Validation of long term equity return models for equity-linked guarantees

Mary R Hardy[*]

R Keith Freeland[†]

Matthew C Till[‡]

Abstract

Since the Canadian Institute of Actuaries and American Academy of Actuaries are now requiring stochastic asset liability modelling for equity-linked life insurance guarantees, a number of models have been proposed for the equity return process. In this paper we present some that have become well known, and discuss the use of residuals to test the fit. After showing that the use of the static, 'actuarial approach' to risk management can result in two models with very similar likelihood and residuals can give very different capital requirements, we devise an extension of the bootstrap to compare all the models, and to determine whether the optimistic or pessimistic view of the long term left tail risk is more consistent with the data.

[*]Mary Hardy is the CIBC Professor of Financial Risk Management, in the Statistics and Actuarial Science Dept of the University of Waterloo, Ontario, Canada.

[†]Keith Freeland is an Assistant Professor in the Statistics and Actuarial Science Dept of the University of Waterloo, Ontario, Canada.

[‡]Matthew Till is a graduate student in the Statistics and Actuarial Science Dept of the University of Waterloo, Ontario, Canada.

1 Introduction

In 2001 the Canadian Institute of Actuaries Task Force on Segregated Funds (CIA(2001)) proposed a stochastic methodology for the calculation of reserves and capital requirements for Segregated Fund contracts. In 2005, the American Academy of Actuaries has followed suit, with its C3 Phase 2 report (AAA, 2005) for variable annuity contracts.

The stochastic valuation of equity-linked insurance, including segregated funds and variable annuities, requires a model for equity returns, which is used to generate a distribution for the future liability under the contract. In both the CIA and AAA reports, the precise nature of the equity return model was not mandated. Instead, actuaries are given the freedom to use any model, provided it can be calibrated to give distribution tails that are sufficiently fat – where 'sufficiently' is defined in the calibration requirements of the respective reports.

At the time that the CIA task force began its deliberations, there were very few models in common use. The lognormal model, the Wilkie model (Wilkie, 1995) and an empirical approach were cited in a survey. One company used a log-stable model.

Since the CIA report was published a plethora of models has been proposed for segregated fund and variable annuity risk management purposes. We will show in this paper that different models that look very similar may give very different results in practice. The objective of this paper is to identify why the differences occur, and to look more closely at the fit of these models to the historical data.

It should be emphasized perhaps that we are dealing entirely in the realm of P-measure, that is, the real world measure. If the risk management of the equity-linked insurance uses financial engineering, a risk neutral measure will be required to determine the hedge strategy. However, there is sufficient un-hedgeable risk involved in these contracts that additional real-world modelling is needed to estimate the distribution of liabilities after taking the hedging strategy into consideration. This is explained in detail in Hardy (2003).

2   Equity Return Models for Embedded Guarantees

In this section we define and briefly discuss the major features of the various models proposed. In all of the following definitions, we let S_t denote the value of an equity index at time t, where S_0 is arbitrarily set at 1.00. As it is often convenient to model the log-return on equities, let Y_t denote

$$Y_t = \log(S_t/S_{t-1})$$

Then, obviously,

$$S_t = S_{t-1} \exp(Y_t) = S_0 \, \exp(Y_1 + Y_2 + ... + Y_t)$$

Use of the lower case indicates an observed value, upper case indicates a random variable.

2.1   The Lognormal Model (ILN)

Under the independent lognormal model, the log-returns Y_t are assumed to have identical Normal distributions, and are independent for $t = 1, 2, ...$, that is

$$Y_t = \mu + \sigma \, z_t \quad \text{where } z_t \text{ are i.i.d., and } z_t \sim \text{N}(0, 1) \quad \forall \, t$$

which means that, given only that $S_0 = 1$, for $k = 1, 2, , ...$

$$S_k \sim \log\text{N}\left(k\mu, \, k\sigma^2\right)$$

This process requires only two parameters, μ and σ. The ILN model is the discretely observed version of geometric Brownian motion, which is one of the original assumptions of the Black-Scholes framework (Black and Scholes, 1973). Short term variations in equity returns often appear consistent with the lognormal model, which is why the basic Black-Scholes framework has proved remarkably robust for short term financial instruments.

However, as we will demonstrate, over longer terms, the ILN model is generally rather thinner tailed than the data, which matters for out-of-the-money options.

In practice it is observed (for example, through implied volatility curves of traded options) that stochastic volatility models are required to adequately model equity returns. By stochastic volatility we mean any model for which the standard deviation of the outcome one or more time units ahead may be modelled stochastically.

2.2 The GARCH(1,1) Model

The ARCH/GARCH family of models, discussed extensively in, for example, Engle (1995). The GARCH(1,1) model is commonly used to represent equity returns in option pricing (eg in Duan (1995)). The auto-regressive conditionally heteroscedastic family of models allow for stochastic volatility more than one period ahead, but, given the full history of the process, the one-period volatility is assumed fixed and known. In particular, the GARCH(1,1) model can be written as:

$$
\begin{aligned}
Y_t | \mathcal{F}_{t-1} &= \mu + \sqrt{h_t} z_t \quad \text{where } z_t \text{ are i.i.d., and } z_t \sim \mathrm{N}(0,1) \quad \forall t \\
h_t &= \alpha_0 + \alpha_1 \left(Y_{t-1} - \mu \right)^2 + \beta \, h_{t-1}
\end{aligned}
\tag{1}
$$

The filtration $\mathcal{F}_{t-1}$ can be thought of as the relevant information for the process up to time $t-1$. There is only one random process here, represented by z_t. Also, given all the information at t, (Y_t and h_t) the volatility at $t+1$ is known, but the volatility at any time from $t+2$ forwards is random through the Y_t process.

If $\alpha_1 + \beta < 1$ then the process is covariance stationary, and we generally require this. In total the process involves four parameters, $\{\mu, \alpha_0, \alpha_1, \beta\}$.

2.3 The MARCH family

Wong and Chan (2005) argue in favour of a mixture of ARCH models, which they call the MARCH family. The MARCH(K; $a_1, ...a_K$; $b_1, ...b_K$) specifies a mixture of K ARCH models, where a_j is the auto-regressive order of the j-th model, and b_j is the ARCH-order

of the j-th model. Wong and Chan specifically identify the MARCH(2;0,0;2,0) model for log-returns, so that the model is a mixture of an ARCH(2) model, and a random walk model – that is,

$$Y_t | \mathcal{F}_{t-1} = \begin{cases} Q_1 & \text{w.p. } q \\ Q_2 & \text{w. p. } (1-q) \end{cases} \tag{2}$$

where

$$
\begin{aligned}
Q_1 | \mathcal{F}_{t-1} &= \mu_1 + \sqrt{h_t} z_t \quad \text{where } z_t \text{ are i.i.d., and } z_t \sim N(0, 1) \quad \forall t \\
h_t &= \alpha_{1,0} + \alpha_{1,1} (Y_{t-1} - \mu_1)^2 + \alpha_{1,2} (Y_{t-2} - \mu_1)^2 \\
Q_2 | \mathcal{F}_{t-1} &= \mu_2 + \alpha_{2,0} z_t
\end{aligned}
$$

Wong and Chan propose the model on the grounds that it offers an improved fit to the higher moments of the historical data, compared with the Regime Switching Log Normal. However, as we shall demonstrate, the overall fit does not appear to be superior.

Not that there are two random processes involved; the mixture random variable and the normal innovation. There are seven parameters, $\{\mu_1, \mu_2, \alpha_{1,0}, \alpha_{1,1}, \alpha{1,2}, \alpha{2,0}, q\}$

2.4 The regime switching lognormal model (RSLN)

The regime switching lognormal model was proposed for this purpose in Hardy(2001), with further discussion in Hardy (2003). The regime switching framework was first introduced by Hamilton (1989). The essence of the lognormal version is that the processes switch between two standard log-normal processes; the switching mechanism is another random process; generally assumed to be Markov. That is, the probability of switching regimes depends only on the current regime, not on the history of the switching process.

