ALTERNATIVE PROCEDURES FOR ESTIMATION OF NONLINEAR REGRESSION PARAMETERS

William J. Price

Bahman Shafii

ALTERNATIVE PROCEDURES FOR ESTIMATION OF NONLINEAR REGRESSION PARAMETERS

William J. Price and Bahman Shafii

Statistical Programs
College of Agriculture
University of Idaho
Moscow, Idaho 83844-2337

ABSTRACT

Biological research data are often represented using nonlinear model specifications that lend themselves to the testing of relevant hypotheses concerning the model parameters. This is typically achieved with classical nonlinear least squares techniques such as Gauss-Newton or Levenberg-Marquardt which allow for both the estimation and inference phases of the analysis. Under some circumstances, however, sensitivity to data or model specifications may lead these methods to fail convergence tests or exhibit nonlinearity in the parameter estimates, which will in turn limit the usefulness of inferential results. In such cases, other estimation methods may present a means of avoiding these problems while providing analogous results. The genetic algorithm combined with bootstrapping and Bayesian estimation are two such alternatives. Genetic algorithms represent a nonparametric approach which, when augmented with bootstrap methods, result in both parameter estimation and approximation of the distribution(s). Bayesian estimation, on the other hand, leads directly to parameter distribution and achieves the required moments. These methods and classical nonlinear least squares are demonstrated utilizing a four- parameter cumulative Weibull function fitted to onion seed germination data.

Keywords: Least Squares Estimation, Genetic Algorithm, Bayesian Techniques

I. INTRODUCTION

Nonlinear models provide a flexible framework for describing biological phenomena. They can often effectively capture the complex patterns and structure present in many biological problems while possessing parameters with relevant biological interpretation. Thus, nonlinear models not only describe general trends within a system, but they may also provide insight into the underlying processes.

In general, the nonlinear model may be given as:

$$\mathbf{Y} = f(X_1, X_2, \ldots X_n; \theta_1, \theta_2, \ldots \theta_p,) + \epsilon \qquad (1)$$

where $\mathbf{Y}$ is the n x 1 response vector, X_i are n x 1 regressor vectors, i=1, 2, ..., k, θ_j are unknown parameters, j=1, 2, ..., p, and ϵ is a nx1 residual vector customarily assumed $\epsilon \sim NID(0, \sigma^2\mathbf{I}_n)$. Parameter estimation is traditionally achieved via iterative least squares techniques such as Gauss-Newton or Levenberg-Marquardt with subsequent statistical inference based on linear approximations assuming asymptotic normality. In some cases, however, these iterative methods may have convergence problems or lead to nonlinearity of the estimation situation which can affect the validity of inferential results. These circumstances may arise from problems associated with model specification or sensitivity to the data being used. Procedures for diagnosing nonlinearity such as profile t-plots and profile pair sketches have been suggested by Bates and Watts (1988), however in practice, the computational requirements for these techniques are often excessive or exhibit instability. One possibility for avoiding nonlinearity problems is to employ alternative estimation techniques. Two procedures considered here are a nonparametric bootstrap genetic algorithm, and a numerically integrated Bayesian method. In both cases, parameter estimation and inference can be developed directly from data derived parameter distributions.

One biological problem which displays complex behavior is the process of seed germination over time. When expressed as cumulative percentages, germination exhibits three distinct phases: a lag phase from an initial time until the onset of germination, an increasing phase where germination accumulates in an approximately linear fashion, and a plateau phase where germination slows to an asymptotic maximum (Figure 1). Nonlinear growth models provide a good basis for describing this process if specified with biologically relevant parameters. Examples of such specifications would be the Logistic, Gompertz, Richards and Weibull growth models (Jansen, 1993; Tipton, 1984; Brown and Mayer, 1988; Torres and Frutes, 1990; Shafii et. al., 1991).

Germination will be represented here by a four-parameter Weibull function expressed as:

$$\mathbf{Y} = M(1 - \exp(-K(\mathbf{x} - L)^C)) \qquad (2)$$

where

$\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n\}$ = n x 1 vector of cumulative germination %,
$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ = n x 1 vector of times,
M = theoretical maximum for cumulative germination,
K = rate of increase,
L = lag time to onset of germination, and
C = shape parameter.

This parameterization allows for the estimation of all three phases of germination: lag phase (L), increasing phase (K), and plateau phase (M). The shape parameter, C, has no direct biological interpretation, but is often necessary to capture the skewed nature of germination data. The four-parameter Weibull model will be used to demonstrate and compare traditional

nonlinear least squares methods with alternative estimation techniques in terms of distributional assumptions, parameter estimation, and inferential results.

