

Model Selection and Averaging in Financial Risk Management

Brian M. Hartman
University of Connecticut

Chris Groendyke
Robert Morris University

June 27, 2013

Abstract

Simulated asset returns are used in many areas of actuarial science. For example, life insurers use them to price annuities, life insurance, and investment guarantees. The quality of those simulations has come under increased scrutiny during the current financial crisis. When simulating the asset price process, properly choosing which model or models to use, and accounting for the uncertainty in that choice, is essential. We investigate how to best choose a model from a flexible set of models. In our regime-switching models the individual regimes are not constrained to be from the same distributional family. Even with larger sample sizes, the standard model-selection methods (AIC, BIC, and DIC) incorrectly identify the models far too often. Rather than trying to identify the best model and limiting the simulation to a single distribution, we show that the simulations can be made more realistic by explicitly modeling the uncertainty in the model-selection process. Specifically, we consider a parallel model-selection method that provides the posterior probabilities of each model being the best, enabling model averaging and providing deeper insights into the relationships between the models. The value of the method is demonstrated through a simulation study, and the method is then applied to total return data from the S&P 500.

Keywords: Asset Simulation, Hidden Markov Models, Latent State Models, GARCH, Stochastic Volatility, Parallel Model Selection.

JEL Classification Codes: C52, C11, C15

1 Introduction

When pricing increasingly more complicated investment guarantees, realistic closed-form solutions for the price are often not available. To estimate the price of the guarantee, the asset value can be simulated multiple times and the price calculated for each simulated stream. The simulated prices form an empirical distribution of the guarantee price. Proper simulation of the asset price is of paramount importance to the accuracy of the guarantee price.

Regime-switching models are rapidly gaining popularity, especially in modeling asset prices. Regime-switching models assume that a discrete process switches between regimes randomly. Each regime is characterized by a different distribution. The process that determines the regime is assumed to be Markovian, i.e., the probabilities of any observation's regime depend only on the regime of the observation immediately prior. In all the current work, the individual regimes are assumed to have the same distributional form and differ only in the parameter values. Added flexibility in the distributional form can have a strong impact, and the need for this additional flexibility might be driven by the underlying economics. For example, in a two-regime model of asset prices, the first regime could occur during a strong economy and the second during a poor economy. Asset prices during a strong economy could be properly modeled using a lognormal distribution, but the prices during a poor economy may need thicker tails—say, an inverse Weibull or generalized Pareto distribution.

In any applied problem including modeling asset prices proper model selection is important. Standard likelihood-based methods, such as the Schwarz-Bayes Criterion (BIC) (Schwarz, 1978) and the Akaike Information Criterion (AIC) (Akaike, 1974), are often used for this task. Unfortunately, these criteria determine only which model is best. If all potential asset streams are simulated from that best model, the simulations do not account for model-selection uncertainty. The simulations implicitly assume that the chosen model is certain to be the only correct model.

Properly applied Bayesian methods give posterior probabilities for each model. Those probabilities allow simulation of the proper proportion of asset streams from each model and account for the model uncertainty. Using the predictive distribution to account for model uncertainty allows for more realistic models and better hedges for variable annuities and other products.

To calculate the posterior model probabilities, the model (and its accompanying regimes) must be treated as a parameter and assumed to be unknown. Assuming that the number of regimes, or even the type of regime, is unknown can be problematic. Robert et al. (2000) were the first to estimate a regime-switching normal model using reversible-jump Markov chain Monte Carlo (RJMCMC) (Green, 1995). They required that the regimes all have a mean of zero. Hardy (2001) and Hardy (2003) showed that the best fit of a regime-switching lognormal model to asset prices (specifically the S&P 500 and TSE 300) has one high-mean, low-variance regime and one low-mean, high-variance regime, implying that constraining the means to be zero, or even equal, is unrealistic.

There is further promise in the machine-learning literature. There are a variety of methods based on the Dirichlet process (e.g. Beal et al., 2002; Teh et al., 2006; Fox et al., 2011). These methods are reviewed and shown to work well in Hartman and Heaton (2011). Unfortunately, these methods require that all the regimes have distributions of the same type. A modeler may want one regime to follow a distribution with fatter tails or more flexibility, while another is more standard. We investigate a method

that can be directly applied to any set of potential models. We focus on various regime-switching models, a GARCH model, and a set of stochastic volatility models.

2 Methodology

2.1 Model-Selection Techniques

2.1.1 AIC, BIC, and DIC

We consider three different likelihood-based criteria that are commonly used as metrics for comparing models. All three seek to find a good-fitting yet parsimonious model by measuring goodness of fit for each model (based on the likelihood) and then imposing a penalty term whose magnitude increases with model complexity. AIC and BIC measure complexities as functions of the number of parameters in the model and set their complexity penalties accordingly. However, it becomes difficult to apply this type of complexity penalty in more complicated hierarchical modeling situations where the number of model parameters cannot be explicitly determined. The deviance information criterion (DIC) was developed by Spiegelhalter et al. (2002) and offers an alternative means of measuring model complexity that can be easily implemented in these situations.

Spiegelhalter et al. (2002) propose that the effective number of parameters in a model (p_D) be calculated as

$$p_D = \overline{D(\theta)} - D(\bar{\theta}),$$

where $D(\theta)$ is defined as the Bayesian deviance:

$$D(\theta) = -2 \log(p(y|\theta)) + 2 \log(f(y)),$$

and $f(y)$ is a function of the data only. We calculate $\overline{D(\theta)} = E^\theta [D(\theta)]$ and $D(\bar{\theta})$ is the Bayesian deviance evaluated at the expectation of θ . The authors note that the intuition for this form for p_D , which is the excess of the mean deviance over the deviance of the means, is analogous to that used in estimating the degrees of freedom for a test. Spiegelhalter et al. (2002) further propose the deviance information criterion as the sum of an estimate of fit and twice the effective number of parameters:

$$\begin{aligned} DIC &= D(\bar{\theta}) + 2p_D \\ &= \overline{D(\theta)} + p_D, \end{aligned} \tag{1}$$

where Equation (1) can be seen to be in the form of a penalized goodness of fit.

