Ann. Inst. Statist. Math.
Vol. 43, No. 2, 391–404 (1991)

A MODIFIED STEEPEST DESCENT METHOD WITH APPLICATIONS TO MAXIMIZING LIKELIHOOD FUNCTIONS

MOHAMED HABIBULLAH[1] AND S. K. KATTI[2]

[1] *Division of Sciences and Mathematics, University of Wisconsin-Superior, Superior, WI 54880, U.S.A.*
[2] *Department of Statistics, University of Missouri-Columbia, Columbia, MO 65211, U.S.A.*

(Received June 28, 1988; revised March 30, 1990)

Abstract. In maximizing a non-linear function $G(\boldsymbol{\theta})$, it is well known that the steepest descent method has a slow convergence rate. Here we propose a systematic procedure to obtain a 1-1 transformation on the variables $\boldsymbol{\theta}$, so that in the space of the transformed variables, the steepest descent method produces the solution faster. The final solution in the original space is obtained by taking the inverse transformation. We apply the procedure in maximizing the likelihood functions of some generalized distributions which are widely used in modeling count data. It was shown that for these distributions, the steepest descent method via transformations produced the solutions very fast. It is also observed that the proposed procedure can be used to expedite the convergence rate of the first derivative based algorithms, such as Polak-Ribiere, Fletcher and Reeves conjugate gradient methods as well.

Key words and phrases: Generalized distributions, log-likelihood functions, steepest descent method, conjugate gradient method.

1. Introduction

In maximizing the likelihood functions $L(\boldsymbol{\theta} \mid \boldsymbol{X})$ or minimizing $G(\boldsymbol{\theta}) = -\ln L(\boldsymbol{\theta} \mid \boldsymbol{X})$ in the domain $\Omega \in R^n$, $n \geq 2$, of parameters, for the class of generalized and mixtures of distributions (Johnson and Kotz (1969)), it is often necessary to use an iterative algorithm. However, for these classes of distributions, $G(\boldsymbol{\theta})$ is generally non-convex and the second derivative $\nabla^2 G(\boldsymbol{\theta})$ is numerically unstable or time consuming. Hence, algorithms that require $\nabla^2 G(\boldsymbol{\theta})$ and its inverse at each iteration are not desirable. But the convergence rates of the steepest descent method or its different modifications, which require only the first derivative, can be slow for the class of non-convex functions.

In this paper, we propose some guidelines for a modification which will improve the convergence rate of the steepest descent method in minimizing the function $G(\boldsymbol{\theta})$. This modification is based on a 1-1 transformation on the parameters $\boldsymbol{\theta}$ that transforms the contours of $G(\boldsymbol{\theta})$ to approximate circles (spheres in higher dimensions) in some special cases. To illustrate the procedure, we minimize $G(\boldsymbol{\theta})$ for the log-zero-Poisson-truncated (l.z.P.t.) distribution (Katti and Rao (1970)) which has two parameters and for the Gegenbauer distribution (Plunkett and Jain (1975)) which has three parameters. Some radical improvements in the convergence rate were from 9425 iterations without our modification to 41 iterations with our modification and from 13010 iterations without our modification to 19 iterations with our modification. In Section 2, we provide a recursion formula for the gradient vector $\nabla G(\boldsymbol{\theta})$ for the l.z.P.t. distribution which we needed for minimizing $G(\boldsymbol{\theta})$ by the steepest descent method. In Section 3, we obtain the solution through the proposed transformation. We also take various recorded data and show that the transformation helps to improve the convergence rate for all these data sets. In Section 4, we consider minimizing $G(\boldsymbol{\theta})$ for the Gegenbauer distribution by using the steepest descent and the Polak-Ribiere (1969) methods. It is observed that the proposed modification improves the convergence rates significantly in both the cases.

2. The l.z.P.t. distribution and the m.l.e. of the parameters

Katti and Rao (1970) developed the l.z.P.t. distribution as a model for count data. The probability generating function (p.g.f) is given by

$$g(z) = \sum_{i=1}^{\infty} z^i P_i = k_1 \ln \frac{(\phi - e^{\lambda z})}{(\phi - 1)} \tag{2.1}$$

where $(\phi, \lambda) \in \Omega = \{(\phi, \lambda) \mid \phi > e^{\lambda}, \lambda > 0\}$, $k_1 = 1/\ln\{(\phi - e^{\lambda})/(\phi - 1)\}$, $P_i(\phi, \lambda) = \Pr(X = i)$, $i = 1, 2, \dots$, and X is l.z.P.t. random variable.

