MODELING LOSSES WITH THE MIXED EXPONENTIAL DISTRIBUTION

CLIVE L. KEATINGE

Abstract

Finding a parametric model that fits loss data well is often difficult. This paper offers an alternative—the semiparametric mixed exponential distribution. The paper gives the reason why this is a good model and explains maximum likelihood estimation for the mixed exponential distribution. The paper also presents an algorithm to find parameter estimates and gives an illustrative example. The paper compares variances of estimates obtained with the mixed exponential distribution with variances obtained with a traditional parametric distribution. Finally, the paper discusses adjustments to the model and other uses of the model.

1. INTRODUCTION

Loss distributions have been a staple of actuarial work for many years. The Casualty Actuarial Society syllabus has included a separate section on the subject since 1985, the year after Hogg and Klugman [5] published *Loss Distributions*. This was the standard actuarial text on the subject until the recent book by Klugman, Panjer, and Willmot [8], *Loss Models: From Data to Decisions*, replaced it. Over the years, numerous authors have published papers dealing with loss distributions. The two books and most papers on the subject emphasize the use of parametric distributions as models for losses. I have found that the set of distributions generally suggested for use is not adequate. Too often, one cannot find a model that fits a data set well. Nonparametric procedures are available, but although they usually produce a good fit to the data, they often do not smooth the data

enough.[1] In this paper, I offer an alternative—the semiparametric mixed exponential distribution.

Statisticians have done quite a bit of work with semiparametric mixture models. Lindsay and Lesperance [12] wrote in their 1995 review of semiparametric mixture models, "There has been a surge of interest in semiparametric mixture models in recent years, as statisticians strive to maintain the efficiencies of parametric methods while incorporating minimal assumptions in their models."[2] I will first explain why the mixed exponential distribution is a good model for losses. I will then discuss the theory underlying maximum likelihood estimation with the mixed exponential distribution. Much of this material has been developed in the statistics literature, but I will highlight the relevant parts of it. Next, I will present an algorithm based on Newton's method to find the maximum likelihood parameter estimates. I will follow with an example of the application of this algorithm to a data set from Klugman, Panjer, and Willmot [8] and with a comparison of the variances of estimates obtained from a mixed exponential distribution and a Pareto distribution, which serves as an example of a traditional parametric distribution. I will then address adjustments that may be necessary when using the mixed exponential distribution, with particular emphasis on how to handle the tail. Finally, I will briefly mention that the mixed exponential distribution is useful for more than just modeling losses.

I will not discuss how to account for loss development before fitting a distribution to a set of data. The actuarial literature has not adequately addressed this very important issue, but it is beyond the scope of this paper. Also, I will assume that all data analyzed has received appropriate trending.

---

[1] Although Klugman, Panjer, and Willmot [8] focus primarily on parametric procedures, they do briefly cover nonparametric procedures in Section 2.11.1.

[2] Lindsay [11] has also written a monograph summarizing much of the recent work in mixture models.

2. MOTIVATION

When working with a set of loss data, we usually want to estimate the underlying probability distribution that describes the process that generated the data. It is generally a plausible assumption that this distribution is reasonably smooth. Thus, smoothing out the data should give a better estimate than simply using the empirical distribution itself. To accomplish such smoothing, we may turn to either parametric or nonparametric procedures. However, a parametric procedure often produces a distribution that does not fit the data well, whereas a nonparametric procedure often produces a distribution that is not smooth enough. What we need is something in between a parametric and a nonparametric procedure—a procedure that will provide a distribution that fits the data well, yet still provides an appropriate amount of smoothing.

We can articulate the amount of smoothing we would like by specifying conditions that the derivatives of the survival function, $S(x)$, should satisfy (where x is the loss size).[3] First, note that $S'(x) = -f(x)$, where $f(x)$ is the probability density function. Clearly, $f(x)$ must not be negative, so we should require that $S'(x) \leq 0$. Next, we would like $f(x)$ to be decreasing, so we require that $S''(x) \geq 0$. Beyond that, we would like $f(x)$ to decrease at a decreasing rate, so we require that $S'''(x) \leq 0$. In general, we would like the derivatives of the survival function to change at a slower and slower rate as the loss size x gets larger and larger and to approach zero asymptotically as x approaches infinity.[4] The mathematical formulation of this requirement is that the survival function should possess derivatives of all orders

[3]The survival function equals one minus the cumulative distribution function. Working with the survival function is more convenient than working with the cumulative distribution function.

[4]These conditions are appropriate for most loss distributions encountered in practice, except perhaps where the loss size x is small. In particular, these conditions are not compatible with a probability density function with a nonzero mode. However, we are assuming that we are not particularly interested in the behavior of the survival function where x is small.

such that

$$(-1)^n S^{(n)}(x) \geq 0, \qquad x > 0.$$

Functions with this alternating derivative property are known as completely monotone functions. There is a beautiful theorem due to Bernstein (1928) which states that a function S on $[0,\infty]$ is completely monotone if and only if it is of the form

$$S(x) = \int_0^\infty e^{-\lambda x} w(\lambda)\, d\lambda,$$

where w is nonnegative. Since we are interested in cases where S is a survival function, we will restrict attention to cases where $S(0) = 1$. This requirement forces w to be a probability function (that may be discrete, continuous, or a combination of the two).[5] In other words, any distribution with the alternating derivative property must be a mixture of exponential distributions, and vice versa.[6]

From now on, I will use a discrete formulation of the mixing distribution w, because as will become clear, we usually deal with mixing distributions that are nonzero at a small number of points. Thus, we have

$$S(x) = \sum_{i=1}^{n} w_i e^{-\lambda_i x}, \qquad w_i > 0, \qquad \sum_{i=1}^{n} w_i = 1,$$

where w_i is the mixing weight corresponding to λ_i. Note that the mean of the ith component distribution of the mixture is $1/\lambda_i$.

One of the distinguishing characteristics of the mixed exponential distribution is that it always has a decreasing failure rate. The failure rate is the probability density function divided by the

[5] Another way of stating this is that S is completely monotone with $S(0) = 1$ if and only if it is the Laplace transform of a probability distribution w. See Feller [3, p. 439] for a proof.

[6] I would like to thank Glenn Meyers for pointing out this equivalence relation, with which he had become familiar through the work of Brockett and Golden [2]. They applied this relation to utility functions just as this paper applies it to loss distributions.

survival function.[7] For the mixed exponential distribution, the failure rate is

$$\sum_{i=1}^{n}\left(\frac{w_i e^{-\lambda_i x}}{\sum_{j=1}^{n} w_j e^{-\lambda_j x}}\right)\lambda_i.$$

This is a weighted average of the λ_i's. As x becomes larger, weight moves away from the larger λ_i's and toward the smaller λ_i's, thus decreasing the failure rate.

Most of the parametric distributions traditionally used to model losses have decreasing failure rates, either throughout the entire distribution or at all but small loss sizes. Some are special cases of the mixed exponential distribution. For example, the Pareto distribution is a mixture of exponential distributions with a gamma mixing distribution. See Appendix A for further discussion of this topic. The advantage that the mixed exponential distribution enjoys over parametric distributions is that the mixed exponential distribution is more general and thus likely to provide a better fit to the data while still providing an appropriate amount of smoothing. It is considered semiparametric because no parametric assumption is made about the form of the mixing distribution. We now turn to the problem of estimating the mixing distribution from a given set of data.

3. MAXIMUM LIKELIHOOD THEORY

Maximum likelihood estimation is the only estimation technique I will cover in this paper. Although other techniques are available, the well-known desirable statistical properties of maximum likelihood estimation usually make it the method of choice.

---

[7] See Section 2.7.2 of Klugman, Panjer, and Willmot [8] for a discussion of failure rates. The failure rate is also known as the hazard rate or the force of mortality. In the context of a loss distribution, "failure" means "loss stoppage." A distribution with a decreasing failure rate has an increasing mean residual lifetime (if it exists).

In this section, I will describe the properties underlying maximum likelihood estimation with the mixed exponential distribution. The proofs are in Appendix B.

I will begin by addressing the situation where no grouping, censoring, or truncation is present in the data. The loglikelihood function is

$$\ln L(w_1, w_2, \ldots) = \sum_{k=1}^{m} \ln f(x_k) = \sum_{k=1}^{m} \ln \left(\sum_{i=1}^{\infty} w_i \lambda_i e^{-\lambda_i x_k} \right),$$

where m is the number of observations. We must find the set of w_i's that maximizes the loglikelihood function, subject to the constraints that each of the w_i's must be greater than or equal to zero and the sum of the w_i's must be one. We consider the λ_i's fixed and arbitrarily close together.

This constrained maximum occurs at the unique point at which the following conditions, known as the Karush–Kuhn–Tucker (KKT) conditions, are satisfied:

$$\frac{\partial \ln L}{\partial w_i} = \sum_{k=1}^{m} \frac{\lambda_i e^{-\lambda_i x_k}}{\sum_{j=1}^{\infty} w_j \lambda_j e^{-\lambda_j x_k}} \leq m, \qquad \text{if} \quad w_i = 0$$

and

$$\frac{\partial \ln L}{\partial w_i} = \sum_{k=1}^{m} \frac{\lambda_i e^{-\lambda_i x_k}}{\sum_{j=1}^{\infty} w_j \lambda_j e^{-\lambda_j x_k}} = m, \qquad \text{if} \quad w_i > 0.$$

The inequality conditions ensure that we cannot increase the loglikelihood by moving a small amount of weight to a λ_i that has zero weight attached to it. The equality conditions ensure that we cannot increase the loglikelihood by moving weight around among the λ_i's that have positive weight attached to them. The number of positive w_i's at this maximum is at most m. None of the corresponding λ_i's can be less than $1/x_m$, where x_m is the largest observation, and none can be greater than $1/x_1$, where x_1

is the smallest observation. The number of positive w_i's tends to increase with the number of observations, but remains below ten in most practical situations.

For grouped data, the loglikelihood function is

$$\begin{aligned}\ln L(w_1, w_2, \ldots) &= a_1 \ln\left(1 - S(b_1)\right) + \sum_{k=2}^{g-1} a_k \ln\left(S(b_{k-1}) - S(b_k)\right) \\ &\quad + a_g \ln\left(S(b_{g-1})\right) \\ &= a_1 \ln\left(\sum_{i=1}^{\infty} w_i (1 - e^{-\lambda_i b_1})\right) \\ &\quad + \sum_{k=2}^{g-1} a_k \ln\left(\sum_{i=1}^{\infty} w_i (e^{-\lambda_i b_{k-1}} - e^{-\lambda_i b_k})\right) \\ &\quad + a_g \ln\left(\sum_{i=1}^{\infty} w_i (e^{-\lambda_i b_{g-1}})\right),\end{aligned}$$

where g is the number of groups, $a_1, \ldots, a_g$ are the number of observations in each group, and $b_1, \ldots, b_{g-1}$ are the group boundaries. We will assume that any adjacent groups that all have zero observations have been combined into one group.

