SOCIETY OF ACTUARIES

Article from:

ARCH 2013.1 Proceedings

August 1- 4, 2012

Yvonne C. Chueh, Paul H. Johnson

Small Sample Stochastic Tail Modeling: Tackling Sampling Errors and Sampling Bias by Pivot-Distance Sampling and Parametric Curve Fitting Techniques

Paul H. Johnson, Jr.[1], Yvonne C. Chueh[2], and Yongxue Qi[3]

Abstract

We describe two original open source software applications that have been developed to aid model efficiency studies: (1) CSTEP (Cluster Sampling for Tail Estimation of Probability) for reducing sampling error through variations of distance sampling and cluster/pivot processes; and (2) AMOOF2 (Actuarial Model Outcome Optimal Fit Version 2.0) for mitigating small sample bias in parametric, time-efficient probability density function fitting. CSTEP uses the scenario reduction method of representative scenarios to sample scenarios from a population of stochastic scenarios to obtain a sample-run distribution of a financial outcome that can be analyzed by AMOOF2 to fit the optimal probability density function.

Introduction

Actuaries often have to analyze the probability distributions of critical financial outcomes arising from a large population of data, such as insurance policies. In order for an actuary's analyses to be in line with their company's risk management criteria, to satisfy regulatory mandates, and to analyze cash flows for the purposes of setting loss reserves and maintaining their company's solvency, these analyses need to be accurately conducted within a relatively short amount of time. A time-effective and accurate analysis of the probability distribution of a financial outcome associated with a large population of data is referred to as model efficiency (Chueh 2002).

Stochastic asset and liability models are often used by actuaries to generate numerous stochastic scenarios, which in turn can generate numerous possible realizations of the desired financial outcome. For example, using an interest rate generator, various stochastic scenarios of one-year forward interest rates, $(i_1, i_2, \ldots, i_h)'$ can be generated. Each stochastic interest rate scenario represents a possible rate path over the next h years. Using a transformation from h-dimensional to one-dimensional space, each stochastic interest rate scenario can generate a financial outcome, such as the surplus for a specific block of business at the end of the next h years. A well-known challenge with this approach is that each transformation from n-dimensional to one-dimensional space can

[1] Assistant Professor; University of Illinois at Urbana-Champaign, Department of Mathematics; 1409 W. Green St.; Urbana, IL 61801; pjohnson@illinois.edu. **Corresponding author.**

[2] Professor; Central Washington University, Department of Mathematics; 400 E. University Way Mail Stop 7424; Ellensburg, WA 98926; chueh@cwu.edu.

[3] Graduate Student; University of Illinois at Urbana-Champaign, Department of Statistics; yqi2@illinois.edu.

take a substantial amount of time. Furthermore, to generate enough realizations of the financial outcome to accurately determine its full-run distribution will come at the cost of an extremely long run time. Obtaining the full-run distribution of the financial outcome for further analysis is often not a model-efficient approach. For this reason, the full-run distribution is often considered unobservable.

One way to mitigate the long run time, and attain model efficiency, is to use a scenario reduction approach (Rosner 2011). An actuary can obtain a sample of scenarios from the population of stochastic scenarios, generate the associated financial outcomes, and analyze the sample-run distribution of the financial outcome in lieu of the full-run distribution of the financial outcome. This approach will have the benefit of greatly reducing run time. However, the use of the sample of scenarios risks introducing sampling error and bias, so that metrics obtained from analysis of the sample-run distribution, such as the mean, the value-at-risk (VaR), or the conditional tail expectation (CTE) (Klugman et al. 2008), may not be close in value to the metrics that would be obtained from analysis of the full-run distribution.

