Milliman White Paper

Solvency II — Accelerating the calculations: Replicating portfolios and model compression

Dominic Clark
Oliver Gillespie
Henny Verheugen

*Contributors include Jeff Courchene, Josh Corrigan,
Avi Freedman, and Ed Morgan*

Solvency II focuses on risk and places great importance on embedding risk management within companies, their processes, and their governance, as well as requiring a similar focus and modernisation within supervisors. In direct contrast to Solvency I, Solvency II is realistic, principles-based, and will involve complex calculations.

In particular, the standard methodology for Solvency II revolves around an economic (i.e. market-consistent) balance sheet, which is then stress tested for various risks. The resulting effects on the balance sheet's net asset value are then combined in order to arrive at the solvency capital requirement. This required capital is compared to available capital (taken from the economic balance sheet).

The calculations involved in constructing the Solvency II economic balance sheet and then performing the stress tests often require the use of Monte Carlo simulations, i.e. stochastic runs. Where internal models are used, they may take similar approaches or may even increase the calculation burden further via the use of 'nested' stochastic techniques (where secondary stochastic simulations are triggered along the main projection of each primary stochastic run).

For large models the runtimes can therefore be considerable, and this reduces the flexibility of the reporting and the depth of insight available, as well as putting pressure on meeting reporting deadlines.

SPEEDING UP THE CALCULATIONS

There are several possible approaches to accelerating the calculations, including:

1. Increase computing power (for example, grid computing–use of a matrix of linked PCs to work in parallel on the calculations, rather than a single machine).

2. Deploy replicating portfolios (construction of a portfolio of assets that will reproduce the behaviour exhibited by a specific set of cash flows under a given set of economic scenarios).

3. Reduce the number of policies used as input to the cash flow projections (for example, via model compression–construction of a small set of 'model points' designed in such a way as to produce the same modelling results as running the cash-flow-projection model over the full policy or asset portfolio).

In this short paper we will consider approaches (2) and (3) in the context of life insurance. More information on approach (1) may be found at http://www.milliman.com/expertise/life-financial/products-tools/c-squared/.

REPLICATING PORTFOLIOS

Replicating portfolios (RP) are effectively an extension of existing asset-liability matching techniques. The RP approach involves modelling the liabilities by reference to their closest asset equivalents (physical and derivative), which can then enable the use of closed-form solutions to replace nested stochastic modelling.

The RP approach allows much faster calculation of the capital requirement for financial risk when full stochastic runs of the liability-cash-flow projection are replaced by the more straightfoward recalculation of RP asset value.

The cash-flow-projection model for a block of insurance business is first constructed using existing tools (for example, Milliman's MG-ALFA). In order to capture financial risk, the cash-flow projection is run many times, each time based on a different economic scenario (taken from a set that has been generated appropriately for the purpose of the calculations at hand). The complete set of results therefore consists of a set of projections of future cash flows, one for each economic scenario tested.

This output then serves as input to the process of constructing a replicating portfolio. An optimisation algorithm is employed to find a suitable set of assets that will closely reproduce the behaviour of the liability cash flows under all the economic scenarios tested. Considerable expertise is usually necessary to identify a suitable shortlist of potential replicating assets likely to achieve a satisfactory fit to the liabilities.

If the replicating assets are chosen so that market information (such as their price or pricing parameters) is readily available, then their value may be derived simply and quickly. Stress testing the liability cash flows for market risk can then be accomplished by simply working with the simplified replicating assets, rather than having to rerun the full liability cash-flow-projection model. This can be very valuable for economic capital calculations, which tend to be complex but do not require absolute accuracy.

The replicating portfolio can also be seen as representing a hedge for the liabilities with respect to financial risk. Where non-financial risk is not material, the Solvency II technical provision for the liabilities can be taken as the market value of the RP. Otherwise, the market value of the RP can serve as the Solvency II best-estimate technical provision, and a suitable risk margin can then be calculated to reflect the non-financial risk.

The RP approach allows much faster calculation of the capital requirement for financial risk when full stochastic runs of the liability-cash-flow projection are replaced by the more straightforward recalculation of RP asset values. This is because such recalculation can typically utilise closed-form formulae under each stress scenario.

A potential problem with this approach may concern dynamic policyholder behaviour. Changes in policyholder behaviour are usually not straightforward to predict or to turn subsequently into assumptions that can be used in modelling and projecting the cash flows. Some models link certain aspects of policyholder behaviour to specific economic conditions. Therefore, each economic scenario used to produce liability cash flows has an associated set of policyholder reactions and behaviours which are used as assumptions in the projection of the cash flows.

Such an approach will necessarily be an approximation of actual observed behaviours, and the accuracy of the final result will depend on the completeness of the set of rules employed. One of the key challenges is to construct an appropriate set of rules applicable for the tails of the economic parameter distributions in extreme scenarios. Given the significant effect that a small movement in policyholder behaviour assumptions can have on the cash flows, it can also be difficult to construct accurate, fit-for-purpose rules and assumptions even in the 'mean' economic scenarios.