We find that for the data we consider (monthly stock index data) a two regime lognormal model works adequately. The regime process is denoted by ρ_t which takes the value 1 for the first regime, and 2 for the second. The log-return process can then be summarized as:

$$Y_t | \rho_t = \mu_{\rho_t} + \sigma_{\rho_t} z_t$$

$$\rho_t | \rho_{t-1} \;\; = \;\; \begin{cases} 1 & \text{w.p. } p_{\rho_{t-1},1} \\ 2 & \text{w.p. } p_{\rho_{t-1},2} = (1 - p_{\rho_{t-1},1}) \end{cases}$$

There are six parameters, $\{\mu_1,\ \mu_2,\ \sigma_1,\ \sigma_2,\ p_{1,2},\ p_{2,1}\}$.

As with the MARCH model, there are two random processes. However, there is a difference between the mixture process and the switching process; under the mixture model, the probability of using either the Q_1 or Q_2 distributions is the same each time period. Using regime switching, the probability depends on which model was used in the previous time period.

2.5   The regime switching draw down model (RSDD)

Panneton(2002) proposes a variation on the regime switching lognormal model. The regime switching draw down model adds a form of autoregression that influences the process when the index S_t falls below a previous high.

The definition of the RSDD process, for the two-regime version, is:

$$\begin{aligned} Y_t | \rho_t \;\; &= \;\; \kappa_{\rho_t} + \phi_{\rho_t}\, D_{t-1} + \sigma_{\rho_t}\, z_t \\ \text{where} \\ D_{t-1} \;\; &= \;\; \min(0, D_{t-2} + Y_{t-1}) \end{aligned}$$

and ρ_t is defined as for the RSLN model above.

The D_t process tracks how far the total log returns have fallen below the previous high. The parameters ϕ_j are generally negative. The intuition is that when returns are far below the previous high, the market exerts pressure for the index to recover. There are open questions here about the effect of survivorship – that one way for the recovery of an index such as the S&P 500 is the replacement of failing stocks with healthier alternatives, giving the index a better return than would be achieved by an investor.

The draw down factor is the only difference between the RSLN and RSDD models, so the RSDD requires eight parameters, $\{\kappa_1,\ \kappa_2,\ \phi_1,\ \phi_2,\ \sigma_1,\ \sigma_2,\ p_{1,2},\ p_{2,1}\}$

Like the RSLN model, this process involves two separate random processes. If the ϕ parameters are set to zero, we recover the RSLN model. In rising markets also, the process follows essentially the RSLN model, as the D_t process will be zero.

2.6 The Regime Switching GARCH model

Combining the popular GARCH and RSLN frameworks, this model was described by Gray(1996) who describes a model which switches between two GARCH(1,1) regimes.

$$
\begin{aligned}
Y_t | \rho_t &= \mu_{\rho_t} + \sqrt{h_{(t)}} \, z_t \\
h_t &= \alpha_{\rho_t,0} + \alpha_{\rho_t,1} \left(Y_{t-1} - \mu_{\rho_t} \right)^2 + \beta_{\rho_t} h_{t-1} \\
\rho_t | \rho_{t-1} &= \begin{cases} 1 & \text{w.p. } p_{\rho_{t-1},1} \\ 2 & \text{w.p. } p_{\rho_{t-1},2} = (1 - p_{\rho_{t-1},1}) \end{cases}
\end{aligned}
$$

This process involves ten parameters: $\{\mu_j, \alpha_{j,0}, \alpha_{j,1}, \beta_j, p_{j,1}\}$ for $j = 1, 2$.

We recover the RSLN model when $\alpha_{j,1}, \beta_j$, are set to zero.

2.7 The AAA Stochastic log-volatility model

The American Academy of Actuaries (AAA, 2005) in its C3 Phase 2 document proposes a 'stochastic log-volatility' model. The basic model description is as follows. The mean and volatility are expressed as annual rates and then adapted for application to the monthly process. This is consistent with the description in the AAA document.

$$
\begin{aligned}
Y_t &= \mu_t/12 + \left(\sigma_t/\sqrt{12} \right) z_{y,t} \\
&\text{where} \\
\mu_t &= A + B\sigma_t + C\sigma_t^2 \\
\log \sigma_t &= \nu_t = (1 - \phi)\nu_{t-1} + \phi \log \tau + \sigma_\nu \, z_{\nu,t}
\end{aligned}
$$

where $(z_{y,t}, z_{\nu,t})$ have a standard bivariate normal distribution, with correlation ρ. Upper and lower bounds to constrain the log-volatility process are omitted for clarity, but have

been incorporated in the estimation processes in the following sections. The lower bound for annual volatility is 3.05%, the upper bound is 79.88% and the upper bound for $\exp((1-\phi)\nu_{t-1} + \phi \log \tau)$ is 30%.

The AAA C3 phase 2 report claims that this model

> ...captures the full benefits of stochastic volatility in an intuitive model suitable for real world projections.

The model is used to set the calibration in the C3 Phase 2 report, and also for pre-recorded scenarios for plug and play use by insurers. The model has six parameters $(\phi, \tau, A, B, C, \rho)$.

3 Maximum likelihood results

We have found (by calculation or estimation) the maximum likelihood parameters for each of the models, and the corresponding value of the maximum log-likelihood based on the S&P 500 total return index from January 1956 to September 2004.

The calculation of the log-likelihood for the ILN, GARCH and RSLN models is given in Hardy(2003). The MARCH process is described in Wong and Chan (2005). The RSGARCH is taken from Gray(1996), and the RSDD is a simple extension of the RSLN process.

The SLV model is a little different. Recall that the maximum log-likelihood is the maximum value over the parameter space of

$$\sum_{t=1}^{n} \log f(y_t | \Theta, y_1, y_2, ..., y_{t-1})$$

where Θ represents the parameter set, and f is the density function for the log-return. Under the SLV model we have a random log-volatility process,

$$\boldsymbol{\nu} = \{\nu_1, \nu_2, \nu_3, ..., \nu_n\}$$

Model	Number of parameters	Maximum log-likelihood
Lognormal	2	1018.0
GARCH(1,1)	4	1030.1
MARCH	7	1039.8
RSLN	6	1042.0
RSDD	8	1047.1
SLV	7	1035.2
	(+ 3 min & max constraints)	
RSGARCH	10	1054.9

Table 1: Maximum log-likelihoods for the seven models.

Given the log-volatility $\boldsymbol{\nu}$, the density function for $y_t|y_1, ..., y_{t-1}, \Theta$ is simply

$$f(y_t|y_1, ..., y_{t-1}, \Theta, \boldsymbol{\nu}) = \frac{1}{\sigma_t/\sqrt{12}} \, \phi \left(\frac{y_t - \mu_t/12.0}{\sigma_t/\sqrt{12}} \right)$$

where $\sigma_t = e^{\nu_t}$.

The volatility process is too complex to calculate this directly, but we can use Monte Carlo simulation to estimate the expectation of $f(y_t|y_1, ..., y_{t-1}, \Theta, \boldsymbol{\nu})$ with respect to $\boldsymbol{\nu}$. That is, generate N volatility paths, each with n values, where n is the length of the data series. Let $\boldsymbol{\nu}_j$ denote the j-th simulated log-volatility path, then the estimated contribution to the log-likelihood from the t-th observation is

$$\sum_{j=1}^{N} \frac{\log f\left(y_t|y_1, ..., y_{t-1}, \Theta, \boldsymbol{\nu}_j\right)}{N}$$

Shephard (2005) describes more sophisticated simulation procedures, using importance sampling, for general stochastic volatility models, but straightforward Monte Carlo appears to work well, here, with very little sampling variability. Using only 1000 simulations, the standard error of the log-likelihood is around 0.2.

The log-likelihoods are given in Table 1. Clearly the lognormal model is an outlier, with a very poor overall fit. The other six models appear comparable.

Likelihood based model selection offers the Akaike Information Criterion, which says that each additional parameter should contribute at least one unit to the log-likelihood, which would suggest that the RSGARCH model is the preferred, followed by the RSDD, RSLN, MARCH, SLV and GARCH. The Bayes information criterion (BIC) is a little more sophisticated, as it takes the sample size into consideration. According to the BIC, each additional parameter should increase the log-likelihood by around 3.1 units. This would indicate that the RSLN is a slightly better fit than the RSDD, the GARCH is preferred to the SLV and that the GARCH and MARCH models are very close.

However, likelihood based selection does not address the central questions – do these models fit the data? Or, the more relevant question – do these models fit the data in the parts of the distribution most critical for the equity-linked capital requirement calculations? How much does the fit quality matter in the final calculation?