In this article, we briefly describe two original open source software applications that have been developed to aid current and future model efficiency studies: (1) CSTEP (Cluster Sampling for Tail Estimation of Probability) for reducing sampling error through variations of distance sampling and cluster/pivot processes (Central Washington University 2011); and (2) AMOOF2 (Actuarial Model Outcome Optimal Fit Version 2.0) for mitigating small sample bias in parametric, time-efficient probability density function fitting (Central Washington University 2012). CSTEP uses the scenario reduction method of representative scenarios to sample scenarios from a population of stochastic scenarios such that metrics of the derived sample-run distribution accurately reflect metrics of the full-run distribution, particularly at the tails of the distribution so that accurate values of metrics like VaR and CTE can be obtained (Chueh 2002, Chueh and Johnson 2012). The sample-run distribution obtained using CSTEP can be analyzed by AMOOF2 to fit the optimal probability density function, the curve with the best goodness-of-fit, out of a large list of single and mixed probability distributions using the Cox and Snell/Cordeiro and Klein (CSCK) method of bias corrected maximum likelihood estimation. Using bias corrected maximum likelihood estimation, estimated parameters of the fitted probability density function are robust to bias that might arise from using a small sample of stochastic scenarios (Cox and Snell 1968, Cordeiro and Klein 1994, Johnson et al. 2012). We assert that the tandem use of CSTEP and AMOOF2 will trade valuable time that an actuary might usually reserve for sample selection and probability density function fitting for time spent on actuarial analytics and model refinement.

CSTEP

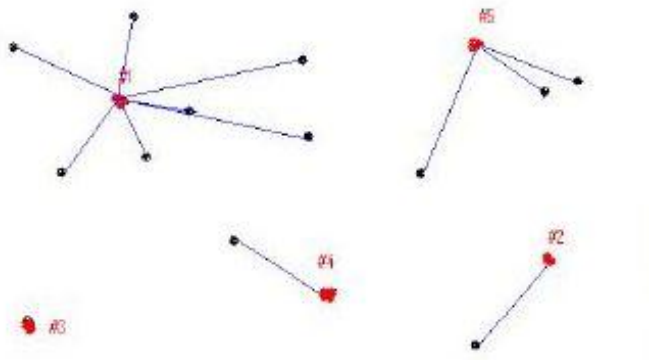
CSTEP is a desktop-based, high computation, open source software that was developed by the Mathematics and Computer Science departments at Central Washington University. CSTEP uses the method of representative scenarios to allow actuaries to sample scenarios from a population of stochastic scenarios (Chueh 2002). Use of CSTEP allows an actuary to construct a sample-run distribution of the financial outcome from the sample of scenarios that closely resembles the full-run distribution of the financial outcome, particularly at the tails of the distribution. This method allows for a model efficient analysis of data. CSTEP also allows actuaries to research more effective sampling techniques for a block of business.

CSTEP was previously discussed at length in Chueh and Johnson (2012); please refer to this reference for more details regarding the program's possible practical uses, sampling algorithms, and CSTEP screen captures. Central Washington University (2011) provides a url that contains the CSTEP program, free to download, with supporting documentation. In this article, we provide only a brief discussion of CSTEP.

CSTEP allows for a population of 8,388,608 stochastic scenarios where each scenario can have up to 4500 time periods. We believe that this is adequate for most real populations of interest to actuaries. CSTEP allows for a flexible sample size of scenarios, as well as reversible and reusable sampling. An actuary can obtain a specific sample of scenarios, save that sample, and then re-sample scenarios to obtain multiple sample-run distributions of the financial outcome. In order to use CSTEP, each stochastic scenario must be a vector of periodic rates such as interest rates, equity returns, or some sort of standardized index. This is not a major limitation, as many financial outcomes of interest can be generated from stochastic rate paths, such as ending surplus or loss reserves.

CSTEP utilizes the method of representative scenarios. Suppose the population consists of N stochastic rate paths. Editable distance formulas similar to Euclidian distance are used to select a sample of n representative, or pivot, scenarios, where n is substantially less than N. This algorithm is rigorously described in Chueh (2002). Informally, one rate path is randomly chosen out of the population of N rate paths. The chosen rate path is designated Pivot 1: the first representative scenario. The "distances" of each of the remaining N – 1 rate paths to Pivot 1 are determined, and the rate path with the largest distance is designated Pivot 2: the second representative scenario. Each of the remaining N – 2 rate paths are assigned to the closest, in terms of distance, of Pivot 1 and Pivot 2 forming two disjoint clusters. This process is repeated to obtain Pivot 3, Pivot 4, …, Pivot n; and thus, n clusters of rate paths. A probability is assigned to each representative (pivot) scenario equal to the number of rate paths in the cluster divided by N. The general representative scenario algorithm just described is pictorially illustrated in Figure 1, with 5 representative (pivot) scenarios denoted as red dots and other scenarios denoted as black dots. Note that this representative scenario algorithm is designed to choose extreme stochastic scenarios, and consequently, extreme financial outcomes as was proven in Chueh (2002).