Huque (1974) developed an expression to compute the probabilities P_i in (2.1). Here, we provide a new simpler recursion formula for the probabilities.

$$P_1 = -k_1\lambda/(\phi - 1),$$
$$P_r = \left\{ \lambda(r-1)P_{r-1} - \frac{1}{k_1}\sum_{i=1}^{r-1} i(r-i)P_iP_{r-i} \right\} \Big/ r(r-1), \qquad r = 2, 3, \dots.$$

The maximum likelihood estimates (m.l.e.) of (ϕ, λ) correspond to the vector $(\hat{\phi}, \hat{\lambda})$ that minimizes $G(\phi, \lambda) = -\sum_{i=1}^{\infty} f_i \ln P_i(\phi, \lambda)$ in Ω. That is, at $(\hat{\phi}, \hat{\lambda})$, the gradient vector $\nabla G(\hat{\phi}, \hat{\lambda}) \approx \mathbf{0}$ and the Hessian matrix $\nabla^2 G(\hat{\phi}, \hat{\lambda})$ is positive definite. For an initial point (ϕ^0, λ^0), the steepest descent method provides the next solution by the relation $(\phi^1, \lambda^1) = (\phi^0, \lambda^0) - \alpha\nabla G(\phi^0, \lambda^0)$ for some constant $\alpha > 0$ which is chosen in such a way that $G(\phi^1, \lambda^1)$ will be significantly less than $G(\phi^0, \lambda^0)$. To determine an appropriate choice for the value of α, a function $h(\alpha) = G((\phi^0, \lambda^0) - \alpha\nabla G(\phi^0, \lambda^0))$ is minimized which is quite difficult since $G(\phi, \lambda)$ can not be expressed in an explicit form in the argument of $(\phi^0, \lambda^0) - \alpha\nabla G(\phi^0, \lambda^0)$. We use a quadratic interpolation formula (Burden (1985)) to obtain a suitable value of α that guarantees $G(\phi^1, \lambda^1) < G(\phi^0, \lambda^0)$. The successive iterations will produce the solutions $(\phi^i, \lambda^i) = (\phi^{i-1}, \lambda^{i-1}) - \alpha \nabla G(\phi^{i-1}, \lambda^{i-1})$ for $i = 1, 2, \dots$, such that $G(\phi^i, \lambda^i) < G(\phi^{i-1}, \lambda^{i-1})$. We stop the iterations when

$$(2.2) \qquad |\nabla G(\hat{\phi}, \hat{\lambda})| < 10^{-5} \quad \text{and} \quad |G(\phi^i, \lambda^i) - G(\phi^{i-1}, \lambda^{i-1})| < 10^{-7}.$$

The gradient vector is $\nabla G(\phi, \lambda) = (\partial G/\partial \phi, \partial G/\partial \lambda) = -(\sum_{i=1}^{\infty}(f_i/P_i)(\partial P_i/\partial \phi), \sum_{i=1}^{\infty}(f_i/P_i)(\partial P_i/\partial \lambda))$, where

$$\frac{\partial P_i}{\partial \phi} = \frac{-(i+1)P_{i+1}}{\lambda(\phi-1)} - \frac{k_1 \lambda^i}{\phi(\phi-1)i!}\sum_{r=0}^{\infty}\frac{r^i}{\phi^r} = cP_i, \quad \text{where} \quad c = \frac{k_1(e^\lambda - 1)}{(\phi-1)(\phi-e^\lambda)},$$

$$\frac{\partial P_i}{\partial \lambda} = \frac{i}{\lambda}P_i + k_2 P_i, \quad \text{where} \quad k_2 = \frac{k_1 e^\lambda}{(\phi - e^\lambda)}.$$

It may be noted that the infinite series $\sum_{r=0}^{\infty}(r^i/\phi^r)$, $\phi > 1$ is a convergent series for all $i = 1, 2, \dots$, and there exists a value M such that $\sum_{r=0}^{\infty}(r^i/\phi^r) - \sum_{r=0}^{M}(r^i/\phi^r) < \epsilon$, where ϵ is a preassigned small number. We investigated for a value of M that satisfies this inequality and found that $M = 60$ serves the purpose for $\epsilon = .00001$. So we replace the infinite series by a finite series using the range for r from 1 to $M = 60$. Hence the components of the gradient vector are

$$\frac{\partial G}{\partial \phi} = \sum_{i=1}^{\infty}\frac{(i+1)f_i P_{i+1}}{\lambda(\phi-1)P_i} + Nc + \frac{k_1}{\phi(\phi-1)}\sum_{i=1}^{\infty}\frac{f_i}{P_i}\frac{\lambda^i}{i!}\sum_{r=0}^{M}\frac{r^i}{\phi^r},$$

$$\frac{\partial G}{\partial \lambda} = -\sum \frac{i}{\lambda} f_i - k_1 \frac{e^\lambda \sum f_i}{(\phi - e^\lambda)}.$$

For data set 1 (see Data on corn borers), we computed the moment estimates of the parameters and used them as initial solutions. Using the steepest descent method we obtained the estimates $(\hat{\phi}, \hat{\lambda})$ that minimized $G(\phi, \lambda)$ after 9425 iterations. In Table 1, we present some iterations and their corresponding solutions. It may be seen that the final solution $(\hat{\phi}, \hat{\lambda})$ satisfies the equation (2.2) and the Hessian matrix $\nabla^2 G(\hat{\phi}, \hat{\lambda})$ is positive definite which indicates that the solution minimized the function $G(\phi, \lambda)$. However, it has taken a large number of iterations. This slow convergence is due to the non-convexity of the function $G(\phi, \lambda)$. To see the behavior of the function G, we computed some points corresponding to $G(\phi, \lambda) \approx 1781.0$, 1784.05, 1786.50 and drew the contours (see Fig. 1). These contour points can be obtained by using SAS GCONTOUR procedure or APL GRAPHPAK. The numbers 1781.0, 1784.05, 1786.50 were chosen arbitrarily close to the value of $G(\phi, \lambda) = 1777.1785$ that corresponds to the moment estimates of (ϕ, λ). The contours are long, thin and non-convex. If a 1-1 transformation could be obtained such that the transformation changes the contours into approximate circles, then the steepest descent method would converge to the solution very fast. So we proceed to find such a transformation in the next section.