The RP can reflect only those specific assumptions of future policyholder behaviour that have been built into the cash-flow-projection runs used to generate the RP. In most cases, the specific pattern of behaviour assumed is therefore locked into the design of the RP, and there is little flexibility to test any deviation from this.

Where changes in policyholder behaviour are difficult to forecast, but at the same time have significant effects on the cash flows, the inflexibility of the RP approach in this regard represents a potential weakness of the methodology. While RP methodology can go a long way towards providing a market-consistent value for a set of liabilities, this will capture financial risk only. Other methods may then be required to analyse and reflect the remaining non-financial risk component.

MODEL COMPRESSION

Instead of focusing on a transformation of the results, an alternative approach to speeding up the calculations is to look at compressing the input data. The very large data sets that form policy portfolio extracts are obvious candidates for this approach.

Techniques for constructing a smaller set of representative model points have been used for many years. More modern approaches to such compression make use of clustering techniques developed in other areas of science to fine-tune the fit of the model points to the original data set.

Methods of clustering essentially work by associating each policy record with a vector and then analysing the distance between all vectors, i.e. the distance between the corresponding points in the vector space. The composition of the vector is set by the user and can include both original data items from the policy record and/or results from a few seriatim calibration runs carried out to obtain key results for each policy (e.g. present value of future profits). Some measure of the importance of each policy record must also be defined (e.g. sum assured or premium).

The clustering process then begins by finding the least important policy/vector that is closest to a neighbour. The first cluster is formed by merging these two policies/vectors. The process then continues recursively until the desired number of clusters has been formed. Each cluster effectively represents a model point and is scaled appropriately such that the resulting (small) set of model points produces results similar to those of the original (large) policy portfolio.

The construction of a sophisticated and reliable set of model points provides the user with a great deal of power, as the results of running complex cash-flow-projection models over numerous economic scenarios can then be obtained reliably and quickly. In this sense, some of the advantages of the RP approach are reproduced. However, the model compression approach is rather more powerful as it also allows fast calculation of sensitivities to non-financial variables (e.g. mortality), including more flexible handling of dynamic policyholder behaviour.

The model-compression approach also benefits from requiring only a handful of calibration scenarios to be run, whereas the RP methodology needs to execute runs over the full set of economic scenarios before the replicating asset set can be constructed.

Milliman has recently developed and implemented sophisticated cluster-modelling techniques to perform model compression. These techniques have generated impressive results.

Milliman has recently developed and implemented sophisticated cluster-modelling techniques to perform model compression. These techniques have generated impressive results, as the following illustration demonstrates.

Example

Cluster modelling was applied to a seriatim block of 210,000 in-force variable annuity (VA) policies to produce a compressed model of just 250 cells. It is important to note that VA business is usually especially difficult to compress to this extent (a ratio of 840-to-1), given the nature of the business–a classic model of this particular portfolio might typically contain, say, 9,000 cells.

The process of model compression begins with splitting the policy portfolio into appropriate segments. Suitable key data items and results are then chosen to make up the vectors that will drive the clustering (here these include, for example, account value, present value of benefits, and present value of profits).

The seriatim model is then run twice (under each of two artificial calibration scenarios) to generate values of the key results for each in-force policy and hence populate the associated vectors. The clustering algorithm is then applied iteratively to the vectors to progressively reduce the model-point set while maintaining a fit of the values of the key results. The compression can be halted at any level and here was stopped once the model had been reduced to 250 cells.

Results produced by running the 250-cell compressed model may be compared to those of a 9,000-cell classic model of the same business. A comparison was made by running each model over 1,000 economic scenarios, and Table 1 sets out the results at various conditional tail expectation (CTE) levels. The particular result being compared is the ending surplus after 30 years.

TABLE 1: 250-CELL CLUSTER MODEL COMPARED TO 9,000-CELL CLASSIC MODEL (USD MILLION)

VA MODEL CTE OF 30-YEAR TERMINAL SURPLUS

CTE LEVEL	250-CELL CLUSTER	9,000-CELL CLASSIC	DIFFERENCE
0%	965	1058	93
50%	359	427	68
65%	159	218	59
70%	-28	21	49
90%	-548	-515	33
95%	-981	-939	42
99%	-1909	-1879	30

Source: Milliman Research Report "Cluster Analysis: A spatial approach to actuarial modelling" (Freedman, Reynolds)

Table 1 shows that the 250-cell model here produces lower results than the 9,000-cell model. Interestingly, however, most of this error is in the 9,000-cell model: running the original seriatim model produces results around $70 million lower than the 9,000-cell model and thus closer to those of the 250-cell model.

It can be noted that the cluster-modelling approach used here required only two calibration runs of the full model in order to generate a compressed model capable of closely reproducing the seriatim model under the full range of scenarios. This can be contrasted with the RP approach where the methodology requires calibration runs over all economic scenarios in order to generate the replicating asset set.

Further details on Milliman's innovative model-compression techniques may be found in the Milliman Research Report "Cluster analysis: A spatial approach to actuarial modelling" (Freedman, Reynolds).

