ROKE, A COMPUTER PROGRAM FOR NONLINEAR LEAST-SQUARES DECOMPOSITION OF MIXTURES OF DISTRIBUTIONS

ISOBEL CLARK

Department of Mineral Resources Engineering, Imperial College of Science and Technology, Prince Consort Road, London SW7 2BP, England

Abstract—The problem of complex distributions resulting from mixtures of different populations is common in most branches of the earth sciences. The program ROKE estimates the parameters of mixtures of normal or lognormal distributions, from data available in histogram form. The full method is described, together with suggestions on the adaptation of the program for other distributions.

Key Words: Histogram frequencies, Least squares, Normal and lognormal distributions, Mixed populations.

INTRODUCTION

The problem of mixtures of distributions is one encountered in many fields, from biometrics (Dick and Bowden, 1973) to mining (Sichel, 1972). The situation arises if a specimen may have derived from one of several populations possessing similar distributions, but perhaps different means and standard deviations. For example, in a biological situation, size measurements will be influenced (usually) by the sex of the subject. In mining or geology the characteristics of a deposit may be modified by reworking or secondary mineralization phases. Specimens drawn from separate phases of mineralization may exhibit different statistical behavior, but may not be distinguishable (or distinguished) by geological or chemical analysis. The task of identifying and quantifying such phases of mineralization by the method presented in this paper is discussed in greater detail in Clark and Garnett (1974).

The program given here is a nonlinear least-squares approach to the solution of mixtures of normal or lognormal component distributions. The method of nonlinear least squares is described in detail, and is applicable to all classes of nonlinear models. The particular application to the situation of mixtures of normal distributions follows, with indications of how the method may be adapted to other distributions, and to mixtures of dissimilar distributions. The existing program is concerned only with normal and two-parameter lognormal distributions, because these were the ones of most interest to the author. However, work is under way to produce adaptations for truncated distributions, and for the three-parameter lognormal.

The nonlinear approach has been used for this problem in the past, the present program having been inspired originally by McCammon's (1%9) work. However, this program ROKE has eliminated approximations inherent in that and other previous approaches, and attempts have been made to produce an accurate and efficient computer program. ROKE will handle a mixture of four normal (or lognormal) component distributions, with data presented in the form of a histogram

containing up to 30 group intervals. These limits may be changed easily within the program. It may be noted that the program does not require that the intervals in the histogram be uniform.

THE METHOD OF NONLINEAR LEAST SQUARES

Observations are available on n independently observed sample points for two variables y and z. A model is postulated in which y is thought to be a function of z, modified by a purely random "error" component, that is

$$y_i = F(z_i; \boldsymbol{\theta}) + \epsilon_i, \qquad\qquad (1)$$

where $\boldsymbol{\theta}' = (\theta_1, \theta_2, \theta_3, \theta_4, \text{- - -}, \theta_k)$ is a vector of k unknown parameters to be estimated, and ϵ_i is the random component of y_i, $i = 1, 2, 3, 4, \ldots, n$. The function $F(z;\theta)$ is a nonlinear function of the θ_j which can not be transformed into a linear function. For example:

$$y_i = \theta_1[1 - \exp(-z_i/\theta_2)]$$

is a nonlinear function of the two parameters θ_1, and θ_2, whereas

$$y_i = \theta_1 \exp(-z_i/\theta_2)$$

is "intrinsically linear" because a logarithmic transformation

$$\log_e y_i = \log_e \theta_1 - \frac{z_i}{\theta_2}$$

allows $\theta'_1 = \log_e \theta$, and $\theta'_2 = 1/\theta$, to be estimated by linear least squares.

To estimate $\boldsymbol{\theta}$ by the method of least squares the criterion

$$S = \sum_{i=1}^{i=n} (\epsilon_i)^2 = \sum_{i=1}^{i=n} [y_i - F(z; \boldsymbol{\theta})]^2 \qquad\qquad (2)$$

must be minimized with respect to each of the θ_j. In linear least squares the result is a system of simultaneous equations which can be solved for the θ_i. In a nonlinear situation a direct solution is not possible, because the resulting equations contain functions and crossproducts of the θ_j. Therefore an iterative or approximation method must be employed.