In this case, the KKT conditions are

$$\begin{aligned}\frac{\partial \ln L}{\partial w_i} &= a_1 \frac{1 - e^{-\lambda_i b_1}}{\sum_{j=1}^{\infty} w_j (1 - e^{-\lambda_j b_1})} + \sum_{k=2}^{g-1} a_k \frac{e^{-\lambda_i b_{k-1}} - e^{-\lambda_i b_k}}{\sum_{j=1}^{\infty} w_j (e^{-\lambda_j b_{k-1}} - e^{-\lambda_j b_k})} \\ &\quad + a_g \frac{e^{-\lambda_i b_{g-1}}}{\sum_{j=1}^{\infty} w_j (e^{-\lambda_j b_{g-1}})} \leq m, \qquad \text{if} \quad w_i = 0\end{aligned}$$

and

$$\frac{\partial \ln L}{\partial w_i} = a_1 \frac{1 - e^{-\lambda_i b_1}}{\sum_{j=1}^{\infty} w_j (1 - e^{-\lambda_j b_1})} + \sum_{k=2}^{g-1} a_k \frac{e^{-\lambda_i b_{k-1}} - e^{-\lambda_i b_k}}{\sum_{j=1}^{\infty} w_j (e^{-\lambda_j b_{k-1}} - e^{-\lambda_j b_k})}$$

$$+ a_g \frac{e^{-\lambda_i b_{g-1}}}{\sum_{j=1}^{\infty} w_j (e^{-\lambda_j b_{g-1}})} = m, \qquad \text{if} \quad w_i > 0.$$

The constrained maximum will occur at a unique point, unless the mixed exponential probabilities for each group are exactly proportional to the number of observations in each group or, in other words, when the data perfectly fits the model. For this situation, we can easily come up with examples where an arbitrarily large number of different mixed exponential distributions, each with an arbitrarily large number of positive w_i's, will maximize the loglikelihood function. However, a perfect fit is highly unlikely unless the number of groups is very small.

When the fit is not perfect, the number of positive w_i's with corresponding λ_i's on $(0,\infty)$ at the maximum is at most $g/2-1$ if g is even and $g/2-1/2$ if g is odd. In addition to the λ_i's on $(0,\infty)$, there may also be λ_i's at zero or infinity (or both) that have positive w_i's. For an exponential distribution with a λ_i of zero (and thus a mean of infinity), the survival function is a constant function of 1. In actuarial terms, the w_i corresponding to a λ_i of zero would indicate the probability that a loss will completely exhaust all layers of coverage, no matter how high. For an exponential distribution with a λ_i of infinity (and thus a mean of zero), the survival function is a constant function of 0. The w_i corresponding to a λ_i of infinity would indicate the probability that a loss will be zero. The number of positive w_i's tends to increase with the number of groups, but remains below ten in most practical situations.

The development for grouped data applies also to censored grouped data, since the censored data is simply in the last group with an upper bound of infinity. For other situations, such as censored ungrouped data (thus partially grouped and partially ungrouped) or data censored at various points or grouped with various boundaries, the logic is similar to that used above, since we can simply sum the appropriate loglikelihood functions.

With ungrouped data truncated (but not shifted) by a deductible d, the loglikelihood function is

$$\ln L(w_1, w_2, \ldots) = \sum_{k=1}^{m} \ln\left(\frac{f(x_k)}{S(d)}\right) = \sum_{k=1}^{m} \ln\left(\frac{\sum_{i=1}^{\infty} w_i \lambda_i e^{-\lambda_i x_k}}{\sum_{j=1}^{\infty} w_j e^{-\lambda_j d}}\right)$$

$$= \sum_{k=1}^{m} \ln\left(\sum_{i=1}^{\infty} w_i^* \lambda_i e^{-\lambda_i (x_k - d)}\right),$$

where

$$w_i^* = \frac{w_i e^{-\lambda_i d}}{\sum_{j=1}^{\infty} w_j e^{-\lambda_j d}}.$$

We can thus convert the problem to a problem without a deductible by subtracting d from each observation. We can then recover the w_i's using the formula

$$w_i = \frac{w_i^* e^{\lambda_i d}}{\sum_{j=1}^{\infty} w_j^* e^{\lambda_j d}}.$$

The same process applies for grouped data with d subtracted from each of the group boundaries instead of the observations. However, the formula to recover the w_i's breaks down if one of the λ_i's with a positive w_i^* is infinity, as quite often occurs with

grouped data. Using the fitted mixed exponential distribution to extrapolate below a deductible is not a good idea.

If a set of data contains several different deductibles, we can subtract the smallest deductible for which a credible amount of data exists from each observation and the higher deductibles. We would have to subtract additional terms from the loglikelihood function to account for these higher deductibles.[8]

4. A MAXIMUM LIKELIHOOD ALGORITHM

I will now present an algorithm that we can use to find the maximum likelihood estimates of the parameters of a mixed exponential distribution. I have based the algorithm on Newton's method, the details of which are in any textbook on numerical analysis. After I present the algorithm, I will comment on alternatives to it. The steps of the algorithm are:

1. Begin with an initial set of positive w_i's and the λ_i's associated with them. The closer these are to the final estimated values, the faster the convergence will be. However, the algorithm will converge regardless of what the initial values are.
2. Assume that the number of parameters is fixed and use Newton's method to find the indicated change in the parameters. I will call this the Newton step. Each λ_i is a parameter, and all but one of the w_i's are parameters. We must set the remaining w_i equal to one minus the sum of the others. Appendix C shows the derivatives needed to find the Newton step.
3. Adjust the parameters by the amount of the Newton step. If all the λ_i's remain positive, if all the w_i's remain between zero and one, and if the loglikelihood function

[8]See Section 2.10 of Klugman, Panjer, and Willmot [8] for a discussion of estimation with incomplete data.

increases, then go to step 4. If the result does not satisfy all these conditions, then try a backward Newton step, then half a forward step, then half a backward step, then a quarter of a forward step, and so on until the result satisfies all the conditions.

4. If one of the λ_i's is approaching zero or infinity (which can happen only with grouped or censored data), go to step 5. If one of the w_i's is approaching zero, go to step 6. If the Newton step is very small, thus indicating convergence, go to step 7. Otherwise, go back to step 2.
5. If one of the λ_i's is approaching zero, then fix that λ_i at a very small value, so it is effectively zero. If one of the λ_i's is approaching infinity, then fix that λ_i at a very large value, so it is effectively infinity. Remove the fixed λ_i, but not its associated w_i, from the Newton iterative process. Go back to step 2.
6. If one of the w_i's is approaching zero, then adjust the parameters by the proportion of the Newton step that makes this w_i exactly zero. Remove it and its associated λ_i as parameters. Often, this λ_i will be approaching one of the other λ_i's. If the eliminated w_i was close enough to zero, its removal should result in an increase in the loglikelihood function. Go back to step 2.
7. If convergence has occurred, then check to see if the result satisfies the KKT conditions. To do this, check the conditions for λ_i's close enough together so that it is clear that if the result satisfies the conditions at the checked λ_i's, the result will also satisfy the conditions at all others in between. If the result satisfies the KKT conditions, then the loglikelihood function has reached its maximum. If the result does not satisfy the conditions, go to step 8.
8. If the result does not satisfy the KKT conditions, then add an additional λ_i and associated w_i as parameters. The new λ_i should be in the vicinity of where the KKT

function is the largest (and thus where a new λ_i is most needed). Give the new w_i a small positive value and proportionately decrease the other w_i's so the sum of the w_i's remains at 1. The value assigned to the new w_i should be small enough so that the loglikelihood function increases from its previous value. (The algorithm will work regardless of the values of the new λ_i and w_i as long as the loglikelihood function increases from its previous value. If it does not increase, the algorithm may lead right back to the point where it was before the new λ_i and w_i were added.) Go back to step 2.

This algorithm will always converge to the maximum likelihood estimates of the parameters, because the loglikelihood function is concave and its value is increasing with each step of the algorithm. The points where Newton's method converges but the result does not satisfy the KKT conditions correspond to local maxima with the number of λ_i's fixed at a specified number. When the result satisfies the KKT conditions, we have reached the global maximum, with no restriction on the number of λ_i's.

With ungrouped data, the fitted mixed exponential mean will always equal the sample mean. This applies at both the global maximum and local maxima with a fixed number of λ_i's. Also, with ungrouped data, the fitted mixed exponential variance will not be less than the sample variance. This applies only at the global maximum. Appendix C gives the proofs of these statements. With grouped data, these relationships cannot hold, because the values of the individual observations are not available.

The variance relationship for ungrouped data results from the smoothing effect of the mixed exponential distribution. Probability from the sample values is effectively spread to surrounding values where no data was observed, thus increasing the variance. Though this produces an upward bias in the variance of the fitted distribution, it reduces the variance of the estimates of the survival probabilities produced by the fitted distribution, as we will see in Section 6.

This variance relationship also holds for nonparametric smoothing procedures. For parametric distributions, the fitted variance can be either larger or smaller than the sample variance, depending on the particular sample. For both the mixed exponential distribution and parametric distributions, as long as the variance of the actual distribution is finite, the ratio of the fitted variance to the sample variance will approach 1 as the sample size goes to infinity, since both will converge to the actual variance of the distribution. If the variance of the actual distribution is infinite, this will be true for the distribution censored at any point.

The given algorithm is certainly not the only one that can be used to maximize the loglikelihood function. I presented it because Newton's method is well-known and it converges very fast once the parameters are in the vicinity of the solution. Step 3 of the algorithm, trying successively smaller forward and backward Newton steps until the loglikelihood increases, is not elegant, but it does work. One could certainly improve the efficiency of the algorithm, but with the ample computing power now available, any improvements would probably be of marginal benefit in most cases.[9]

One could use a "canned" optimization program (which may use Newton's method with approximations of the derivatives) to maximize the loglikelihood function. Such programs can work well, but one must take care to ensure that the program does not stop before reaching the solution. Also, since the λ_i's are generally of very different magnitudes, a scaling adjustment may be helpful.

5. AN EXAMPLE

I will now illustrate how the algorithm works. I will use some grouped general liability data taken from Table 2.27 of Klugman,

[9]Bohning [1] reviewed several maximum likelihood algorithms that have been proposed for use with semiparametric mixture models.

Panjer, and Willmot [8]. The first three columns of Table 1 show the data. The loss amounts shown are the group boundaries.

We begin by fixing the number of means at one (though we need not begin with one). Instead of referring to the λ_i's associated with a mixed exponential distribution, throughout this example I will refer to the means (the reciprocals of the λ_i's). Regardless of the initial value we select, we will obtain rapid convergence to a mean of 51,190. The second column of Table 2 shows this result. The third column shows the value of the KKT function

$$h(\lambda) = a_1 \frac{1 - e^{-\lambda b_1}}{\sum_{j=1}^{\infty} w_j (1 - e^{-\lambda_j b_1})} + \sum_{k=2}^{g-1} a_k \frac{e^{-\lambda b_{k-1}} - e^{-\lambda b_k}}{\sum_{j=1}^{\infty} w_j (e^{-\lambda_j b_{k-1}} - e^{-\lambda_j b_k})}$$
$$+ a_g \frac{e^{-\lambda b_{g-1}}}{\sum_{j=1}^{\infty} w_j (e^{-\lambda_j b_{g-1}})}$$

for a number of means. As it must, $h(\lambda)$ has a value of 336 (the number of observations) at 51,190, but the function is larger than this everywhere else. Thus, we have not reached the maximum.