CONCLUSION

The computational implications of Solvency II are likely to be significant, especially for those companies that have not yet taken steps to adapt their systems to carry out the type of calculations required. New and more extensive data are demanded (e.g. for policy information and setting assumptions), and extraction of these data from more traditional administration systems can often prove to be problematic.

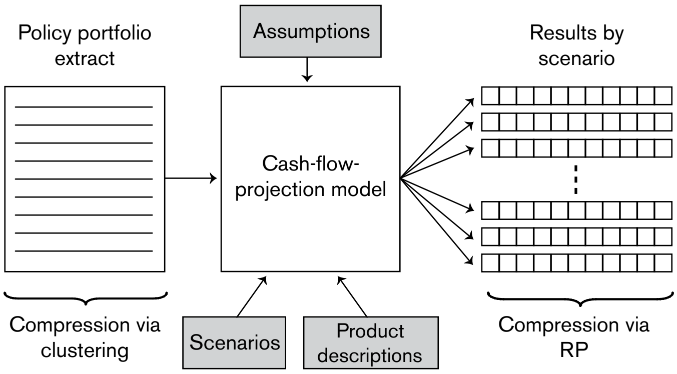
Above all, the projection of best-estimate cash flows, risk margins, and capital requirements can require sophisticated software tools, seriatim policy data, and stochastic (or nested stochastic) projections, giving rise to impractical runtimes. This short paper has given a brief overview of two software techniques that are being developed (along with others) to face the computational performance challenges of Solvency II and similar financial projection applications (e.g. market-consistent embedded value) to allow fast production of reliable results and sensitivities.

Replicating-portfolio techniques are enjoying rising popularity and offer both reduced runtimes and a direct link with the Solvency II categorisation of liabilities as hedgeable or non-hedgeable. However, the resulting replicating-portfolio asset set locks in a particular view of non-financial variables (and in particular, policyholder behaviour). Also, a calibration overhead exists whereby the original model must be run over all economic scenarios to generate the input to the optimisation algorithm that constructs the replicating portfolio.

Model-compression techniques based on clustering also offer significantly reduced runtimes. However, by focusing on a compression of the policy portfolio data, the sensitivity to non-financial elements can be retained and hence tested and analysed. Also, highly compressed models providing an excellent fit to original portfolios require only a handful of calibration runs to be performed.

Figure 1 illustrates how the clustering approach, by operating at the input stage to the cash-flow-projection model, can offer greater flexibility for the rapid testing of non-financial sensitivities. RP techniques operate at the opposite (results) end of the process and thus incorporate a particular view of non-financial variables. A fresh set of stochastic runs is therefore needed for in-depth testing of these variables.

FIGURE 1: APPLICATION OF COMPRESSION TECHNIQUES



It is interesting to note that these techniques are not incompatible with each other, given that in effect they apply at distinct stages of the calculation process. In fact, both approaches could potentially be deployed in tandem, with the cash-flow projections generated by a compressed model then serving as input to the process of construction of a replicating portfolio. In this way, advantage might be taken of the strengths of both approaches. We also note that both RP techniques and cluster analysis are applicable not only to policy (i.e. liability) portfolios, but also to asset portfolios, further reducing runtimes.

Last, both approaches are specific examples of a more general set of techniques for compressing models. All modelling carries the risk that the final model may miss real risk factors (as has been dramatically illustrated by recent events in the financial markets), and compressing models will tend to exacerbate this possibility. In this sense, clustering analysis has some advantages thanks to the flexibility of the approach in setting the level of detail to be retained.

All modelling carries the risk that the final model may miss real risk factors (as has been dramatically illustrated by recent events in the financial markets), and compressing models will tend to exacerbate this possibility. In this sense, clustering analysis has some advantages thanks to the flexibility of the approach in setting the level of detail to be retained.

Milliman has wide experience in helping companies across Europe make the change from the realm of more traditional insurance calculations and systems to modern economic risk-and value-based approaches for reserving capital and measuring performance. This change brings many benefits, not only easing the transition to Solvency II, but also providing management with much greater insight into the risks, rewards, and opportunities of the business, thereby optimising competitiveness.

For more information, contact your local Milliman consultant.

UK:

Life – Oliver Gillespie at +44 (0)20 7847 1541

P&C – Gary Wells at +44 (0)20 7847 1607

Netherlands and Belgium:

Life – Henny Verheugen at +31 (0)6 10149938

P&C – Joël van der Vorst at +31 (0)6 51 39 60 49

Germany:

P&C – Jeff Courchene at +49 (0)89 5908 2392

Southern Europe:

Life – Dominic Clark at +34 91 789 3470

The materials in this document represent the opinion of the authors and are not representative of the views of Milliman, Inc. Milliman does not certify the information, nor does it guarantee the accuracy and completeness of such information. Use of such information is voluntary and should not be relied upon unless an independent review of its accuracy and completeness has been performed. Materials may not be reproduced without the express consent of Milliman.

Copyright © 2008 Milliman, Inc.

Paseo de la Castellana, 141 P. 18-20
28046 Madrid, Spain
+34 91 789-3470

www.europe.milliman.com