INTERPRETATIONS OF SEMI-PARAMETRIC MIXTURE MODELS, UNBIASED ESTIMATORS OF ULTIMATE VALUE FOR INDIVIDUAL CLAIMS AND CONDITIONAL PROBABILITY APPLICATIONS TO CALCULATE BULK RESERVES

by Rajesh Sahasrabuddhe, FCAS, MAAA

Abstract

Semi-parametric mixture models have well documented technical advantages for modeling loss distributions. These technical advantages are documented in papers that focus on the estimation of the parameters of semi-parametric models.

This paper assumes that the parameters have already been determined and then provides an interpretation of the results of the parameter estimation. This interpretation is intended to make semi-parametric models intuitively appealing. If we accept this interpretation of the parameters, then we can use conditional probability concepts to calculate bulk reserves either deterministically or in a stochastic framework.

1. Introduction

A recent paper by Keatinge[1] discussed the virtues of semi-parametric mixture models vis-à-vis (fully) parametric models and non-parametric (empirical) models. The advantages discussed in Keatinge focus on the attractive compromise between smoothing and data responsiveness offered by semi-parametric models. Semi-parametric models have the following density function:

$$f(x) = w_1 \times f_1(x) + w_2 \times f_2(x) + \ldots + w_n \times f_n(x)$$

where:

i. $f(x)$ represents the probability density function for the mixture model,
ii. $f_i(x)$ represents the probability density function for the i-th component of the mixture, and
iii. w_i represents the mixing weight corresponding to i-th component of the mixture.

Furthermore, the mixing weights are subject to the constraints that:

i. $w_i > 0$
ii. $\sum w_i = 1$.

[1] Keatinge, Clive L., "*Modeling Losses with the Mixed Exponential Distribution,*" Proceedings of the Casualty Actuarial Society 1999 Vol: LXXXVI Page(s): 654-698, Casualty Actuarial Society: Arlington, Virginia

The mixture is considered semi-parametric since each component of the mixture is a parametric model but the distribution of mixing weights is model free. Also, it should be noted that there is no restriction that the components of the mixture have the same model form (e.g. exponential, lognormal, Pareto) or, for that matter, any specific model form.

The remainder of this paper assumes that model forms and parameters have already been determined.

2. Interpreting Semi-parametric Models

While Keatinge's arguments are certainly persuasive, there may be a more important argument supporting the use of semi-parametric models: they are intuitively appealing.

Specifically, it is reasonable to assume that loss experience is comprised of observations from a discrete number of underlying loss processes. The table below provides some examples:

	Coverage			
	Auto Liability	**Workers Compensation**	**Medical Malpractice**	**Homeowners**
Loss Processes	Property Damage Only	Indemnity (Lost Time)	Nuisance	Theft
	Bodily Injury and Property Damage	Indemnity and Medical	Expense Only (Dismissals)	Partial Property Loss
		Death	Settlement without litigation	Total Loss of Property and Contents
			Settlements with litigation	

Under this assumption, it is then reasonable to interpret the mixing weights (w_i) as the percentage of total claims that are generated by each loss process. Logically, the named loss processes must therefore be exhaustive[2]. That is, all claims must fall into one of these categories and the sum of the probabilities associated with the loss processes must equal 1. It would then follow that the components of the mixture would be interpreted as describing the distribution of claims amounts resulting from each loss process.

Although in many cases the loss process is coded in the claims record[3], this is not always the case. The table above is meant to provide examples of types of multiple loss processes that might produce the observed claim distribution That is, we assume that the

[2] . Since we can define the loss process as narrowly or broadly as we desire, we are not concerned that a mixture model would be required to describe a single loss process nor are we concerned that multiple loss processes would be described by distributions that are not significantly different.

[3] For example, auto liability loss records will often indicate whether the loss is for bodily injury or property damage.

mixture model identifies all significantly different underlying loss processes[4]. In addition, it should be noted that there is no requirement that we identify the loss processes by name. In fact it may not be possible to identify each and every loss process.

3. Using Conditional Probability Concepts to Estimate Claim Level Bulk Reserves

The term "bulk reserve" is used to represent the reserve for development on known (or reported) losses and is often referred to as the "incurred but not enough reported" ("IBNER") reserve. As is common in analysis of claims-made coverages, the bulk reserve is often estimated in the aggregate for a body of claims. For occurrence coverages, the bulk reserve and the reserve for unreported claims are often estimated on a combined basis.