One notable advantage of this DIC criterion is that all of the required quantities can easily be

computed using Markov chain Monte Carlo (MCMC) output: $\overline{D(\theta)}$ as the sample average of the $D(\theta)$ values over the samples of θ , and $D(\bar{\theta})$ as D evaluated at the sample average value of θ . When using this criterion to compare models, we will tend to prefer those models having smaller values of DIC.

2.1.2 Parallel Model Selection

Thus far we have mentioned model-comparison procedures that use metrics such as AIC, BIC, or DIC for each individual model, and then we choose the model with the best value of the metric. A different approach considers the various candidate models simultaneously. In a Bayesian setting, model selection could be accomplished by exploring the joint space of models and model parameters; over the course of the simulation, evidence is gathered for the various models, allowing for the comparison of the posterior probabilities of all models under consideration. Several methods for calculating these probabilities have been suggested: RJMCMC and the saturation method (Carlin and Chib, 1995) are two such methods.

Congdon (2006) suggests a different method, which samples from all candidate models separately (in parallel) and then compares the evidence for each at the end. This method differs from the RJMCMC and saturation methods, which accumulate evidence for the various models by jumping between models in the combined model and parameter space. Congdon's method is Bayesian in nature in that it incorporates prior information about the parameters (and potentially about the models as well) into the procedure.

The essential idea of this method is that the parameters θ_j (the model parameters under model j) are indifferent (flat) under all models $k \neq j$ so that $P(\theta_{j \neq k} | M = k) = 1$. Then assuming the independence among θ_j given the model M yields $P(\theta, M = k) = P(\theta_k | M = k)$. If we assume a prior distribution giving equal weight to each model under consideration, we can calculate

$$P(M = k | Y, \theta) = \frac{P(Y | M = k, \theta_k) P(\theta_k | M = k)}{\sum_j P(Y | M = j, \theta_j) P(\theta_j | M = j)} \\ \propto P(Y | M = k, \theta_k) P(\theta_k | M = k),$$

which can be approximated through MCMC samples. $P(Y | M = k, \theta_k)$ is the likelihood of the data given both the model and the current values of the model parameters. The parameters from each iteration of the MCMC chain are used to calculate the joint likelihood. The denominator is simply the sum of those likelihoods from all possible models at that same iteration.

Because the model label (parameter) is given a distribution, this is necessarily a Bayesian method. If a frequentist method is required, a potentially viable alternative is Akaike weights (Burnham and Anderson, 2002). Parallel model selection has been used in actuarial science and finance. Chen et al. (2011) use the method to propose a generalized CAPM relationship and Peters et al. (2009) use it in claims reserving.

Table 1: Models of Interest

Model #	Model Type
1	Regime-switching gamma-gamma
2	Regime-switching gamma-lognormal
3	Regime-switching gamma-Weibull
4	Regime-switching lognormal-lognormal
5	Regime-switching lognormal-Weibull
6	Regime-switching Weibull-Weibull
7	Independent gamma
8	Independent lognormal
9	Independent Weibull

2.2 Models of Interest

For the regime-switching models, we examine three separate distributions: gamma, lognormal, and Weibull. These three component distributions yield varying levels of tail thickness and skewness. Considering a one- or two-regime structure and the three distributions, we have nine total regime-switching models to compare (see table 1).

2.3 Maximum Likelihood Parameter Estimation Using the EM Algorithm

In this study the returns are assumed to come from a regime-switching model. Under a regime-switching model we observe all of the returns, but we do not observe the sequence of regimes underlying the return process. This missing information is a major challenge when performing inference for this type of model. We can deal with this problem by considering the unknown sequence of regimes as missing data, allowing us to apply methods commonly used in missing-data problems.

The expectation-maximization (EM) algorithm (Dempster et al., 1977) is a standard tool for calculating the likelihood function in situations involving missing data; this method works best where it is relatively easy to compute the complete data likelihood (i.e., the likelihood including all data), but more difficult to compute the likelihood of only the existing data. It is an iterative algorithm that alternates between E (expectation) and M (maximization) steps until some convergence criterion is satisfied. In general, the E step consists of taking the expectations of the missing data, given the current parameter estimates, while the M step involves maximizing the complete data likelihood (or perhaps the complete data log-likelihood) with respect to the parameter values.

In Chapter 4 of Zucchini and MacDonald (2009), a general method is described for applying the EM algorithm to a regime-switching (hidden Markov) model. In this case, the E step consists of estimating

the conditional probabilities of being in particular regimes and transitioning to other regimes, based on the observations and the current parameter estimates. For the M step, the goal is to maximize the complete data likelihood with respect to (1) the initial distribution, (2) the transition probability matrix, and (3) the parameters of the underlying distribution. The details are contained in the appendix.

2.4 Bayesian Estimation Using MCMC

When estimating regime-switching models from a Bayesian perspective, the Gibbs sampler (Gelfand and Smith, 1990) fits the model by drawing each parameter sequentially from its full conditional distribution. The full conditional distribution is the posterior distribution of each parameter, holding all the other parameters fixed. Additionally, using data augmentation, the state vector is sampled directly and included as a parameter. With an estimate of the state vector, the observations in a given regime are simply independent and identically distributed observations. Then we are able to use the standard estimation methods for the parameters of that distribution.

Originally, we used the conjugate prior distributions where we could (normal-gamma for the lognormal distribution and gamma for the scale parameter of the gamma distribution), but we quickly found that the model-selection techniques and the parameter estimates were highly sensitive to the hyperparameter selections. The results were more robust when uniform priors and Metropolis-Hastings (Hastings, 1970; Metropolis et al., 1953) steps were used. Please see the appendix for further details.

3 Simulation Study

The model-selection methods in section 2.1 are compared in this section. As the sample size becomes larger, all the model-selection methods should correctly identify the distributions more often. Ideally, the methods should be consistent (meaning that as the sample size increases to infinity, the proportion of correctly identified models tends to one).

First, the models were fitted to monthly total return data from the S&P 500 from February 1956 to October 2010 (657 observations) using the EM algorithm described in section 2.3. The data were obtained from Yahoo! finance (Yahoo! Inc., 2010). The resulting parameter estimates are presented in Table 2.