4   Residual Analysis

For each of these models there is a sequence of residuals (or combination of residuals for the regime switching models) which should be an i.i.d. Normally distributed sample according to the assumptions of the model. If we examine the residuals and they are far away from the Normal model, that's a signal that the fit is not adequate.

We use two methods for exploring the residuals. The q-q plot shows the model residuals quantiles against the Normal distribution quantiles. If the model assumptions hold, the residuals should, broadly, lie on the central diagonal.

In Figures 1 and 2 we show the q-q plots for all seven models. If the residuals are normal they would lie on the diagonal shown.

The second item is the Jarque-Bera test of normality – a statistical test often used for residual analysis. Failing this test indicates that the residuals are not normal, which is an indication that the model is not consistent with the data.

For all the residual analysis we have used maximum likelihood estimates of all parameters.

4.1 ILN

For the fitted values of μ and σ, and the observed values of the log-returns, y_t, the ILN residuals are simply

$$r_t = \frac{y_t - \hat{\mu}}{\hat{\sigma}}$$

which under the ILN assumptions are i.i.d. N(0,1) distributed.

The q-q plot is shown in Figure 1. The fit fails, predominantly, in the left tail, where the residuals are much fatter tailed than the Normal.

The failure of the fit on the left tail causes the failure of the Jarque Bera test of normality, with a p-value of around 10^{-40}.

4.2 GARCH

The GARCH residuals are

$$r_t = \frac{y_t - \hat{\mu}}{\sqrt{h_t}}$$

where

$$h_t = \hat{a}_0 + \hat{a}_1(y_{t-1} - \hat{\mu})^2 + \hat{\beta}\, h_{t-1}$$

The q-q plot is shown in Figure 1.

Again, the left tail fit is poor, and this is reflected in a Jarque Bera test p-value, once again, around 10^{-40}.
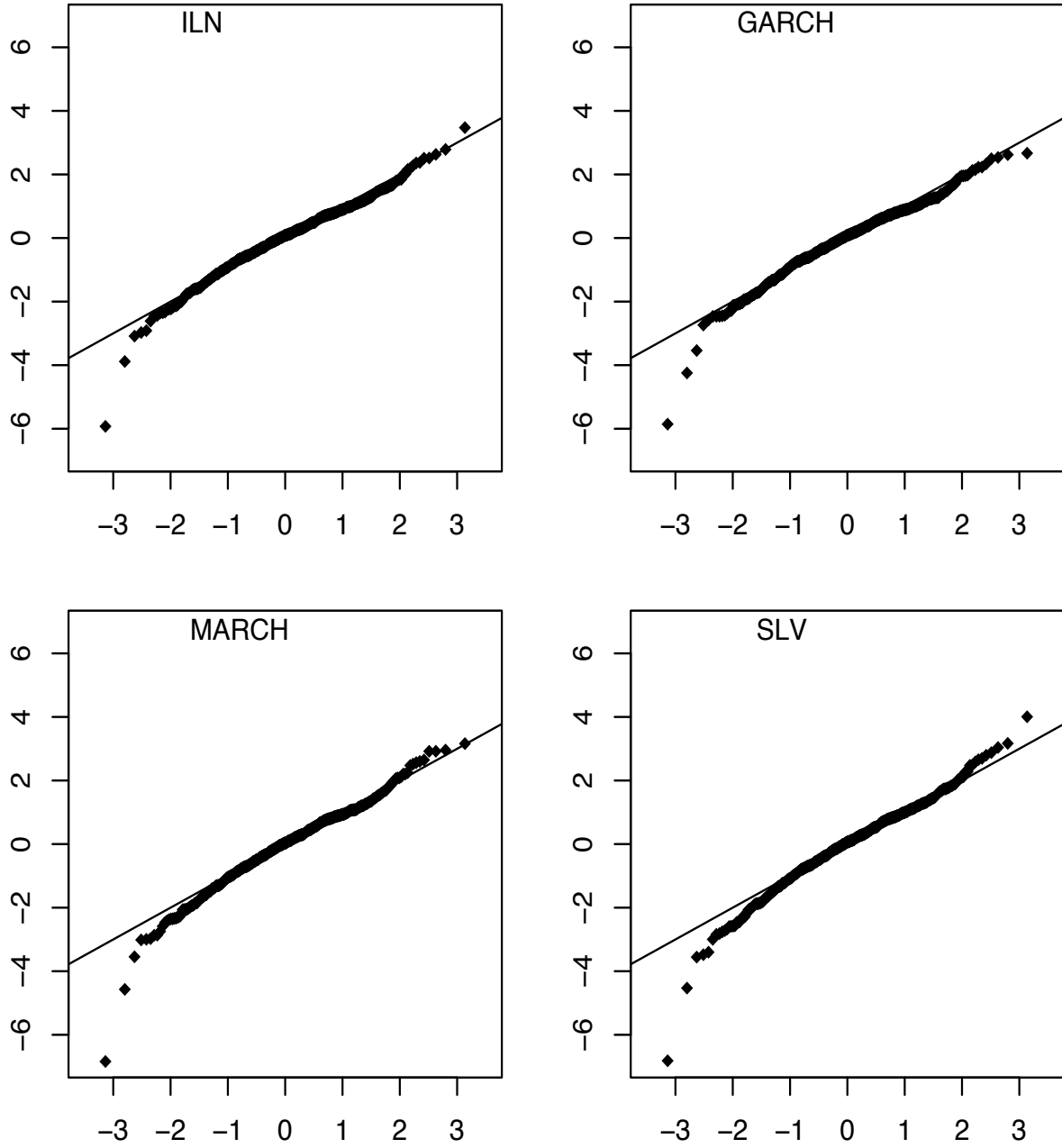
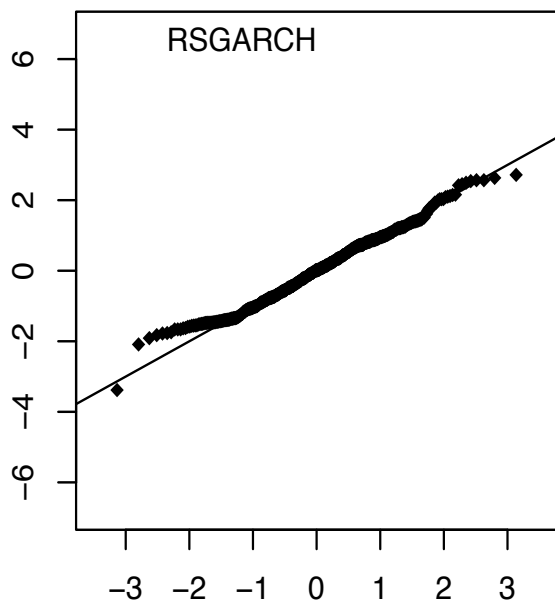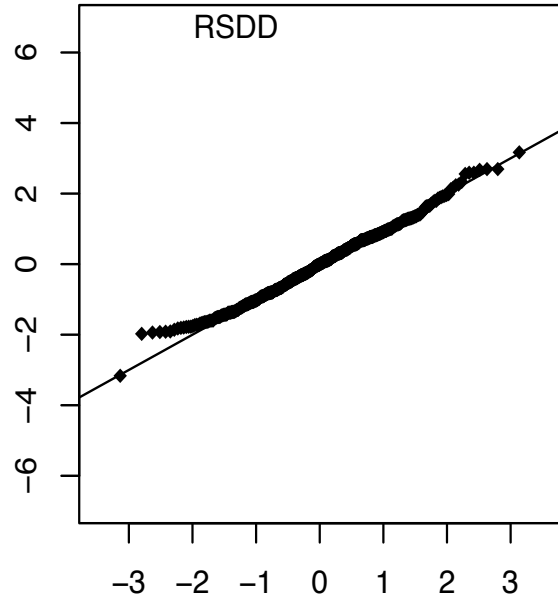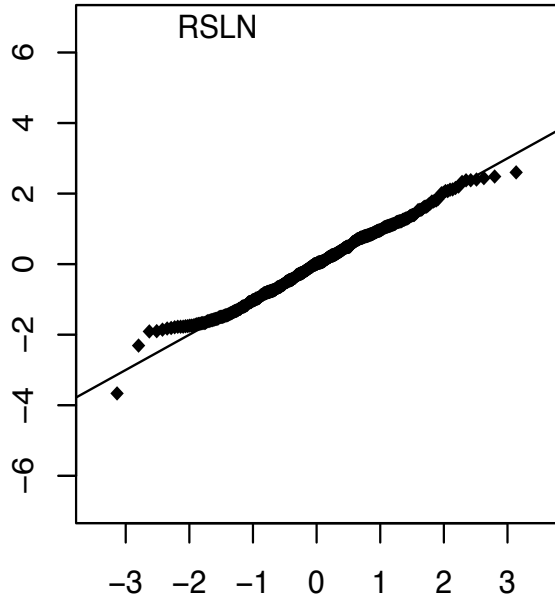Figure 1: q-q plot of residuals under the ILN, GARCH, MARCH and SLV models.