The mixture model represents the overall distribution of claim values *without any prior knowledge*. However for reported claims, we will have some knowledge about each claim. This "knowledge" will generally include amounts paid to date and case basis reserves.

For purposes of this discussion, it should be assumed that an unbiased estimator for the ultimate value of each claim is available. (We will return to this assumption in the next section.) This unbiased estimator will be denoted *U.*

We now focus on the likelihood of the various loss processes generating a claim of size *U.* What we are really concerned with is not the absolute probabilities but rather the relative probabilities. Recall the conditional probability relationship:

$$\Pr(A_i \mid B) = \frac{\Pr(A_i \cap B)}{\Pr(B)} \text{ where}$$

A_i represents the event that loss process i underlies the claim and
B represents event that the unbiased estimate of the claim is equal to U.

The relative probabilities i.e. Pr $(A_1|B)$, Pr $(A_2|B)$ ... Pr $(A_n|B)$ are proportional to:

$$\Pr(A_i \cap B) = \Pr(A_i \mid B) \times \Pr(B) = \Pr(B \mid A_i) * \Pr(A_i).$$

In this case we decide to use the second expression. Under the interpretation offered in Section 2:

$\Pr(A_i) = w_i$ and
$\Pr(B|A_i)$ is proportional to f_i(U).

So we can restate the mixture model with adjusted mixing weights defined as follows:

$$\tilde{w}_i = \Pr(A_i) \times \Pr(B \mid A_i) \propto w_i \times f_i(U). \tag{2.1}$$

[4] To the extent that those differences are represented in the data, of course.

where the ~ indicates that the parameter (or distribution) is applicable to an individual claim and has been adjusted to consider the unbiased estimator of the ultimate value of that claim.

Normalizing mixture weights, we obtain:

$$\tilde{w}_i = w_i \times f_i(U) / \sum_i w_i \times f_i(U) . \tag{2.2}$$

The conditional density function is then equal to:

$$f(x \mid U) = \tilde{f}(x) = \tilde{w}_1 \times f_1(x) + \tilde{w}_2 \times f_2(x) + \ldots + \tilde{w}_n \times f_n(x) . \tag{2.3}$$

The table below provides an example of how mixing weights adjust for various U values for a mixture of 3 component lognormal models.

i	1	2	3	Total
mu$_i$	9	10	12	
sigma$_i$	1.5	1.75	1.5	
Initial Mixing Weight (w_i)	50%	20%	30%	100%
Mean	24,959	101,849	501,320	183,246
Standard Deviation	72,716	459,801	1,460,532	
Example #1				
Unbiased Estimator (U)	10,000			
$f_i(U)$	2.63E-05	2.06E-05	4.72E-06	
Adjusted Mixing Weight	70%	22%	8%	100%
Mean	24,959	101,849	501,320	77,943
Example #2				
Unbiased Estimator (U)	150,000			
$f_i(U)$	2.67E-07	8.33E-07	1.77E-06	
Adjusted Mixing Weight	16%	20%	64%	100%
Mean	24,959	101,849	501,320	344,701
Example #3				
Unbiased Estimator (U)	750,000			
$f_i(U)$	3.73E-09	3.98E-08	2.11E-07	
Adjusted Mixing Weight	3%	11%	87%	100%
Mean	24,959	101,849	501,320	445,681

In particular, readers should observe how the mixing weights shift given the unbiased estimator for the claim and the means and standard deviations of the components of the mixture.

Readers will also note that the mean value for each claim is not equal to the unbiased estimator. This is because the process is designed to produce individual distributions that, taken together, describe the distribution of a portfolio of claims. The process is not necessarily appropriate for any individual claim. (As discussed in the following section, determination of an unbiased estimator for a claim may also not be possible.) That is, as with most actuarial techniques, the predictive value of the results requires sufficient credibility of the claims portfolio being modeled. Finally, after reviewing the relative stability[5] of the distribution parameters and the stability of the unbiased estimator, it may be appropriate to balance the results of the two indications.

4. Calculation of Bulk Reserves

With this adjusted density function, we can calculate bulk reserves either deterministically or stochastically. The deterministic estimate is simply calculated by integrating and subtracting the current reported value (denoted R) of the claim as follows:

$$BULK = \frac{\int x \times \tilde{f}(x)dx}{\int \tilde{f}(x)dx} - R. \qquad (3.1)$$

It should be noted that the limits of integration are not included in the formula above. A possible adjustment would be to truncate the distribution from below at the paid to date value of the claim or to truncate the distribution from above at the maximum probable loss. Without these adjustments, the denominator of the first term equals 1 and is not necessary.