All but one of the simulations exhibit state persistence, the tendency for a model to remain in its current state rather than switch to another state. More precisely, if p_{ij} is the probability of moving from state i to state j then a model would exhibit state persistence if $p_{ii} > p_{ij} \forall i, j \neq i$. This property is commonly seen in regime-switching models.

In all regime-switching cases, the majority of the data fall into one regime, and the second regime

Table 2: Maximum likelihood parameter estimates

Model	Regime 1	Regime 2	Transition Matrix
1	GA(980.1,967.5)	GA(267.8,268.6)	$\begin{bmatrix} 0.954 & 0.046 \\ 0.119 & 0.881 \end{bmatrix}$
2	GA(267.3,268.2)	LN(0.013,0.001)	$\begin{bmatrix} 0.878 & 0.122 \\ 0.046 & 0.954 \end{bmatrix}$
3	GA(996.1,985.8)	WB(19.1,1.03)	$\begin{bmatrix} 0.966 & 0.034 \\ 0.069 & 0.931 \end{bmatrix}$
4	LN(0.013,0.001)	LN(-0.008,0.004)	$\begin{bmatrix} 0.955 & 0.045 \\ 0.143 & 0.857 \end{bmatrix}$
5	LN(0.001,0.001)	WB(19.16,1.03)	$\begin{bmatrix} 0.966 & 0.034 \\ 0.068 & 0.932 \end{bmatrix}$
6	WB(32.66,1.02)	WB(22.11,1.07)	$\begin{bmatrix} 0.857 & 0.143 \\ 0.905 & 0.095 \end{bmatrix}$
7	GA(544.5,539.9)	-	-
8	LN(0.008,0.002)	-	-
9	WB(19.12,1.03)	-	-

Table 3: Comparison of Gamma and Lognormal Distributions

Distribution	Mean	Mode	Variance	Skewness	Kurtosis
GA(980.1, 967.5)	1.0130	1.0120	0.0010	0.0639	0.0061
LN(0.013,0.001)	1.0136	1.0121	0.0010	0.0949	0.0160
GA(267.8, 268.6)	0.9970	0.9933	0.0037	0.1222	0.0224
LN(-0.008, 0.004)	0.9940	0.9881	0.0040	0.1902	0.0644

helps to model the tails. Model 6, the Weibull-Weibull model, is the only model where the rarer regime does not significantly help model the left tail. In that model the first regime solidly describes the lower tail of the distribution and the second regime covers the unusual, large returns (greater than about 8% in one month). When a single Weibull distribution is used to model the entire distribution of returns, it performs rather poorly. The independent gamma and lognormal distributions fit better, but not with anywhere near the quality of the regime-switching fits. The three regime-switching models that include only the gamma and lognormal distributions (1: GA-GA, 2: GA-LN, and 4: LN-LN) all look very similar. This is true throughout our study. The model-selection methods struggled to distinguish between the gamma and lognormal distributions, making the calculation of posterior model probabilities all the more important. When the gamma parameters are large, the gamma and lognormal distributions are similar. Table 3 compares a few summary statistics of the fitted distributions in our study. The distributions are especially close in the first two moments and differ slightly beyond that, to be expected with two-parameter models fit to essentially the same data.

For our simulation study, we examined six sample sizes: 20, 50, 100, 200, 500, and 1000. For each model, 120 data sets of each sample size are simulated using the S&P 500 parameter estimates from the previous section. Generating the data in this manner provides two benefits: (1) because the parameter estimates come from fits to the same data, the generated data sets should be relatively similar, making the model-selection task more difficult, and (2) using the stock index data puts the simulation study in a practical and applicable context. AIC and BIC are then calculated using the EM algorithm and DIC and parallel model probabilities are calculated from the MCMC chain. Even though these criteria are calculated using different methods, they can still be compared because they are all trying to choose which model best describes the data.

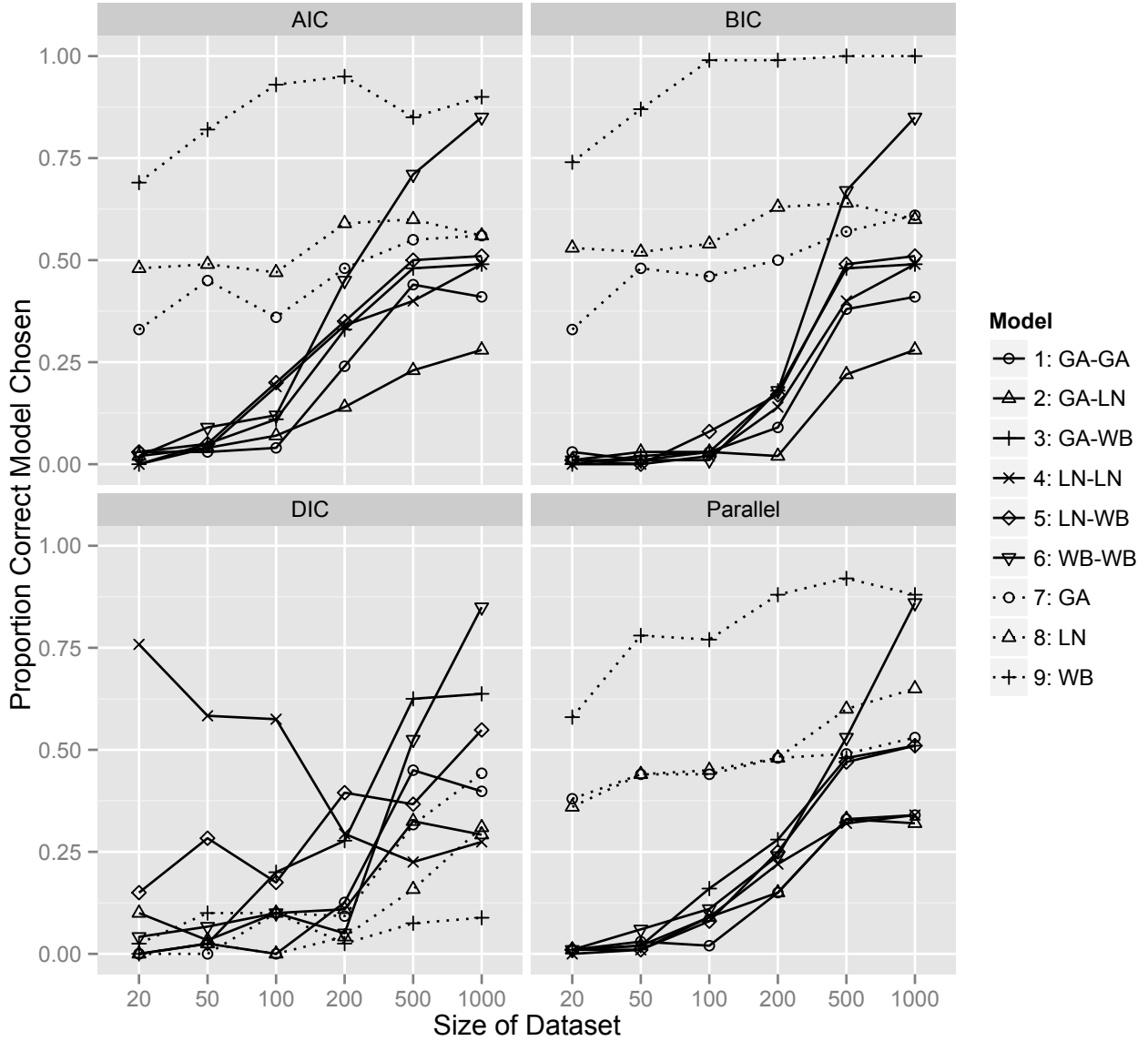
For AIC, BIC, and DIC the proportions of correctly identified data sets are presented in figure 1. Parallel model selection provides probabilities of each model being the correct model. To compare this method directly to the previous three methods, figure 1 also presents the posterior model probabilities for the correct model.