Figure 2: q-q plot of residuals under the Regime Switching models.

4.3 MARCH

The MARCH is a mixture of two models each with N(0,1) i.i.d. residuals under the model assumption. Thus, the mixed residuals should also be i.i.d. N(0,1), so that

$$r_t = \hat{q}\,\frac{y_t - \hat{\mu}_1}{\sqrt{h_t}} + (1 - \hat{q})\frac{y_t - \hat{\mu}_2}{\sqrt{\hat{\alpha}_{2,0}}}$$

where

$$h_t = \hat{\alpha}_{1,0} + \hat{\alpha}_{1,1}\left(y_{t-1} - \hat{\mu}\right)^2 + \hat{\alpha}_{1,2}\left(y_{t-2} - \hat{\mu}\right)^2$$

The q-q plot is given in Figure 1. We see the same poor left tail fit, and in fact the Jarque Bera statistic is even worse than the GARCH and ILN models, at 10^{-45}.

4.4 SLV

The residuals for the SLV model depend on the random volatility process. Using the same technique as for the log-likelihood, we can estimate the residuals by Monte Carlo simulation of the log-volatility paths, and, as with the log-likelihood, the sampling variability is small. Using N simulations of the log-volatility path $\boldsymbol{\nu}$, such that $\sigma_{j,t}$ is the t-th simulated volatility from the j-th simulated path, the residuals are estimated from

$$\tilde{r}_t = \frac{1}{N}\sum_{j=1}^{N}\frac{y_t - \hat{\mu}_t/12}{\hat{\sigma}_{j,t}/\sqrt{12}}$$

The q-q plot is given in the lower right plot of Figure 1. The fit is not very good in either tail, and the Jarque Bera test fails with a p-value of 10^{-40}.

14

4.5 RSLN

For the RSLN model, we can calculate two residual values at each data point, one assuming the process is in the first regime, and one assuming it is in the second. That is:

$$(r_t|(\rho_t = 1)) = \frac{y_t - \hat{\mu}_1}{\hat{\sigma}_1}$$

and similarly for r_t given $\rho_t = 2$.

Although we do not know which regime the process is in, we can assign probabilities, based on information from the observations. That is, given the model parameters, we know p, say the probability that $\rho_t = 1|y_1, y_2, ..., y_t$, and $1 - p$, the probability that $\rho_t = 2|y_1, y_2, ..., y_t$. This is actually a step in the likelihood calculation process, and is described in Hardy (2003). The unconditional expected residual series

$$r_t = p\,(r_t|(\rho_t = 1)) + (1 - p)\,(r_t|(\rho_t = 2))$$

is the series plotted in q-q form in Figure 2. The overall fit is better than the models in Figure 1, and this is supported by the Jarque Bera test which is passed, with a p-value of 0.45, indicating that the residuals are consistent with a normal distribution.

4.6 RSDD

The RSDD residuals were calculated similarly to the RSLN, with

$$(r_t|\rho_t = 1) = \frac{y_t - \hat{\kappa}_1 + \hat{\phi}_1\,d_{t-1}}{\hat{\sigma}_1}$$

where $d_{t-1} = \min(0, d_{t-2} + y_{t-1})$, and similarly for $r_t|\rho_t = 2$. We assume $d_0 = 0$.

The q-q plot shows the residuals are slightly closer to the normal quantiles than the RSLN model in the tails, though the distribution is slightly less close in the centre. The residuals pass the Jarque Bera test with a p-value of 0.18.

4.7 RSGARCH

Similarly to the other RSDD models, we calculate the residuals for the two regimes, for example:

$$(r_t|(\rho_t = 1)) = \frac{y_t - \hat{\mu}_1}{\sqrt{h_1(t)}}$$

and then weight by the probabilities associated with the two regimes. The results, in Figure 2, are similar to the other regime switching models, and the residuals pass the Jarque Bera test with a p-value of 0.15.

4.8 Summary

Only the regime switching models pass the Jarque Bera test, and the reason is clear from the q-q plots, which show that all the others fail in the left tail of the residuals. The SLV model also does not appear to offer a very good fit of the right tail either.

There is no way to tell from the residuals analysis whether any of the regime switching models is better than the others.

5 How much does it matter?

All of these models produce comparable maximum likelihood figures; indicating an acceptable overall fit. The residuals show that the models may be separated into two or three groups – the regime switching models appear to offer a reasonable fit, the others, less so, though the poor fit for, say, the GARCH model arises from a very small number of outlying data points.

If there is little difference in the capital requirements generated using the different models, then it would not matter too much which is used in practice. In this section we apply all of the models of the previous section to an sample variable annuity type contract. We do not calibrate the models to the AAA table; instead we use the maximum likelihood

parameter estimates in the scenario generation. The reason is that we are interested in whether the calibration requirements are justifiable, or whether any model producing similar maximum log-likelihood is, essentially, just as valid as any other.

We will look at the implications for capital requirements based on a traditional static actuarial approach, and, separately, using a simple delta hedge.

5.1 Example contract

To illustrate the effect of model uncertainty we use a 20-year single premium guaranteed minimum accumulation benefit, (GMAB) issued to a life age 50. Some of the assumptions are simplistic, the example is used solely for illustration. The single premium is invested in the policyholder's fund, F_t, which is assumed to be invested in the S&P 500 index, with dividends reinvested. The benefit on death or maturity is the greater of the accumulated investment proceeds and a guarantee. The guarantee is set at issue equal to the amount of the single premium. After 10 years, if the policy is still in force, the guarantee is reset to the accumulated investment proceeds, if that is greater than the guarantee. If the guarantee exceeds the fund then a cash payment equal to the difference is paid into the fund at that time.

We assume the policyholder is age 50 at issue, and that mortality follows the rates given in the appendix of Hardy(2003). Lapses are assumed level at 8% per year (for simplicity). A management charge of 300 basis points (bps) (3%) per year is deducted from the policyholder's fund, of which 20 basis points is used to fund the guarantee.

There are two common approaches to the risk management of segregated fund and variable annuity contracts of which this is a simplified example. The 'actuarial approach', which may be described as a static partial hedge, uses Monte Carlo simulation to project the liabilities, using a real-world distribution, (that is, not risk-neutral) discounted at the risk-free rate, assumed here to be 5%, to give a distribution of the present value of the liabilities if no hedging strategy is adopted. We apply a risk measure to the distribution to ascertain the total economic capital required for the contract.

The other approach is to use dynamic hedging to limit the risk. The simplest version of this is to assume a simple Black-Scholes hedge is established at the contract inception,

and rebalanced monthly. The hedge will not be perfect for a number of reasons, including discrete hedging error (a perfect hedge requires continuous rebalancing), transactions costs on the rebalancing of the hedge assets, and model error, because the log-normal model of Black-Scholes is, as we have seen, not a particularly good fit to the historical equity returns data. These hedge imperfections constitute an additional liability, which we can estimate by Monte Carlo simulation under a real-world measure. The process is described in more detail in Hardy(2003). We assume a risk free rate of interest of 5% per year, and a volatility for hedging purposes of 18% per year, which is broadly consistent with the volatility in the models and data.

The risk measure used us the Conditional Tail Expectation (CTE), which is the basis of both the CIA segregated fund report (CIA, 2001) and the AAA (2005) C3 Phase 2 report. The CTE at α, $0 \leq \alpha \leq 1$ is estimated as the average of the worst $100(1 - \alpha)\%$ of the simulations. For all the simulations summarized below we have used 10,000 projections.

