DISCUSSION OF PAPER PUBLISHED IN VOLUME LXXX

MINIMUM DISTANCE ESTIMATION OF LOSS DISTRIBUTIONS

STUART A. KLUGMAN AND A. RAHULJI PARSA

DISCUSSION BY CLIVE L. KEATINGE

1. INTRODUCTION

Klugman and Parsa have introduced the theory underlying minimum distance estimation with parametric distributions. In this review, I develop their ideas further to provide a more complete view of the characteristics of minimum distance estimation. I conclude that minimum distance estimation can be more efficient than the authors imply—but that there is little basis for using it in place of maximum likelihood estimation.

2. THEORY

The objective function that Klugman and Parsa consider is

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{k} w_i [G(c_i;\boldsymbol{\theta}) - G_n(c_i)]^2, \tag{2.1}$$

where G is the model functional, G_n is the corresponding empirical functional, $c_1 < c_2 < \cdots < c_k$ are arbitrarily selected values, and $w_1, w_2, \ldots, w_k > 0$ are arbitrarily selected weights. The functionals that Klugman and Parsa consider are the limited expected value function and the cumulative distribution function. The minimum distance estimate is the value of $\boldsymbol{\theta}$ that minimizes $Q(\boldsymbol{\theta})$. From here on, I will follow the authors' convention of writing $G(c_i;\boldsymbol{\theta})$ as G_i and $G_n(c_i)$ as $G_{n,i}$.

A necessary condition for $Q(\boldsymbol{\theta})$ to be at a minimum is for the p functions

$$\partial Q/\partial\theta_j = 2\sum_{i=1}^{k} w_i[G_i - G_{n,i}]G_i^{(j)} \tag{2.2}$$

to be equal to zero, where $G_i^{(j)}$ is the partial derivative of the model functional with respect to θ_j evaluated at c_i, and p is the number of elements in the parameter vector $\boldsymbol{\theta}$. Another necessary condition for $Q(\boldsymbol{\theta})$ to be at a minimum is for the $p \times p$ matrix with jlth element

$$\partial^2 Q/\partial\theta_j\partial\theta_l = 2\sum_{i=1}^{k} w_i G_i^{(j)} G_i^{(l)} + 2\sum_{i=1}^{k} w_i[G_i - G_{n,i}]G_i^{(j,l)} \tag{2.3}$$

to be positive semidefinite (which includes the positive definite case).

As the sample size goes to infinity, there will be a solution that satisfies these two conditions if and only if the $p \times p$ matrix with jlth element

$$E[\partial^2 Q/\partial\theta_j\partial\theta_l] = 2\sum_{i=1}^{k} w_i G_i^{(j)} G_i^{(l)} \tag{2.4}$$

is positive semidefinite, where the derivatives are evaluated at the true parameter values. If all the weights are positive, this matrix must be positive semidefinite. Though having some negative weights is counterintuitive, the theory does not rule them out as long as the matrix is positive semidefinite. For smaller sample sizes, the more negative weights there are, and the larger they are in magnitude, the less likely it is that a solution that satisfies the two necessary conditions above will exist.

Klugman and Parsa state that the minimum distance estimator is consistent and asymptotically unbiased with asymptotic covariance matrix $n^{-1}\mathbf{A}^{-1}\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'\mathbf{A}^{-1}$ if $\mathbf{A}$ is positive definite, where

A is the matrix defined by Equation (2.3), **B** is the $p \times k$ matrix with jlth element $\partial^2 Q/\partial\theta_j \partial G_{n,l} = -2w_l G_l^{(j)}$, and $n^{-1}\boldsymbol{\Sigma}$ is the asymptotic covariance matrix of the empirical functional. This is correct, except that, as noted in Benichou and Gail [2], one should use the asymptotic expectation of the empirical functional instead of the observed value in Equation (2.3). Making this correction causes the second term of Equation (2.3) to vanish, thus yielding Equation (2.4). Luong and Thompson [4] show this result in a more general setting.

An issue that Klugman and Parsa do not address is identifying the sets of weights that will produce the minimum asymptotic variance for the estimators of the parameters or of functions of the parameters. A set of weights $w_1, w_2, \ldots, w_k$ will produce the minimum asymptotic variance for the estimator of a function $h(\boldsymbol{\theta})$ if **A** is positive definite and

$$w_i = \frac{(\boldsymbol{\Sigma}^{-1}\mathbf{D}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{d})_i}{(\mathbf{D}'\mathbf{v})_i}, \tag{2.5}$$

where **D** is the $p \times k$ matrix with jlth element $G_l^{(j)}$, **d** is the vector of length p with jth element $\partial h/\partial\theta_j$, and **v** is an arbitrary nonzero vector of length p. The minimum asymptotic variance is $n^{-1}\mathbf{d}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{d}$. The proof is in the appendix. Defining $h(\boldsymbol{\theta})$ to be θ_j yields weights that produce the minimum asymptotic variance for the estimator of the parameter θ_j itself. The main diagonal of $n^{-1}(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}$ gives the minimum asymptotic variances for the estimators of the θ_js.

In general, the asymptotic variances cannot be minimal for all of the parameters at the same time. However, the asymptotic variances can be minimal simultaneously if the definition of the objective function is expanded to

$$Q^*(\boldsymbol{\theta}) = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}[G_i - G_{n,i}][G_j - G_{n,j}]. \tag{2.6}$$

The equation uses an entire matrix of weights instead of one weight for each c_i. The appendix gives the minimum asymptotic variance condition. The most obvious matrix of weights that satisfies this condition is $\boldsymbol{\Sigma}^{-1}$. Luong and Thompson [4] show this result in a more general setting.

When estimating parameters, it is not possible to find an optimal set of weights, since the true values of the parameters are unknown. A reasonable requirement is that the weights used to estimate the parameters be asymptotically optimal, or at least close to asymptotically optimal, under the assumption that the estimated parameter values are the true parameter values. Finding an acceptable set of weights by trial and error is one option. Alternatively, a systematic procedure that often works, hereafter called Procedure 1, is to estimate the parameters using any reasonable set of weights, optimize the weights using the estimated parameter values, estimate the parameters again using the new set of weights, and so on, until the process converges. Yet it is possible that the process will not converge.