Data on corn borers.

	(from Martin and Katti (1965))					McGuire *et al.*
Count i of	Observed frequencies f_i in data					
borers	set#1	set#2	set#6	set#7	set#8	frequency
0	187	117	19	24	43	62
1	185	87	12	16	35	121
2	200	50	18	16	17	132
3	164	38	18	18	11	105
4	107	21	11	15	5	74
5	68	7	12	9	4	42
6	49	2	7	6	1	17
7	39	2	8	5	2	11
8	21	0	4	3	2	8
9	12	1	4	4		5
10	11		1	3		0
11	2		0	0		0
12	5		1	1		1
13	2		1			0
14	3		0			
15	1		1			
16			0			
17			1			
18			0			
19			1			
20			0			
21			0			
22			0			
23			0			
24			0			
25			0			
26			1			

f_i = # of stalks with i corn borers.

Table 1. Solutions at different iterations by steepest descent method for data set 1.

Iteration	$\hat{\phi}$	$\hat{\lambda}$	$G(\hat{\phi}, \hat{\lambda})$
0	12.25	1.90	1777.178564
50	12.31311887	1.951440897	1775.29471
5000	15.73691187	2.102903981	1773.489471
9400	17.16316015	2.154415336	1773.241316
9424	18.36834031	2.193903385	1773.182911
9425	18.36836028	2.193905391	1773.182911

$(\hat{\phi}, \hat{\lambda}) = (18.36834028, 2.193904391)$ and $\nabla^2 G(\hat{\phi}, \hat{\lambda})$ has eigenvalues $\tau_1 = 1157.3731$, τ_2 07442002.

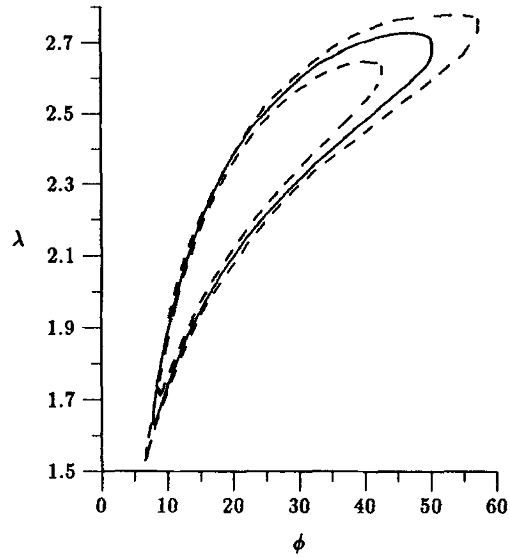

Fig. 1. Approximate contours for data set #1 of (i) $G(\phi, \lambda) \approx 1781$, (ii) $G(\phi, \lambda) \approx 1784.05$, (iii) $G(\phi, \lambda) \approx 1786.5$.

3. Minimization of $G(\phi, \lambda)$ with transformation

To convert the contours of $G(\phi, \lambda)$ for data set 1 to approximate circles, we use the structural restriction that $\phi > e^{\lambda}$ in Ω and make a transformation

$$(\phi_1, \lambda_1)^T = (\ln \phi - \lambda, \lambda)^T \tag{3.1}$$

with inverse transformation

$$(\phi, \lambda)^T = (e^{\phi_1 + \lambda_1}, \lambda_1)^T. \tag{3.2}$$

The Jacobian of the transformation is $J_1 = \begin{bmatrix} e^{\phi_1+\lambda_1} & e^{\phi_1+\lambda_1} \\ 0 & 1 \end{bmatrix}$ and the gradient in (ϕ_1, λ_1) space is $\nabla G(\phi_1, \lambda_1) = \nabla G(\phi, \lambda) J_1$. We use equation (3.1) on the points of the contours in Fig. 1 and obtain the points in (ϕ_1, λ_1) space. Note that the contours in Fig. 2 obtained by using these points look convex. In an attempt to find a transformation that will convert these contours into approximate circles, we take the points corresponding to $G(\phi_1, \lambda_1) \approx 1784.05$ and fit a quadratic form

$$(\phi_1, \lambda_1) A_2 (\phi_1, \lambda_1)^T + (\phi_1, \lambda_1)(E, F)^T + 1 = 0 \tag{3.3}$$

where A_2, E, F are estimated by minimizing $\sum((\phi_1, \lambda_1)A_2(\phi_1, \lambda_1)^T + (\phi_1, \lambda_1)(E, F)^T + 1)^2$ by least square method. The summation is taken on all points (ϕ_1, λ_1) considered. The estimated matrix is $\hat{A}_2 = \begin{bmatrix} .5643 & -.3986 \\ -.3986 & .4158 \end{bmatrix}$ with eigenvalues

$\tau_1 = .895545$, $\tau_2 = 0.0846$ and the matrix of the eigenvectors is $P = \begin{bmatrix} .769 & .639 \\ -.639 & .769 \end{bmatrix}$