In Figure 3 we show the results of using each model for the actuarial approach. The x-axis represents the CTE parameter. At $\alpha = 0$ we have the mean loss for the contract. As α moves near to one, the curve tends to the maximum simulated losses. The vertical line lies at $\alpha = 0.95$, corresponding to the standard used for total balance sheet requirement for segregated funds. In Table 2 we show explicitly the 95% CTEs for the different models.

The graph shows that the results fall into three groups. The RSGARCH and RSLN models generate similar CTE results, with the highest values for the CTEs for all values of α. The next group contains the GARCH, MARCH, SLV and ILN models. The 95% CTEs for this group are more than 1.5% less than the top group. At the bottom, offering the most optimistic view of the loss distribution is the RSDD model, which generates a 95% CTE more than \$3 smaller for every \$100 premium, than the other regime switching models. This difference could be very significant; suppose we assume the RSDD model best represents the future, but in fact the RSLN proves to be more accurate. For every contract on the cohort, the economic capital would be inadequate by 3% of the single premium, even after allowing for all margin offset income.

The split of the results, with the regime switching models generating both the most optimistic and the most pessimistic views of the loss distribution is somewhat surprising, as the analysis of likelihoods and residuals indicated that the regime switching models offered a very similar fit to the data. Moreover, the CTEs using the RSDD model are
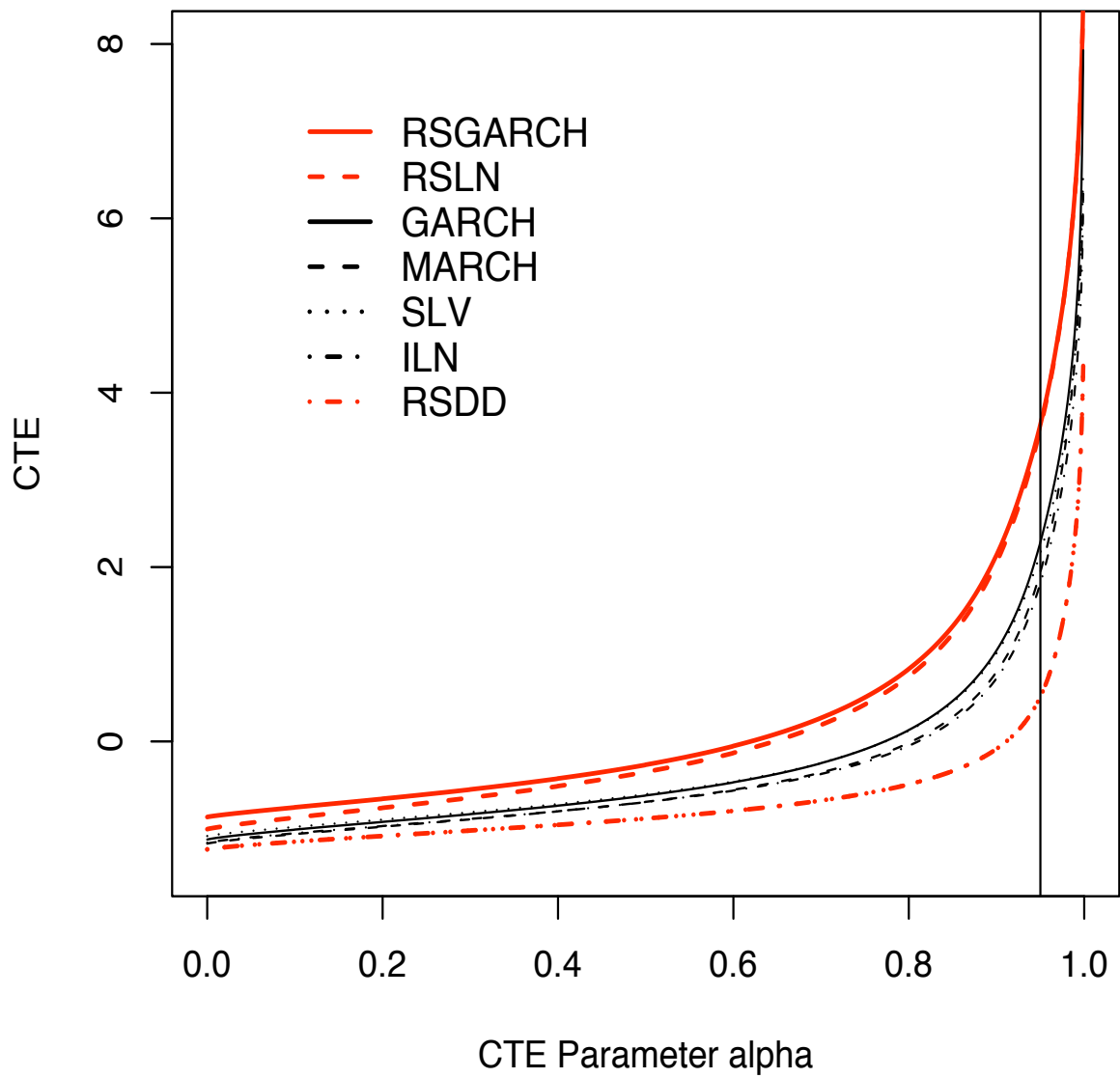
18

Figure 3: CTEs using the actuarial approach.

Model	95% CTE
RSGARCH	3.64%
RSLN	3.59%
GARCH	2.29%
MARCH	1.94%
SLV	2.16%
ILN	1.81%
RSDD	0.51%

Table 2: 95% CTE for GMAB, % of single premium, actuarial risk management.

substantially lower even than the ILN model, which offers a very poor fit to the data, in particular in the crucial left tail area of the returns distribution.

In Figure 4 we show the CTE results using the dynamic hedging approach, where the different models are used to project the hedge to estimate the distribution of the hedge costs and unhedged liabilities. The Figure is plotted on the same scale as the actuarial approach in Figure 3. The 95% CTEs are given in Table 3. We notice here that the hedge clearly achieves a lot of risk mitigation – the maximum losses are much lower than under the actuarial approach, and, the 95% CTE is lower using the dynamic hedging approach than it is under the actuarial approach for all models except the RSDD. The right tail protection is achieved at a cost, however, since the mean outcome is a loss using the hedging approach compared with the actuarial (which indicates that the margin offset is too low, if the other assumptions are appropriate).

So, it appears that model selection is not so critical using the dynamic hedging approach, but may have significant impact using the actuarial approach. An explanation is that the models all look fairly similar in the monthly returns. For the dynamic hedging approach, where the month to month equity index movement is the key factor, this means that the results are robust with respect to model variation. However, the actuarial approach depends on the long term accumulations under the models. As we have seen, the RSDD model appears to behave very similarly to the RSLN (for example) on a month to month basis, but the differences over long accumulation periods are significant.