With Equation (2.1), finding the optimal set of weights at each iteration of the process is problematical, since minimum asymptotic variance can be achieved for only one parameter at a time. One possible solution is to consider the sets of weights defined by Equation (2.5) with $h(\boldsymbol{\theta})$ defined to be one particular parameter θ_j, look at the ratios of the diagonal elements of $n^{-1}\mathbf{A}^{-1}\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'\mathbf{A}^{-1}$ to the corresponding diagonal elements of the minimum asymptotic covariance matrix $n^{-1}(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}$, and then search for a $\mathbf{v}$ that minimizes the sum of these ratios. There is no easy way to do this, since, as the components of $\mathbf{v}$ vary, there are many local minima for the sum. The best one can practically do is to systematically try a number of values for the components of $\mathbf{v}$ and use those that yield the smallest values for the sum as starting values in an optimization routine. An additional potential problem is that the minimum for the sum could occur at a point where $\mathbf{A}$ is not positive definite. In that case, modifications to the procedure would be necessary.

With Equation (2.6), the easiest matrix of weights to use at each iteration of the process is $\boldsymbol{\Sigma}^{-1}$. If the functional is the cumulative distribution function, then the process, if it converges, yields the grouped maximum likelihood estimate. The appendix shows this result. Of course, it would be much easier just to find the grouped maximum likelihood estimate directly.

Another possible procedure, hereafter called Procedure 2, is to use $\boldsymbol{\Sigma}^{-1}$ as the matrix of weights in Equation (2.6) and to treat it as a function of the parameters, instead of fixed, when minimizing the objective function. This procedure produces an estimate for each of the parameters directly, instead of a series of estimates that might or might not converge. If the functional is the cumulative distribution function, the result is the minimum chi-square estimate. The appendix shows this result.

Moore [5] shows that the asymptotic covariance matrix of both the grouped maximum likelihood estimator and the minimum chi-square estimator is $n^{-1}(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}$, with the cumulative distribution function and the true parameter values used to evaluate the expression. If one uses Procedure 1 with Equation (2.6), or if one uses Procedure 2, with a functional other than the cumulative distribution function, similar reasoning reveals that the asymptotic covariance matrix will also be $n^{-1}(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}$, with the selected functional and the true parameter values used to evaluate the expression. If one uses Procedure 1 with Equation (2.1), the asymptotic covariance matrix will be $n^{-1}\mathbf{A}^{-1}\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'\mathbf{A}^{-1}$, with weights optimized using the selected functional and the true parameter values.

In the last section of their paper, Klugman and Parsa provide results of a simulation study they conducted with samples of size 500 to investigate how well the asymptotic estimates perform. They make the point that averaging the values in estimated asymptotic covariance matrices considerably overstates the values in the true asymptotic covariance matrix when

using minimum distance estimation with the Pareto distribution.

This phenomenon occurs because the distribution of estimates is skewed to the right. The overstatement is not a feature specific to minimum distance estimation. It also occurs with other estimation methods, including maximum likelihood estimation.

Right skewness also causes the sample covariance matrix of parameter estimates to tend to be larger than the asymptotic covariance matrix, though Klugman and Parsa did not note this because of an errant asymptotic covariance matrix. They showed this matrix to be

$$\begin{bmatrix} 0.6640 & 120.3 \\ 120.3 & 21{,}830 \end{bmatrix}$$

when it should have been

$$\begin{bmatrix} 0.3595 & 75.68 \\ 75.68 & 16{,}794 \end{bmatrix}.$$

Their sample covariance matrix of parameter estimates was

$$\begin{bmatrix} 0.5133 & 108.8 \\ 108.8 & 24{,}150 \end{bmatrix}.$$

This right skewness of estimates is a feature of the Pareto distribution. Other distributions may exhibit different behavior.

3. EXAMPLES

I will now illustrate results from the previous section using examples that Klugman and Parsa use. I will also discuss each of the examples. Where I show numerical values that differ slightly from what Klugman and Parsa show in their paper, I have used values that I believe are more accurate.

Example One—Improving the Efficiency of the Minimum Distance Estimator

The first example involves a Pareto distribution fit to 6,656 general liability claims. Klugman and Parsa show 10,000,000 as the largest c_i. However, they actually used 100,000,000. I will use that value here. Grouped maximum likelihood estimation yields parameter estimates of $\hat{\alpha} = 1.4826$ and $\hat{\lambda} = 705.79$. The asymptotic covariance matrix for these parameter values is

$$\begin{bmatrix} 0.0020472 & 1.3679 \\ 1.3679 & 1{,}090.4 \end{bmatrix}.$$

Klugman and Parsa then use minimum distance estimation with the limited expected value function and weights of 1 at all c_is. This yields parameter estimates of $\hat{\alpha} = 1.3388$ and $\hat{\lambda} = 590.33$. The authors give the asymptotic covariance matrix for these parameter values as

$$\begin{bmatrix} 0.034751 & 33.571 \\ 33.571 & 32{,}765 \end{bmatrix}.$$

Making the correction to the matrix **A** noted in the previous section yields a corrected asymptotic covariance matrix of

$$\begin{bmatrix} 0.036691 & 35.603 \\ 35.603 & 34{,}880 \end{bmatrix}.$$

These values are substantially higher than the maximum likelihood values. A fairer comparison would be to compare the asymptotic covariance matrices at the same parameter values. The minimum distance asymptotic covariance matrix at the maximum likelihood parameter estimates is

$$\begin{bmatrix} 0.023518 & 21.819 \\ 21.819 & 20{,}579 \end{bmatrix}.$$

These values are still substantially higher than the maximum likelihood values.

The procedures described in the previous section can do better. Using Procedure 1 and Equation (2.1), we can start with the authors' estimates of $\hat{\alpha} = 1.3388$ and $\hat{\lambda} = 590.33$. We define $h(\boldsymbol{\theta})$ to be the parameter α and proceed with the iterative process until it converges to estimates of $\hat{\alpha} = 1.4753$ and $\hat{\lambda} = 704.94$, quite similar to the maximum likelihood estimates.