Suppose a close approximation to the real values of $\boldsymbol{\theta}$ is made, say $\boldsymbol{\theta_0}$, where $\theta_0 = \{\theta_{01}, \theta_{02}, \ldots \theta_{0k}\}$. Then the Gauss-Newton method (Draper and Smith, 1967) may be used to produce more accurate approximations to 0.

Using Taylor's expansion:

$$F(z; \boldsymbol{\theta}) = F(z; \boldsymbol{\theta_0}) + (\theta_1 - \theta_{01}) \frac{\partial F(z; \boldsymbol{\theta})}{\partial \theta_1}\bigg|_{\boldsymbol{\theta} = \boldsymbol{\theta_0}}$$

$$+ (\theta_2 - \theta_{02}) \frac{\partial F(z; \boldsymbol{\theta})}{\partial \theta_2}\bigg|_{\boldsymbol{\theta} = \boldsymbol{\theta_0}}$$

$$+ \cdots + (\theta_k - \theta_{0k}) \frac{\partial F(z; \boldsymbol{\theta})}{\partial \theta_k}\bigg|_{\boldsymbol{\theta} = \boldsymbol{\theta_0}}$$

$$+ \text{ terms of higher order in } (\theta_j - \theta_{0j}). \qquad (3)$$

If θ_0 is close enough to θ the higher order terms may be neglected so that:

$$F(z; \boldsymbol{\theta}) = F(z; \boldsymbol{\theta}_0) + \Delta\theta_1 \frac{\partial F(z; \boldsymbol{\theta}_0)}{\partial\theta_1} + \Delta\theta_2 \frac{\partial F(z; \boldsymbol{\theta}_0)}{\partial\theta_2}$$
$$+ \ldots + \Delta\theta_k \frac{\partial F(z; \boldsymbol{\theta}_0)}{\partial\theta_k}, \qquad (4)$$

where

$$\frac{\partial F(z; \boldsymbol{\theta}_0)}{\partial\theta_j}$$

denotes

$$\left. \frac{\partial F(z; \boldsymbol{\theta})}{\partial\theta_j} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

Equation (2) then can be differentiated with respect to each θ_j to produce a set of simultaneous equations resulting in a solution for $\Delta\theta$.

$$D \cdot \Delta\boldsymbol{\theta} = g, \qquad (5)$$

where

$$g_i = \sum_{i=1}^{i=n} \frac{\partial F(z_i; \boldsymbol{\theta}_0)}{\partial\theta_j} [y_i - F(z_i; \boldsymbol{\theta}_0)] \qquad (6)$$

$$j = 1, 2, 3, 4, \ldots, k$$

and D is a k by k matrix defined as

$$d_{jl} = \sum_{i=1}^{i=n} \frac{\partial F(z_i; \boldsymbol{\theta}_0)}{\partial\theta_j} \times \frac{\partial F(z_i; \boldsymbol{\theta}_0)}{\partial\theta_l}, \qquad (7)$$

where

$$j=1,2,3,4,\ldots,k, \qquad l=1,2,3,4,\ldots,k.$$

A new set of approximations to θ are produced, that is $\theta_1 = \theta_0 + \Delta\theta$. The procedure is repeated with θ_1, θ_2 and so on until no further improvement in the sum of squares can be achieved. If θ_0 is not close to θ the result may be a local minimum and not the optimum. If θ_0 is at a great distance from θ, the usual result is that no improvement of the estimates can be determined, and fresh approximations must be made.

Faster convergence to the optimum may be obtained by adding fractions of $\Delta\theta$ to θ_0, and choosing the values which give the lowest sum of squares. The author has determined inverse powers of 3 most useful in this respect, that is 1/3, 1/9, 1/27 and 1/81, and these have been incorporated into the computer program.

MIXTURES OF SEVERAL NORMAL DISTRIBUTIONS

Suppose a model is postulated which is a mixture of m normal (Gaussian) distributions. This implies that if a single specimen is taken from the overall population it must have derived from one of m component normal distributions. The probability density function for the value of such a specimen would be:

$$q(x; \boldsymbol{\theta}) = \sum_{l=1}^{l=m-1} \alpha_l \phi\left(\frac{x - \mu_l}{\sigma_l}\right) + \left(1 - \sum_{l=1}^{l=m-1} \alpha_l\right) \phi\left(\frac{x - \mu_m}{\sigma_m}\right),$$

(8)

where α_j is the proportion of the overall population deriving from component distribution l; μ_l, is the mean of the lth component distribution; σ_l is the standard deviation of the lth component; and $\varphi(z)$ is the probability density function of the standard normal distribution.