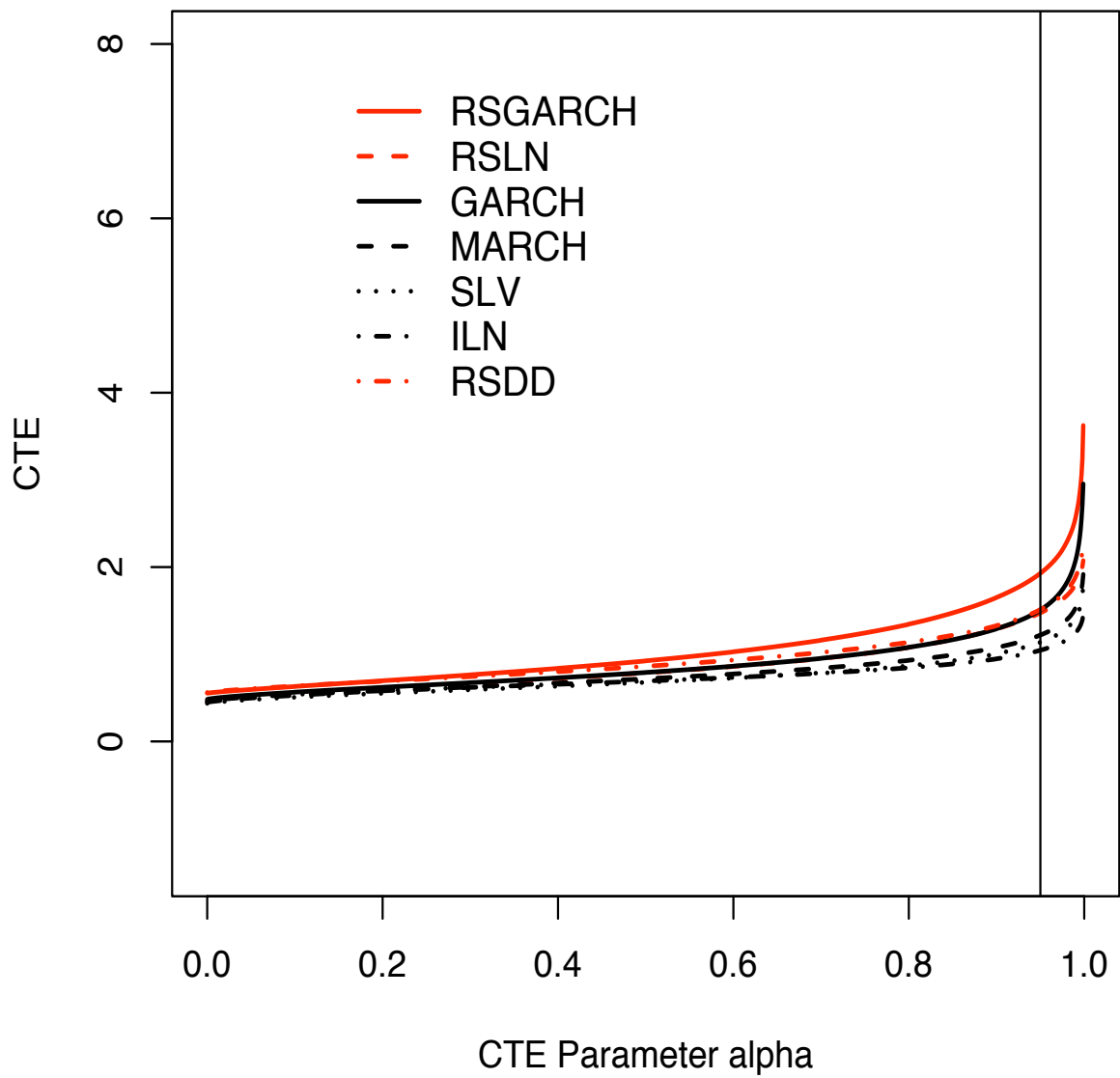The actuarial approach to risk management is the dominant method in Canada and

Figure 4: CTEs using the dynamic hedging approach.

Model	95% CTE
RSGARCH	1.93%
RSLN	1.48%
GARCH	1.51%
MARCH	1.22%
SLV	1.14%
ILN	1.05%
RSDD	1.51%

Table 3: 95% CTE for GMAB, % of single premium, dynamic hegding risk management.

possibly also in the USA. It therefore matters whether the RSLN ar the RSDD is a better model. If it is the RSDD and we assume the RSLN, we waste resources with unnecessary solvency capital. Vice versa, and we risk significant capital shortage in the event of poor market returns. In the following sections we will try to extract more information from the data using bootstrap techniques.

6 Bootstrapping the Data

6.1 Bootstrapping time series

The bootstrap is a technique for exploring the relationship between a sample and an underlying population. The essential idea is that of resampling from the sample – that is, creating a new sample, the same size as the original, by drawing with replacement from the original sample. The relationship between the new, 'pseudo-sample' and the original sample, mirrors in many ways the relationship between the original sample and the underlying population. Hence, the bootstrap can offer a non-parametric (or parametric) estimate of the uncertainty of statistics such as quantiles estimated from samples. Efron and Tibshirani (1993) describe the basic techniques.

The application of bootstrap techniques to time series is a more recent development. With time series, simply drawing with replacement will not replicate the original distribution,

as it will lose the serial dependency in the data. To retain dependency, we draw blocks of data rather than individual values. The optimal block size is not always evident. If blocks are too short, dependency in the original sample will be lost. If the correlations are, broadly positive, then the pseudo-sample will be too thin tailed, as the blocks will not adequately pick up periods of prolonged poor returns. If the time series is essentially uncorrelated, then short blocks will not affect the results. If the correlations are negative, then block sizes that are too short will cause the pseudo-sample to be too fat-tailed, as the original smoothing of results from consecutive observations will not be reflected in the pseudo-sample. If the block sizes are too large, we lose tail information in the data; the blocks effectively reduce the sample size. This will cause the sample to be thinner tailed than the original sample. After some exploration, we used 6 month block sizes; the results were not sensitive to block sizes of between 2 and 12 months.

6.2   Accumulation factors

The n-year accumulation factor is the accumulation after n years of an investment of \$1. In terms of the log-return random variable, Y_t, we are interested in the sums of consecutive values, which gives the log of the accumulation factor. The CIA (2001) and AAA (2005) requires models to be calibrated to tables of accumulation factors for one, five, ten and (for the AAA) twenty years. The CIA sets standards for the left 2.5%, 5% and 10% quantiles. The AAA sets standards for both the left and right tails, up to the lower and upper 20% quantiles.

The index data we use covers 48 years. So, for the 1-year factor we have 48 non-overlapping observations. We may consider the minimum value observed, which (using January to January data) is 0.770, as an estimate of the 1/49 or 2.04% quantile; the second smallest value, 0.835 is an estimate of the 2/49 or 4.08% quantile. Interpolating between these values gives an estimate of 0.7844 for the 2.5% quantile. Similarly, we have an estimate for the 5% quantile of 0.836, and for the 10% quantile we have 0.859.

In fact, even before we bootstrap, the data has more to tell us, since we get a slightly different story by looking at the years from February to February, or March to March etc. The different starting points make little difference to the 10% quantile estimate, but are more significant for the 5% and 2.5% quantiles. The range for the 2.5% estimate is

23

Bootstrap	2.5%-ile	5%-ile	10%-ile
90% Confidence Interval	(0.67, 0.87)	(0.76, 0.91)	(0.84, 0.97)
Model	**2.5%-ile**	**5%-ile**	**10%-ile**
ILN	0.829	0.868	0.916
GARCH	0.812	0.857	0.912
MARCH	0.823	0.868	0.918
SLV	0.825	0.868	0.915
RSLN	0.764	0.829	0.908
RSDD	0.768	0.831	0.901
RSGARCH	0.792	0.847	0.910

Table 4: Model and bootstrap quantiles for one-year accumulation factors.

(0.638,0.836).

Using standard bootstrap techniques we generate a 10,000 samples of 48 1-year accumulation factors by drawing with replacement in four month block. We calculate the estimate of the 2.5%, 5% and 10% quantiles for each pseudo sample just as we did for the original data. We can use the bootstrap sample to construct a confidence interval for the true quantiles. We compare this with the parametric estimates from the models. Any model which generates quantiles outside the bootstrap confidence interval is considered to have failed the test of fit.

The results for the 1-year factors are shown in Table 4. We see that, even though the regime switching models are fatter tailed than the other models, none of the models fails the bootstrap test.

We repeat the exercise for the 10-year accumulation factors. Here, with 48 years of data we only have four non-overlapping observations. The minimum value is an estimate of the 20% quantile, which is as far into the tail as we can explore by standard methods. The empirical estimate ranges from around 1.2 to 2.2. The bootstrap 90% confidence interval, and model values for the 20th percentiles are given in Table 5. Again, we find that none of the models generates quantiles outside the bootstrap interval – which is very wide.

	20%-ile
Bootstrap 90% Confidence Interval	(0.95, 2.83)
Model	
ILN	1.841
GARCH	1.847
MARCH	1.909
SLV	1.788
RSLN	1.773
RSDD	1.953
RSGARCH	1.660

Table 5: Model and bootstrap quantiles for ten-year accumulation factors.

7   Oversampling

The bootstrap, conventionally applied, did not help to identify whether any of the models are more consistent with the data. In this section we consider the effect of extending the time series bootstrap. For the 10-year accumulation factor, the bootstrap method requires the construction of sets of only four values. What are the implications of continuing to sample from the data, and using the resulting distribution? That is, suppose we sample 10,000 10-year accumulation factors from the data, in blocks of 4 or 6 months. How would that distribution compare with the underlying population, on average?

The answer is that, if the data is essentially zero or positively correlated, in the sums as well as the individual months, oversampling will give a distribution that is thinner tailed, on average, than the original population.