In addition, the numerator can readily be modified to consider the effect of policy limits. This adjustment is left to the interested reader.

More powerfully however, we can use the adjusted mixing weights to simulate a range of ultimate values for each claim. This is done in a two step process. In the first step, we draw from a Discrete (*x*, *p*) distribution where the loss processes are the *x* values and the adjusted mixing weights are the *p* values. This step determines the loss model that describes the distribution of ultimate values of the claim. In the second step, we draw a loss value from the loss model from the first step. This amount represents the simulated ultimate value of the claim. Commercial simulation software can then be used to develop both mean estimates of the bulk reserve as well as bulk reserve estimates at various confidence levels.

5. The Unbiased Estimator

The discussion above assumes that an unbiased estimator of the ultimate value of each claim is available. As we know this is almost never the case. (If it were, there would be much less need for actuaries.) However, we should recognize a biased estimator is usually available and an adjustment factor can be applied to this estimator to remove the

[5] "Stability" here refers to the change in these items resulting from incremental (marginal) increases in the data underlying their estimation.

bias. That biased estimator is the reported (paid plus case reserve) value of the claim and the adjustment factor is related to the loss development factor.

The loss development factor would have to be adjusted to remove the distorting influences of closed claims and unreported claims. That is, a cumulative reported loss development factor at age (maturity) *j* could be written as follows[6]:

$$RLDF_j = \frac{\text{Paid on Closed}_j + \text{Ultimate on Open}_j + \text{Ultimate on Unreported}_j}{\text{Paid on Closed}_j + \text{Reported on Open}_j} \qquad (4.1)$$

It should be noted that for this purpose, the loss development factors merely need to be for the same type of claim as the bulk reserve being estimated. That is, we are simply trying to develop the adjustment factor for known claims and liability for known claims exists regardless of whether coverage is written on a claims-made or occurrence basis. Therefore, we could use claims-made factors in this exercise to develop bulk reserves for occurrence basis coverage.

Rewriting the numerator of equation 4.1 as Ultimate Loss and taking reciprocals, we arrive at the following:

$$RLDF_j^{-1} = \frac{\text{Paid on Closed}_j + (\text{Paid on Open}_j - \text{Paid on Open}_j) + \text{Reported on Open}_j}{\text{Ultimate Loss}}$$

$$RLDF_j^{-1} = PLDF_j^{-1} - \frac{\text{Paid on Open}_j}{\text{Ultimate Loss}} + \frac{\text{Reported on Open Claims}_j}{\text{Ultimate Loss}} \qquad (4.2)$$

where "*PLDF*$_j$" denotes the cumulative paid loss development factor at age *j*.

We can rewrite the second term of the right hand side of the equation using the following relationship:

$$\frac{\text{Paid on Open}_j}{\text{Ultimate Loss}} \times \frac{\text{Total Paid}_j}{\text{Total Paid}_j} = PLDF_j^{-1} \times \frac{\text{Paid on Open}_j}{\text{Total Paid}_j} \qquad (4.3)$$

and we then rewrite equation 4.2 as:

$$RLDF_j^{-1} = PLDF_j^{-1} - PLDF_j^{-1} \times \frac{\text{Paid on Open}_j}{\text{Total Paid}_j} + \frac{\text{Reported on Open Claims}_j}{\text{Ultimate Loss}} \; . \qquad (4.4)$$

Rearranging equation 4.4, we obtain:

$$RLDF_j^{-1} - PLDF_j^{-1} \times \left(1 - \frac{\text{Paid on Open}_j}{\text{Total Paid}_j}\right) = \frac{\text{Reported on Open Claims}_j}{\text{Ultimate Loss}} . \qquad (4.5)$$

[6] Assumes no reserve is required for reopened claims.

Taking reciprocals, we have:

$$\left[RLDF_j^{-1} - PLDF_j^{-1} \times \left(1 - \frac{\text{Paid on Open}_\text{j}}{\text{Total Paid}_\text{j}}\right)\right]^{-1} = \frac{\text{Ultimate Loss}}{\text{Reported on Open Claims}_\text{j}}. \tag{4.6}$$

Focusing on the right side of this equation we have the following:

$$\begin{aligned}&\frac{\text{Ultimate Loss}}{\text{Reported on Open}_\text{j}}\\ &= \frac{\text{Paid on Closed}_\text{j} + \text{Ultimate on Open}_\text{j} + \text{Ultimate on Unreported}_\text{j}}{\text{Reported on Open}_\text{j}}.\end{aligned} \tag{4.7}$$

For convenience, we will refer to the three terms of the right hand side of this equation as ***F***, ***G***, and ***H***. We should recognize that the middle term (***G***) is the bias adjustment that we need to convert the reported value to an unbiased estimator of ultimate loss (***U***).