Since $h(\lambda)$ is largest at large means, we move a small amount of weight to a large mean. The actual value of this mean or the amount of weight we place on it is not important as long as the loglikelihood increases. With two means, the algorithm converges to means of 13,570 and 176,638 with weights of 0.7566 and 0.2434, respectively. From Table 2, we see that we still have not reached the maximum.

Since $h(\lambda)$ is again largest at large means, we move a small amount of weight to a large mean and proportionately scale back the weights on the existing two means (checking to be sure that the loglikelihood increases). With three means, the algorithm converges to means of 10,598, 73,440, and 686,632 with weights

TABLE 1

COMPARISON OF FITTED DISTRIBUTIONS

Empirical			Mixed Exponential			Transformed Beta			Pareto			Lognormal		
			Mean	Weight										
			0	0.0526	θ	21,239								
			12,336	0.5999	α	0.9102								
			77,922	0.3102	γ	1.1998	θ	14,679	μ	9.4812				
			712,302	0.0373	τ	0.6427	α	1.0758	σ	1.7162				
			Loglikelihood	−818.26	Loglikelihood	−820.16	Loglikelihood	−820.78	Loglikelihood	−821.33				
Loss Amt	No.> Loss Amt	Survival Prob	No.> Loss Amt	Survival Prob	Diff From Empirical	No.> Loss Amt	Survival Prob	Diff From Empirical	No.> Loss Amt	Survival Prob	Diff From Empirical	No.> Loss Amt	Survival Prob	Diff From Empirical
0	336	1.0000	336.00	1.0000	0.00%	336.00	1.0000	0.00%	336.00	1.0000	0.00%	336.00	1.0000	0.00%
2,500	278	0.8274	278.00	0.8274	0.00%	278.46	0.8288	0.17%	283.70	0.8443	2.05%	279.84	0.8329	0.66%
7,500	217	0.6458	216.79	0.6452	−0.10%	215.59	0.6416	−0.65%	215.53	0.6415	−0.68%	210.86	0.6276	−2.83%
12,500	180	0.5357	174.25	0.5186	−3.20%	175.36	0.5219	−2.58%	173.19	0.5155	−3.78%	171.72	0.5111	−4.60%
17,500	144	0.4286	144.26	0.4293	0.18%	147.11	0.4378	2.16%	144.42	0.4298	0.29%	145.55	0.4332	1.08%
22,500	122	0.3631	122.74	0.3653	0.61%	126.25	0.3757	3.48%	123.64	0.3680	1.34%	126.50	0.3765	3.69%
32,500	92	0.2738	95.10	0.2830	3.37%	97.66	0.2907	6.16%	95.69	0.2848	4.01%	100.27	0.2984	8.99%
47,500	73	0.2173	72.65	0.2162	−0.48%	72.17	0.2148	−1.13%	71.10	0.2116	−2.60%	76.14	0.2266	4.30%
67,500	58	0.1726	56.06	0.1668	−3.34%	53.00	0.1577	−8.62%	52.67	0.1568	−9.19%	57.06	0.1698	−1.61%
87,500	47	0.1399	45.15	0.1344	−3.95%	41.60	0.1238	−11.49%	41.67	0.1240	−11.34%	45.14	0.1344	−3.95%
125,000	29	0.0863	31.47	0.0937	8.51%	29.37	0.0874	1.27%	29.77	0.0886	2.65%	31.73	0.0944	9.43%
175,000	22	0.0655	20.82	0.0620	−5.34%	20.89	0.0622	−5.04%	21.42	0.0637	−2.64%	22.02	0.0655	0.09%
225,000	15	0.0446	14.94	0.0445	−0.42%	16.11	0.0479	7.38%	16.65	0.0496	11.01%	16.41	0.0488	9.37%
325,000	9	0.0268	9.54	0.0284	6.04%	10.94	0.0326	21.57%	11.44	0.0341	27.15%	10.32	0.0307	14.61%
475,000	7	0.0208	6.66	0.0198	−4.82%	7.30	0.0217	4.31%	7.72	0.0230	10.30%	6.13	0.0182	−12.49%
675,000	5	0.0149	4.87	0.0145	−2.55%	5.00	0.0149	0.09%	5.34	0.0159	6.83%	3.64	0.0108	−27.26%
1,000,000	3	0.0089	3.08	0.0092	2.54%	3.27	0.0097	9.07%	3.53	0.0105	17.53%	1.94	0.0058	−35.30%

TABLE 2

KARUSH–KUHN–TUCKER FUNCTION

Mean = $\frac{1}{\lambda}$	One Mean		Two Means		Three Means		Four Means	
	$\frac{1}{\lambda}$	$h(\lambda)$	$\frac{1}{\lambda}$	$h(\lambda)$	$\frac{1}{\lambda}$	$h(\lambda)$	$\frac{1}{\lambda}$	$h(\lambda)$
							0	**336.000**
1,000		1173.337		432.190		396.167		335.881
2,000		1060.099		402.882		373.030		334.870
3,000		968.445		381.032		356.735		333.555
4,000		898.142		366.497		347.053		333.025
5,000		842.334		356.811		341.618		333.218
6,000		796.360		350.199		338.684		333.773
7,000		757.342		345.567		337.168		334.415
8,000		723.472		342.262		336.435		334.993
9,000		693.574		339.891		336.118		335.442
10,000		666.848		338.213		336.012		335.748
					10,598	**336.000**		
							12,336	**336.000**
			13,570	**336.000**				
20,000		496.647		340.263		336.150		335.188
30,000		410.371		353.183		336.079		334.770
40,000		360.872		363.845		336.007		335.056
50,000		336.444		369.889		335.970		335.455
	51,190	**336.000**						
60,000		389.835		371.973		335.979		335.775
70,000		971.560		371.194		335.998		335.958
					73,440	**336.000**		
							77,922	**336.000**
80,000		3,995		368.495		335.992		335.997
90,000		14,647,843		364.604		335.944		335.915
100,000		43,187,502		360.067		335.856		335.745
			176,638	**336.000**				
200,000		6,191,258		338.944		334.801		333.922
300,000		32,692,464		414.348		334.964		334.227
400,000		75,160,236		558.705		335.441		335.019
500,000		123,867,653		729.785		335.788		335.598
600,000		172,830,341		903.455		335.960		335.903
					686,632	**336.000**		
700,000		219,258,668		1068.732		335.999		335.999
							712,302	**336.000**
800,000		262,097,414		1221.452		335.950		335.956
900,000		301,125,152		1360.665		335.845		335.827
1,000,000		336,492,659		1486.842		335.708		335.645
2,000,000		554,645,524		2264.678		334.206		333.527
3,000,000		655,188,015		2622.692		333.230		332.125
4,000,000		712,101,357		2825.202		332.618		331.242
5,000,000		748,596,059		2955.002		332.205		330.645
6,000,000		773,959,261		3045.185		331.909		330.217
7,000,000		792,600,348		3111.454		331.688		329.896
8,000,000		806,875,259		3162.194		331.516		329.646
9,000,000		818,155,499		3202.285		331.378		329.447
10,000,000		827,293,145		3234.758		331.266		329.284

of 0.6270, 0.3340, and 0.0390, respectively. Again, we have not reached the maximum.

The KKT function is now largest below the first mean of 10,598. We move a small amount of weight to a small mean and proportionately scale back the weights on the existing three means. When we resume iterating, this smallest mean heads toward zero. We then fix it at a small value (for example, 25, 1% of the first group boundary). Effectively, we assign all the probability associated with this mean to the first group. We resume iterating, and the algorithm converges to the values shown at the top of Table 1. The table shows the first mean as zero, because that is its true value. As the last column of Table 2 shows, the KKT function now never exceeds 336. We have thus reached the maximum likelihood estimates of the mixed exponential parameters.

Table 1 shows the fitted survival probabilities. The fitted and empirical probabilities match exactly at the first group boundary. This will always occur when a mean of zero has a positive weight in the final parameter set, since this is the only way the KKT function can be equal to the number of observations when λ_i is infinity. Likewise, anytime a mean of infinity has a positive weight in the final parameter set, the survival probabilities will match exactly at the last group boundary.

If the data includes various deductibles, attachment points, or policy limits, we can obtain the empirical distribution using the Kaplan–Meier Product-Limit estimator. This estimator provides empirical survival probabilities that take into account the effect of unobserved losses below deductibles and attachment points as well as losses capped by policy limits. Klugman, Panjer, and Willmot [8] cover this estimator briefly. It has historically been used extensively in survival analysis, and Klein and Moeschberger [7] and London [14] cover the subject in more detail.

For comparison, Table 1 also shows the fits for three distributions other than the mixed exponential. The parameterizations

of the transformed beta and the Pareto are the same as those that Klugman, Panjer, and Willmot [8] use. See Appendix A for details. The lognormal parameterization is the standard one. The transformed beta provides the best fit, as measured by the loglikelihood, of the distributions used by Klugman, Panjer, and Willmot [8]. The Pareto is a special case of both the mixed exponential and the transformed beta. As expected, the mixed exponential provides the best fit.

We would prefer the mixed exponential distribution if our hypothesis is that the actual distribution has the alternating derivative property, which is a much weaker hypothesis than one that states that the actual distribution follows a particular parametric form. In most situations, I have found little or no justification for a stronger parametric hypothesis.

The usual way to evaluate a hypothesis is to perform a test such as the chi-square goodness-of-fit test. When the parameters are estimated from the data, this test is not appropriate with the mixed exponential distribution, since the mixed exponential does not have a fixed number of parameters. However, with most loss data I have encountered in practice, the appropriateness of the mixed exponential will be evident from a comparison of the fitted and empirical distributions.

For the other three distributions in Table 1, we can perform chi-square goodness-of-fit tests. We will combine the last three groups, and the two groups before the last three, so there are at least five losses in each of the resulting 14 groups. The results are as follows:

Distribution	Chi-square Statistic	Degrees of Freedom	p-value
Transformed Beta	9.24	9	0.41
Pareto	10.55	11	0.48
Lognormal	11.12	11	0.43

Another way to evaluate the Pareto hypothesis would be to use a likelihood ratio test. Since the Pareto distribution is a special case of the transformed beta distribution, under the Pareto hypothesis, twice the difference of the maximum values of the loglikelihoods of the Pareto and transformed beta has approximately a chi-square distribution with two degrees of freedom (the difference in the number of parameters). In this case, $2 \times ((-820.16) - (-820.78)) = 1.24$, which yields a p-value of 0.54. Thus, the Pareto distribution would not be rejected in favor of the transformed beta distribution.[10]

In this example, none of the distributions shown in Table 1 would be rejected as possible models for the actual distribution. However, as I mentioned above, hypothesizing a particular parametric distribution is dubious in most cases I have encountered. In general, the larger the data set, the more evident this becomes.