Now, we define the following transformation

$$(\phi^*, \lambda^*)^T = J_2^{-1}(\phi_1, \lambda_1)^T \tag{3.4}$$

with the Jacobian of transformation $J_2 = PD^{-1/2}$, where D is a diagonal matrix with τ_1 and τ_2 as diagonal elements respectively. The inverse transformation is

$$(\phi_1, \lambda_1)^T = J_2(\phi^*, \lambda^*). \tag{3.5}$$

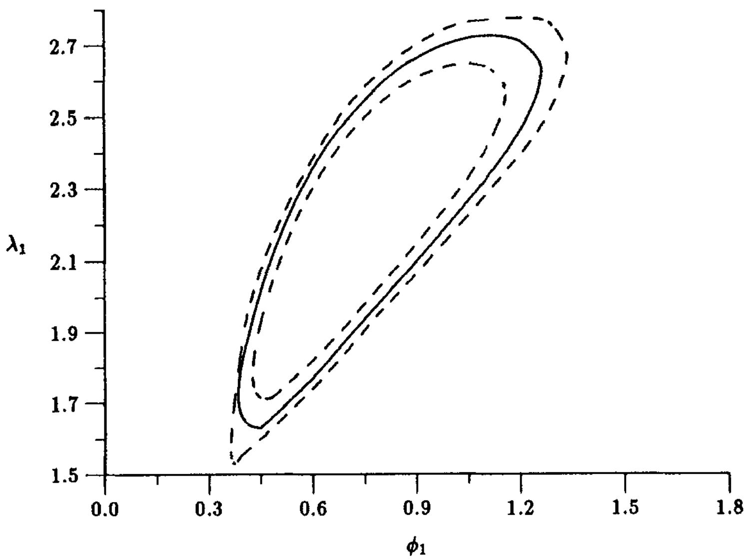

Fig. 2. Contours of $G(\phi_1, \lambda_1) \approx 1781, 1784.05, 1786.5$, where (ϕ_1, λ_1) are obtained through equation (3.1).

The quadratic form in (ϕ^*, λ^*) is $(\phi^*, \lambda^*)I_2(\phi^*, \lambda^*)^T + (\phi^*, \lambda^*)J_2^T(E, F)^T + 1 = 0$ which does not have a product term. The contours in (ϕ^*, λ^*) space are shown (see Fig. 3) to be approximate circles. The gradient vector is $\nabla G(\phi^*, \lambda^*) = \nabla G(\phi, \lambda)J_1J_2$. Note that the transformation in (3.4) could have been based on the contour points corresponding to other values of $G(\phi_1, \lambda_1)$ as well. However, Fig. 3 shows that this transformation converted all contours into approximate circles. We use the steepest descent method to find $(\hat{\phi}^*, \hat{\lambda}^*)$ which minimizes $G(\phi^*, \lambda^*)$ in (ϕ^*, λ^*) space. The inverse transformations in (3.5) and (3.2) give the estimates $(\hat{\phi}, \hat{\lambda})$ of (ϕ, λ) in Ω. In Table 2, we present some iterations and the corresponding solutions in (ϕ, λ) space obtained through the transformations. The final solution $(\hat{\phi}, \hat{\lambda})$ satisfies the condition in equation (2.2) and $\nabla^2 G(\hat{\phi}, \hat{\lambda})$ is found to be positive definite. It may be mentioned that the convergence occurred within 9 iterations while it took 9425 iterations without transformation.

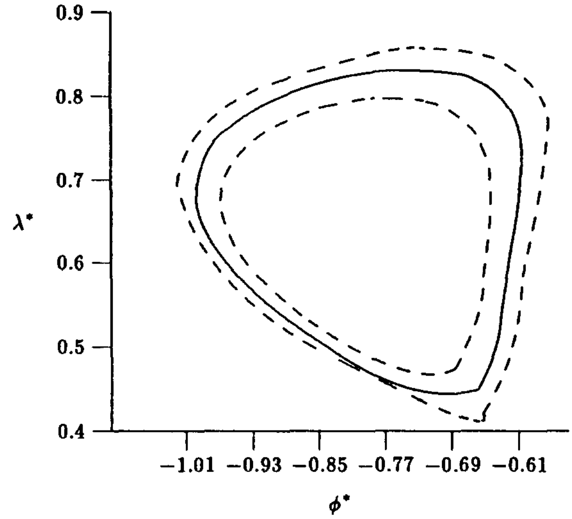

Fig. 3. Contours of $G(\phi^*, \lambda^*) \approx 1781, 1784.05, 1786.5$, where (ϕ^*, λ^*) are obtained through equation (3.4).

Table 2. Solutions through transformations for data set 1.

Iteration	ϕ	λ	$G(\phi, \lambda)$
0	12.25	1.90	1777.178564
1	15.2998139	2.097696583	1773.709921
7	18.36780527	2.19388844	1773.182911
8	18.36834033	2.193903397	1773.182911
9	18.36836385	2.193905176	1773.182911

$(\hat{\phi}, \hat{\lambda}) = (18.36836385, 2.193905176)$, $\nabla^2 G(\hat{\phi}, \hat{\lambda})$ has eigenvalues $(\tau_1, \tau_2) = (1157.372, 0.07442)$.