Using Procedure 1 and Equation (2.6), we can also start with the authors' estimates of $\hat{\alpha} = 1.3388$ and $\hat{\lambda} = 590.33$. With $\boldsymbol{\Sigma}^{-1}$ as the matrix of weights, the process converges to estimates of $\hat{\alpha} = 1.4752$ and $\hat{\lambda} = 705.51$, again quite similar to the maximum likelihood estimates.

Using Procedure 2 produces estimates of $\hat{\alpha} = 1.4431$ and $\hat{\lambda} = 684.63$, somewhat removed from the maximum likelihood estimates, but still much closer to them than to the authors' minimum distance estimates.

To compare the different estimators directly, we will examine the asymptotic covariance matrices for each of the estimators at the maximum likelihood parameter estimates. Both Procedure 1 with Equation (2.6) and Procedure 2 have asymptotic covariance matrices of

$$\begin{bmatrix} 0.0020356 & 1.3594 \\ 1.3594 & 1{,}083.6 \end{bmatrix}.$$

These are slightly smaller than the maximum likelihood values. Procedure 1 with Equation (2.1) has an asymptotic covariance matrix of

$$\begin{bmatrix} 0.0020356 & 1.3594 \\ 1.3594 & 1{,}083.7 \end{bmatrix}.$$

The asymptotic variances of $\hat{\alpha}$ and the asymptotic covariances of $\hat{\alpha}$ with $\hat{\lambda}$ are identical in the preceding two matrices. This must

be true, as the appendix shows. The asymptotic variance of $\hat{\lambda}$ is only very slightly higher in the second matrix than in the first.

Table 1 shows the weights that underlie the second matrix, generated by a **v** of

$$\begin{bmatrix} 0.000001 \\ 0.0007093 \end{bmatrix}.$$

One could multiply this vector by a nonzero factor without affecting the variances or covariance. However, the factor has to be positive to keep **A** positive definite. Note that several of the weights are negative, but that is not a problem here. Adjacent positive weights offset the two largest negative weights. Of course, assigning the weight within each pair to just one of the two adjacent values would yield virtually the same result.

The optimized weights decrease substantially as the c_is increase, in contrast to the uniform weights that Klugman and Parsa use. Clearly, the poor performance of the uniform weights results from an excessive amount of weight in the tail of the distribution. If one were simply to remove the weight at 100,000,000, the asymptotic covariance matrix at the maximum likelihood parameter estimates would improve to

$$\begin{bmatrix} 0.0069132 & 5.8391 \\ 5.8391 & 5{,}199.8 \end{bmatrix}.$$

Example One—Discussion

Table 2 shows the empirical limited expected values along with the fitted limited expected values for the maximum likelihood parameter estimates and the original minimum distance estimates with uniform weights. I confine attention to these here, since the other minimum distance estimates obtained are fairly close to the maximum likelihood estimates.

I suspect that most modelers would prefer the original minimum distance parameter estimates, since they provide a much closer fit in the tail at a modest cost in terms of the fit low in

TABLE 1

EXAMPLE 1 WEIGHTS

Limit	Original	Optimized
50	1	5509
100	1	1279
150	1	537
200	1	300
250	1	145
300	1	183
400	1	130
500	1	71
600	1	49
700	1	32
800	1	30
900	1	−4
1,000	1	52
1,500	1	36
2,000	1	18
2,500	1	−4
3,000	1	2.0
3,500	1	1.7
4,000	1	1.4
4,500	1	1.1
4,999	1	−76
5,000	1	77
6,000	1	1.3
7,500	1	1.3
9,999	1	160
10,000	1	−159
12,000	1	0.48
15,000	1	0.51
20,000	1	0.31
25,000	1	0.25
35,000	1	0.21
50,000	1	0.15
75,000	1	0.05
100,000	1	0.11
250,000	1	0.054
500,000	1	0.0011
1,000,000	1	0.0236
100,000,000	1	0.00087

TABLE 2

EXAMPLE 1 LIMITED EXPECTED VALUES

		Pareto		Mixed Exponential	
Limit	Empirical	Maximum Likelihood	Min Dist Unif Wts	Maximum Likelihood	Min Dist Unif Wts
50	48	48	47	48	48
100	92	91	90	91	91
150	133	130	129	131	131
200	170	166	164	168	168
250	203	199	196	202	202
300	235	230	226	233	233
400	291	285	280	288	289
500	338	333	327	336	337
600	379	376	368	378	379
700	415	414	406	415	416
800	448	448	439	447	449
900	477	479	469	477	478
1,000	504	507	497	503	505
1,500	610	619	607	607	610
2,000	686	698	687	682	686
2,500	745	758	748	740	744
3,000	792	806	797	787	792
3,500	831	844	838	826	831
4,000	864	877	873	860	865
4,500	893	905	903	889	894
4,999	919	929	929	914	919
5,000	920	929	929	914	919
6,000	962	969	973	957	963
7,500	1,014	1,015	1,025	1,008	1,014
9,999	1,079	1,069	1,087	1,071	1,078
10,000	1,079	1,069	1,087	1,071	1,078
12,000	1,117	1,100	1,124	1,110	1,118
15,000	1,163	1,135	1,168	1,156	1,164
20,000	1,222	1,176	1,219	1,213	1,221
25,000	1,264	1,204	1,256	1,254	1,263
35,000	1,318	1,242	1,308	1,308	1,319
50,000	1,367	1,277	1,357	1,353	1,367
75,000	1,408	1,309	1,406	1,393	1,408
100,000	1,433	1,329	1,437	1,416	1,430
250,000	1,511	1,377	1,518	1,481	1,514
500,000	1,587	1,401	1,565	1,540	1,592
1,000,000	1,662	1,418	1,602	1,592	1,650
100,000,000	1,662	1,458	1,713	1,618	1,669

the distribution. This is in spite of the fact that the original minimum distance estimator has much greater asymptotic variances than the maximum likelihood estimator. If one makes this judgment, then one is implicitly acknowledging that the assumption that the data comes from a Pareto distribution is not appropriate here. Otherwise, one would prefer the estimator with the smaller asymptotic variances.