6. VARIANCE

With parametric distributions, we can obtain the asymptotic variances and covariances of the maximum likelihood estimators of the parameters by calculating the covariance matrix. We can then use the covariance matrix to find the asymptotic variances of the maximum likelihood estimators of functions of the parameters that are of interest, such as survival probabilities and limited expected values.[11]

This approach does not work with the semiparametric mixed exponential distribution. Tierney and Lambert [16] obtained a result that implies that the asymptotic variance of the maximum likelihood estimator of a function of mixed exponential parameters is equal to the variance of the empirical estimator for ungrouped data. For a survival probability, the empirical estimator is the sample proportion of observations that exceeds the loss

[10] See Section 2.9 of Klugman, Panjer, and Willmot [8] for a more thorough discussion of these tests.

[11] See Section 2.5 of Klugman, Panjer, and Willmot [8] for a discussion.

amount under consideration. This has a binomial distribution that approaches a normal distribution as the number of observations approaches infinity. This result means that, asymptotically, we do not reduce the variance of our survival probability estimates, or any other estimates based on the mixed exponential parameters, by using the fitted distribution instead of the empirical distribution.

In practice, we do not have infinite samples. To see what happens with finite samples, we must resort to simulation. Tables 3A, 3B, and 3C show the results of simulations using sample sizes of 10, 50, and 250, respectively. In each case, the simulated distribution is the Pareto distribution from Table 1. I used a Pareto distribution to facilitate comparison of the variances of estimates obtained using the mixed exponential distribution with the variances of estimates obtained using the Pareto distribution. The Pareto distribution serves as an example of a parametric distribution with a fixed number of parameters. These tables show estimates of the bias and variance of survival probability estimates based on 10,000 simulations, for a mixed exponential fit without grouping the data, and both a mixed exponential and a Pareto fit with data grouped using the boundaries from Table 1. The tables display bias as a percentage of the actual survival probability, and variance as a ratio to the variance of the empirical estimator. Table 3C also shows the asymptotic variance for the Pareto distribution. I focus on the survival function because any other function of interest can be expressed in terms of the survival function. For example, the limited expected value is simply the integral of the survival function from zero to the limit being considered.

The grouped mixed exponential results are close to the ungrouped results in the middle of the distribution, but are dramatically worse at small loss amounts and in the tail. The reason for this is that the grouped data provides virtually no information about the distribution either below the first group boundary of 2,500 or above the last group boundary of 1,000,000. There-

TABLE 3A

SIMULATION RESULTS—10 OBSERVATIONS

			Ungrouped Mixed Exponential		Grouped Mixed Exponential		Grouped Pareto	
Loss Amt	Survival Probability	10 Times Empirical Variance	Bias	Ratio to Empirical Variance	Bias	Ratio to Empirical Variance	Bias	Ratio to Empirical Variance
10	0.9993	0.00073	−0.25%	1.09	−3.79%	85.35	−0.02%	0.05
100	0.9927	0.00722	−1.20%	0.79	−3.89%	8.34	−0.13%	0.12
1,000	0.9316	0.06376	−3.61%	0.56	−4.58%	0.94	−0.66%	0.37
2,500	0.8443	0.13142	−4.45%	0.56	−4.96%	0.64	−0.72%	0.54
7,500	0.6415	0.22999	−3.99%	0.64	−4.28%	0.65	0.36%	0.72
12,500	0.5155	0.24976	−3.15%	0.69	−3.37%	0.69	1.06%	0.79
17,500	0.4298	0.24508	−2.63%	0.72	−2.79%	0.72	1.14%	0.82
22,500	0.3680	0.23257	−2.40%	0.74	−2.49%	0.74	0.76%	0.83
32,500	0.2848	0.20368	−2.47%	0.75	−2.47%	0.76	−0.70%	0.82
47,500	0.2116	0.16683	−3.13%	0.75	−3.04%	0.75	−3.32%	0.79
67,500	0.1568	0.13219	−4.16%	0.73	−4.03%	0.74	−6.17%	0.74
87,500	0.1240	0.10863	−5.05%	0.71	−4.89%	0.72	−7.94%	0.70
125,000	0.0886	0.08075	−6.23%	0.68	−6.02%	0.68	−8.98%	0.64
175,000	0.0637	0.05968	−7.13%	0.65	−6.72%	0.65	−7.68%	0.61
225,000	0.0496	0.04710	−7.60%	0.63	−6.79%	0.64	−4.98%	0.59
325,000	0.0341	0.03290	−7.90%	0.61	−5.65%	0.63	1.75%	0.58
475,000	0.0230	0.02245	−7.66%	0.58	−1.60%	0.67	12.25%	0.58
675,000	0.0159	0.01564	−7.00%	0.56	6.94%	0.78	25.39%	0.59
1,000,000	0.0105	0.01038	−6.00%	0.54	26.22%	1.02	44.47%	0.61
2,000,000	0.0050	0.00499	−4.31%	0.51	108.32%	1.93	92.25%	0.68
3,000,000	0.0033	0.00324	−3.39%	0.51	205.80%	2.94	131.11%	0.75
5,000,000	0.0019	0.00188	−1.82%	0.51	417.16%	5.05	195.67%	0.85
10,000,000	0.0009	0.00089	2.63%	0.54	982.05%	10.63	321.95%	1.04
20,000,000	0.0004	0.00042	11.91%	0.60	2178.28%	22.38	514.73%	1.31
30,000,000	0.0003	0.00027	19.65%	0.65	3423.15%	34.60	672.44%	1.52
50,000,000	0.0002	0.00016	31.42%	0.69	6002.45%	59.92	937.99%	1.86
100,000,000	0.0001	0.00008	46.04%	0.72	12761.24%	126.28	1469.57%	2.47

10 (Sample Size) Times Empirical Variance

$$= 10 \cdot \frac{\text{Surv Prob} \cdot (1 - \text{Surv Prob})}{10} = \text{Surv Prob} \cdot (1 - \text{Surv Prob})$$

$$\text{Bias} = \frac{\text{Average Simulated Fitted Survival Probability} - \text{Survival Probability}}{\text{Survival Probability}}$$

Ratio to Empirical Variance

$$= \frac{\text{Variance of Simulated Fitted Survival Probabilities}}{\text{Empirical Variance}}$$

TABLE 3B

SIMULATION RESULTS—50 OBSERVATIONS

			Ungrouped Mixed Exponential		Grouped Mixed Exponential		Grouped Pareto	
Loss Amt	Survival Probability	50 Times Empirical Variance	Bias	Ratio to Empirical Variance	Bias	Ratio to Empirical Variance	Bias	Ratio to Empirical Variance
10	0.9993	0.00073	−0.15%	0.87	−1.94%	98.53	0.00%	0.00
100	0.9927	0.00722	−0.59%	0.55	−1.95%	9.50	−0.02%	0.03
1,000	0.9316	0.06376	−1.45%	0.51	−1.98%	0.97	−0.10%	0.27
2,500	0.8443	0.13142	−1.50%	0.59	−1.75%	0.68	−0.09%	0.50
7,500	0.6415	0.22999	−0.63%	0.72	−0.73%	0.74	0.29%	0.75
12,500	0.5155	0.24976	−0.11%	0.77	−0.15%	0.78	0.57%	0.77
17,500	0.4298	0.24508	0.02%	0.79	0.04%	0.80	0.66%	0.75
22,500	0.3680	0.23257	−0.04%	0.80	0.01%	0.80	0.61%	0.73
32,500	0.2848	0.20368	−0.36%	0.80	−0.29%	0.81	0.27%	0.69
47,500	0.2116	0.16683	−0.83%	0.81	−0.75%	0.81	−0.42%	0.68
67,500	0.1568	0.13219	−1.22%	0.81	−1.15%	0.82	−1.22%	0.68
87,500	0.1240	0.10863	−1.46%	0.81	−1.40%	0.82	−1.76%	0.69
125,000	0.0886	0.08075	−1.85%	0.81	−1.85%	0.81	−2.21%	0.69
175,000	0.0637	0.05968	−2.40%	0.80	−2.56%	0.80	−2.09%	0.68
225,000	0.0496	0.04710	−2.97%	0.78	−3.29%	0.78	−1.57%	0.66
325,000	0.0341	0.03290	−3.99%	0.76	−4.27%	0.75	−0.03%	0.62
475,000	0.0230	0.02245	−5.12%	0.72	−3.80%	0.74	2.58%	0.57
675,000	0.0159	0.01564	−6.08%	0.69	0.28%	0.78	5.98%	0.52
1,000,000	0.0105	0.01038	−6.92%	0.65	12.80%	0.93	10.95%	0.47
2,000,000	0.0050	0.00499	−7.59%	0.60	71.75%	1.64	22.91%	0.40
3,000,000	0.0033	0.00324	−7.36%	0.57	143.09%	2.50	32.00%	0.36
5,000,000	0.0019	0.00188	−6.20%	0.56	301.35%	4.36	45.95%	0.31
10,000,000	0.0009	0.00089	−2.70%	0.55	733.42%	9.21	70.11%	0.27
20,000,000	0.0004	0.00042	2.41%	0.55	1653.60%	19.39	101.68%	0.23
30,000,000	0.0003	0.00027	5.82%	0.56	2611.72%	29.98	124.40%	0.21
50,000,000	0.0002	0.00016	10.34%	0.56	4596.97%	51.92	158.47%	0.19
100,000,000	0.0001	0.00008	15.23%	0.57	9799.13%	109.42	216.68%	0.17

50 (Sample Size) Times Empirical Variance

$$= 50 \cdot \frac{\text{Surv Prob} \cdot (1 - \text{Surv Prob})}{50} = \text{Surv Prob} \cdot (1 - \text{Surv Prob})$$

$$\text{Bias} = \frac{\text{Average Simulated Fitted Survival Probability} - \text{Survival Probability}}{\text{Survival Probability}}$$

Ratio to Empirical Variance

$$= \frac{\text{Variance of Simulated Fitted Survival Probabilities}}{\text{Empirical Variance}}$$