The intuition is straightforward; if we oversample we are exploring regions of the distribution outside of the range of the data, on average. For example, suppose we have a random sample of four values from a Uniform(0,1) distribution. On average, the minimum value will be around 0.2. No matter how much we oversample from this set of four observations, we will not, on average, sample at all from the bottom 20% of the distribution. Thus, we will end up with a thinner tailed distribution than the underlying population.

However, we know that for negatively correlated distributions, it is possible for blocking to fatten tails if the block sizes are not big enough. For positively correlated distributions blocking with small blocks will thin the tails. So the dual effect of oversampling and blocking will result in thinner tails than the underlying population for zero and positively correlated data, but may combine for thinner or fatter tails, or neither, for negatively correlated data.

In Figure 5 we plot the partial autocorrelation function for both the original log-returns, and for the sums of the log-returns. The graphs show that there is no evidence of negative correlation, over short or long term accumulations. Hence, we expect oversampling to generate a thinner tail for the accumulation factors than the underlying population.

We use this result by repeating the bootstrap test. This time, we estimate the left tail quantiles of the oversampled distribution for the 10-year accumulation factors. We compare these quantile estimates with those from the different models proposed. The oversampled distribution should be thinner tailed (on average) than the population. Thus, the left tail quantiles from the oversampled distribution should be larger than the underlying population. A model with smaller left tail quantile values than the oversampled distribution passes the test. Larger left tail quantiles imply a thinner tail than the oversampled distribution, and such a model is less likely to generate the data than the fatter tailed, and so fails the test. The results will apply for quantiles outside the range of the original data – which means that the test will be more useful for comparing the ten-year accumulation factors than for the one-year. Once we move to the quantiles that are within the range of the data, oversampling gives broadly the same information as the regular bootstrap.

The left tail quantiles for the oversampled distribution and the model ten year accumulation factors are given in Table 6. We generated 1,000 values for the 10-year accumulation factor, sampling from the original values in 6-month blocks.

The table shows that only two models pass this test for left tail weight for all three quantiles – the RSLN and RSGARCH models. The relevant values are bolded in the table. The SLV model lies close to the oversample distribution. All the rest are significantly thinner tailed than the oversample distribution. The relative weights are illustrated in Figure 6, which shows the left tails of the density functions for the oversample distribution and four of the model pdfs.
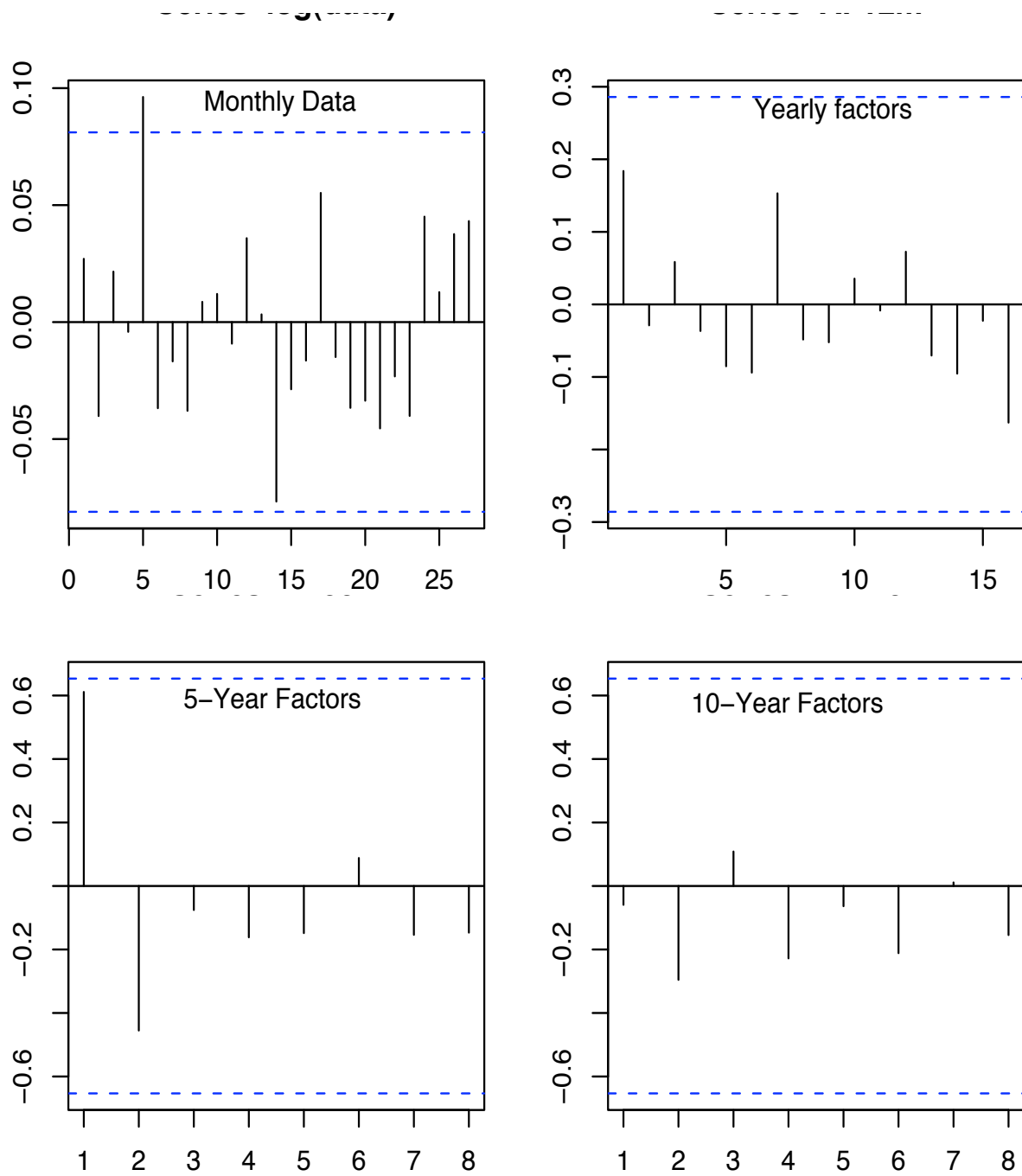
26

Figure 5: Partial autocorrelation functions for the monthly, yearly, five-yearly and ten-yearly accumulation factors.
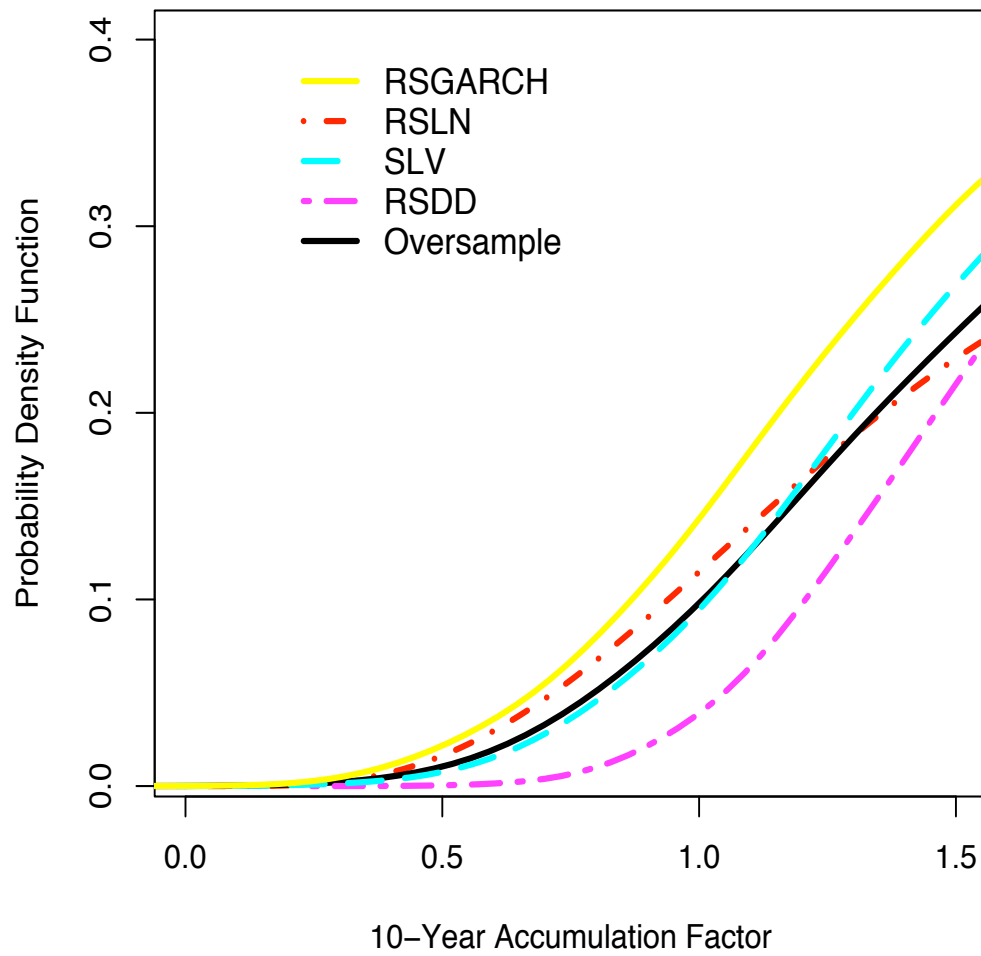
Figure 6: Left tail of the distribution of 10-year accumulation factors.