Figure 1: CSTEP: Pictorial Illustration of the General Representative Scenario Algorithm



CSTEP employs many distance formulas to determine the distance between pairs of stochastic scenarios. These distance formulas are provided and discussed in Chueh (2002) and Chueh and Johnson (2012). For example, under the significance method, there is only one representative (pivot) scenario: a scenario in which all of the rates are equal to zero. The distance from each stochastic scenario in the population to this zero scenario is determined, and is called the "significance". All N scenarios are then sorted by significance in ascending order, and a sample of n scenarios is uniformly chosen from the sorted list of population scenarios. Thus, the scenarios corresponding to evenly marked percentiles will be chosen, resulting in central metrics of the sample-run distribution (mean, median, variance) closely matching the corresponding central metrics of the full-run distribution. Another distance measure is based on economic present value. This method defines a more general distance between each pair of

stochastic scenarios. The economic present value method utilizes cash flows for each time period which allow actuaries to incorporate the asset runoff speed of a specified block of business into the distance measure. By studying the asset runoff to assign proper values to the cash flows, it is more likely that extreme stochastic scenarios with be associated with extreme financial outcomes, resulting in tail metrics of the sample-run distribution (VaR, CTE) closely matching the corresponding tail metrics of the full-run distribution. Actuaries can try various cash flow values in the economic present value distance formula to find the precise association with the model financial outcome distribution.

AMOOF2

AMOOF2 is high computation, open source software that is currently being developed by the Mathematics and Computer Science departments at Central Washington University and the Mathematics department at the University of Illinois at Urbana-Champaign. AMOOF2 uses a method of bias corrected maximum likelihood estimation to allow actuaries to fit the optimal probability density function to a sample-run distribution of a financial outcome obtained using CSTEP. Use of AMOOF2 allows an actuary to parametrically calculate important metrics of the sample-run distribution of the financial outcome, such as the mean, VaR, and CTE. AMOOF2 is designed to aid in efficient stochastic modeling of financial outcomes, in areas such as capital determination, reserve setting, risk analysis, and product pricing.

We will briefly discuss the key features of AMOOF2 in this article. Central Washington University (2012) provides a url that contains the current release of the AMOOF2 program itself, free to download. It should be noted that AMOOF2 is still a work in progress; we are currently refining the program to output the most accurate and efficient possible results for use in model efficient stochastic modeling.

AMOOF2 is a continuation of the Actuarial Model Optimal Outcome Fit Project that was originally discussed in Chueh and Curtis (2005). AMOOF2 is a stand-alone desktop suite that is linked to Microsoft Excel (Microsoft 2010) and also incorporates formulas obtained using Mathematica 8.0 (Wolfram Research Inc. 2010). AMOOF2 contains several tabs to guide an actuary through analysis of a sample-run distribution of a financial outcome. Figure 2 shows the first tab available to an actuary upon initialization of AMOOF2, the “Access Data” tab. Using this tab, an actuary can calculate various empirical raw moments of the sample-run distribution, and obtain values of VaR and CTE at various percentages/security levels. A histogram of the sample-run distribution is also generated on this tab. The actuary will also have the option of further adjusting the inputted sample data prior to further analysis, such as by removing negative values or applying a transformation to each data point.

Figure 3 shows the “Model Selection” tab. An actuary can fit candidate probability density functions to the sample-run distribution of the financial outcome using any of the 22 single probability distributions shown in Figure 3. All probability distribution functions follow the parameterizations of Klugman et al. (2008). In future releases of AMOOF2, it will be possible for an actuary to fit a probability density function that is a two-point mixture of any of the 22 single probability distributions.