The vector of parameters to be estimated is therefore:

$$\boldsymbol{\theta}' = \{\mu_1, \sigma_1, \alpha_1, \mu_2, \sigma_2, \alpha_2, \ldots, \mu_m, \sigma_m\}.$$

The cumulative distribution function of x would be given by:

$$Q(x; \boldsymbol{\theta}) = \int_{\infty}^{x} q(v; \boldsymbol{\theta})\, \mathrm{d}v$$

$$= \sum_{l=1}^{l=m-1} \alpha_l \Phi\left(\frac{x - \mu_l}{\sigma_l}\right)$$

$$+ \left(1 + \sum_{l=1}^{l=m-1} \alpha_l\right) \Phi\left(\frac{x - \mu_m}{\sigma_m}\right),$$

(9)

where $\varphi(z)$ is the cumulative distribution function of the standard normal curve.

In an analysis for the components of such mixtures of distributions, data are available usually in the form of a histogram-or can be converted easily to this form. We shall denote the endpoints of the groups in such a histogram by x_1, x_2, x_3, ... x_n, where n is the number of groups in the histogram. The 'frequency in group i' refers to the number of samples with values lying between x_{i-1} and x_i. The 'observed proportion' of samples in group i is given by the frequency in that group divided by the total number of samples taken. This will correspond to y_i in equation (1).

According to the model of a mixture of n components, the "expected" or theoretical proportion of samples determined in group i would be given by:

$$F(z_i; \boldsymbol{\theta}) = Q(x_i; \boldsymbol{\theta}) - Q(x_{i-1}; \boldsymbol{\theta}) \quad i = 1, 2, 3, 4, \ldots, n.$$

(10)

The partial derivatives of $F(z;\theta)$ then become:

$$\frac{\partial F(z_i; \boldsymbol{\theta})}{\partial \theta_j} = \frac{\partial Q(x_i; \boldsymbol{\theta})}{\partial \theta_j} - \frac{\partial Q(x_{i-1}; \boldsymbol{\theta})}{\partial \theta_j}$$

$$i = 1, 2, 3, \ldots, n, \quad j = 1, 2, 3, \ldots, 3m - 1. \quad (11)$$

To implement the nonlinear least-squares solution for θ, the partial derivatives must be determined for the *3m - I* parameters. The partial derivatives $Q(x;\theta)$ are as follows:

(a) Proportion parameters

$$\frac{\partial Q(x;\theta)}{\partial \theta_l} = \Phi\left(\frac{x-\mu_l}{\sigma_l}\right) - \Phi\left(\frac{x-\mu_m}{\sigma_m}\right) \quad l = 1, 2, 3, \ldots, m-1.$$

(12)

(b) Means of components

$$\frac{\partial Q(x;\theta)}{\partial \mu_l} = \frac{\alpha_l}{\sigma_l}\phi\left(\frac{x-\mu_l}{\sigma_l}\right) \quad l = 1, 2, 3, \ldots, m.$$

(13)

(c) Standard deviations

$$\frac{\partial Q(x;\theta)}{\partial \sigma_l} = -\frac{\alpha_l(x-\mu_l)}{\sigma_l^2}\phi\left(\frac{x-\mu_l}{\sigma_l}\right) \quad l = 1, 2, 3, \ldots, m.$$

(14)

MIXTURES OF OTHER DISTRIBUTIONS

(a) Lognormal

By definition, if a variable is lognormal its natural logarithm has a normal distribution. Thus, if a histogram is formed from the logarithms of the sample values, this may be analyzed as a mixture of normal distributions. Once the means and standard deviations of these components have been determined, the corresponding parameters of the "actual" lognormal components can be determined from:

$$\lambda_l = \exp\left(\mu_l + \tfrac{1}{2}\sigma_l^2\right)$$
$$\omega_l^2 = \lambda_l^2[\exp(\sigma_l^2) - 1] \quad l = 1, 2, 3, \ldots, m.$$

The relative proportions of the mixtures are unaffected by the transformation. If the data are only available in the form of a histogram, it is necessary to take logarithms of the endpoints of the groups, and continue the analysis as described.