Substituting equation 4.7 into equation 4.6 and solving for G, we have:

$$G_j = \left(RLDF_j^{-1} - PLDF_j^{-1} \times \left(1 - \frac{\text{Paid on Open}_\text{j}}{\text{Total Paid}_\text{j}}\right)\right)^{-1} - F_j - H_j \tag{4.8}$$

$$\begin{aligned}G_j = &\left(RLDF_j^{-1} - PLDF_j^{-1} \times \left(1 - \frac{\text{Paid on Open}_\text{j}}{\text{Total Paid}_\text{j}}\right)\right)^{-1} - \\ &\frac{\text{Paid on Closed}_\text{j}}{\text{Reported on Open}_\text{j}} - \frac{\text{Ultimate on Unreported}_\text{j}}{\text{Reported on Open}_\text{j}}\end{aligned}. \tag{4.9}$$

The author recognizes that (1) $\frac{\text{Paid on Open}_\text{j}}{\text{Total Paid}_\text{j}}$, (2) $\frac{\text{Paid on Closed}_\text{j}}{\text{Reported on Open}_\text{j}}$ and (3) $\frac{\text{Ultimate on Unreported}_\text{j}}{\text{Reported on Open}_\text{j}}$ are statistics that are not "natural" and are not generally readily available.

However, it is the author's opinion that (1) and (2) should be straightforward to compile from a loss database since they are based strictly on reported values. Therefore it should not be inordinately more difficult to develop these statistics than it is to develop loss development factors.

The third statistic should also be straightforward to determine and in many cases it may not be necessary. That is, with respect to the numerator of this statistic, we can use a frequency / severity approach. We already have a severity model, *f(x)*, and unreported frequency should be readily estimated using a claim count development factors. The denominator is based only on reported data.

It should also be noted that for many lines of business substantially all claims are reported within two years so this term would be unnecessary after 24 months. For longer-tailed lines of business, loss development factors are generally available on a claims-made basis. If we are using these claims-made loss development factors, this term is, by definition, equal to 0 and is therefore not necessary.

Finally we should recognize that statistics (1) and (2) should be expected to approach zero as claims mature. Statistic (3) should be expected to become large as claims mature. In fact since (3) may become unstable at late maturities, the actuary may simply want to use the reported value of the claim (without adjustment) at late maturities. This is not altogether unwarranted since at late maturities the unknown facts associated with a claim will decrease and reported value of the claim is more likely to be an unbiased estimator of ultimate value.

Using $R_{j,k}$ to denote the reported value of the j-th claim at age k, and U_j to denote the unbiased estimator of the ultimate value of the j-th claim, we can now state the following:

$$U_j = R_{j,k} \times G_k . \tag{4.10}$$

Readers will note that G is based on aggregate statistics such as loss development factors. These statistics only consider the age of a claim and therefore ignore many other factors[7] that would influence development of a given claim. Determining an appropriate G for any single claim is extremely difficult. The framework described in this paper therefore provides an attractive compromise between the unbiased estimator and the *a priori* average claim size.

While this may seem like a significant effort, the reward is equally significant. Namely, the actuary now has insight into the average level of misstatement in case reserves. The actuary should recognize this as particularly important information in evaluating reserves.

6. Conclusion and Summary

Using the procedure above, we can transform a semi-parametric mixture model from its generic form of

$$f(x) = w_1 \times f_1(x) + w_2 \times f_2(x) + \ldots + w_n \times f_n(x)$$

to a form that may be used to describe the distribution of the ultimate value of a known claim:

$$f(x \mid U) = \tilde{f}(x) = \tilde{w}_1 \times f_1(x) + \tilde{w}_2 \times f_2(x) + \ldots + \tilde{w}_n \times f_n(x)$$

[7] Such as claim size.

where the mixing weights are adjusted based on an unbiased estimator of the ultimate value of the claim. This unbiased estimator can be calculated as a function of the reported value of the claim.

As actuarial analyses move from deterministic frameworks to stochastic frameworks, the distributions of ultimate values for known claims will gain in importance.

* * * * *

Acknowledgements

The author would like to acknowledge and thank the following individuals who reviewed this paper: David R. Lesieur, Paul Braithwaite and Don Doty

As always, any errors that remain are the responsibility of the author.