3.1 *Transformation based on adjustments*

The transformation in (3.1) required the eigenvalues τ_1, τ_2 to be positive. This was accomplished by the transformation (3.1) that converted the non-convex contours into convex contours. However, a transformation with this property may not be available. Here we suggest a modification which is free from a trial and error transformation and takes care of negative eigenvalues (if any). We recommend the following procedure and explain it using data set 1. We fit a quadratic function of the form $(\phi, \lambda)A_2(\phi, \lambda)^T + (\phi, \lambda)(E, F)^T + 1 = 0$ on the points of the contour for $G(\phi, \lambda) \approx 1784.05$ in (ϕ, λ) space. The least square estimate of A_2 is $\hat{A}_2 = \begin{bmatrix} .00016 & -.007796 \\ -.007796 & .343609 \end{bmatrix}$ with eigenvalues $(\tau_1, \tau_2) = (.343786, -.0000106)$, one of which is negative. Thus the transformation (3.4) is not feasible. Therefore, we take a modified matrix L as:

$$L = \hat{A}_2 + \mu I, \qquad \text{with} \quad \mu = k|\operatorname{Min}\{\tau_1, \tau_2\}|. \tag{3.6}$$

Any number $k > 1$ will make the matrix L a positive definite matrix. For example, for (i) $k = 1.1$, the eigenvalues of L are (.34379835, .000001068), (ii) $k = 3$, the eigenvalues are (.343818, .00002136). However, the eigenvectors of L and $\hat{A}_2$ for all values of k are the same. We have chosen $k = 3$ and considered the transformation

$$(\phi^*, \lambda^*)^T = J_3^{-1}(\phi, \lambda)^T \tag{3.7}$$

where $J_3 = P_1 D_1^{-1/2}$, and $P_1 = \begin{bmatrix} .0226 & -.9997 \\ -.9997 & -.0226 \end{bmatrix}$ is the matrix of the eigenvectors and D_1 is a diagonal matrix with the eigenvalues (.3438, .0000213) of L in the diagonal. Hence the transformation (3.7) based on the eigenvalues and the eigenvectors of the adjusted matrix L is feasible. The quadratic form in (ϕ^*, λ^*) will be $(\phi^*, \lambda^*)B(\phi^*, \lambda^*)^T + (\phi^*, \lambda^*)J_3^T(E, F)^T + 1 = 0$, where B is a diagonal matrix. However, the transformation in (3.7) will not necessarily convert the contours in Fig. 1 into approximate circles but it will rotate the principal axes of the contour parallel to the co-ordinate axes (see Fig. 4). We call the transformation in (3.7) as the transformation based on adjustments. In Table 3, we present some solutions at different iterations using this transformation.

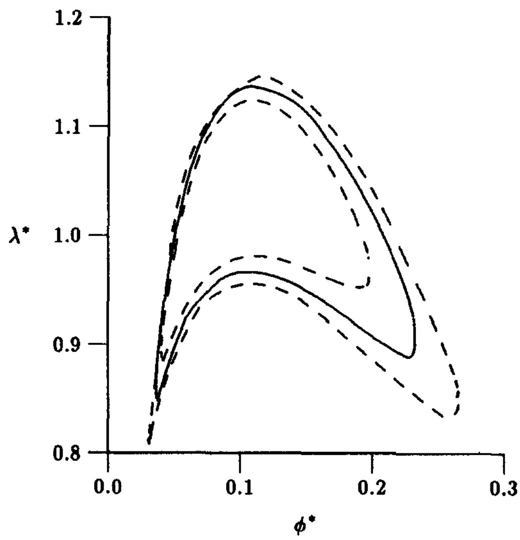

Fig. 4. Contours of $G(\phi^*, \lambda^*) \approx 1781, 1784.05, 1786.5$, where (ϕ^*, λ^*) are obtained by transformation based on adjustments (equation (3.7)).

Table 3. Solutions through the transformation based on adjustments ($k = 3$) for data set 1.

Iteration	ϕ	λ	$G(\phi, \lambda)$
0	12.25	1.90	1777.178564
10	18.03671562	2.183006653	1773.187168
40	18.36821013	2.193900089	1773.182911
41	18.3683262	2.193902156	1773.182911

$(\hat{\phi}, \hat{\lambda}) = (18.3682962, 2.193902156)$, $\nabla^2 G(\hat{\phi}, \hat{\lambda})$ has eigenvalues $(\tau_1, \tau_2) = (1157.374303, 0.0744)$.

Table 3 shows that the transformation based on the adjusted matrix has produced the solution after 41 iterations as opposed to 9425 iterations without any transformations. We also checked the effect of different values of k in equation (3.7) on the convergence rate to the solution. This is shown in Table 4.

Table 4. Number of iterations required to converge to the solution through transformation based on adjustment for different k (data set 1).

k	No. of iterations needed to converge	k	No. of iterations needed to converge	k	No. of iteration needed to converge
1.1	571	6.0	16	50	7
1.8	110	8.0	12	70	19
2.0	88	10.0	20	80	60
2.5	62	15.0	20	90	100
3.0	41	18.0	25	100	40
5.0	26	40.0	8	1000	225

Thus for moderate values of k, the transformation based on adjustments is found to reduce the number of iterations significantly.

3.2 *Applications of the transformation to other data sets*

We used data sets 2, 6, 7, 8 (see Data on corn borers) to obtain the m.l.e. $(\hat{\phi}, \hat{\lambda})$ by minimizing $G(\phi, \lambda)$. At first, we used the steepest descent method (without transformation) using moment estimates as initial solutions as in Section 2 and checked how many iterations it took to get the solutions using the stopping rule in (2.2). We then used the steepest descent method in the transformed spaces. For transformation through matrix adjustment, we used $k = 3$. Table 5 shows the solutions and the number of iterations taken to minimize $G(\phi, \lambda)$.