TABLE 3C

SIMULATION RESULTS—250 OBSERVATIONS

			Ungrouped Mixed Exponential		Grouped Mixed Exponential		Grouped Pareto		Grouped Pareto Asymptotic Variance
Loss Amt	Survival Probability	250 Times Empirical Variance	Bias	Ratio to Empirical Variance	Bias	Ratio to Empirical Variance	Bias	Ratio to Empirical Variance	Ratio to Empirical Variance
10	0.9993	0.00073	−0.08%	0.59	−1.06%	134.65	0.00%	0.00	0.00
100	0.9927	0.00722	−0.27%	0.40	−1.05%	12.75	0.00%	0.03	0.03
1,000	0.9316	0.06376	−0.58%	0.49	−0.92%	1.10	−0.04%	0.25	0.25
2,500	0.8443	0.13142	−0.48%	0.62	−0.64%	0.73	−0.06%	0.48	0.49
7,500	0.6415	0.22999	−0.03%	0.78	−0.06%	0.80	−0.05%	0.75	0.77
12,500	0.5155	0.24976	0.02%	0.82	0.02%	0.84	−0.03%	0.77	0.78
17,500	0.4298	0.24508	−0.12%	0.84	−0.11%	0.85	−0.05%	0.75	0.75
22,500	0.3680	0.23257	−0.29%	0.86	−0.28%	0.86	−0.08%	0.72	0.72
32,500	0.2848	0.20368	−0.52%	0.89	−0.52%	0.89	−0.18%	0.69	0.68
47,500	0.2116	0.16683	−0.60%	0.90	−0.61%	0.91	−0.34%	0.68	0.66
67,500	0.1568	0.13219	−0.55%	0.90	−0.55%	0.91	−0.51%	0.69	0.67
87,500	0.1240	0.10863	−0.54%	0.90	−0.50%	0.91	−0.63%	0.71	0.68
125,000	0.0886	0.08075	−0.66%	0.89	−0.57%	0.90	−0.74%	0.72	0.70
175,000	0.0637	0.05968	−0.91%	0.88	−0.83%	0.89	−0.75%	0.72	0.70
225,000	0.0496	0.04710	−1.14%	0.87	−1.13%	0.88	−0.67%	0.71	0.69
325,000	0.0341	0.03290	−1.52%	0.87	−1.61%	0.89	−0.39%	0.68	0.66
475,000	0.0230	0.02245	−1.94%	0.86	−1.73%	0.89	0.14%	0.62	0.60
675,000	0.0159	0.01564	−2.39%	0.85	−0.68%	0.88	0.86%	0.56	0.54
1,000,000	0.0105	0.01038	−3.04%	0.83	3.36%	0.96	1.94%	0.48	0.47
2,000,000	0.0050	0.00499	−4.61%	0.75	27.42%	1.78	4.60%	0.36	0.33
3,000,000	0.0033	0.00324	−5.63%	0.71	61.92%	2.85	6.62%	0.30	0.27
5,000,000	0.0019	0.00188	−6.62%	0.66	145.14%	5.12	9.66%	0.23	0.20
10,000,000	0.0009	0.00089	−6.62%	0.60	383.10%	11.02	14.73%	0.16	0.12
20,000,000	0.0004	0.00042	−5.55%	0.55	900.71%	23.32	20.95%	0.11	0.08
30,000,000	0.0003	0.00027	−4.78%	0.53	1442.30%	36.06	25.17%	0.09	0.06
50,000,000	0.0002	0.00016	−3.64%	0.51	2565.35%	62.44	31.15%	0.06	0.04
100,000,000	0.0001	0.00008	−2.43%	0.49	5508.39%	131.56	40.53%	0.04	0.02

250 (Sample Size) Times Empirical Variance

$$= 250 \cdot \frac{\text{Surv Prob} \cdot (1 - \text{Surv Prob})}{250} = \text{Surv Prob} \cdot (1 - \text{Surv Prob})$$

$$\text{Bias} = \frac{\text{Average Simulated Fitted Survival Probability} - \text{Survival Probability}}{\text{Survival Probability}}$$

Ratio to Empirical Variance

$$= \frac{\text{Variance of Simulated Fitted Survival Probabilities}}{\text{Empirical Variance}}$$

fore, the fitted distribution often contains means of either zero or infinity or both.

Because the Pareto is less flexible than the mixed exponential, the Pareto usually provides survival probability estimates with a smaller variance. This effect is most notable at small loss amounts and in the tail. However, this fact illustrates the problem with using the Pareto or other parametric distributions with a fixed number of parameters. If we knew that the actual distribution were a Pareto, we would of course prefer to fit a Pareto instead of a mixed exponential. However, the assumption that the distribution is a Pareto is virtually never valid. If our data set is small, the fit may appear to be good, but the tail is simply a function of the assumption that the distribution is a Pareto. The fitted tail may or may not be anywhere close to the actual tail. If our data set is large, then unless we really do have a Pareto, we will probably observe a poor fit in the tail because the Pareto is not flexible enough. Thus, though the Pareto provides estimates with smaller variance than the mixed exponential, these estimates may be significantly biased if the actual distribution is not a Pareto.

For the ungrouped mixed exponential, as the number of observations increases, the bias gradually disappears, and the ratio of the variance to the empirical variance eventually approaches 1. This process takes longer at small loss amounts and in the tail. For the grouped mixed exponential, the results are similar except that outside the layer boundaries, the estimator remains poor. Note that an empirical estimate of the survival probability is not an option outside the layer boundaries, since an empirical estimator is only available at the layer boundaries. For the Pareto, with 250 observations, the variance is very close to the asymptotic variance, but there is still some significant bias in the tail.

I have displayed results for only one distribution. The most notable feature that differs by distribution is that, generally

speaking, for a given number of observations and a given survival probability, the thinner the tail of a distribution, the smaller the variance. Roughly, this is because there is less spread in the mixing distribution of mixed exponential distributions with thinner tails than in those with thicker tails.

7. ADJUSTMENTS AND OTHER USES

In this section, I will first address the issue of estimating the tail of a distribution. Table 1 showed only survival probabilities up to 1,000,000. Table 4 shows survival probabilities up to 100,000,000. The first distribution in the table is the mixed exponential that we fit previously. The second distribution is the mixed exponential that results when we move one claim from the 675,000–1,000,000 group to the 475,000–675,000 group. The survival probabilities are very close to one another except in the tail. When we move one claim, we acquire a mean of infinity with a small positive weight. The survival function now approaches the value of this weight, instead of zero, as the loss amount approaches infinity. For comparison, Table 4 also shows the Pareto and lognormal distributions from Table 1. If we were to move this same claim and then fit a Pareto or lognormal distribution, the tails would be very close to those from Table 1. However, we have no way to tell from the available data whether either of them is anywhere close to the actual tail. The tails of the Pareto and lognormal distributions are between the two mixed exponential tails, and are also very different from one another.

Thus we see that we cannot reliably use the mixed exponential distribution or any parametric distribution to extrapolate beyond the available data. However, an advantage of the mixed exponential is that if other data is available to assist in estimating the tail, or if we simply use judgment, we can find a mixed exponential distribution that both fits the available data and produces the desired tail. For example, suppose we believe that the tail is likely to have a shape like the Pareto tail. We may base this belief on data we have from a similar source or simply judgment. We can

TABLE 4

TAIL COMPARISON

	Mixed Exponential		Mix Exp–1 Clm Moved		Pareto		Lognormal		Mix Exp–Pareto Tail	
	Mean	Weight	Mean	Weight					Mean	Weight
	0	0.0526	0	0.0525					0	0.0526
	12,336	0.5999	12,260	0.5950					12,303	0.5980
	77,922	0.3102	72,792	0.2962	θ	14,679	μ	9.4812	76,185	0.3063
	712,302	0.0373	326,741	0.0497	α	1.0758	σ	1.7162	437,233	0.0340
			Infinity	0.0066					2,216,890	0.0073
									10,459,111	0.0014
									74,727,807	0.0003
Loss Amt	No.> Loss Amt	Survival Probability	No.> Loss Amt	Survival Probability	No.> Loss Amt	Survival Probability	No.> Loss Amt	Survival Probability	No.> Loss Amt	Survival Probability
0	336.00	1.0000	336.00	1.0000	336.00	1.0000	336.00	1.0000	336.00	1.0000
2,500	278.00	0.8274	278.00	0.8274	283.70	0.8443	279.84	0.8329	278.00	0.8274
7,500	216.79	0.6452	216.76	0.6451	215.53	0.6415	210.86	0.6276	216.77	0.6452
12,500	174.25	0.5186	174.24	0.5186	173.19	0.5155	171.72	0.5111	174.24	0.5186
17,500	144.26	0.4293	144.27	0.4294	144.42	0.4298	145.55	0.4332	144.26	0.4293
22,500	122.74	0.3653	122.77	0.3654	123.64	0.3680	126.50	0.3765	122.75	0.3653
32,500	95.10	0.2830	95.13	0.2831	95.69	0.2848	100.27	0.2984	95.12	0.2831
47,500	72.65	0.2162	72.63	0.2162	71.10	0.2116	76.14	0.2266	72.65	0.2162
67,500	56.06	0.1668	55.99	0.1666	52.67	0.1568	57.06	0.1698	56.03	0.1667
87,500	45.15	0.1344	45.07	0.1341	41.67	0.1240	45.14	0.1344	45.10	0.1342
125,000	31.47	0.0937	31.48	0.0937	29.77	0.0886	31.73	0.0944	31.44	0.0936
175,000	20.82	0.0620	20.98	0.0624	21.42	0.0637	22.02	0.0655	20.85	0.0621
225,000	14.94	0.0445	15.13	0.0450	16.65	0.0496	16.41	0.0488	14.99	0.0446
325,000	9.54	0.0284	9.54	0.0284	11.44	0.0341	10.32	0.0307	9.57	0.0285
475,000	6.66	0.0198	6.26	0.0186	7.72	0.0230	6.13	0.0182	6.60	0.0196
675,000	4.87	0.0145	4.34	0.0129	5.34	0.0159	3.64	0.0108	4.81	0.0143
1,000,000	3.08	0.0092	3.00	0.0089	3.53	0.0105	1.94	0.0058	3.26	0.0097
2,000,000	0.76	0.0022	2.25	0.0067	1.69	0.0050	0.57	0.0017	1.60	0.0048
3,000,000	0.19	0.0006	2.22	0.0066	1.09	0.0033	0.26	0.0008	1.10	0.0033
5,000,000	0.01	0.0000	2.22	0.0066	0.63	0.0019	0.09	0.0003	0.65	0.0019
10,000,000	0.00	0.0000	2.22	0.0066	0.30	0.0009	0.02	0.0001	0.30	0.0009
20,000,000	0.00	0.0000	2.22	0.0066	0.14	0.0004	0.00	0.0000	0.14	0.0004
30,000,000	0.00	0.0000	2.22	0.0066	0.09	0.0003	0.00	0.0000	0.09	0.0003
50,000,000	0.00	0.0000	2.22	0.0066	0.05	0.0002	0.00	0.0000	0.05	0.0002
100,000,000	0.00	0.0000	2.22	0.0066	0.03	0.0001	0.00	0.0000	0.03	0.0001

add eight more group boundaries as shown in Table 4 to increase the number of groups to 25. We can then allocate the three claims above 1,000,000 to the nine groups above 1,000,000 so that the empirical survival probabilities above 1,000,000 match those of the Pareto distribution. We can then find a maximum likelihood estimate based on these 25 groups. The last two columns of Table 4 show the resulting distribution. The mixed exponential distribution is flexible enough so that we can append whatever tail we think appropriate while affecting the fit in the lower portion of the distribution very little.

In the example above, we adjusted the data before fitting to produce an appropriate tail. We may need to adjust the data for other reasons. For example, we may have to adjust for loss development. I will not discuss this issue further in this paper. However, such adjustments would change the empirical distribution to which we fit.