Figure 2. AMOOF2: Access Data

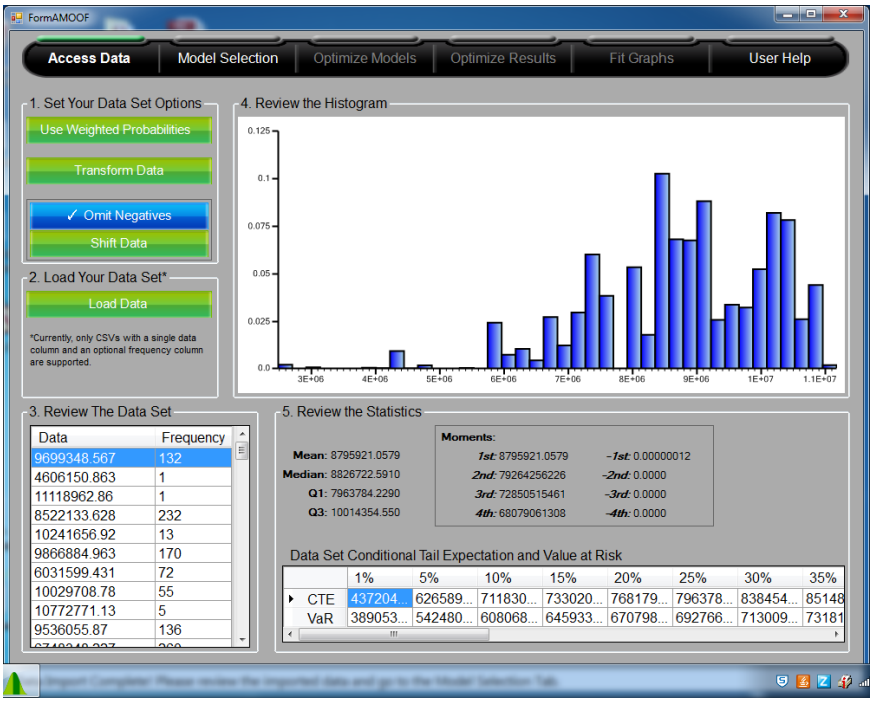
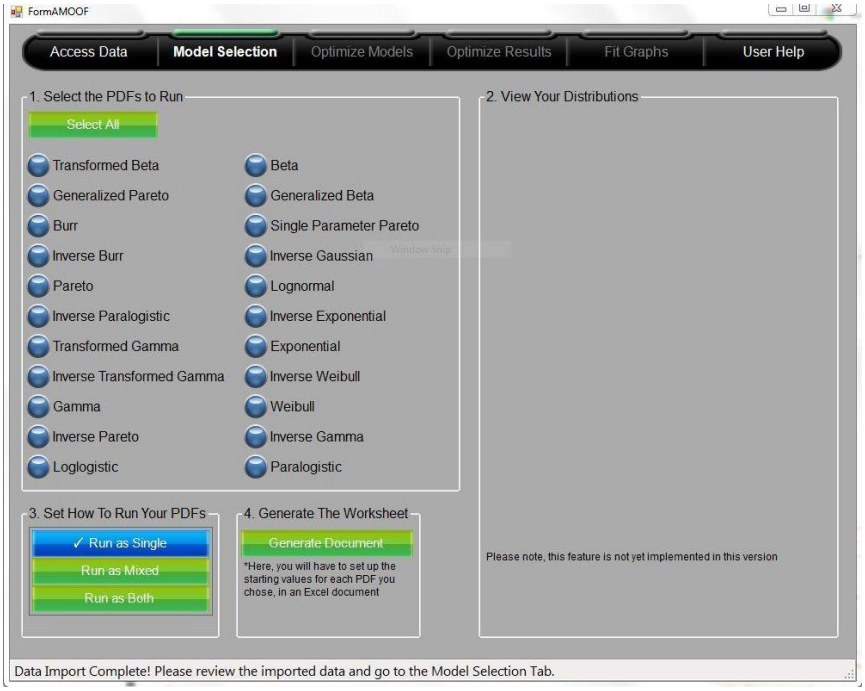


Figure 3. AMOOF2: Model Selection



The parameters of a fitted probability density function are initially estimated using the method of maximum likelihood. For distributions where there are analytic formulas for the maximum likelihood estimators, such as the lognormal distribution, those formulas are used to obtain maximum likelihood estimates (MLEs). For distributions where there are no closed form solutions for the maximum likelihood estimators, such as the transformed gamma distribution, Microsoft Excel's Solver function is used to maximize the loglikelihood function and obtain the MLEs, where the method of moments estimates of the parameter values are used as starting values for Solver.