Table 5. M.L.E. of the parameters by steepest descent method for different data sets.

Data	Moment estimates		m.l.estimates		# of iterations needed		
	ϕ_m	λ_m	$\phi_{m.l.e.}$	$\lambda_{m.l.e.}$	Without transform	With transform a*	b*
1	12.25	1.90	8.36836	.19390	9425	41	9
2	15.25	1.47	4.15148	.44655	75	25	19
6	8.50	1.85	8.66853	.83531	4291	20	10
7	10.75	1.88	7.52613	.79454	2569	61	9
8	7.75	1.33	2.68443	.70586	3564	220	143

a*: Solutions with transformatin in (3.7), b*: Solutions with transformation in (3.4).

Table 5 shows that for most of the data sets, the steepest descent method without transformation takes a large number of iterations to converge to the solution. It is apparent from this table that if a transformation is available which converts the non-convex contours into convex contours, then we may use it before the transformation based on the eigenvalues and eigenvectors is used. Otherwise, use the transformation of the form (3.7) based on eigenvalues and eigenvectors of the adjusted matrix.

4. Maximization of the likelihood function for Gegenbauer distribution

Plunkett and Jain (1975) developed the Gegenbauer distribution which has the p.g.f.

$$g(z) = (1-\alpha-\beta)^{\lambda}(1-\alpha z-\beta z^2)^{-\lambda} \tag{4.1}$$

where $(\alpha,\, \beta,\, \lambda) \in \Omega = \{(\alpha,\, \beta,\, \lambda) \mid 0 < \alpha,\, \beta < 1,\, \alpha+\beta < 1,\, \lambda > 0\}$. Borah (1983) gave the recurrence relation

$$\begin{aligned} P_0 &= (1-\alpha-\beta)^{\lambda}, \qquad P_1 = \alpha\lambda(1-\alpha-\beta)^{\lambda}, \\ P_{r+1} &= \frac{1}{r+1}\{\alpha(\lambda+r)P_r + \beta(2\lambda+r-1)P_{r-1}\}, \qquad r \geq 1 \end{aligned}$$

for probabilities and computed the moment estimates of α, β, λ for the data set on European corn borers (McGuire *et al.* (1957)). We attempt to minimize $G(\alpha,\, \beta,\, \lambda) = -\sum f_i \ln P_i(\alpha,\, \beta,\, \lambda)$ in Ω to obtain the m.l.e. of $(\alpha,\, \beta,\, \lambda)$ by using the steepest descent method for the same data set (see Data on corn borers). We use Borah's moment estimates as the initial solution and stop at the k-th iteration when

$$\begin{aligned} &|\nabla G(\alpha^k,\, \beta^k,\, \lambda^k)| < 10^{-5} \quad \text{and} \\ &|G(\alpha^k,\, \beta^k,\, \lambda^k) - G(\alpha^{k-1},\, \beta^{k-1},\, \lambda^{k-1})| < 10^{-7}. \end{aligned} \tag{4.2}$$

To compute $\nabla G(\alpha, \beta, \lambda)$, we need the partial derivatives $\partial P_i/\partial\alpha$, $\partial P_i/\partial\beta$, $\partial P_i/\partial\lambda$, where

$$\frac{\partial P_0}{\partial\alpha} = -AP_0, \quad A = -\frac{\lambda}{1-\alpha-\beta}, \quad \frac{\partial P_1}{\partial\alpha} = \lambda\left(P_0 + \alpha\frac{\partial P_0}{\partial\alpha}\right),$$

$$\frac{\partial P_i}{\partial\alpha} = \left\{\alpha\frac{\partial P_{i-1}}{\partial\alpha} + \beta\frac{\partial P_{i-2}}{\partial\alpha}\right\} + (\lambda + \alpha A)P_{i-1} + A(\beta P_{i-2} - P_i), \qquad \text{for } i = 2, 3, \ldots,$$

$$\frac{\partial P_0}{\partial\beta} = -AP_0, \quad \frac{\partial P_1}{\partial\beta} = \alpha\lambda\frac{\partial P_0}{\partial\beta},$$

$$\frac{\partial P_i}{\partial\beta} = \left\{\alpha\frac{\partial P_{i-1}}{\partial\beta} + \beta\frac{\partial P_{i-2}}{\partial\beta}\right\} + \lambda P_{i-2} - A(P_i - \alpha P_{i-1} - \beta P_{i-2}) \qquad \text{for } i = 2, 3, \ldots,$$

$$\frac{\partial P_0}{\partial\lambda} = P_0\ln(1-\alpha-\beta), \quad \frac{\partial P_1}{\partial\lambda} = \alpha\left(P_0 + \lambda\frac{\partial P_0}{\partial\lambda}\right),$$

$$\frac{\partial P_i}{\partial\lambda} = \left\{\alpha(\lambda+i-2)\frac{\partial P_{i-1}}{\partial\lambda} + \beta(2\lambda+i-3)\frac{\partial P_{i-2}}{\partial\lambda} + 2\beta P_{i-2}\right\}\Big/(i-1) \qquad \text{for } i = 2, 3, \ldots.$$

Table 6 shows solutions at different iterations through the steepest descent method.