Just as we may adjust the data, we may also need to adjust the fitted distribution. The best fitting distribution, which satisfies the KKT conditions, will not, in all cases, be the most appropriate estimate to use. When conditions warrant, we may set any of the means and weights at fixed values before fitting. For example, despite any data adjustments we have made, if the best fitting distribution contains a mean of infinity, we may fix the largest mean and possibly its weight at a value that yields a tail that we feel is more appropriate. As another example, if we are fitting a number of distributions as part of the same project, we may find it convenient to use the same fixed means for each distribution. If the means are not too far apart, the resulting distributions are likely to fit almost as well as if we had not fixed the means. We could also impose constraints on the relationships among the means and weights through the use of Lagrange multipliers. Also, we could, through trial and error, simply select a distribution that visually fits the data well.

We can use the mixed exponential distribution for more than modeling losses. We can use the mixed exponential to model

anything where we expect a function with alternating derivatives. For example, I have found it useful in modeling the probability that a claim does not have any allocated loss adjustment expense attached to it as a function of the claim size. This is not a probability function, so we cannot use maximum likelihood estimation. However, we can use a least squares procedure to fit the distribution to the data.

8. CONCLUSION

In this paper, I have tried to provide the background needed for an actuary to begin using the mixed exponential distribution in his or her work. I believe that the combination of flexibility and smoothness that the mixed exponential provides makes it an extremely useful actuarial modeling tool.

REFERENCES

[1] Bohning, Dankmar, "A Review of Reliable Maximum Likelihood Algorithms for Semiparametric Mixture Models," *Journal of Statistical Planning and Inference* 47, 1/2, October 1995, pp. 5–28.

[2] Brockett, Patrick L., and Linda L. Golden, "A Class of Utility Functions Containing All the Common Utility Functions," *Management Science* 33, 8, August 1987, pp. 955–964.

[3] Feller, William, *An Introduction to Probability Theory and Its Applications*, Volume II, Second Edition, New York: John Wiley & Sons, 1971.

[4] Hillier, Frederick S., and Gerald J. Lieberman, *Introduction to Operations Research*, Sixth Edition, New York: McGraw-Hill, 1995.

[5] Hogg, Robert V., and Stuart A. Klugman, *Loss Distributions*, New York: John Wiley & Sons, 1984.

[6] Jewell, Nicholas P., "Mixtures of Exponential Distributions," *The Annals of Statistics* 10, 2, June 1982, pp. 479–484.

[7] Klein, John P., and Melvin L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer, 1997.

[8] Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot, *Loss Models: From Data to Decisions*, New York: John Wiley & Sons, 1998.

[9] Lindsay, Bruce G., "Properties of the Maximum Likelihood Estimator of a Mixing Distribution," *Statistical Distributions in Scientific Work* 5, editors C. Taillie, G. Patil and B. Baldessari, Boston: D. Reidel, 1981, pp. 95–109.

[10] Lindsay, Bruce G., "The Geometry of Mixture Likelihoods: A General Theory," *The Annals of Statistics* 11, 1, March 1983, pp. 86–94.

[11] Lindsay, Bruce G., *Mixture Models: Theory, Geometry and Applications*, Hayward, California: Institute of Mathematical Statistics, 1995.

[12] Lindsay, Bruce G., and Mary L. Lesperance, "A Review of Semiparametric Mixture Models," *Journal of Statistical Planning and Inference* 47, 1/2, October 1995, pp. 29–39.

[13] Lindsay, Bruce G., and Kathryn Roeder, "Uniqueness of Estimation and Identifiability in Mixture Models," *The Canadian Journal of Statistics* 21, 2, June 1993, pp. 139–147.

[14] London, Dick, *Survival Models and Their Estimation*, Third Edition, Winsted, Connecticut: ACTEX Publications, 1997.

[15] Polyá, George, and Gabor Szegö, *Problems and Theorems in Analysis*, Volume II (revised and enlarged translation by Claude E. Billigheimer of *Aufgaben und Lehrsätze aus der Analysis*, Volume II, Fourth Edition, 1971), Berlin: Springer-Verlag, 1976.

[16] Tierney, Luke, and Diane Lambert, "Asymptotic Efficiency of Estimators of Functionals of Mixed Distributions," *The Annals of Statistics* 12, 4, December 1984, pp. 1380–1387.

APPENDIX A

In this appendix, I will address the issue of which of the parametric distributions generally used to model losses have completely monotone density functions and are thus special cases of the mixed exponential distribution. I will use the same parameterizations that are used in Klugman, Panjer, and Willmot [8].

The transformed beta distribution has probability density function

$$f(x) = \frac{\Gamma(\alpha+\tau)}{\Gamma(\alpha)\Gamma(\tau)} \frac{\gamma(x/\theta)^{\gamma\tau}}{x[1+(x/\theta)^{\gamma}]^{\alpha+\tau}}.$$

If $\gamma\tau > 1$, then $f(x)$ is not completely monotone because it has a nonzero mode.

If $\gamma\tau \leq 1$ and $\gamma \leq 1$, then $f(x)$ is completely monotone. To see this, note that, ignoring factors not involving x, we can write $f(x)$ as the product of $x^{\gamma\tau-1}$ and $[1+(x/\theta)^{\gamma}]^{-\alpha-\tau}$. The first factor is clearly completely monotone. We can use induction with the product rule for differentiation to show that the second factor is completely monotone. Similarly, we can use induction to show that the product of the two factors is also completely monotone. Feller [3, p. 441] gives a short proof of the fact that the product of completely monotone functions is also completely monotone.

Notable special cases of the transformed beta distribution that are also special cases of the mixed exponential distribution are the Pareto (which has γ and τ fixed at 1) and the Burr (which has τ fixed at 1) with $\gamma \leq 1$.

The set of parameters for which $f(x)$ is completely monotone when $\gamma\tau \leq 1$ and $\gamma > 1$ is an open question. If γ is too large, then $f(x)$ will not be completely monotone, but I could not find a proof that would definitively determine the status of all distributions with parameters in this region.

The transformed gamma distribution has probability density function

$$g(x) = \frac{\tau(x/\theta)^{\alpha\tau} e^{-(x/\theta)^{\tau}}}{x\Gamma(\alpha)}.$$

If $\alpha\tau > 1$, then $g(x)$ is not completely monotone because it has a nonzero mode.

If $\tau > 1$, then $g(x)$ is not completely monotone because it has an increasing failure rate in the tail.

If $\alpha\tau \leq 1$ and $\tau \leq 1$, then $g(x)$ is completely monotone. To see this, note that, ignoring factors not involving x, we can write $g(x)$ as the product of $x^{\alpha\tau-1}$ and $e^{-(x/\theta)^{\tau}}$. These are both completely monotone, so their product is completely monotone.

Notable special cases of the transformed gamma distribution that are also special cases of the mixed exponential distribution are the gamma (which has τ fixed at 1) with $\alpha \leq 1$ and the Weibull (which has α fixed at 1) with $\tau \leq 1$.

The inverse transformed gamma, lognormal, and inverse Gaussian distributions are never completely monotone, since they always have nonzero modes.

All of the distributions mentioned, except for the transformed gamma with certain parameters ($\tau > 1$ or $\tau = 1$, $\alpha \geq 1$), have decreasing failure rates in the tail.

APPENDIX B

In this appendix, I will provide proofs of the key properties underlying maximum likelihood estimation with the mixed exponential distribution—first for ungrouped data, then for grouped data.

Ungrouped Data

The loglikelihood function is

$$\ln L(w_1, w_2, \ldots) = \sum_{k=1}^{m} \ln f(x_k) = \sum_{k=1}^{m} \ln \left(\sum_{i=1}^{\infty} w_i \lambda_i e^{-\lambda_i x_k} \right),$$

where m is the number of observations. We must find the set of w_i's that maximizes the loglikelihood function, subject to the constraints that each of the w_i's must be greater than or equal to zero and the sum of the w_i's must be one. From now on, when I refer to maximizing the loglikelihood function, I mean maximizing the loglikelihood function subject to these constraints. We consider the λ_i's fixed and arbitrarily close together. Thus, the only parameters are the w_i's.

The ln function is strictly concave and the sum of strictly concave functions is also strictly concave.[12] This fact allows us to conclude that if more than one set of w_i's maximizes the loglikelihood function, each set must yield identical values of $\sum_{i=1}^{\infty} w_i \lambda_i e^{-\lambda_i x_k}$ for each x_k. If two sets of w_i's yielding different values of $\sum_{i=1}^{\infty} w_i \lambda_i e^{-\lambda_i x_k}$ maximized the loglikelihood function, each set of w_i's on the line segment between them (which would satisfy the constraints) would yield a value of the loglikelihood function greater than the maximum (since $\sum_{i=1}^{\infty} w_i \lambda_i e^{-\lambda_i x_k}$ is a linear function of the w_i's). Clearly, this cannot be.

[12] See Appendix 2 of Hillier and Lieberman [4] for a discussion of concavity and convexity.

We can view maximizing the loglikelihood function as a convex programming problem, since the loglikelihood function is concave and the constraints are linear (and thus convex). The theory of convex programming gives us a set of necessary and sufficient conditions, the Karush–Kuhn–Tucker (KKT) conditions, for the loglikelihood function to be at a maximum. For ungrouped data, these conditions are

$$\frac{\partial \ln L}{\partial w_i} = \sum_{k=1}^{m} \frac{\lambda_i e^{-\lambda_i x_k}}{\sum_{j=1}^{\infty} w_j \lambda_j e^{-\lambda_j x_k}} \leq u, \qquad \text{if} \quad w_i = 0$$

and

$$\frac{\partial \ln L}{\partial w_i} = \sum_{k=1}^{m} \frac{\lambda_i e^{-\lambda_i x_k}}{\sum_{j=1}^{\infty} w_j \lambda_j e^{-\lambda_j x_k}} = u, \qquad \text{if} \quad w_i > 0$$

for some number u. If we sum the KKT conditions, giving weight w_i to each element of the sum, we have

$$u = \sum_{i=1}^{\infty} w_i u = \sum_{i=1}^{\infty} w_i \frac{\partial \ln L}{\partial w_i} = \sum_{i=1}^{\infty} \sum_{k=1}^{m} \frac{w_i \lambda_i e^{-\lambda_i x_k}}{\sum_{j=1}^{\infty} w_j \lambda_j e^{-\lambda_j x_k}}$$

$$= \sum_{k=1}^{m} \frac{\sum_{i=1}^{\infty} w_i \lambda_i e^{-\lambda_i x_k}}{\sum_{j=1}^{\infty} w_j \lambda_j e^{-\lambda_j x_k}} = m.$$

Thus, we see that u must be equal to m, the number of observations.[13]

[13] See Chapter 13 of Hillier and Lieberman [4] for an introductory treatment of convex programming. Jewell [6] gave a direct derivation of the Karush–Kuhn–Tucker conditions for the mixed exponential case.