For all the models, the proportion of correct identifications by AIC increases as the sample size increases. All of the independent models (7, 8, and 9) are much easier to identify at the smaller sample sizes. More complicated models will require more data before the improved fit compensates for the complexity penalty. Model 9 (independent Weibull) performs especially well, likely due to the large difference between that model and the other eight.

Because the data sets are common among the model-selection techniques, the likelihood portion of the metric will be the same for both AIC and BIC. However, the penalty term will be different. For all of the sample sizes in our study $\log(n) > 2$, so BIC will penalize the two-regime models more than AIC. That is apparent when looking at the correctly identified proportions. For the three one-regime models, the proportions for BIC are all greater than or equal to the proportions for AIC. The opposite is true for the two-regime models.

DIC was the worst performer. It had an especially hard time identifying model 9, where (with $n = 1000$) it identified 58% of the data sets coming from model 5 and another 33% coming from model 6 (both regime-switching models with the Weibull distribution as at least one of the regimes). The DIC results for model 4 are also very interesting. The proportion correctly identified is high for the small sample sizes and decreases as the sample size increases. This is again due to the similarity between the three models with gamma and lognormal regimes. As the sample size increases, the proportion of correctly identified data sets tends to $1/3$ for models 1, 2, and 4, implying difficulty in discerning between the three models with gamma and lognormal regimes. Additionally, when the sample size is small, model 4 is preferred regardless of whether the data were generated using model 1, 2, or 4 (selected about 80% of the time in all three).

Figure 1: Correctly identified data sets



Note: Proportion of correctly identified data sets (for the AIC, BIC, and DIC model-selection criteria) and posterior probabilities of identifying the correct model (for the parallel model-selection technique). For each model-selection criterion and model, results are presented as a function of sample size. Solid lines represent the regime-switching models (models 1-6), while dotted lines indicate the iid models (models 7-9).

Like the proportions of AIC, BIC, and DIC, the probabilities of the correct model found through parallel model selection grow slowly toward one.

While all of the proportions and probabilities for the different methods are increasing (outside of

Monte Carlo error), only models 6 and 9 approach one with any speed. With a sample size of 1000, only those two models have proportions greater than 0.65. Less than two out of three is not good enough when business decisions will be based on the results. One of the strengths of the parallel model-selection procedure is that it provides probabilities for each model, and examining those probabilities provides a more comprehensive picture of the strengths and weaknesses of the model-selection process. Table 4 provides the posterior model probabilities when the sample size is equal to 1000. One theme is immediately apparent: the technique (as was the case with AIC, BIC, and DIC) has difficulty differentiating between the gamma and lognormal models. For example, in models 1, 2, and 4 nearly all the probability is evenly spread between those three models. The sampler is sure that the model is regime-switching, but it cannot tell whether the first regime is gamma or lognormal, nor whether the second regime is gamma or lognormal. Similarly, for models 3 and 5 one of the regimes is definitely Weibull, but it is difficult to determine whether the other regime is gamma or lognormal. Finally, the independent lognormal and gamma models are hard to differentiate (models 7 and 8). Without gamma or lognormal elements, the sampler performs very well, giving 0.85 probability to the correct regime-switching WB-WB model and 0.99 probability to the correct independent Weibull model. For all models, parallel model selection does a good job determining whether the model has one or two regimes. This was also true of AIC, BIC, and DIC.