Maximum likelihood estimators possess highly desirable properties such as asymptotic unbiasedness, consistency, and asymptotic normality. However, many of these properties, such as unbiasedness, may not be valid for small sample sizes. While "small" is a subjective term, we use it to refer to data with 100 or fewer observations. This might be an issue for the method of representative scenarios discussed in the CSTEP section if an actuary collects a small sample of representative scenarios to obtain the sample-run distribution of the financial outcome.

To mitigate the impact of small sample bias on the quality of MLEs obtained using AMOOF2, we allow an actuary to implement the Cox and Snell/Cordeiro and Klein (CSCK) method of bias corrected maximum likelihood estimation (Cox and Snell 1968, Cordeiro and Klein 1994). This method is described in detail in Johnson et al. (2012). Essentially, the CSCK method is a corrective approach to bias adjustment of MLEs obtained using small sample data. Initial, potentially biased, MLEs are obtained using the usual method of maximum likelihood. Then, a CSCK analytic MLE bias term, that is a function of both the initial MLEs and several cumulants of the probability distribution, is subtracted from each MLE to obtain a bias corrected maximum likelihood estimate (BMLE). Therefore, for a specific distributional parameter: BMLE = MLE – CSCK Analytic MLE Bias Term.

Future releases of AMOOF2 will contain CSCK Analytic MLE Bias terms for 15 out of the 22 probability distributions in AMOOF2. These CSCK Analytic MLE Bias terms were obtained using a Mathematica 8.0 program that was developed by the Mathematics department at the University of Illinois at Urbana-Champaign called "CSCK MLE Bias Calculation"; this program is available in Johnson et al. (2012). For 7 out of the 22 probability distributions in AMOOF2, we could not obtain CSCK Analytic MLE Bias terms, as Mathematica 8.0 could not calculate the required cumulants. However, when possible, an actuary will be able to use AMOOF2 and the provided CSCK Analytic MLE Bias terms to calculate BMLEs for their fitted distributions, and obtain unbiased parameter estimates for a sample-run distribution generated from a small sample of stochastic scenarios.

Once all of the candidate probability density functions have been fit, a Microsoft Word file is generated that summarizes key statistics for each probability distribution. These statistics include the differences between the fitted distribution's first four positive and negative parametric raw moments and the empirical raw moments of the sample-run distribution of the financial outcome, as well as the maximized loglikelihood value for each candidate distribution.

An actuary can fit the optimal probability density function using the "Optimize Models" and "Optimize Results" tabs, provided in Figure 4 and Figure 5, respectively. The red lights in Figure 4 indicate that the optimization process failed for the associated candidate probability density functions and that the solution diverges, while the green lights indicate the optimal solutions from Excel's Solver have successfully been found. Candidate probability density functions can be evaluated for the best goodness-of-fit by comparing maximized loglikelihood values, or the probability model's parametric raw moments to the empirical raw moments of the sample-run distribution. Parametric values of VaR and CTE can be determined for each candidate probability distribution using high precision Riemann sums (Gaussian Quadrature Integration) that can be compared to the empirical VaR and CTE of the sample-run distribution. Actuaries may input their own distributional parameter values to recalculate and

compare the new tail metrics after the bias correction is made. The interactive fitted density curve will show how well the data fits the candidate probability distribution. The green integral button in Figure 5 can be clicked anytime to calculate the parametric distribution metrics for each candidate probability density function on the list.