This situation is quite common with parametric distributions. They often are not flexible enough to provide a high quality fit over the entire range of the data. In this case, there are alternatives to using minimum distance estimation with weights selected to trade off the quality of fit in one part of the distribution for another. One option would be to fit a parametric distribution to the upper section of the data only and to use the empirical distribution below that. Another option would be to use the semiparametric mixed exponential distribution, which is more flexible and thus better able to provide a good fit over the entire distribution in many situations. The mixed exponential distribution generally works very well with loss distributions, as I discussed in detail in Keatinge [3].

Table 2 shows the fitted limited expected values for mixed exponential distributions fit using maximum likelihood estimation and minimum distance estimation with uniform weights. The means and weights of the exponential distributions in each mixture are as follows:

Maximum Likelihood		Minimum Distance Uniform Weights	
Mean	Weight	Mean	Weight
398	0.659077	394	0.648798
1,326	0.215884	1,405	0.259393
3,097	0.088849	4,446	0.067363
12,285	0.030721	18,513	0.023568
36,128	0.004935	356,076	0.000878
445,785	0.000535		

Each of these provides an excellent fit over the entire range of the data. In this comparison of limited expected values, the minimum distance estimator provides a closer fit than the maximum likelihood estimator because it uses the empirical limited expected values directly, whereas the maximum likelihood estimator uses the number of losses that fall in each interval. Since the mixed exponential distribution is flexible enough to provide a good fit over the entire distribution, it is not very sensitive to the choice of weights.

Example Two—Improving the Efficiency of the Minimum Distance Estimator

The second example involves 463 medical malpractice claim report lags truncated from above and fit to a Burr distribution. Grouped maximum likelihood estimation yields parameter estimates of $\hat{\alpha} = 0.40274$, $\hat{\lambda} = 34.224$, and $\hat{\tau} = 3.1181$. The asymptotic covariance matrix for these parameter values is

$$\begin{bmatrix} 0.017336 & 0.57436 & -0.035566 \\ 0.57436 & 20.6558 & -1.21351 \\ -0.035566 & -1.21351 & 0.10703 \end{bmatrix}.$$

Klugman and Parsa then use minimum distance estimation with the cumulative distribution function. They use weights of 4 where the empirical cumulative distribution function is less than 0.5, and the reciprocal of the empirical variance where the cumulative distribution function is greater than 0.5. (The weight at lag 162 is set equal to the weight at lag 156, since the empirical cumulative distribution function at lag 162 is 1.) This might or might not produce good results, but there is no particular theoretical justification for it, since it does not take into account the correlation among the values of the empirical functional. These weights yield parameter estimates of $\hat{\alpha} = 0.48800$, $\hat{\lambda} = 36.989$, and $\hat{\tau} = 2.9495$. The authors give the asymptotic covariance matrix for these parameter values

as

$$\begin{bmatrix} 0.081077 & 2.6655 & -0.16625 \\ 2.6655 & 89.507 & -5.5313 \\ -0.16625 & -5.5313 & 0.33525 \end{bmatrix}.$$

Making the correction to the matrix **A** noted in the previous section yields a corrected asymptotic covariance matrix of

$$\begin{bmatrix} 0.055717 & 1.8069 & -0.10330 \\ 1.8069 & 60.474 & -3.3933 \\ -0.10330 & -3.3933 & 0.22450 \end{bmatrix}.$$

These values are substantially higher than the maximum likelihood values. A fairer comparison would be to compare the asymptotic covariance matrices at the same parameter values. The minimum distance asymptotic covariance matrix at the maximum likelihood parameter estimates is

$$\begin{bmatrix} 0.039702 & 1.3327 & -0.10037 \\ 1.3327 & 46.371 & -3.4111 \\ -0.10037 & -3.4111 & 0.29501 \end{bmatrix}.$$

These values are still significantly higher than the maximum likelihood values.

We now try the procedures described in the previous section. Using Procedure 1 and Equation (2.1), we can start with the authors' estimates of $\hat{\alpha} = 0.48800$, $\hat{\lambda} = 36.989$, and $\hat{\tau} = 2.9495$. We define $h(\boldsymbol{\theta})$ to be the parameter α and proceed with the iterative process until it converges to estimates of $\hat{\alpha} = 0.40253$, $\hat{\lambda} = 34.205$, and $\hat{\tau} = 3.1270$, quite similar to the maximum likelihood estimates.

Procedure 1, with Equation (2.6) and $\boldsymbol{\Sigma}^{-1}$ as the matrix of weights, does not converge. If it did, it would yield the maximum likelihood estimates. Procedure 2 produces minimum chi-square estimates of $\hat{\alpha} = 0.36995$, $\hat{\lambda} = 33.702$, and $\hat{\tau} = 2.8685$.

Procedure 1 with Equation (2.1), at the maximum likelihood parameter estimates, has an asymptotic covariance matrix of

$$\begin{bmatrix} 0.017336 & 0.57436 & -0.035566 \\ 0.57436 & 20.6564 & -1.21355 \\ -0.035566 & -1.21355 & 0.10704 \end{bmatrix}.$$

As must be true, the asymptotic variance of $\hat{\alpha}$ and the asymptotic covariances of $\hat{\alpha}$ with $\hat{\lambda}$ and $\hat{\tau}$ are identical to the maximum likelihood (and minimum chi-square) values. The other entries are only very slightly higher than the maximum likelihood (and minimum chi-square) values. Table 3 shows the weights that underlie the matrix, generated by a $\mathbf{v}$ of

$$\begin{bmatrix} 1 \\ 34.25 \\ -2.229 \end{bmatrix}.$$

Example Two—Discussion

Table 4 shows the empirical cumulative distribution function along with the fitted cumulative distribution function for the maximum likelihood parameter estimates, the original minimum distance estimates, and the minimum chi-square estimates. If one believes that a Burr distribution is appropriate, then one should prefer the maximum likelihood or minimum chi-square estimators, since they have smaller asymptotic variances.