Table 4: Posterior model probabilities using parallel model selection, N = 1000

		Chosen Model								
		1	2	3	4	5	6	7	8	9
True Model	1: GA-GA	0.33	0.31	0.01	0.34	0.01	-	-	-	-
	2: GA-LN	0.31	0.32	0.01	0.36	0.01	-	-	-	-
	3: GA-WB	-	-	0.51	-	0.49	-	-	-	-
	4: LN-LN	0.32	0.32	-	0.36	-	-	-	-	-
	5: LN-WB	-	-	0.49	-	0.51	-	-	-	-
	6: WB-WB	0.03	0.03	0.02	0.03	0.03	0.85	-	-	-
	7: GA	-	-	-	-	-	-	0.50	0.50	-
	8: LN	-	-	-	-	-	-	0.31	0.69	-
	9: WB	-	-	0.01	-	-	0.01	-	-	0.99

In addition to the improved understanding of the data, these probabilities can be used for model averaging. For example, assume the data set of interest (say asset returns) provided the same probabilities as model 1 in table 4. To simulate future asset streams, 33% can be drawn from model 1, 32% from model 2, 1% from model 3, 36% from model 4, and 1% from model 5. This will allow the simulated returns to better account for the uncertainty inherent in the model choice. Thus, one of the main advantages to using the parallel model-selection technique is that, unlike AIC, BIC, or DIC, it provides probabilities

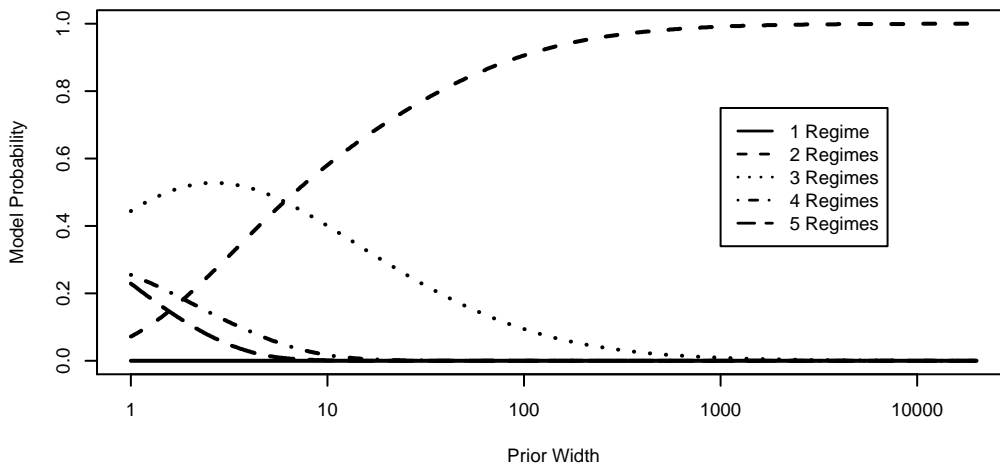
for use in model averaging.

4 Applications

4.1 Number of Regimes in an RSLN Model

Parallel model selection can be used to select the number of regimes in an RSLN model. Parallel model selection penalizes complex models by including the prior density of the parameters. As long as that density is smaller than one, more parameters decrease the model probability, assuming that the likelihood stays the same. In our case, we assume the prior distributions are uniform with length l implying a prior density of $1/l$. The wider the prior distributions, the simpler the preferred model. If $l < 1$ then instead of a complexity penalty, there is a complexity premium which does not make much sense. If $l = 1$ then the model choice is indifferent to complexity. Because the models are nested, when $l \leq 1$ the highest number of regimes will be selected, though Monte Carlo error and limited benefit to additional regimes can make its probability less than one. In the case of our S&P 500 data, the model probabilities are plotted against the prior width in figure 2. While we checked values of the prior width all the way up to 20,000, the parameters are means and variances of a lognormal distribution for monthly stock returns; they do not need to be terribly wide. Notice that for any reasonable prior width either the 2- or 3-regime RSLN model is preferred, confirming the results in Hardy (2001) and Hartman and Heaton (2011). The model with a single regime is never preferred, even with a large complexity penalty (wide priors).

Figure 2: Posterior Probability of the Number of Regimes in an RSLN Model



4.2 Model Selection and Averaging

To show the importance of posterior model probabilities in model selection, we continue to examine the set of monthly total return data from the S&P 500. While the simulation study focused on regime-switching models, to truly apply the parallel model selection to financial data we focus on the LN and LN-LN models and add several stochastic volatility models and a GARCH model.

4.2.1 GARCH Model

One class of model commonly used to describe or simulate financial time series is the GARCH (generalized autoregressive conditionally heteroskedastic) model (Bollerslev, 1986). The GARCH(p, q) model assumes that the variance h_t of a data point y_t is a function (in particular, an ARMA process) of the previous q data points and previous p variances, specifically

$$h_t = \alpha_0 + \sum_{i=1}^q \alpha_i y_{t-i}^2 + \sum_{j=1}^p \beta_j h_{t-j}$$

We used the GARCH(1,1) model to fit the variance of the log-returns from the S&P 500 data set. This model has the advantage of being both parsimonious and flexible, making it suitable to model financial data, particularly when heteroskedasticity is suspected or observed. Following Ardia and Hoogerheide (2010), we represent the log returns y_t under the GARCH(1,1) model with Student- t innovations as

$$y_t = \epsilon_t \left(\frac{\nu - 2}{\nu} \varpi_t h_t \right)^{1/2}$$

where ϖ_t has an inverse gamma distribution with parameters $(\frac{\nu}{2}, \frac{\nu}{2})$ and ϵ_t has a standard normal distribution (see Geweke (1993) for details). We used the R package **bayesGARCH** (Ardia, Ardia; R Core Team, 2012) to fit this model to the data. For α_0, α_1 , and β_1 , we employed diffuse normal priors, truncated to \mathbb{R}^+ ; for ν , a translated exponential prior distribution was used.

4.2.2 Stochastic Volatility Models

In its most basic form, log returns follow a normal distribution with log-variance following a first-order auto-regressive model, formally (Kim et al., 1998)

$$y_i | \tau_i \sim N(0, \exp\{\tau_i\}) \tag{2}$$

$$\tau_i | \tau_{i-1} \sim N(\alpha + \beta \tau_{i-1}, \omega^2) \tag{3}$$

$$(\alpha, \beta | \omega^2) \sim N(b_0, \omega^2 B_0) \tag{4}$$

$$\omega^2 \sim IG(c_0, d_0). \tag{5}$$

In this case, b_0, B_0, c_0 , and d_0 are all specified hyperparameters. Alternatively, following Lopes and Tsay (2011) the log-returns could follow a t-distribution with ν degrees of freedom, location 0, and scale $\exp\{\tau_i\}$

