\}_{j=1}^R$ from the approximate posterior (for a given setting of the parameters), we approximate the expectation with $\sum_{j=1}^R f(\widehat{a}_0^j, ..., \widehat{a}_M^j)/R$.

Note that we are not sampling from the true posterior, but from the approximate posterior for some setting of the variational parameters. This makes the sampling very straightforward. Because we assume the posteriors are standard distributions, we can use highly optimized sampling routines. Because we assume that the parameters are independent in the posterior, the sampling is done only in low-dimensional spaces.

By using Monte Carlo sampling in this way, we retain many of the attractive properties of the VB framework: convergence of the variational parameters can still be easily monitored, and the result is a compact representation of the approximate posterior. We also inherit one of the features of Monte Carlo techniques: The approximation becomes very straightforward, with no need to derive and evaluate new bounds for every problematic term in the variational free energy. In addition, this method makes it possible to apply the VB framework to models for which deriving suitable analytic bounds may not be

possible. The free energy will not be deterministic, although with a large enough sample the difference will be minimal. Again, because we only sample from low-dimensional standard distributions, the loss in speed is small.

5 Experimental Results

5.1 Synthetic Data

First, we tested each variational Bayesian ARCH model on synthetic data generated from the same type (normal or Student-t) of ARCH model. The purpose here is to see if the variational free energy can be used to determine the model order, on data for which we know the true model order. Model order $N_{\text{true}} = 3$, and coefficients $\{a_0, ..., a_3\} = \{0.1, 0.5, 0.2, 0.1\}$ were used for both normal-ARCH and t-ARCH models. Each test run consisted of first generating 512 data points from the model, and then fitting a VB model of order N to the data. Fifty test runs were executed for each model type, for values $N = \{0, 1, 2, 3, 4, 5, 6\}$. The priors were set to $\alpha_i = \beta = 1$. For the t-ARCH model, the degrees of freedom was $\nu = 5$. The variational parameters were optimized using a constrained Newton's method.[2] Fitting a model took between a few seconds and approximately five minutes, depending on model order, on an Intel Pentium 1GHz processor.

Figure 1 shows the negative free energy as a function of model order for each model type. For both, the free energy is maximized at $N = 3$. For this set of experiments, the variational free energy correctly indicates the true model order.

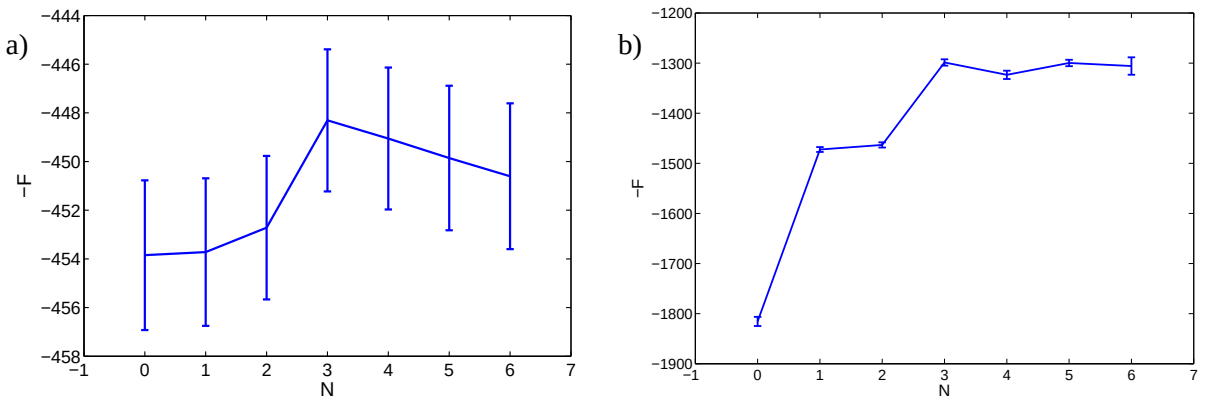

Figure 1: Negative free energy as a function of model order. Both (a) the normal-ARCH and (b) the t-ARCH maximize the free energy at $\widehat{N} = 3$, which is the true model order.