None of the distributions provides a particularly good fit very low in the distribution. If one does not believe that a Burr distribution is appropriate over the entire range of the data, one could fit that distribution only above a certain point and use an empirical distribution below that. The mixed exponential distribution always has a mode at zero, and since the data clearly shows a mode significantly greater than zero, the mixed exponential would not fit well over the entire range of the data. However, one could fit the mixed exponential to the section of the distribution to the right of the mode.

TABLE 3

EXAMPLE 2 WEIGHTS

Lag	Original	Optimized
6	4.0	2195
12	4.0	309
18	4.0	112
24	4.0	59.4
30	4.0	42.0
36	4.0	31.1
42	4.0	24.7
48	4.0	20.5
54	4.1	17.6
60	4.2	13.6
66	4.5	15.2
72	4.9	14.0
78	5.5	13.3
84	6.2	13.0
90	6.8	12.8
96	7.4	12.8
102	8.1	13.1
108	9.7	13.5
114	10.6	14.2
120	11.0	15.2
126	12.7	16.6
132	21.2	18.6
138	26.8	21.5
144	47.3	25.9
150	58.9	33.4
156	232.5	48.6
162	232.5	94.5

At the conclusion of the second example, Klugman and Parsa show numbers implying that an approximate 95% confidence interval for the number of claims that will be reported after Lag 168 is $72 + / - 15$ for the maximum likelihood estimator, and $59 + / - 20$ for the minimum distance estimator with their set of weights. These are incorrect. The actual confidence intervals should be $72 + / - 57$ and $59 + / - 61$, respectively. For the minimum chi-square estimator, the confidence interval is $102 + / - 89$. The lengths of these confidence intervals indicate

TABLE 4

EXAMPLE 2 CUMULATIVE DISTRIBUTION FUNCTIONS

		Burr			Weibull
Lag	Empirical	Maximum Likelihood	Min Dist Original Wts	Minimum Chi-square	Maximum Likelihood
6	0.0086	0.0020	0.0026	0.0032	0.0159
12	0.0216	0.0173	0.0194	0.0226	0.0513
18	0.0389	0.0574	0.0604	0.0672	0.1000
24	0.1210	0.1257	0.1276	0.1365	0.1585
30	0.2181	0.2142	0.2139	0.2212	0.2235
36	0.2959	0.3101	0.3079	0.3102	0.2924
42	0.4298	0.4025	0.3998	0.3955	0.3628
48	0.5011	0.4860	0.4838	0.4729	0.4327
54	0.5637	0.5585	0.5576	0.5411	0.5005
60	0.6156	0.6207	0.6212	0.6006	0.5649
66	0.6631	0.6736	0.6754	0.6521	0.6250
72	0.7149	0.7188	0.7216	0.6968	0.6802
78	0.7603	0.7574	0.7611	0.7357	0.7301
84	0.7970	0.7907	0.7949	0.7698	0.7746
90	0.8207	0.8195	0.8241	0.7998	0.8138
96	0.8402	0.8447	0.8493	0.8263	0.8478
102	0.8553	0.8668	0.8714	0.8498	0.8770
108	0.8834	0.8863	0.8907	0.8709	0.9019
114	0.8942	0.9036	0.9077	0.8898	0.9227
120	0.8985	0.9190	0.9229	0.9069	0.9401
126	0.9136	0.9329	0.9363	0.9224	0.9544
132	0.9503	0.9454	0.9484	0.9365	0.9660
138	0.9611	0.9567	0.9592	0.9494	0.9754
144	0.9784	0.9669	0.9690	0.9612	0.9829
150	0.9827	0.9763	0.9778	0.9721	0.9889
156	0.9957	0.9849	0.9859	0.9821	0.9936
162	1.0000	0.9927	0.9933	0.9914	0.9972
168	1.0000	1.0000	1.0000	1.0000	1.0000

that the volume of data is not sufficient to provide a reliable estimate of the number of claims that will be reported after Lag 168, even if one accepts the assumption that a Burr distribution is appropriate for this data.

Accomando and Weissner [1] suggest using a Weibull distribution for this data. Maximum likelihood estimation yields

parameter estimates of $\hat{\theta} = 67.3$ and $\hat{\tau} = 1.71$, with the cumulative distribution function expressed as $F(x) = 1 - e^{-(x/\theta)^{\tau}}$. Table 4 shows the fitted cumulative distribution function. The approximate 95% confidence interval for the number of claims that will be reported after Lag 168 is $4 + / - 3$. The reason that this is so different from the Burr confidence intervals is that the confidence intervals depend on the assumption that a particular distribution is appropriate over the entire range of the distribution, including the portion for which we do not yet have data. There is no way to tell whether a Burr distribution, a Weibull distribution, or some other distribution is most appropriate beyond the range of the data. Attempting to extrapolate from the data to obtain the number of unreported claims, without reference to other experience for which claims after Lag 168 have been observed, is likely to lead to a very unreliable estimate.

The data in this example is truncated at a single point, and though that makes the data of limited use for estimation beyond the truncation point, adjusting for the truncation in the minimum distance estimation procedure is straightforward. Likewise, data with a single censorship point does not present difficulties. However, with data that contains multiple truncation or censorship points on the left or the right, constructing the empirical distribution becomes more complicated. The most logical approach is to use the Kaplan-Meier Product-Limit estimator.