Figure 4. AMOOF2: Optimize Models

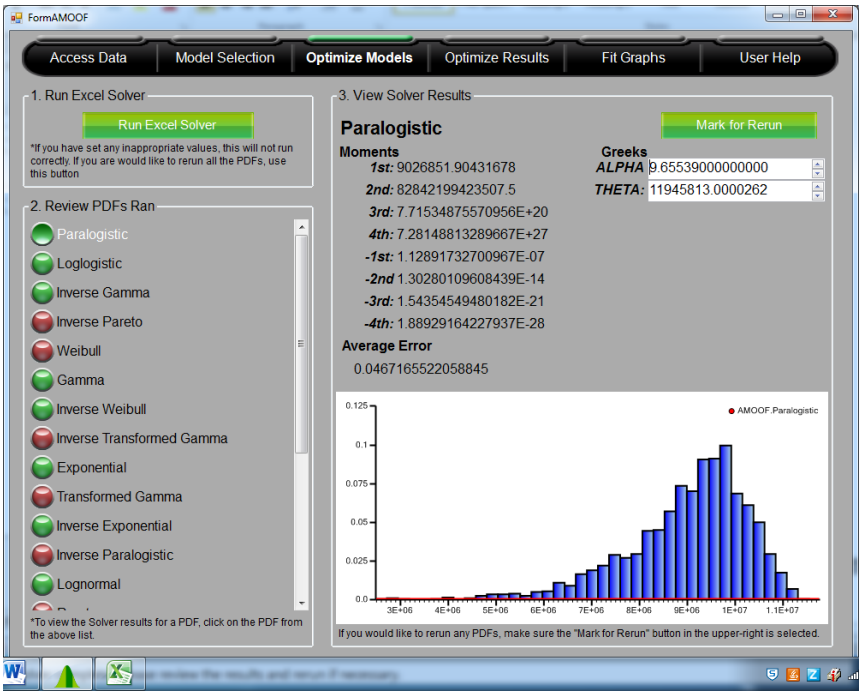
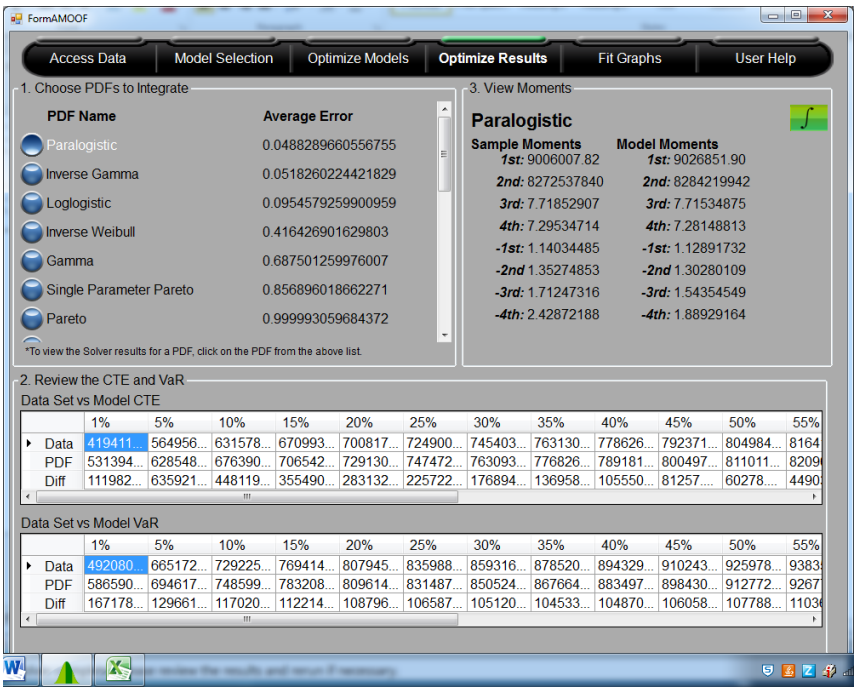


Figure 5. AMOOF2: Optimize Results



Concluding Comments

We have briefly described two original open source software applications that have been developed to aid model efficiency studies: (1) CSTEP for reducing sampling error and (2) AMOOF2 for mitigating small sample bias. CSTEP automates the method of representative scenarios, allowing an actuary to quickly obtain a sample-run distribution of the financial outcome of interest with metrics that closely resemble those of the full-run distribution of the financial outcome, particularly at the tails. The sample-run distribution can then be further analyzed via AMOOF2, which quickly calculates metrics of the sample-run distribution and fits various candidate probability density functions, allowing an actuary to determine the optimal parametric probability model.

There are many advantages to using CSTEP to obtain samples of scenarios. CSTEP uses the method of representative (pivot) scenarios which captures extreme scenarios that replicate the tail of a probability distribution, making modeling more efficient. CSTEP allows for the use of negative interest rates and expands the distance formulas for actuaries to adjust for actual stochastic cash flows, such as by using an original approach: the economic present value method. CSTEP also contains an enhanced graphical user interface and extensive packaged online documentation to guide actuaries intuitively. Finally, CSTEP is not simply a sampling tool; it is designed for additional research as well. Actuaries can edit the distance formulas between scenarios to gain additional insight regarding the nature of the sample-run (and full-run) distribution of the financial outcome.