We now examine the function

$$h(\lambda) = \sum_{k=1}^{m} \frac{\lambda e^{-\lambda x_k}}{\sum_{j=1}^{\infty} w_j \lambda_j e^{-\lambda_j x_k}}, \qquad 0 \le \lambda \le \infty.$$

To satisfy the KKT conditions, this function must have a maximum of m that occurs at the points corresponding to where w_i is greater than zero. We first note that $h(0) = h(\infty) = 0$, so the w_i's corresponding to λ_i's of zero and infinity must be zero. Taking the derivative of $h(\lambda)$ gives

$$\frac{dh}{d\lambda} = \sum_{k=1}^{m} \frac{(-\lambda x_k + 1)e^{-\lambda x_k}}{\sum_{j=1}^{\infty} w_j \lambda_j e^{-\lambda_j x_k}}.$$

Polyá and Szegö [15] showed that an exponential polynomial of the form $\sum_{k=1}^{m} p_k(\lambda) e^{-\lambda x_k}$ that is not zero everywhere, where p_k is a real ordinary polynomial of degree d_k, has at most $\sum_{k=1}^{m} (d_k + 1) - 1$ zeros.[14] Thus $dh/d\lambda$ has at most $2m - 1$ zeros. When the KKT conditions are satisfied, $dh/d\lambda$ must be zero where $h(\lambda)$ assumes the value m on $(0, \infty)$. Since maxima must alternate with minima (where $dh/d\lambda$ must also be zero), $h(\lambda)$ can assume the value m at no more than m points on $(0, \infty)$. Since the w_i's corresponding to λ_i's of zero and infinity are zero, the number of positive w_i's at the point that the loglikelihood function is at its maximum is at most m, the number of observations.[15] We can also see that none of the corresponding λ_i's can be less than $1/x_m$, where x_m is the largest observation, since every term of the expression for $dh/d\lambda$ is positive for λ less than $1/x_m$. Likewise, none of the λ_i's can be greater than $1/x_1$, where x_1 is the smallest observation, since every term of the expression for $dh/d\lambda$ is negative for λ greater than $1/x_1$.

---

[14] See Part Five, Problem 75 of Polyá and Szegö [15].

[15] Using a more general technique, Lindsay [10] showed that this is true for mixtures of any type of distribution.

We will now determine whether the loglikelihood can attain its maximum at more than one set of w_i's. We do know that if more than one set yielded the maximum, each set would have to give the same value of $\sum_{i=1}^{n} w_i \lambda_i e^{-\lambda_i x_k}$ for each x_k. Let $\lambda_1, \ldots, \lambda_n$ be the points at which the w_i's are positive where the loglikelihood is at its maximum. If more than one set of w_i's gave the same value of $\sum_{i=1}^{n} w_i \lambda_i e^{-\lambda_i x_k}$ for each x_k, then the function $\sum_{i=1}^{n} (w_i - w_i^*) \lambda_i e^{-\lambda_i x}$ would have at least m zeros, one for each x_k. From Polyá and Szegö's result, this function can have no more than $n-1$ zeros. Since we have already determined that $n \leq m$, we have a contradiction. We thus conclude that the loglikelihood attains its maximum at a unique set of w_i's.[16]

Grouped Data

The loglikelihood function is

$$\begin{aligned}
\ln L(w_1, w_2, \ldots) &= a_1 \ln (1 - S(b_1)) + \sum_{k=2}^{g-1} a_k \ln (S(b_{k-1}) - S(b_k)) \\
&\quad + a_g \ln (S(b_{g-1})) \\
&= a_1 \ln \left(\sum_{i=1}^{\infty} w_i (1 - e^{-\lambda_i b_1}) \right) \\
&\quad + \sum_{k=2}^{g-1} a_k \ln \left(\sum_{i=1}^{\infty} w_i (e^{-\lambda_i b_{k-1}} - e^{-\lambda_i b_k}) \right) \\
&\quad + a_g \ln \left(\sum_{i=1}^{\infty} w_i (e^{-\lambda_i b_{g-1}}) \right),
\end{aligned}$$

where g is the number of groups, $a_1, \ldots, a_g$ are the number of observations in each group, and $b_1, \ldots, b_{g-1}$ are the group boundaries. We will assume that any adjacent groups that all have zero observations have been combined into one group. The development is analogous to that for ungrouped data down to where we

[16]The reasoning in this and the previous paragraph is taken from Jewell [6].

examine the function

$$h(\lambda) = a_1 \frac{1 - e^{-\lambda b_1}}{\sum_{j=1}^{\infty} w_j (1 - e^{-\lambda_j b_1})} + \sum_{k=2}^{g-1} a_k \frac{e^{-\lambda b_{k-1}} - e^{-\lambda b_k}}{\sum_{j=1}^{\infty} w_j (e^{-\lambda_j b_{k-1}} - e^{-\lambda_j b_k})}$$

$$+ a_g \frac{e^{-\lambda b_{g-1}}}{\sum_{j=1}^{\infty} w_j (e^{-\lambda_j b_{g-1}})}, \qquad 0 \le \lambda \le \infty.$$

We note that $h(0)$ and $h(\infty)$ are not necessarily equal to zero, so the w_i's corresponding to λ_i's of zero and infinity are not necessarily equal to zero. Taking the derivative of $h(\lambda)$ gives

$$\frac{dh}{d\lambda} = a_1 \frac{b_1 e^{-\lambda b_1}}{\sum_{j=1}^{\infty} w_j (1 - e^{-\lambda_j b_1})} + \sum_{k=2}^{g-1} a_k \frac{-b_{k-1} e^{-\lambda b_{k-1}} + b_k e^{-\lambda b_k}}{\sum_{j=1}^{\infty} w_j (e^{-\lambda_j b_{k-1}} - e^{-\lambda_j b_k})}$$

$$+ a_g \frac{-b_{g-1} e^{-\lambda b_{g-1}}}{\sum_{j=1}^{\infty} w_j (e^{-\lambda_j b_{g-1}})}$$

$$= \sum_{k=1}^{g-1} \left[\frac{a_k}{\sum_{j=1}^{\infty} w_j (e^{-\lambda_j b_{k-1}} - e^{-\lambda_j b_k})} - \frac{a_{k+1}}{\sum_{j=1}^{\infty} w_j (e^{-\lambda_j b_k} - e^{-\lambda_j b_{k+1}})} \right]$$

$$\times b_k e^{-\lambda b_k},$$

where $b_0 = 0$ and $b_g = \infty$.

We may now apply Polyá and Szegö's result, except if all of the $g - 1$ coefficients in the above equation are zero. This will occur only when the mixed exponential probabilities for each group are exactly proportional to the number of observations in

each group or, in other words, when the data perfectly fits the model. For this situation, we can easily come up with examples where an arbitrarily large number of different mixed exponential distributions, each with an arbitrarily large number of positive w_i's, will maximize the loglikelihood function. However, a perfect fit is highly unlikely unless the number of groups is very small.

When the fit is not perfect, Polyá and Szegö's result ensures that $dh/d\lambda$ has at most $g-2$ zeros. Thus, when the KKT conditions are satisfied, $h(\lambda)$ can assume the value m on $(0,\infty)$ at no more than $g/2-1$ points if g is even and no more than $g/2-1/2$ points if g is odd. This places a bound on the number of positive w_i's with corresponding λ_i's on $(0,\infty)$ at the point that the loglikelihood function is at its maximum. In addition, it is possible that the w_i's corresponding to λ_i's of zero and infinity may be positive.

We now move to the proof of uniqueness. Let $\lambda_1,\ldots,\lambda_n$ be the points at which the w_i's are positive where the loglikelihood is at its maximum. If more than one set of w_i's maximized the loglikelihood, each would have to give the same value of $\sum_{i=1}^{n} w_i(e^{-\lambda_i b_{k-1}} - e^{-\lambda_i b_k})$ for each group with a nonzero number of observations (where b_{k-1} and b_k are the group boundaries). Since adjacent groups with zero observations have been combined, the minimum number of such groups will be $g/2$ if g is even and $g/2-1/2$ if g is odd. Therefore, $\sum_{i=1}^{n}(w_i - w_i^*)(e^{-\lambda_i b_{k-1}} - e^{-\lambda_i b_k})$ has to be zero for each of these groups. This implies that, for each group, the function $\sum_{i=1}^{n}(w_i - w_i^*)e^{-\lambda_i x}$ has the same value at both b_{k-1} and b_k. Thus the derivative of this function must be zero somewhere between b_{k-1} and b_k. Therefore, the function $\sum_{i=1}^{n}(w_i - w_i^*)\lambda_i e^{-\lambda_i x}$ must have at least $g/2$ zeros if g is even and at least $g/2-1/2$ zeros if g is odd. From Polyá and Szegö's result, this function can have no more than n^*-1 zeros, where n^* is the number of λ_i's at which the w_i's are positive, excluding λ_i's of zero and infinity (since these terms drop out of the function). Since

we have already determined that $n^* \leq g/2 - 1$ if g is even and $n^* \leq g/2 - 1/2$ if g is odd, we have a contradiction. We thus conclude that the loglikelihood attains its maximum at a unique set of w_i's.[17]

[17]Using a more general technique, Lindsay and Roeder [13] derived similar results to those for grouped data shown here. Those results apply to mixtures of a broader class of distributions.

APPENDIX C

Use of Newton's method requires calculation of the gradient vector of first partial derivatives and the Hessian matrix of second partial derivatives of the loglikelihood function.

In the derivatives that follow, w_1 is not a real parameter, but we set w_1 equal to one minus the sum of the other w_i's.[18]

$$\left(\frac{\partial \ln L}{\partial \lambda_i}\right)_k \qquad \text{and} \qquad \left(\frac{\partial \ln L}{\partial w_i}\right)_k$$

refer to the terms of the first partial derivatives corresponding to the kth observation (for ungrouped data) or kth group (for grouped data).