$$y_i|\tau_i \sim t_\nu(0, \exp\{\tau_i\}). \tag{6}$$

4.2.3 Comparing SV Models

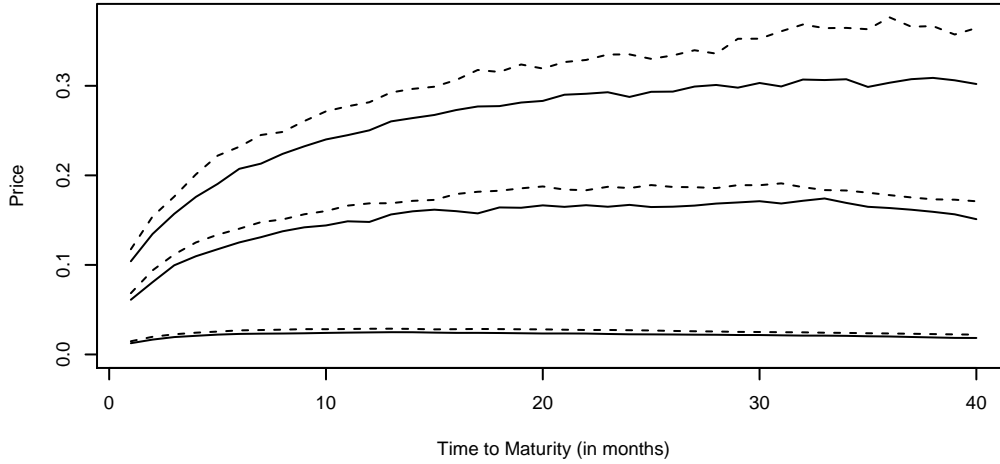
As an example, we use parallel model selection to compare the many stochastic volatility models (either normal or t with ν degrees of freedom). The results are contained in table 5. Using the stochastic volatility

Table 5: Model-Selection Statistics for S&P 500 Data

Distribution	Probability	Distribution	Probability
Normal	0.001	t(16)	0.054
t(2)	0.000	t(17)	0.025
t(3)	0.004	t(18)	0.003
t(4)	0.064	t(19)	0.016
t(5)	0.110	t(20)	0.013
t(6)	0.095	t(21)	0.012
t(7)	0.053	t(22)	0.009
t(8)	0.080	t(23)	0.033
t(9)	0.063	t(24)	0.009
t(10)	0.114	t(25)	0.010
t(11)	0.041	t(26)	0.012
t(12)	0.050	t(27)	0.002
t(13)	0.050	t(28)	0.004
t(14)	0.021	t(29)	0.003
t(15)	0.048	t(30)	0.001

models, we will price a return-of-premium option of an investment product following the S&P 500. This is financially equivalent to a put option with a strike price equal to the original investment. We chose a simple option to focus on the impact of model selection, but the method can be easily extended to more complicated products. Again for simplicity, we assume an interest rate of zero and an initial investment of one. Under AIC, BIC, and DIC, only the t(5) model would be chosen and all return paths would be simulated from that distribution. Using parallel model selection, simulations use different distributions according to the computed probabilities. The impact on the price and risk management can be large. Figure 3 plots the mean price and the 0.95 and 0.99 VaR for this option. The solid lines are from the t(5) model and the dashed lines are from the averaged model.

Figure 3: Comparison of the cost of a return-of-premium option



The averaged model has a very similar mean, but the tail risk measures are greatly affected by the choice of model. Using only the $t(5)$ model may lead to inadequate capital.

4.2.4 Modeling Stock Returns

More generally, one can compare any number of models using parallel model selection. We now combine all the models discussed in this paper (RSLN, SV, and GARCH) to fit the same option as in the SV example. The model selection results are displayed in table 6. Similar to the stochastic volatility example,

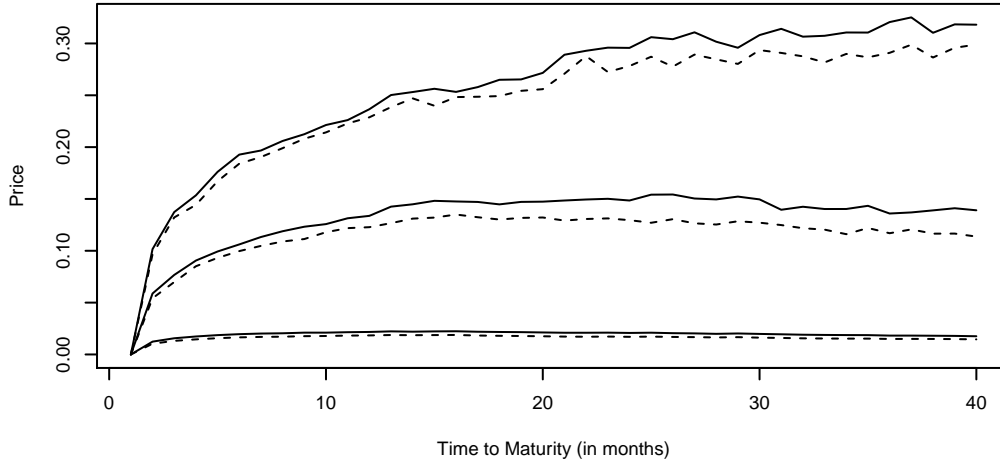
Table 6: Model-Selection Statistics for S&P 500 Data

Distribution	Probability
LN	0.000
RSLN-2	0.004
RSLN-3	0.065
RSLN-4	0.003
RSLN-5	0.018
GARCH	0.910
SV (All)	0.000

under AIC, BIC, and DIC, only the GARCH model would be used. Averaging the models affects the prices and risk management of the product (see figure 4). Again, the solid lines are the single model (only GARCH) and the dashed lines are the averaged model.

In this case the RSLN models have thinner tails than the GARCH model. That is why the averaged

Figure 4: Comparison of the cost of a return-of-premium option (all models)



model has smaller risk measures than the GARCH model. In the previous example, the single model required too little risk capital. Conversely, this one requires too much. Either way, it is important to properly account for the model uncertainty.