4. GOODNESS-OF-FIT TESTS

Klugman and Parsa propose a goodness-of-fit test using the statistic

$$(\mathbf{G}_n - \mathbf{G})'\mathbf{W}^{1/2}\{n^{-1}\mathbf{W}^{1/2}[\mathbf{I} - \mathbf{D}'(\mathbf{D}\mathbf{W}\mathbf{D}')^{-1}\mathbf{D}\mathbf{W}] \times \boldsymbol{\Sigma}[\mathbf{I} - \mathbf{W}\mathbf{D}'(\mathbf{D}\mathbf{W}\mathbf{D}')^{-1}\mathbf{D}]\mathbf{W}^{1/2}\}^{-}\mathbf{W}^{1/2}(\mathbf{G}_n - \mathbf{G}),$$

where $\mathbf{W}$ is a matrix of the weights and "$^-$" indicates a generalized inverse. If the distribution being fit is the correct one,

this statistic has an asymptotic chi-square distribution with $k - p$ degrees of freedom. The statistic

$$(\mathbf{G}_n - \mathbf{G})'\{n^{-1}[\mathbf{I} - \mathbf{D}'(\mathbf{D}\mathbf{W}\mathbf{D}')^{-1}\mathbf{D}\mathbf{W}] \times \boldsymbol{\Sigma}[\mathbf{I} - \mathbf{W}\mathbf{D}'(\mathbf{D}\mathbf{W}\mathbf{D}')^{-1}\mathbf{D}]\}^{-}(\mathbf{G}_n - \mathbf{G})$$

also has an asymptotic chi-square distribution with $k - p$ degrees of freedom. This statistic does not contain the square root of $\mathbf{W}$, which could be messy when $\mathbf{W}$ is not a diagonal matrix. In their proof, Klugman and Parsa use a vector $\mathbf{V}_n = \mathbf{W}^{1/2}(\mathbf{G}_n - \mathbf{G})$ and a matrix $\mathbf{R} = \mathbf{W}^{1/2}\mathbf{D}'$. By leaving off the $\mathbf{W}^{1/2}$ in this vector and matrix, one can use the same reasoning to obtain the alternate statistic.

If one uses Procedure 1 with Equation (2.6) and $\boldsymbol{\Sigma}^{-1}$ as the matrix of weights, or if one uses Procedure 2, $n(\mathbf{G}_n - \mathbf{G})' \cdot \boldsymbol{\Sigma}^{-1}(\mathbf{G}_n - \mathbf{G})$ has an asymptotic chi-square distribution with $k - p$ degrees of freedom. This follows either from using $n\mathbf{I}$ as the generalized inverse in the authors' statistic or $n\boldsymbol{\Sigma}^{-1}$ as the generalized inverse in the alternate statistic. If G is the cumulative distribution function, this is the standard chi-square goodness-of-fit statistic.

With $n\mathbf{W}^{-1/2}\boldsymbol{\Sigma}^{-1}[\mathbf{I} - \mathbf{D}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{D}\boldsymbol{\Sigma}^{-1}]\mathbf{W}^{-1/2}$ as the generalized inverse in the authors' statistic (the Moore-Penrose inverse) or $n\boldsymbol{\Sigma}^{-1}[\mathbf{I} - \mathbf{D}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{D}\boldsymbol{\Sigma}^{-1}]$ as the generalized inverse in the alternate statistic, one finds that

$$n(\mathbf{G}_n - \mathbf{G})'\{\boldsymbol{\Sigma}^{-1}[\mathbf{I} - \mathbf{D}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{D}\boldsymbol{\Sigma}^{-1}]\}(\mathbf{G}_n - \mathbf{G})$$

has an asymptotic chi-square distribution with $k - p$ degrees of freedom, which is independent of the weights used. If G is the cumulative distribution function, this is the statistic given by Moore [5, p. 90] as applicable with the maximum likelihood estimator and the minimum chi-square estimator, among others.

Regardless of which generalized inverse one uses, the tests in this section are valid as long as the weights in the test statistic

are consistent with the weights used in fitting. One may fix the weights beforehand, or derive them as in Procedure 1 or Procedure 2. The tests are valid even if the weights are suboptimal. If the test statistic exceeds its critical value, that indicates a problem with the selected distribution, not, as the authors imply, with the weights. The weights may indeed be poorly chosen and thus give large asymptotic variances, but that does not affect the validity of the test.

5. CONCLUSION

Minimum distance estimation has some interesting properties, but as a practical matter, I see little reason to prefer it to maximum likelihood estimation. The main purported advantage of minimum distance estimation is that, through adjustment of the weights, it can provide a closer fit to the parts of the distribution that are of the most interest. This leads to an estimator with a larger variance than the maximum likelihood estimator, however. And, if one believes that the model one is using is appropriate, one should prefer the estimator with the smaller variance.

Minimum distance estimation is a clumsy remedy for a model that is not flexible enough. Instead of resorting to minimum distance estimation, I believe one would be better off addressing the inadequacies of the model itself. One possible option is to fit a parametric distribution to the upper section of the data only and to use the empirical distribution below that. Another possible option is to use the semiparametric mixed exponential distribution.

Minimum distance estimation performed with weights selected to achieve high efficiency generally produces parameter estimates close to the maximum likelihood estimates. But maximum likelihood estimation is usually somewhat easier to implement than minimum distance (or minimum chi-square) estimation, especially if the data contains multiple truncation or

censorship points. Minimum distance estimation would be most useful in situations where maximum likelihood estimation is not feasible, such as when limited expected values are the only data available.

REFERENCES

[1] Accomando, Frank W., and Edward W. Weissner, "Report Lag Distributions: Estimation and Application to IBNR Counts," *Transcripts of the 1988 Casualty Loss Reserve Seminar*, 1988, pp. 1038–1133.

[2] Benichou, Jacques, and Mitchell H. Gail, "A Delta Method for Implicitly Defined Random Variables," *The American Statistician* 43, no. 1 (February 1989), pp. 41–44.

[3] Keatinge, Clive L., "Modeling Losses with the Mixed Exponential Distribution," *PCAS* LXXXVI, 1999, pp. 654–698.

[4] Luong, A., and M. E. Thompson, "Minimum-distance Methods based on Quadratic Distances for Transforms," *The Canadian Journal of Statistics* 15, no. 3 (September 1987), pp. 239–251.

[5] Moore, David S., "Chi-square Tests," *Studies in Statistics*, editor Robert V. Hogg, The Mathematical Association of America, 1978, pp. 66–106.