For ungrouped data, the required derivatives are

$$\frac{\partial \ln L}{\partial \lambda_i} = \sum_{k=1}^{m} \left(\frac{\partial \ln L}{\partial \lambda_i}\right)_k = \sum_{k=1}^{m} \frac{w_i(1-\lambda_i x_k)e^{-\lambda_i x_k}}{\sum_{j=1}^{n} w_j \lambda_j e^{-\lambda_j x_k}},$$

$$i = 1,\ldots,n,$$

$$\frac{\partial \ln L}{\partial w_i} = \sum_{k=1}^{m} \left(\frac{\partial \ln L}{\partial w_i}\right)_k = \sum_{k=1}^{m} \frac{\lambda_i e^{-\lambda_i x_k} - \lambda_1 e^{-\lambda_1 x_k}}{\sum_{j=1}^{n} w_j \lambda_j e^{-\lambda_j x_k}},$$

$$i = 2,\ldots,n,$$

$$\frac{\partial^2 \ln L}{\partial \lambda_i^2} = \sum_{k=1}^{m} \left[\frac{w_i x_k(\lambda_i x_k - 2)e^{-\lambda_i x_k}}{\sum_{j=1}^{n} w_j \lambda_j e^{-\lambda_j x_k}} - \left(\left(\frac{\partial \ln L}{\partial \lambda_i}\right)_k\right)^2\right],$$

$$i = 1,\ldots,n,$$

[18] An alternative way to formulate the problem would be to keep w_1 as a parameter and use a Lagrange multiplier to ensure that the sum of the w_i's is one.

$$\frac{\partial^2 \ln L}{\partial \lambda_i \partial \lambda_l} = \sum_{k=1}^{m} \left[-\left(\frac{\partial \ln L}{\partial \lambda_i}\right)_k \left(\frac{\partial \ln L}{\partial \lambda_l}\right)_k \right],$$

$$i = 1,\ldots,n, \quad l = 1,\ldots,n, \quad i \neq l,$$

$$\frac{\partial^2 \ln L}{\partial w_i \partial w_l} = \sum_{k=1}^{m} \left[-\left(\frac{\partial \ln L}{\partial w_i}\right)_k \left(\frac{\partial \ln L}{\partial w_l}\right)_k \right],$$

$$i = 2,\ldots,n, \quad l = 2,\ldots,n,$$

$$\frac{\partial^2 \ln L}{\partial \lambda_1 \partial w_i} = \sum_{k=1}^{m} \left[\frac{-(1-\lambda_1 x_k)e^{-\lambda_1 x_k}}{\sum_{j=1}^{n} w_j \lambda_j e^{-\lambda_j x_k}} - \left(\frac{\partial \ln L}{\partial \lambda_1}\right)_k \left(\frac{\partial \ln L}{\partial w_i}\right)_k \right],$$

$$i = 2,\ldots,n,$$

$$\frac{\partial^2 \ln L}{\partial \lambda_i \partial w_i} = \sum_{k=1}^{m} \left[\frac{(1-\lambda_i x_k)e^{-\lambda_i x_k}}{\sum_{j=1}^{n} w_j \lambda_j e^{-\lambda_j x_k}} - \left(\frac{\partial \ln L}{\partial \lambda_i}\right)_k \left(\frac{\partial \ln L}{\partial w_i}\right)_k \right],$$

$$i = 2,\ldots,n,$$

and $$\frac{\partial^2 \ln L}{\partial \lambda_i \partial w_l} = \sum_{k=1}^{m} \left[-\left(\frac{\partial \ln L}{\partial \lambda_i}\right)_k \left(\frac{\partial \ln L}{\partial w_l}\right)_k \right],$$

$$i = 2,\ldots,n, \quad l = 2,\ldots,n, \quad i \neq l.$$

For grouped data, the required derivatives are

$$\frac{\partial \ln L}{\partial \lambda_i} = \sum_{k=1}^{g} a_k \left(\frac{\partial \ln L}{\partial \lambda_i}\right)_k = \sum_{k=1}^{g} a_k \frac{w_i(-b_{k-1}e^{-\lambda_i b_{k-1}} + b_k e^{-\lambda_i b_k})}{\sum_{j=1}^{n} w_j (e^{-\lambda_j b_{k-1}} - e^{-\lambda_j b_k})},$$

$$i = 1,\ldots,n,$$

$$\frac{\partial \ln L}{\partial w_i} =$$

$$\sum_{k=1}^{g} a_k \left(\frac{\partial \ln L}{\partial w_i} \right)_k = \sum_{k=1}^{g} a_k \frac{(e^{-\lambda_i b_{k-1}} - e^{-\lambda_i b_k}) - (e^{-\lambda_1 b_{k-1}} - e^{-\lambda_1 b_k})}{\sum_{j=1}^{n} w_j (e^{-\lambda_j b_{k-1}} - e^{-\lambda_j b_k})},$$

$$i = 2, \ldots, n,$$

$$\frac{\partial^2 \ln L}{\partial \lambda_i^2} = \sum_{k=1}^{g} a_k \left[\frac{w_i (b_{k-1}^2 e^{-\lambda_i b_{k-1}} - b_k^2 e^{-\lambda_i b_k})}{\sum_{j=1}^{n} w_j (e^{-\lambda_j b_{k-1}} - e^{-\lambda_j b_k})} - \left(\left(\frac{\partial \ln L}{\partial \lambda_i} \right)_k \right)^2 \right],$$

$$i = 1, \ldots, n,$$

$$\frac{\partial^2 \ln L}{\partial \lambda_i \partial \lambda_l} = \sum_{k=1}^{g} a_k \left[- \left(\frac{\partial \ln L}{\partial \lambda_i} \right)_k \left(\frac{\partial \ln L}{\partial \lambda_l} \right)_k \right],$$

$$i = 1, \ldots, n, \quad l = 1, \ldots, n, \quad i \neq l,$$

$$\frac{\partial^2 \ln L}{\partial w_i \partial w_l} = \sum_{k=1}^{g} a_k \left[- \left(\frac{\partial \ln L}{\partial w_i} \right)_k \left(\frac{\partial \ln L}{\partial w_l} \right)_k \right],$$

$$i = 2, \ldots, n, \quad l = 2, \ldots, n,$$

$$\frac{\partial^2 \ln L}{\partial \lambda_1 \partial w_i} = \sum_{k=1}^{g} a_k \left[\frac{(b_{k-1} e^{-\lambda_1 b_{k-1}} - b_k e^{-\lambda_1 b_k})}{\sum_{j=1}^{n} w_j (e^{-\lambda_j b_{k-1}} - e^{-\lambda_j b_k})} - \left(\frac{\partial \ln L}{\partial \lambda_1} \right)_k \left(\frac{\partial \ln L}{\partial w_i} \right)_k \right],$$

$$i = 2, \ldots, n,$$

$$\frac{\partial^2 \ln L}{\partial \lambda_i \partial w_i} = \sum_{k=1}^{g} a_k \left[\frac{(-b_{k-1} e^{-\lambda_i b_{k-1}} + b_k e^{-\lambda_i b_k})}{\sum_{j=1}^{n} w_j (e^{-\lambda_j b_{k-1}} - e^{-\lambda_j b_k})} - \left(\frac{\partial \ln L}{\partial \lambda_i} \right)_k \left(\frac{\partial \ln L}{\partial w_i} \right)_k \right],$$

$$i = 2, \ldots, n,$$

$$\text{and} \qquad \frac{\partial^2 \ln L}{\partial \lambda_i \partial w_l} = \sum_{k=1}^{g} a_k \left[-\left(\frac{\partial \ln L}{\partial \lambda_i} \right)_k \left(\frac{\partial \ln L}{\partial w_l} \right)_k \right],$$
$$i = 2, \ldots, n, \quad l = 2, \ldots, n, \quad i \neq l.$$

The Newton step is the inverse of the Hessian matrix multiplied by the negative of the gradient vector. To remove one of the parameters from the iterative process without reconstructing the entire gradient and Hessian, set that parameter's component of the gradient to zero, its diagonal element of the Hessian matrix to one, and the off-diagonal elements of its row and column of the Hessian matrix to zero.

With ungrouped data, the fitted mixed exponential mean will always equal the sample mean at both the global maximum and at local maxima. To see this, first note that each of the $\partial \ln L / \partial w_i$ values must be zero, so the KKT equalities are satisfied. We have seen that this implies that

$$\sum_{k=1}^{m} \frac{\lambda_i e^{-\lambda_i x_k}}{\sum_{j=1}^{n} w_j \lambda_j e^{-\lambda_j x_k}} = m, \qquad i = 1, \ldots, n.$$

Since each of the $\partial \ln L / \partial \lambda_i$ values must be zero, we may sum over them to obtain

$$\sum_{i=1}^{n} \sum_{k=1}^{m} \frac{w_i (1 - \lambda_i x_k) e^{-\lambda_i x_k}}{\sum_{j=1}^{n} w_j \lambda_j e^{-\lambda_j x_k}} = \sum_{i=1}^{n} \left[w_i \frac{1}{\lambda_i} \sum_{k=1}^{m} \frac{\lambda_i e^{-\lambda_i x_k}}{\sum_{j=1}^{n} w_j \lambda_j e^{-\lambda_j x_k}} \right]$$
$$- \sum_{k=1}^{m} \left[x_k \sum_{i=1}^{n} \frac{w_i \lambda_i e^{-\lambda_i x_k}}{\sum_{j=1}^{n} w_j \lambda_j e^{-\lambda_j x_k}} \right] = m \sum_{i=1}^{n} w_i \frac{1}{\lambda_i} - \sum_{k=1}^{m} x_k = 0.$$

Since $1/\lambda_i$ is the mean of the ith exponential distribution in the mixture, we can see that the mixed exponential mean must indeed be equal to the sample mean.

Also, with ungrouped data, the fitted mixed exponential variance will not be less than the sample variance at the global maximum. To see this, first note that at each of the λ_i's with positive weight attached, $d^2h/d\lambda^2$ must be less than or equal to zero. We may sum over these second derivatives, giving weight w_i to each element of the sum, to obtain

$$\sum_{i=1}^{n}\sum_{k=1}^{m}\frac{w_i(\lambda_i x_k^2 - 2x_k)e^{-\lambda_i x_k}}{\sum_{j=1}^{n} w_j\lambda_j e^{-\lambda_j x_k}} = \sum_{k=1}^{m}\left[x_k^2\sum_{i=1}^{n}\frac{w_i\lambda_i e^{-\lambda_i x_k}}{\sum_{j=1}^{n} w_j\lambda_j e^{-\lambda_j x_k}}\right]$$

$$-\sum_{i=1}^{n}\left[w_i\frac{2}{\lambda_i}\sum_{k=1}^{m}\frac{\lambda_i x_k e^{-\lambda_i x_k}}{\sum_{j=1}^{n} w_j\lambda_j e^{-\lambda_j x_k}}\right]$$

$$=\sum_{k=1}^{m}x_k^2 - \sum_{i=1}^{n}\left[w_i\frac{2}{\lambda_i}\sum_{k=1}^{m}\frac{e^{-\lambda_i x_k}}{\sum_{j=1}^{n} w_j\lambda_j e^{-\lambda_j x_k}}\right]$$

$$=\sum_{k=1}^{m}x_k^2 - \sum_{i=1}^{n}\left[w_i\frac{2}{\lambda_i^2}\sum_{k=1}^{m}\frac{\lambda_i e^{-\lambda_i x_k}}{\sum_{j=1}^{n} w_j\lambda_j e^{-\lambda_j x_k}}\right]$$

$$=\sum_{k=1}^{m}x_k^2 - m\sum_{i=1}^{n}w_i\frac{2}{\lambda_i^2} \le 0.$$

To get from the term in the second line above to the second term in the third line, we use the fact that each of the $\partial \ln L/\partial \lambda_i$ values must be zero. Since $2/\lambda_i^2$ is the second moment of the ith exponential distribution in the mixture, and since we know that the mixed exponential mean must be equal to the sample mean, we can see that the mixed exponential variance cannot be less than the sample variance.[19]

[19]Lindsay [9] showed that these moment relationships hold for mixtures of a broader class of distributions.