Table 6. Solutions at different iterations through steepest descent method.

Iteration	α	β	λ	$G(\alpha, \beta, \lambda)$
0	.2404	.0087	7.6609	1136.863469
1000	.2450479215	.005239659917	7.644936011	1136.764084
10000	.2511696073	.004344014233	7.465097115	1136.754083
13000	.2528185998	.004096439394	7.418086948	1136.751475
13009	.2528210333	.004092625331	7.417969661	1136.751467
13010	.2528237527	.00409566154	7.417940969	1136.751466

$(\hat\alpha, \hat\beta, \hat\lambda)$ = (.2528237527, .00409566154, 7.417940969), $\nabla G(\hat\alpha, \hat\beta, \hat\lambda)$ = (.2057, .3887, .0606).

It may be noted that the solution $(\hat\alpha, \hat\beta, \hat\lambda)$ does not satisfy the condition (4.2). Thus, we could improve on the solution if we would continue the iterations. However, we stopped the iterations.

4.1 *Minimization of the function $G(\alpha, \beta, \lambda)$ through transformation*

Now, we use the transformation explained in Section 3. In this case, we are dealing with the three-dimensional parameter space. To develop the transformation, we need some contour points corresponding to a fixed value of $G(\alpha, \beta, \lambda)$.

For data set of McGuire *et al.* (see Data on corn borers), we decided to compute some points on a contour in Ω for an arbitrary value $G = 1136.52$, which is chosen to be a smaller value than $G = 1136.863468$ that corresponds to the moment estimates of $(\alpha, \beta, \lambda) = (.2404, .0087, 7.6609)$. To get some approximate contour points, we fixed the value of α in Ω, and computed points $(\beta, \lambda) \in \Omega$ such that $1136.7515 \leq G \leq 1136.7525$. Then we changed the value of α in its range in Ω and repeated search for different (β, λ) in the subset of Ω. Similarly, we fixed β at different values in Ω and then again computed points (α, β, λ) such that $G \in (1136.7515, 1136.7525)$. We used 62 points on this approximate contour. These points may also be obtained by using Swell's algorithm (1988*a*, 1988*b*).

For unification purposes, we denote these points of (α, β, λ) by the vector $U^T = (u_1, u_2, u_3)$ and fit the quadratic form $U^T AU + B^T U + 1 = 0$, where A and B are obtained by minimizing $Q = \sum_{i=1}^{62}(U^T AU + B^T U + 1)^2$ by least square method. The estimate of A is given by $\hat{A} = \begin{bmatrix} 3.348 & .4157 & .133 \\ .4157 & 1.94 & .0169 \\ .133 & .0169 & .0053 \end{bmatrix}$ with eigenvalues $\tau_1 = 3.4675$, $\tau_2 = 1.8345$, $\tau_3 = .000046$. The corresponding matrix of the eigenvectors is $P = \begin{bmatrix} .9637 & .2639 & -.039 \\ .2641 & -.964 & -.0002 \\ .0383 & .01027 & .9992 \end{bmatrix}$.

Our recommended transformation from U to V is: $V = D^{1/2}P^{-1}U$, where D is a diagonal matrix with the diagonal elements as τ_1, τ_2, τ_3 respectively. The Jacobian of the transformation is $J = PD^{-1/2}$. Hence, $\nabla G(v_1, v_2, v_3) = \nabla G(u_1, u_2, u_3)J$. After the solution in the transformed space is obtained, the inverse transformation $U = PD^{-1/2}V$ is used to get the solution in the original space. Table 7 shows the solutions $(\hat{\alpha}, \hat{\beta}, \hat{\lambda})$ at different iterations through transformations.

Table 7. Solutions at different iterations by steepest descent method after transformation.

Iteration	α	β	λ	$G(\alpha, \beta, \lambda)$
0	.2404	.0087	7.6608	1136.863469
5	.258822003	.003484880929	7.233541182	1136.742094
17	.28000028127	.00004963279	6.697799515	1136.713482
18	.2801850243	.000024341265	6.693286003	1136.713170
19	.2801876584	.000000492846	6.693220862	1136.713170

$(\hat{\alpha}, \hat{\beta}, \hat{\lambda}) = (.280187658, .00000049285, 6.69322086)$, $\nabla G = (-.00000039, .0000036, .0000008)$, $\nabla^2 G(\hat{\alpha}, \hat{\beta}, \hat{\lambda})$ has eigenvalues $\tau_1 = 1.9861$, $\tau_2 = 3.4398$, $\tau_3 = 4.8745$.

By comparing Table 6 with Table 7, it may be seen that while the solution was not achieved by 13010 iterations without transformation, it was achieved by only 19 iterations through the transformation.