Figure 2 shows a plot of the true coefficients and the variational approximate posterior for one run of each of the two model types.

5.2 Value at Risk

Value at Risk (VaR) estimation is an important problem in econometrics. Financial institutions such as banks are required by law to retain a certain portion of their assets to set against possible investment losses. To do this, banks must compute the maximum value that they might lose on their investments, over a given time window τ, within a given confidence interval ϵ. For example, the Basel Capital Accord [Basel Committee on Banking Supervision 2003] dictates the capital reserves which European banks must keep on hand, as a function of the banks' internal VaR models. If the bank can accurately estimate its VaR, and thus its minimal capital requirements, capital is available to be invested. Overestimating VaR results in opportunity costs, since too much capital is held in reserve. Because of the many different

[2] All experiments were implemented in Matlab. The builtin Matlab constrained function minimizer `fmincon` was used.

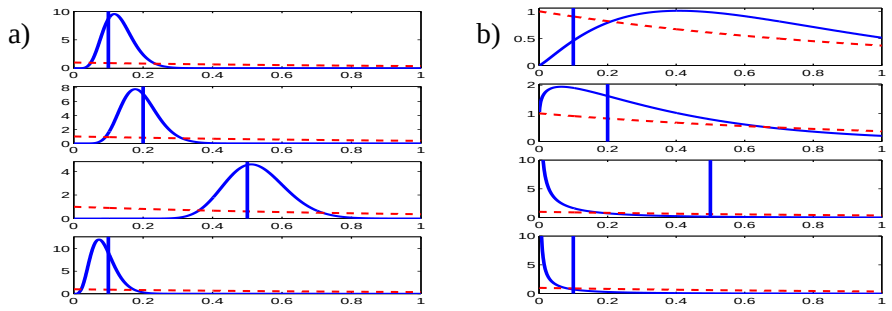

Figure 2: True parameters, priors and posteriors for (a) the normal-ARCH (left column) and (b) the t-ARCH (right column) models. The parameters are shown $\{a_3, a_2, a_1, a_0\}$ from top to bottom. The dashed line shows the prior, and the solid line shows the variational posterior. The vertical line is the true parameter value. In general, the t-ARCH model has greater parameter uncertainty, although the true model order can still be inferred.

investments made by the bank, and the high-frequency of financial data, many VaR models must be fit in a short period of time.

Given the value of an investment p_t at time t, the value at risk is defined as the value V where:

$$P(p_{t+\tau} - p_t \leq -V) \leq 1 - \epsilon \tag{14}$$

In words, the probability that the investor will lose more than V over the next τ time periods is less than $1 - \epsilon$. Typically, ϵ is set to 0.95 or 0.99, and τ is on the order of 10 business days.

The ARCH family of models has been widely used for VaR estimation. Given a model of volatility, and assuming that the investment has zero mean return over the window of interest, we can predict the VaR by computing the variance and ϵ confidence intervals at time $t + \tau$. Given a posterior distribution over model parameters, we can compute the expected VaR.

We used the daily closing price of the Dow Jones industrial index to test the VB ARCH models on VaR estimation. Given the daily closes p_t from to February 11, 1983 to May 14, 2002, we computed the scaled daily log-returns as:

$$r_t = 100 * \log(p_t/p_{t-1}) \tag{15}$$

This resulted in a time series of 4860 points. We then subdivided this into a training set of 3645 points and a test set of 1215 points.

Normal-ARCH and t-ARCH models were fit to the training set for each model order from $N = 0, ..., 20$. By monitoring the negative free energy, it was found that the optimal model order for the normal-ARCH was $N = 12$, and for the t-ARCH was $N = 18$. The variational parameters were optimized using Newton's method. In all cases, the priors were $\alpha_i = \beta = 1$, and for the t-ARCH model, $\nu = 5$.

Figure 3 shows the average absolute error between the absolute returns and the predicted volatility at the 95% confidence value, for the two models at their optimal model order. We also show the error of the same models where the parameters were found using maximum likelihood with a model order selected by the Bayesian Information Criterion [Schwartz 1978]; and where the distribution over model parameters was found by drawing 1000 samples using importance sampling with the variational posterior as a proposal distribution.