There are many advantages to using AMOOF2 to fit probability density functions to data. AMOOF2 analyzes unequal frequencies of one variable data sets produced by highly complicated real world company cash flow models using both maximum likelihood estimation and the method of moments. More key information on the parametric probability tail distribution behavior can be obtained using parametric integration calculus and numerical analysis that used to be time prohibitive. As was previously discussed, AMOOF2 allows an actuary the option of using small sample analytic CSCK MLE bias correction formulas to calculate BMLEs, which are robust to small sample bias. Finally, AMOOF2 implements a new graphical user interface and Integrate PARI/GP, a computer algebra system with the main aim of efficiently facilitating number theory computations.

CSTEP and AMOOF2 are both copyrighted, but are available as free downloadable, open-source software tools for research and implementation. We believe that CSTEP and AMOOF2 can further facilitate research collaborations between both industry and academic actuaries, and advance the development of even more powerful, sophisticated model efficiency methods and tools to empower stochastic modeling of critical financial outcomes for applications such as competitive pricing, innovative product design, setting loss reserves, testing cash flows, evaluating asset adequacy, and budgeting risk capital.

Acknowledgments

The authors would like to acknowledge the Central Washington University Computer Science department, its chair Dr. James Schwing, and all the senior computer science student teams who contributed their software engineering expertise and timely support to the authors' software projects. The authors also acknowledge Mathematica expert Dr. Dan Curtis for his help in starting the Actuarial Model Optimal Outcome Fit Project. Yvonne would also like to acknowledge the University of Connecticut for providing privileged educational and research opportunities, her former professors and advisors Dr. Charles Vinsonhaler, ASA and Dr. Jeyaraj Vadiveloo, FSA, MAAA, CFA, for their professional inspiration and research vision along with Alastair Longley-Cook. They motivated the engineering of CSTEP and AMOOF2. Special thanks to modeling consultant Edward F. Cowman, FSA, MAAA, for his

extraordinary real model data contribution and the valuable testing assistance. All three authors especially appreciate the funding awarded by The Actuarial Foundation to support their challenging team coordination and timeline.

References

1. Central Washington University. 2011. "Cluster Sampling for Tail Estimation of Probability (CSTEP)". http://www.cwu.edu/~chueh/.

2. Central Washington University. 2012. "Actuarial Model Optimal Outcome Fit 2.0 (AMOOF2)". http://amoof.amp-software.net/download.php.

3. Chueh, Y. C. 2002. "Efficient Stochastic Modeling for Large and Consolidated Insurance Business: Interest Sampling Algorithms". *North American Actuarial Journal* 6(3): 88-103.

4. Chueh, Y. C. and W. D. Curtis. 2005. "Optimal Probability Density Function Models for Stochastic Model Outcomes: Parametric Model Fitting on Tail Distributions". *International Mathematica Conference*: Banff, CA.

5. Chueh, Y. C. and P. H. Johnson Jr. 2012. "CSTEP: A HPC Platform for Scenario Reduction Research on Efficient Stochastic Modeling -- Representative Scenario Approach". *Actuarial Research Clearing House* 2012.1: 1-12.

6. Cordeiro, G.M. and R. Klein. 1994. "Bias Correction in ARMA Models". *Statistics and Probability Letters* 19: 169- 176.

7. Cox, D.R. and E. J. Snell. 1968. "A General Definition of Residuals". *Journal of the Royal Statistical Society* B(30): 248- 275.

8. Johnson Jr., P. H., Y. Qi, and Y. C. Chueh. 2012. "CSCK MLE Bias Correction." Fully functional Mathematica report, part of the AMOOF2 project between Central Washington University and the University of Illinois at Urbana-Champaign: http://www.math.uiuc.edu/~pjohnson/CSCK_MLE_Bias_Calculation_063012.pdf.

9. Klugman, S. A., H. H. Panjer, and G. E. Willmot. 2008. "Loss Models: From Data to Decisions, Third Edition". Hoboken, NJ: Wiley.

10. Microsoft. 2010. "Microsoft Excel [computer software]". Redmond, WA: Microsoft.

11. Rosner, B.B. 2011. "Model Efficiency Study Results". Research Projects – Life Insurance, Society of Actuaries.

12. Wolfram Research Inc. 2010. "Mathematica Edition: Version 8.0 [computer software]". Champaign, IL: Wolfram Research, Inc.