APPENDIX

Here I find the sets of weights that minimize the asymptotic variance of $h(\hat{\boldsymbol{\theta}})$, given by

$$n^{-1}\mathbf{d}'\mathbf{A}^{-1}\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'\mathbf{A}^{-1}\mathbf{d} = n^{-1}\mathbf{d}'(\mathbf{DWD}')^{-1}\mathbf{DW}\boldsymbol{\Sigma}\mathbf{WD}'(\mathbf{DWD}')^{-1}\mathbf{d}. \tag{A.1}$$

The matrix $\mathbf{W}$ is a symmetric matrix of weights, which may or may not be a diagonal matrix. One can express any set of weights as a symmetric matrix by allocating the weight assigned to each off-diagonal term equally to both sides of the diagonal. I assume that $\boldsymbol{\Sigma}$ has rank k, $\mathbf{D}$ has rank p, and $\mathbf{d}$ has at least one nonzero element.

The first step is to take the derivative with respect to each entry in $\mathbf{W}$. The derivative of $(\mathbf{DWD}')(\mathbf{DWD}')^{-1}$ is zero. Therefore, by the product rule for differentiation, $(\mathbf{DWD}')$ times the derivative with respect to a particular entry in $\mathbf{W}$ within $(\mathbf{DWD}')^{-1}$ must be equal to the negative of the derivative with respect to that entry within $(\mathbf{DWD}')$ times $(\mathbf{DWD}')^{-1}$. Thus, using the product rule and the symmetry of (A.1), the derivative with respect to the ijth entry in $\mathbf{W}$ is

$$\begin{aligned} &n^{-1}\mathbf{d}'(\mathbf{DWD}')^{-1}\mathbf{D}(\mathbf{1}_{ij} + \mathbf{1}_{ji}) \\ &\quad \times [\mathbf{I} - \mathbf{D}'(\mathbf{DWD}')^{-1}\mathbf{DW}]\boldsymbol{\Sigma}\mathbf{WD}'(\mathbf{DWD}')^{-1}\mathbf{d}, \end{aligned}$$

where $\mathbf{1}_{ij}$ indicates a $k \times k$ matrix with the ijth entry equal to $\mathbf{1}$ and the remaining entries equal to 0.

Since the derivative with respect to all entries in $\mathbf{W}$ must be 0 for (A.1) to be at a minimum, and since the expression in brackets is idempotent with a nullspace consisting of the p columns of $\mathbf{D}'$, $\boldsymbol{\Sigma}\mathbf{WD}'(\mathbf{DWD}')^{-1}\mathbf{d}$ must be in the column space of $\mathbf{D}'$ or $\boldsymbol{\Sigma}\mathbf{WD}'(\mathbf{DWD}')^{-1}\mathbf{d} = \mathbf{D}'\mathbf{u}$, where $\mathbf{u}$ is a vector of length p. Multiplying both sides by $(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{D}\boldsymbol{\Sigma}^{-1}$ yields $\mathbf{u} = (\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{d}$. Thus,

$$\mathbf{WD}'(\mathbf{DWD}')^{-1}\mathbf{d} = \boldsymbol{\Sigma}^{-1}\mathbf{D}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{d}. \tag{A.2}$$

Substituting (A.2) into (A.1) shows that the minimum asymptotic variance of $h(\hat{\boldsymbol{\theta}})$ is $n^{-1}\mathbf{d}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{d}$.

This also shows that if the asymptotic variance of $h(\hat{\boldsymbol{\theta}})$ is at its minimum value, then the asymptotic covariance of $h(\hat{\boldsymbol{\theta}})$ with any other function $h^*(\hat{\boldsymbol{\theta}})$ is $n^{-1}\mathbf{d}^{*\prime}(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{d}$, where $\mathbf{d}^*$ is the vector of length p with jth element $\partial h^*/\partial\theta_j$. However, this does not imply that the asymptotic variance of $h^*(\hat{\boldsymbol{\theta}})$ is $n^{-1}\mathbf{d}^{*\prime}(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{d}^*$.

Multiplying both sides of (A.2) by $(\mathbf{W}\mathbf{D}')^-$, a left-inverse of $\mathbf{W}\mathbf{D}'$, yields $(\mathbf{D}\mathbf{W}\mathbf{D}')^{-1}\mathbf{d} = (\mathbf{W}\mathbf{D}')^-\boldsymbol{\Sigma}^{-1}\mathbf{D}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{d}$. Substituting this into (A.2) yields $[\mathbf{I} - \mathbf{W}\mathbf{D}'(\mathbf{W}\mathbf{D}')^-]\boldsymbol{\Sigma}^{-1}\mathbf{D}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{d} = \mathbf{0}$. The expression in brackets is idempotent with a nullspace consisting of the p columns of $\mathbf{W}\mathbf{D}'$, so $\boldsymbol{\Sigma}^{-1}\mathbf{D}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{d}$ must be in the column space of $\mathbf{W}\mathbf{D}'$ or $\boldsymbol{\Sigma}^{-1}\mathbf{D}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{d} = \mathbf{W}\mathbf{D}'\mathbf{v}$, where $\mathbf{v}$ is a vector of length p.

If $\mathbf{W}$ must be a diagonal matrix, then

$$w_i = (\boldsymbol{\Sigma}^{-1}\mathbf{D}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}\mathbf{d})_i/(\mathbf{D}'\mathbf{v})_i,$$

where $\mathbf{v}$ is an arbitrary nonzero vector. Thus, the equation can be satisfied for weights associated with a space of dimension p. In general, a set of weights cannot satisfy this equation for all functions $h(\boldsymbol{\theta})$, unless $k = p + 1$. It is possible in this case because the intersection of p spaces of dimension p within a space of dimension $p + 1$ has a dimension of at least 1.

If $\mathbf{W}$ may be a full symmetric matrix, then there are many $\mathbf{W}$s that will satisfy this equation. If $\mathbf{W}\mathbf{D}'$ has the same column space as $\boldsymbol{\Sigma}^{-1}\mathbf{D}'(\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')^{-1}$, then $\mathbf{W}$ can satisfy this equation for all functions $h(\boldsymbol{\theta})$. The most obvious choice for $\mathbf{W}$ with this property is $\boldsymbol{\Sigma}^{-1}$.