5 Conclusion

Fully understanding and accounting for model uncertainty is essential when modeling or simulating asset returns, claims experience, or any other business process. Standard methods of model selection (AIC, BIC, and DIC) determine which model is best and give only a rough idea about how close the other models are to the best one. That rough idea is not enough to decide how to use the other models when making decisions. When one model is dramatically better than the others, only knowing the best model will be sufficient. Far too often, the potential models are very similar in their fit. In that case, a simulation should account for that model uncertainty by drawing a proportion of the simulations from each of the models that fit the data well. Under the standard methods, the proper proportions are unknown.

Parallel model selection provides the posterior probabilities for each model being the best. This method is easier to implement than RJMCMC and more flexible than methods based on the Dirichlet process. A simulation that draws samples from each model according to the posterior probabilities will properly account for the model uncertainty implicit in any modeling problem. This was readily apparent in the analysis of the S&P data, where many of the model probabilities were similar. That analysis also

showed that failing to account for the model uncertainty underestimates the downside risk, exposing the writer to more risk than accounted.

6 Acknowledgements

This work was supported by a generous grant from The Actuarial Foundation. The authors would like to thank an anonymous reviewer, whose comments and suggestions greatly increased the quality of this paper. The authors would also like to thank the attendees at the Statistical Society of Canada Annual Meeting in Guelph, the Actuarial Research Conference in Winnipeg, the Montreal Seminar of Actuarial and Financial Mathematics, and the statistics colloquium at Brigham Young University for their insightful comments and questions, namely Paul Marriott, Daniel Alai, Jed Frees, Saeed Ahmadi, and Mary Hardy.

A Estimation Methods

A.1 Maximum Likelihood Estimation Using the EM Algorithm

If the state vector is known, regime-switching models have a straightforward likelihood. Because in reality the state vector is unknown, it can be treated as missing data and estimated using the EM algorithm. In the E step, we calculate the conditional expectation of the state vector given all the regime-specific parameters and the transition matrix. In the M step we maximize the likelihood with respect to the regime-specific parameters and the transition matrix, assuming the conditional probabilities calculated in the E step.

In order to describe the details of each step, we first define a few terms. The transition probability matrix is π and the individual probability of moving from regime j to regime k is defined as p_{jk} . The density of the observation y_i , given it is in regime r , is denoted $f_r(y_i)$. The densities from both regimes are put into a matrix $\mathbf{P}(y_i)$ as

$$\mathbf{P}(y_i) = \begin{bmatrix} f_1(y_i) & 0 \\ 0 & f_2(y_i) \end{bmatrix}.$$

Using that matrix, the forward probabilities are defined as

$$\alpha_i = \nu \mathbf{P}(y_i) \prod_{s=2}^i \pi \mathbf{P}(y_s)$$

where π is the transition matrix and ν is the stationary transition probability vector ($\nu\pi = \nu$). α_i will have as many elements as there are regimes. The j^{th} element, $\alpha_i(j)$, is a joint probability, $Pr(Y_1 =$

$y_1, Y_2 = y_2, \dots, Y_i = y_i, X_i = j$). Additionally, the backward probabilities are defined as

$$\beta_i = \begin{cases} \left(\prod_{s=i+1}^N \pi \mathbf{P}(y_s) \right) \mathbf{1}^T & \text{if } i < N \\ \mathbf{1} & \text{if } i = N. \end{cases}$$

If the state of each individual observation is known, the log-likelihood can be written as

$$\begin{aligned} \log(\Pr(\mathbf{y}, \mathbf{x})) &= \log \left(\nu_{x_1} \prod_{i=2}^N p_{x_{i-1}x_i} \prod_{i=1}^N f_{x_i}(y_i) \right) \\ &= \log(\nu_{x_1}) + \sum_{i=2}^N \log(p_{x_{i-1}x_i}) + \sum_{i=1}^N \log(f_{x_i}(y_i)). \end{aligned}$$

Define two indicator functions as $u_j(i) = 1\{x_i = j\}$ and $v_{jk}(i) = 1\{x_{i-1} = j, x_i = k\}$, then

$$\log(\Pr(\mathbf{y}, \mathbf{x})) = \sum_{r=1}^R u_r(1) \log(\nu_{x_1}) + \sum_{j=1}^R \sum_{k=1}^R \left[\sum_{i=2}^N v_{jk}(i) \log(p_{x_{i-1}x_i}) \right] + \sum_{r=1}^R \sum_{i=1}^N u_r(i) \log(f_{x_i}(y_i))$$

For the E step, we replace the two indicator functions with their expectations.

$$\begin{aligned} \hat{u}_j(i) &= \Pr(x_i = j | \mathbf{y}) = \frac{\alpha_i(j) \beta_i(j)}{\sum_{r=1}^R \alpha_i(r) \beta_i(r)} \\ \hat{v}_{jk}(i) &= \Pr(x_{i-1} = j, x_i = k | \mathbf{y}) = \frac{\alpha_{i-1}(j) p_{jk} f_k(y_i) \beta_i(k)}{\sum_{r=1}^R \alpha_i(r) \beta_i(r)} \end{aligned}$$

For the M step, we maximize the log-likelihood with the two indicator functions replaced by their expectations. This maximization can be done in two steps. The first two terms are only a function of the transition probability matrix. Because the stationary distribution is a function of the transition probability matrix, those terms need to be maximized numerically. The third term only depends upon the regime-specific parameters. The estimates for the lognormal distribution have the following forms:

$$\begin{aligned} \hat{\mu}_j &= \frac{\sum_{i=1}^N \hat{u}_j(i) \log(y_i)}{\sum_{i=1}^N \hat{u}_j(i)} \\ \hat{\sigma}_j^2 &= \frac{\sum_{i=1}^N \hat{u}_j(i) (\log(y_i) - \hat{\mu}_j)^2}{\sum_{i=1}^N \hat{u}_j(i)} \end{aligned}$$

The parameters of both the Weibull and the Gamma distributions will need to be estimated numerically with each observation weighted by its $u_j(i)$ term.