(b) Other continuous distributions

Any distribution for which the cumulative distribution function and the probability density function can be calculated (numerically or analytically) may be used in such an analysis. The sole requirement of the method is that the terms $\partial Q/\partial \theta_i$ must be calculable.

(c) Mixtures of different distributions

There seems to be no theoretical restriction on analyzing for mixtures of dissimilar distributions. However, strong practical justification would need to be present, and good indications to be searched for in the types of distributions.

NUMBER OF COMPONENTS

There seems to be no theoretical maximum to the number of components postulated. The formulation of both model and technique is general for m. A practical constraint is that n must be at least as large as $3m$, so that D is not singular, and the goodness of fit may be tested. Obviously the more parameters to be estimated, the higher the dimensions being searched for a minimum. It has been determined in practice that as m increases, the difference between θ_0 and θ must decrease if the optimum is to be determined. That is, the higher m is, the closer θ_0 must be to the true value.

GOODNESS OF FIT

Once the optimum solution for θ has been determined, it is desirable to check the goodness of fit of the model to the data. The author has found the χ^2 test most useful for this purpose, although various alternatives have been suggested.

The statistic used is:

$$S = \sum_{i=1}^{i=n} \frac{[f_i - NF(z_i; \boldsymbol{\theta})]^2}{NF(z_i; \boldsymbol{\theta})}, \tag{15}$$

where f is the observed frequency in group i, and N is the total number of samples used.

Under the null hypothesis that the samples were taken from a population which consists of a mixture of m distributions as postulated, S should have an approximate χ^2 distribution with $n - 3m$ degrees of freedom.

THE PROGRAM ROKE

The solution for the parameters of the mixed population was simplified slightly for the purposes of computation. Rather than compare the observed proportion in each group with its expected value, it was decided to use the cumulative proportion up to the endpoint of each group, and compare this with its expected value. That is, we redefine y_i as the observed proportion of samples below endpoint x_i, and replace equation (10) by:

$$F(z_i; \boldsymbol{\theta}) = Q(x_i; \boldsymbol{\theta}) \tag{16}$$

to give the expected proportion of samples below endpoint x_i. Equations (12)-(14) then give the partial derivatives for $F(z;\theta)$, and these are substituted into the nonlinear least squares method as described by equations (2)–(7).

This modification gives a significant increase in speed of computation, and seems (in practice) to be more stable than the full method. It also is analogous exactly to the standard graphical methods (Hald, 1952) which attempt to fit a model to the cumulative curves.

It should be noted that the original definition of $F(z;\theta)$ in equation (10) must be used for the χ^2 test described in equation (15).

The information supplied to the program must be in the form of a histogram — although this may be modified easily — and a set of initial estimates for the unknown parameters of the particular model to be fitted. These initial estimates may be made by eye (for the experienced user) or by

graphical methods. The author has found probability plots of the cumulative curve particularly useful for estimating standard deviations.

The program, as it stands, will solve for normal components for lognormal components as specified. It is possible to estimate other distributions by changing the FUNCTION segments ATUAN and ENLAD, and if necessary the conditions imposed on the parameters within IFFISH.

Program sections

ROKE   Main Program Segment. This controls input
        and output of the program, and the calling of the various routines.

ANDRAD This subroutine performs a χ^2 test between the model specified by the
        parameters, and the histogram from the data. It also provides a visual
        comparison between the model and the data.

IFFISH This is the nonlinear least-squares routine, which solves for the "best" estimates
        of the various parameters. Incorporated in the subroutine is a slightly modified
        IBM SSP routine for solving sets of simultaneous equations (lines 209-238 in
        Appendix II).

ENLAD   This function supplies the partial differential of $F(X;\theta)$ at value X with respect to
        parameter J, as given by equations (12)-(14)

ATUAN   This function gives the probability of a sample from a population with MC
        components, and the parameter values stored in PARS, taking a value below X.
        That is, function ATUAN provides values of $Q(X;\theta)$ as given by equation (9).

HAVNOR This is a numerical approximation function for the standard normal cumulative
        distribution function $\Phi(x)$.

OSSKIL is a numerical approximation function for the standard normal density function
        $\varphi(x)$.

Core requirements

Program ROKE was developed and tested on the CDC 6400 installation at Imperial College, London. The compiler used was the Minnesota FORTRAN Compiler MNF, which required 20,100 words of core to handle this program.