4.2 *Minimization of the function $G(\alpha, \beta, \lambda)$ by Polak-Ribiere method*

In this section, we use the Polak-Ribiere conjugate gradient method (1969) to minimize the function $G(\alpha, \beta, \lambda)$ starting wiht Borah's moment estimates as initial solution for the data set of McGuire *et al.* According to this technique, the solution at the $(k+1)$-th iteration is given by the relation

$$(\alpha^{k+1}, \beta^{k+1}, \lambda^{k+1}) = (\alpha^k, \beta^k, \lambda^k) + \theta_k s_k \tag{4.3}$$

where $s_k = -\nabla G(\alpha^k, \beta^k, \lambda^k) + d_k s_{k-1}$, $d_k = \nabla G(\alpha^k, \beta^k, \lambda^k)^T[\nabla G(\alpha^k, \beta^k, \lambda^k) - \nabla G(\alpha^{k-1}, \beta^{k-1}, \lambda^{k-1})]/\|\nabla G(\alpha^{k-1}, \beta^{k-1}, \lambda^{k-1})\|^2$ and θ_k is a value of θ that minimizes $G((\alpha^k, \beta^k, \lambda^k) + \theta s_k)$, $\theta \geq 0$. We stop the iterations when (4.2) is satisfied. Table 8 shows the solutions at different iterations.

Table 8. Solutions at different iterations by Polak-Ribiere method without transformation.

Iteration	α	β	λ	$G(\alpha, \beta, \lambda)$
0	.2404	.0087	7.6608	1136.863469
1000	.2520259804	.004211622612	7.440607057	1136.752723
3000	.2558610543	.003620181761	7.333120685	1136.724677
4000	.2739761228	.000693017350	6.863254771	1136.721061
4500	.2760161047	.000343361446	6.814076093	1136.718413
4825	.280186730	.0000125708152	6.69326011	1136.713174
4826	.280186942	.0000005175371	6.693211926	1136.713171

$(\hat{\alpha}, \hat{\beta}, \hat{\lambda}) = (.280186942, .0000005175371, 6.693211926)$, $\nabla G = (-.00000041, .00000369, .00000082)$, $\nabla^2 G(\hat{\alpha}, \hat{\beta}, \hat{\lambda})$ has eigenvalues $\tau_1 = 1.9861$, $\tau_2 = 3.4399$, $\tau_3 = 4.8746$.

We then used the Polak-Ribiere conjugate gradient method on the transformed space. Table 9 shows the solutions $(\hat{\alpha}, \hat{\beta}, \hat{\lambda})$ at different iterations through transformations.

Table 9. Solutions at different iterations by Polak-Ribiere method after transformation.

Iteration	α	β	λ	$G(\alpha, \beta, \lambda)$
0	.2404	.0087	7.6608	1136.863469
5	.2410794738	.007328078408	7.65206931	1136.785283
20	.256013348	.00359362946	7.305228941	1136.75221
30	.2609120073	.00314457272763	7.303228941	1136.739232
34	.2798652971	.000027268808	6.701765125	1136.713476
35	.2801781777	.000000475538	6.6940130126	1136.713183

$(\hat{\alpha}, \hat{\beta}, \hat{\lambda}) = (.2801781777, .000000475538, 6.6940130126)$, $\nabla G = (.000000589, .000004338, .000000205)$, $\nabla^2 G(\hat{\alpha}, \hat{\beta}, \hat{\lambda})$ has eigenvalues $\tau_1 = 1.9862$, $\tau_2 = 3.4398$, $\tau_3 = 4.8740$.

It is seen from Tables 8 and 9 that the transformation has improved the convergence rate of the Polak-Ribiere conjugate gradient method. We also used the Fletcher and Reeves gradient method (1964) to minimize the function $G(\alpha, \beta, \lambda)$ and observed that 5258 iterations were taken to minimize the function $G(\alpha, \beta, \lambda)$ in the original space while it took 25 iterations to converge in the transformed space. In fact, any first derivative based minimization technique could be used in the transformed space to expedite its convergence rate.

References

Borah, M. (1983). A study of certain discrete probability distributions, Doctoral Dissertation, Gauhati University, Gauhati, India.

Burden, R. (1985). *Numerical Analysis*, Prindle, Weber and Schmidt, Boston.

Fletcher, R. and Reeves, C. M. (1964). Function minimization by conjugate gradients, *Comput. J.*, **7**, 149–154.

Huque, F. (1974). Adaptive estimators for the log-zero-Poisson truncated distribution, Doctoral Dissertation, University of Missouri-Columbia, U.S.A.

Johnson, N. L. and Kotz, S. (1969). *Discrete Distributions*, Haughton Mifflin, Boston.

Katti, S. K. and Rao, A. V. (1970). The log-zero-Poisson distribution, *Biometrics*, **26**, 801–803.

Martin, D. C. and Katti, S. K. (1965). Fitting of certain contagious distributions to some available data by maximum likelihood method, *Biometrics*, **21**, 34–48.

McGuire, J. V., Brinkley, T. A. and Bancroft, T. A. (1957). The distribution of European corn borer larvae Pyraustra Nubilalis (HBN) in field corn, *Biometrics*, **13**, 65–78.

Plunkett, I. G. and Jain, G. C. (1975). Three generalized negative binomial distributions, *Biom. Zeit.*, **17**, 286–302.

Polak, E. and Ribiere, G. (1969). Note sur la convergence de methods de directions conjugees, *Rev. Franc Informat Recherche Operationelle*, **16**, 35–43.

Swell, G. (1988*a*). Plotting contour surfaces of a function of three variables, *ACM Trans. on Math. Software*, **14**, 33–41.

Swell, G. (1988*b*). Software for plotting contour surfaces of a function of three variables, *ACM Trans. on Math. Software*, **14**, 42–44.