For t-ARCH, the variational Bayesian model does as well as the importance sampled model, and both do significantly better than the maximum likelihood model. Note that the maximum likelihood model also had the advantage of being fit at the optimal model order, according to the variational free energy. For the Normal-ARCH models, there is no significant difference between the three models. The normal models do significantly worse than the t-ARCH models. Significance was tested using a

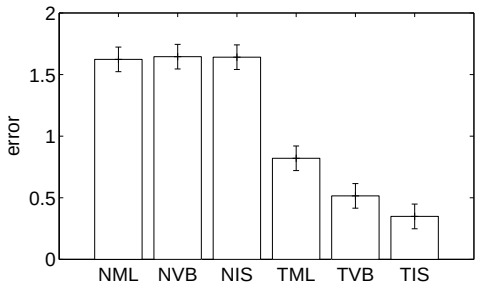

Figure 3: Model performance on predicting the volatility of the Dow Jones industrial average. Average error per time step and 95% confidence intervals are shown. Smaller error is better. The models are labeled as: N==normal-ARCH, T=t-ARCH; ML==maximum likelihood, VB==variational Bayes; IS==importance sampling.

non-parametric Kruskal-Wallis test. Because the test assumes that the samples are independent, we first subsampled the errors at an interval of 20 to remove autocorrelations.

6 Discussion

In our tests, the normal-ARCH model did not benefit greatly from the Bayesian treatment. In contrast, the VB-t-ARCH performed significantly better than the maximum likelihood version. The overall superior performance of the t-ARCH over normal-ARCH models is consistent with other studies of ARCH models on financial data. In general, the t-ARCH model seems to be more problematic to fit, with larger parameter uncertainty. This may explain the better performance of the Bayesian versions over the ML version.

Acknowledgments

The author would like to thank Tatiana Miazhynskaia and Achim Lewandowski for useful discussions. This work was funded by the Austrian Science Fund (FWF) under grant SFB#010: "Adaptive Information Systems and Modeling in Economics and Management Science". The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Education, Science and Culture and by the Austrian Federal Ministry for Transport, Innovation and Technology.

References

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. of Uncertainty in AI (UAI-99)*, pp. 21–30.

Attias, H. (2000). A variational bayesian framework for graphical models. See Solla, Leen, and Müller [2000].

Basel Committee on Banking Supervision (2003, April). The new basel capital accord: Third consultative paper. Technical report, Bank for International Settlements.

Beal, M. (2002). Variational Bayesian quick reference sheet. last revised 01/04/2002, available at `http://www.gatsby.ucl.ac.uk/~beal/papers.html`.

Beal, M. and Z. Ghahramani (2002). The variational bayesian EM algorithm for incomplete data: with application to scoring graphical model. *Bayesian Statistics 7*. to appear.

Bera, A. and M. Higgins (1993). Arch models: Properties, estimation and testing. *Journal of Economic Surveys 7*(4), 307–366.

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

Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance 1*(2), 223–236.

de Freitas, N., P. Højen-Sørensen, M. I. Jordan, and S. Russell (2001). Variational MCMC. In J. Breese and D. Koller (Eds.), *Proc. of Uncertainty in AI (UAI-01)*.

Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica 50*, 987–1006.

Ghahramani, Z. and M. Beal (2000). Variational inference for Bayesian mixtures of factor analysers. See Solla, Leen, and Müller [2000], pp. 449–455.

Kaufmann, S. and S. Frühwirth-Schnatter (2002). Bayesian analysis of switching ARCH-models. *Journal of Time Series Analysis 23*(4), 425–458.

Penny, W. and S. Roberts (2002). Bayesian multivariate autoregressive models with structured priors. *IEEE Proceedings on Vision, Signal and Image Processing 149*(1), 33–41.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Solla, S. A., T. K. Leen, and K.-R. Müller (Eds.) (2000). *Advances in Neural Information Processing Systems*, Volume 12. The MIT Press, Cambridge.