Once the program was compiled, the version presented here required 11,200 words of core to run. A version which could handle ten components and a 75-group histogram required 11,600 words. However, this would be an extreme situation, and the existence of ten components would require a great amount of justification.

A slightly modified version of this program runs on a 32K (8 bit word) minicomputer installation within the Department of Mineral Resources Engineering at Imperial College. The program uses Double Precision throughout, and gives satisfactory accuracy of results.

Run timings

The program compiled under MNF in 1850 milliseconds. A typical run of a three component analysis on a 25-group histogram took 4.3 sec, and a run of a seven component analysis on a 67-group histogram took approximately 80 sec.

Timings depend not only on the number of components in the model and the number of groups in the histogram, but also on how many iterations are required to determine the solution. There is no

limitation placed on the number of iterations within the program. It is the author's experience that a solution seldom runs to more than twenty iterations, and rarely to more than thirty. The example quoted of a three-component analysis took 16 iterations.

If it is desired to limit the number of iterations, to, say, 25 simply add after line 189 of Appendix II

$$\text{IF (ITER.GT.25) GO TO 10}$$

Input to program

The form for the input of the data has been made as flexible as possible. Titles and Variable Formats are read in alphameric form, and used for the various input data. All input is read in one section of the main program, so that alterations may be made with ease.

Card 1:	TTL	A title card for job identification. Up to 80 characters may be used.
Card 2:	Columns	Variable
	1-2	NG	The number of groups in the data histogram
	3	MC	The number of components to be included in the population model
	4	NO	N for normally distributed components, L for lognormal components, (default is normal).
Card 3:	FMT	The format for the cards on which the frequencies in each histogram group have been punched.
Cards 3a,b:	FREQ	The histogram frequencies, in the format specified by Card 3.
Card 4:	FMT	The format for reading the upper endpoints of the groups in the histogram. Note that no upper end-point should be supplied for the last group.
Cards 4a,b:	EPTS	The upper endpoints of the histogram groups, in the format specified by Card 4.
Card 5:	FMT	The format for reading the initial estimates of the component parameters.
Cards 5a,b:	PARS	The initial estimates of the component parameters, in the order: mean of component 1 standard deviation of component 1 proportion of samples coming from component 1 mean of 2 standard deviation of 2 proportion of 2 mean of 3, etc. No proportion should be supplied for the last component.

An example of input cards is given in Appendix la.

Output from program

The output has been designed to present, simply and concisely, the input data, the initial estimated model, and the final model, incorporating both quantitative tests and visual comparisons of the fit of models to data.

The output is in two pages, and the longest line printed is 120 characters long. An example of the printout is given in Appendix lb.

Page 1:   Title for job
Initial parameters
χ^2 goodness of fittest of the "initial" population to the data histogram
A visual comparison of the data or "observed" histogram to the estimated or "expected" histogram.

Page 2:   Title for job
Number of iterations taken
Root mean square deviation of estimated probabilities from the observed proportions
χ^2 goodness of fit test of the "final" population to the data histogram
A visual comparison of the observed histogram and the final model.

REFERENCES

Clark, I., and Garnett, R. H. T., 1974, Identification of multiple mineralisation phases by statistical methods: Trans. Inst. Min. Metall., v. 83, no. 809, p. A43-A52.

Dick, N. P., and Bowden, D. C., 1973, Maximum likelihood estimation for mixtures of two normal distributions: Biometrics v. 29, p. 781-790.

Draper, N. R., and Smith, H., Jr., 1967, Applied regression analysis: John Wiley & Sons, London and New York, 407 p. Hald, A., 1952, Statistical theory with engineering applications: John Wiley & Sons, London and New York, 783 p.

McCammon, R. B., 1969, FORTRAN IV program for nonlinear estimation: Kansas Geol. Survey Computer Contr. 34, 20 p.

Sichel, H. S., 1972, Statistical evaluation of diamondiferous deposits: 10th Inter. Sym. Appl. of Computer Methods in the Mineral Industry, South African Inst. Min. Metall., Johannesburg, Paper No. 5, 9 p.

Please note that, due to the inefficiency of the OCR program, Appendices have been omitted from this copy. If I cannot guarantee that it works, I won't hand it out. If you really want a copy of 25 year old Fortran code, please e-mail geoecosse@kriging.com