For a set of weights to produce minimum asymptotic variance, $\mathbf{A} = \mathbf{D}\mathbf{W}\mathbf{D}'$ must be positive definite, whether $\mathbf{W}$ is a diagonal or a full symmetric matrix. If $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$, then since $\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}'$ is

positive definite, minimum asymptotic variance is achieved for all functions $h(\boldsymbol{\theta})$.

Results specific to the cumulative distribution function

If G is the cumulative distribution function, then $\boldsymbol{\Sigma}^{-1}$ is a tridiagonal matrix with

$$\boldsymbol{\Sigma}^{-1}_{ii} = \frac{G_{i+1} - G_{i-1}}{(G_i - G_{i-1})(G_{i+1} - G_i)} \qquad \text{and}$$

$$\boldsymbol{\Sigma}^{-1}_{i,i-1} = \boldsymbol{\Sigma}^{-1}_{i-1,i} = \frac{-1}{G_i - G_{i-1}},$$

where $G_0 = 0$ and $G_{k+1} = 1$.

* * * * *

If the matrix of weights is $\boldsymbol{\Sigma}^{-1*}$, based on a given cumulative distribution function, then the objective function is proportional to

$$\begin{aligned}
&n(\mathbf{G}_n - \mathbf{G})'\boldsymbol{\Sigma}^{-1*}(\mathbf{G}_n - \mathbf{G}) \\
&= n\left[\sum_{i=1}^{k} \frac{(G^*_{i+1} - G^*_{i-1})(G_{n,i} - G_i)^2}{(G^*_i - G^*_{i-1})(G^*_{i+1} - G^*_i)}\right. \\
&\qquad \left. -2\sum_{i=2}^{k} \frac{(G_{n,i} - G_i)(G_{n,i-1} - G_{i-1})}{G^*_i - G^*_{i-1}}\right] \\
&= n\left[\sum_{i=1}^{k} \frac{(G_{n,i} - G_i)^2}{G^*_i - G^*_{i-1}} + \sum_{i=2}^{k+1} \frac{(G_{n,i-1} - G_{i-1})^2}{G^*_i - G^*_{i-1}}\right. \\
&\qquad \left. -2\sum_{i=2}^{k} \frac{(G_{n,i} - G_i)(G_{n,i-1} - G_{i-1})}{G^*_i - G^*_{i-1}}\right] \\
&= n\sum_{i=1}^{k+1} \frac{[(G_{n,i} - G_{n,i-1}) - (G_i - G_{i-1})]^2}{G^*_i - G^*_{i-1}}.
\end{aligned}$$

If one treats the matrix of weights as a function of the parameters as in Procedure 2 from Section 2, then $\mathbf{G}$ and $\mathbf{G}^*$ are identical, and the objective function is proportional to the chi-square function.

If the matrix of weights is fixed, then one finds the minimum of the objective function by taking the derivative of the numerator of each of the terms with respect to each of the parameters and finding the point at which all of the derivatives are equal to zero. The derivative with respect to the jth parameter is

$$-2n\sum_{i=1}^{k+1}\frac{[(G_{n,i}-G_{n,i-1})-(G_i-G_{i-1})](G_i^{(j)}-G_{i-1}^{(j)})}{G_i^*-G_{i-1}^*}.$$

With Procedure 1 from Section 2, $\mathbf{G}$ and $\mathbf{G}^*$ must be identical at the final parameter estimates. At that point, the expression reduces to

$$-2n\sum_{i=1}^{k+1}(G_{n,i}-G_{n,i-1})\frac{(G_i^{(j)}-G_{i-1}^{(j)})}{G_i-G_{i-1}},$$

which is proportional to the derivative of the grouped log-likelihood function

$$n\sum_{i=1}^{k+1}(G_{n,i}-G_{n,i-1})\ln(G_i-G_{i-1})$$

with respect to the jth parameter.

* * * * *

The jlth entry of the inverse of the minimum asymptotic covariance matrix is

$$(n\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}')_{jl}$$
$$= n\left[\sum_{i=1}^{k}\frac{(G_{i+1}-G_{i-1})G_i^{(j)}G_i^{(l)}}{(G_i-G_{i-1})(G_{i+1}-G_i)}-\sum_{i=2}^{k}\frac{G_i^{(j)}G_{i-1}^{(l)}+G_i^{(l)}G_{i-1}^{(j)}}{G_i-G_{i-1}}\right]$$

$$= n\left[\sum_{i=1}^{k}\frac{G_i^{(j)}G_i^{(l)}}{G_i-G_{i-1}}+\sum_{i=2}^{k+1}\frac{G_{i-1}^{(j)}G_{i-1}^{(l)}}{G_i-G_{i-1}}-\sum_{i=2}^{k}\frac{G_i^{(j)}G_{i-1}^{(l)}+G_i^{(l)}G_{i-1}^{(j)}}{G_i-G_{i-1}}\right]$$

$$= n\sum_{i=1}^{k+1}\frac{(G_i^{(j)}-G_{i-1}^{(j)})(G_i^{(l)}-G_{i-1}^{(l)})}{G_i-G_{i-1}}.$$

Then, to confirm that this is identical to the grouped maximum likelihood value, the second derivative with respect to the jth and lth parameters of the grouped loglikelihood function $n\sum_{i=1}^{k+1}(G_{n,i}-G_{n,i-1})\ln(G_i-G_{i-1})$ is

$$n\sum_{i=1}^{k+1}(G_{n,i}-G_{n,i-1})\left[\frac{G_i^{(j,l)}-G_{i-1}^{(j,l)}}{G_i-G_{i-1}}-\frac{(G_i^{(j)}-G_{i-1}^{(j)})(G_i^{(l)}-G_{i-1}^{(l)})}{(G_i-G_{i-1})^2}\right],$$

and its negative expectation is

$$n\sum_{i=1}^{k+1}\frac{(G_i^{(j)}-G_{i-1}^{(j)})(G_i^{(l)}-G_{i-1}^{(l)})}{G_i-G_{i-1}}.$$