A.2 Bayesian Estimation Algorithm

The Bayesian estimation algorithm is very similar to the EM algorithm. The prior distribution over the model space is uniform, implying that all models are equally likely *a priori*. Additionally, the prior

distribution of the individual state assignments is also uniform, implying the same ignorance about each observation's regime. The prior distributions for the gamma, lognormal, and Weibull distributions are not as straightforward. The choice of the prior can have a large effect on the performance of both DIC and the parallel model selection. We first chose conjugate priors when we could (gamma for a gamma parameter, normal-inverse gamma for the lognormal parameters, and inverse gamma for a Weibull parameter). Under those priors, DIC did not perform well. The model selected depended almost entirely on the hyperparameters, not on the actual data. We then used a uniform prior for all parameters in each model. While that choice requires Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) steps because the priors are no longer conjugate, DIC performed much better. When a prior is conjugate, the full conditional distribution of the parameter is available in a known distributional form. Without the conjugacy, the posterior distribution is only known to a proportionality constant. As such, the parameters must be updated by first proposing a new parameter value and then dividing its posterior density by the density of the current parameter value. In that way, the proportionality constants will cancel. The ratio becomes the acceptance probability of the proposed value. Each row of the transition matrix is given a Dirichlet prior ($\pi_r \sim Dir(1, 1)$). We did not assume a preference for state persistence, but that is possible through this prior distribution.

The MCMC algorithm includes the following steps:

1. Initialize all parameters. We randomly assigned each observation to a regime and then calculated the maximum likelihood estimates of the regime-specific parameters and the transition matrix.
2. Draw the state vector (x_j) one randomly selected observation at a time from the following equation:

$$\Pr(x_i | \boldsymbol{\theta}, \mathbf{y}, \mathbf{x}_{1:i-1}, \mathbf{x}_{i+1:n})$$

which reduces through the Markov property to

$$\Pr(x_i | \boldsymbol{\theta}, y_i, x_{i-1}, x_{i+1}) \propto \begin{cases} \nu_{x_1} p_{x_1, x_2} f_{x_1}(y_1) & \text{if } i = 1 \\ p_{x_{i-1}, x_i} p_{x_i, x_{i+1}} f_{x_i}(y_i) & \text{if } 1 < i < N \\ p_{x_{N-1}, x_N} f_{x_N}(y_N) & \text{if } i = N \end{cases}$$

3. Draw each row of the transition probability matrix from

$$\pi_r \sim Dir(1 + n_{r1}, 1 + n_{r2})$$

where $n_{jk} = \sum_{i=2}^N v_{jk}(i)$.

4. Draw the regime-specific parameters using only the observations assigned to that regime. Because the prior distributions are uniform, the posterior distributions are proportional to the individual likelihood functions. If no observations were assigned to the regime, draw the parameters using the entire sample.
5. Continue steps 2-4 until convergence. Discard those observations and then continue steps 2-4 until a strong picture of the posterior distributions emerges.

References

- Akaike, H. (1974). A new look at the statistical identification model. IEEE transactions on Automatic Control 19(6), 716–723.
- Ardia, D. bayesgarch: Bayesian estimation of the garch (1, 1) model with student-t innovations in r, 2007. URL <http://CRAN.R-project.org/package=bayesGARCH>.
- Ardia, D. and L. Hoogerheide (2010). Bayesian estimation of the garch (1, 1) model with student-t innovations. The R Journal 2(2), 41–47.
- Beal, M., Z. Ghahramani, and C. Rasmussen (2002). The infinite hidden Markov model. Advances in Neural Information Processing Systems 1, 577–584.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. Journal of econometrics 31(3), 307–327.
- Burnham, K. P. and D. R. Anderson (2002). Model selection and multi-model inference: a practical information-theoretic approach. Springer Verlag.
- Carlin, B. and S. Chib (1995). Bayesian model choice via markov chain monte carlo methods. Journal of the Royal Statistical Society. Series B (Methodological), 473–484.
- Chen, C. W., R. H. Gerlach, and A. M. Lin (2011). Multi-regime nonlinear capital asset pricing models. Quantitative Finance 11(9), 1421–1438.
- Congdon, P. (2006). Bayesian model choice based on monte carlo estimates of posterior model probabilities. Computational statistics & data analysis 50(2), 346–357.
- Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1–38.
- Fox, E., E. Sudderth, M. Jordan, and A. Willsky (2011). A Sticky HDP-HMM with Application to Speaker Diarization. Annals of Applied Statistics.
- Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85(410), 398–409.
- Geweke, J. (1993). Bayesian treatment of the independent student-t linear model. Journal of Applied Econometrics 8(S1), S19–S40.
- Green, P. J. (1995, December). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82(4), 711–732.
- Hardy, M. (2001). A regime-switching model of long-term stock returns. North American Actuarial Journal 5(2), 41–53.
- Hardy, M. (2003). Investment Guarantees: Modeling and Risk Management for Equity Linked Life Insurance. John Wiley and Sons.

- Hartman, B. M. and M. J. Heaton (2011). Accounting for regime and parameter uncertainty in regime-switching models. Insurance: Mathematics and Economics 49(3), 429 – 437.
- Hastings, W. K. (1970, April). Monte Carlo methods using Markov chains and their applications. Biometrika 57(1), 97–109.
- Kim, S., N. Shephard, and S. Chib (1998). Stochastic volatility: likelihood inference and comparison with arch models. The Review of Economic Studies 65(3), 361–393.
- Lopes, H. F. and R. S. Tsay (2011). Particle filters and bayesian inference in financial econometrics. Journal of Forecasting 30(1), 168–209.
- Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. Journal of Chemical Physics 21, 1087–1091.
- Peters, G. W., P. V. Shevchenko, and M. V. Wüthrich (2009). Model uncertainty in claims reserving within tweedie’s compound poisson models. arXiv preprint arXiv:0904.1483.
- R Core Team (2012). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- Robert, C., T. Ryden, and D. Titterton (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62(1), 57–75.
- Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics 6(2), 461–464.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van der Linde (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64(4), 583–639.
- Teh, Y., M. Jordan, M. Beal, and D. Blei (2006). Hierarchical dirichlet processes. Journal of the American Statistical Association 101(476), 1566–1581.
- Yahoo! Inc. (2010, December). Yahoo! finance.
- Zucchini, W. and I. MacDonald (2009). Hidden Markov models for time series: an introduction using R, Volume 110. Chapman & Hall/CRC.