Model	2.5%	5%	10%
Oversample Distribution	**1.041**	**1.228**	**1.478**
ILN	1.096	1.269	1.501
GARCH	1.074	1.259	1.509
MARCH	1.131	1.297	1.558
SLV	1.087	1.249	**1.467**
RSLN	**0.914**	**1.105**	**1.378**
RSDD	1.277	1.439	1.653
RSGARCH	**0.905**	**1.086**	**1.315**

Table 6: Model and oversample quantiles for ten-year accumulation factors.

This test gives strong support that the RSLN and RSGARCH results for capital requirements in Table 2 are more reliable than the RSDD, despite the apparent similarity in the distributions in Section 4.

The same arguments that we have applied to the left tail also apply to the right tail. In Figure 7 we compare the same four distributions with the oversample distribution. We see that only one distribution is fatter tailed on the right side – the RSLN.

8 Conclusion

While maximum likelihood arguments can assess whether one model has a better overall fit to the data than another, they do not tell us whether either of the models provides an adequate fit. We have shown that the residuals give some further information on this, and for the monthly S&P data set, the regime switching models appear to provide a substantially better fit than the ARCH type or stochastic log volatility models. However, this does not necessarily help to determine capital requirements since we show that, using the actuarial approach, two very similar models, the RSLN and RSDD, can give very different capital requirements for a simple example variable annuity benefit.

Using an extension of the bootstrap has given us a new tool to explore the tail fit for the accumulation factors, by extracting more information from the data. Here we show
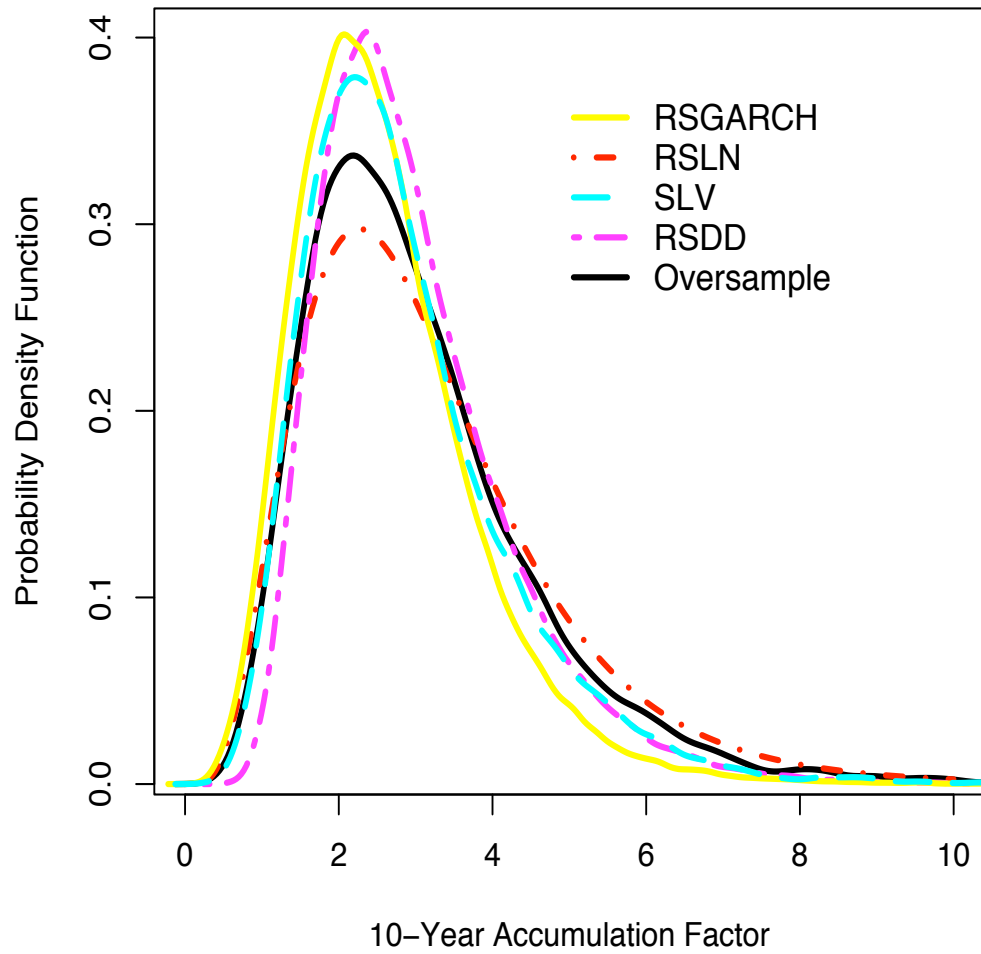
Figure 7: Full distribution of 10-year accumulation factors.

again that the RSLN and RSGARCH models appear to offer a satisfactory fit to the left tail for the 10-year accumulation factors, while the RSDD, along with the MARCH, ILN, GARCH models do not. The SLV is borderline on this test.

9 Acknowledgements

References

American Academy of Actuaries (AAA) (2005) Recommended Approach for Setting Regulatory Risk-Based Capital Requirements for Variable Annuities and Similar Products, Presented by the American Academy of Actuaries Life Capital Adequacy Subcommittee to the National Association of Insurance Commissioners Capital Adequacy Task Force (C3 Phase 2 Report).
http://www.actuary.org/pdf/life/c3_june05.pdf

Black, Fischer and Scholes, Myron (1973) The pricing of options and corporate liabilities. *Journal of Political Economy* **81** 637-654.

Canadian Institute of Actuaries (CIA) (2001) *Report of the Task Force on Segregated Fund Investment Guarantees* Canadian Institute of Actuaries; available from
http://www.actuaries.ca/publications/2002/202012e.pdf

Duan Jun Chuan (1995) The GARCH Option Pricing Model. *Mathematical Finance* 5, 13-32.

Efron Bradley and Tibshirani, Robert J. (1993) *An Introduction to the Bootstrap.* Chapman and Hall/CRC, Boca Raton, Florida.

Engle Robert F (Ed) (1995)ARCH: Selected Readings. Oxford University Press.

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

Hamilton James D (1989) A new approach to the economic analysis of non-stationary time series *Econometrica* **57** 357-384.

Hardy Mary R (2001) A regime switching model of long term stock returns. North American Actuarial Journal **5 2** 41-53.

Hardy Mary R (2003) *Investment Guarantees; Modeling And Risk Management For Equity Linked Life Insurance.* Wiley, New Jersey.

Panneton Christian-Marc (2002) Mean-Reversion in Equity Models in the Context of Actuarial Provisions for Segregated Fund Investment Guarantees. *Segregated Funds Symposium Proceedings.* Canadian Institute of Actuaries.

Shephard Neil (2005) *Stochastic Volatility: Selected Readings.* Oxford University Press.

Wilkie A. D.(1995) More on a Stochastic Asset Model for Actuarial Use,*British Actuarial Journal*, **1**, V,777-964.

Wong Albert C. S. and Wai-Sum Chan (2005) Mixture Gaussian Time Series Modelling of Long-Term Market Returns. *North American Actuarial Journal*, 9(4).