II. METHODS

Iterative Least Squares

The form of the Weibull model used with iterative least squares is:

$$\mathbf{Y} = M(1 - \exp(-K(\mathbf{x} - L)^C)) + \boldsymbol{\varepsilon} \tag{3}$$

where the variables and parameters are those defined in (2) and $\boldsymbol{\varepsilon}$ is a gaussian error term, $\boldsymbol{\varepsilon} \sim NID(0,\sigma^2\mathbf{I_n})$. Least squares estimation is achieved by an iterative optimization algorithm such as Levenberg-Marquardt, Gauss-Newton or similar which minimize a sum squared residual loss function through a matrix of partial derivatives. Techniques for inferential results are analogous to that of linear least squares with an additional assumption of asymptotic normality. For example a $(1-\alpha)$ joint confidence region is given by:

$$(\theta - \hat{\theta})'\hat{\mathbf{V}}'\hat{\mathbf{V}}(\theta - \hat{\theta}) \leq ps^2F(p, n-p;\alpha) \tag{4}$$

where θ and $\hat{\theta}$ are the parameter and estimate vectors, respectively, $\hat{\mathbf{V}}$ is the partial derivative matrix evaluated at $\hat{\theta}$, and s^2 is the sample variance. This region defines an ellipsoid in p space centered at $\hat{\theta}$ which can be reduced to

$$\hat{\theta}_j \pm t_{(\alpha/2,\ n-p)}SE_j \tag{5}$$

for inferences concerning a single parameter, $\hat{\theta}_j$, with associated standard error, SE_j.

Bootstrap Genetic Algorithm

The genetic algorithm is an optimization technique which draws on the concepts of biological natural selection for its methods. The Weibull model assumed is:

$$\mathbf{Y}_{(g)} = M(1 - \exp(-K(\mathbf{x} - L)^C)) \tag{6}$$

where the subscript g indicates final estimation and is determined from a finite number of bootstrap samples. The genetic algorithm is a nonparameteric procedure and therefore no inferential assumptions are required.

Estimation with a genetic algorithm is an iterative technique which proceeds as outlined in Figure 2. The parameters are defined in vector form as “genomes”, i.e. [M, L, K, C] and may be coded in either binary format (Davis, 1991) or floating point format (Michalewicz, 1992). A set or population of parameter genomes is initialized with random

values. Each member of the initial population is then evaluated against the data by means of a predefined loss function (fitness). Although the loss function can be defined in any manner, squared or absolute residual sums are customary (Davis, 1991). After evaluation, the most fit genomes, i.e. those with smallest loss function, are retained. This subset is subjected to a "reproduction" phase where a new genome population is created for the next evaluation cycle. The reproduction process is guided by two controlling mechanisms (operators) of mutation and cross-over. Mutation is defined as a random permutation in the off-spring of one or more genome elements (parameters). Cross-over occurs when two genomes trade parameter elements to create new genomes (Davis, 1991). The rate of mutation and cross-over operations are determined by preset probabilities. Once reproduction is completed, the new population is evaluated and the most "fit" subset again selected. This will continue until minimal difference among the selected genomes is reached according to a previously determined tolerance limit. Convergence will be affected by the size of the populations as well as the values given for mutation and cross-over probabilities. Determination of these settings is specific to both the problem type and the exact algorithm employed and therefore, must be optimized on a case by case basis. Generally, however, cross-over operators yield better convergence than mutation operators and thus, cross-over probabilities are generally set higher than those for mutation. Population size must be large enough to capture population variability without becoming excessively large leading to slower computation.

The process described above is a deterministic one. At the completion of the final iteration, the single most fit genome is selected as the solution. Additional methods must therefore be carried out to obtain a measure of variability for the estimates. The bootstrap simulation technique provides a good mechanism to achieve this objective.

In a bootstrap simulation, a large number of samples are taken from the data with replacement. Each sample is subjected to the genetic algorithm resulting in a collection of parameter estimates. If G is the cumulative distribution of θ, percentile intervals (Efron and Tibshirani, 1993) can be constructed as:

$$[\hat{\theta}_{\%L}, \hat{\theta}_{\%U}] \approx [\hat{G}^{-1}(\alpha/2), \hat{G}^{-1}(1-\alpha/2)] \tag{7}$$

where $\hat{G}^{-1}$ is based on the collection of sample based parameter estimates. For a finite number of bootstrap samples, B, and given $G^{-1}(\alpha/2) = \theta(\alpha/2)$, a $(1-\alpha)$ percentile interval is given by:

$$[\hat{\theta}_{\%L}, \hat{\theta}_{\%U}] \approx [\hat{\theta}_{B}(\alpha/2), \hat{\theta}_{B}(1-\alpha/2)]. \tag{8}$$

These percentile intervals may then be used to make inferences concerning the parameter estimates.

Bayesian Estimation

Bayesian estimation is based on the Weibull model

$$\mathbf{Y}^* = M(1 - \exp(-K(\mathbf{x} - L)^C)) \tag{9}$$

where $\mathbf{Y}^*$ is the predicted response according to the most probable parameter values of the posterior probability distribution. This distribution is developed as follows.

If the response, percent germination, is assumed gaussian,

$$\mathbf{Y} \sim NID(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n) \text{ for } \boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots, \mu_n\} = M(1 - \exp(-K(\mathbf{x} - L)^C)), \text{ and} \tag{10}$$

the parameter vector is

$$\{\theta, \sigma\} = \{M, L, K, C, \sigma\} \tag{11}$$

then the likelihood for $\{\theta, \sigma\}$ given $\mathbf{Y}$ is:

$$\mathcal{L}(\theta, \sigma|\mathbf{Y}) \propto \{\sigma^{-n}\}\exp\{1/2\textstyle\sum_{i=1}(Y_i - \mu_i)^2/(\sigma^2)\} \tag{12}$$

Further, the general form of the prior distribution of $\{\theta, \sigma\}$ is:

$$p(\theta, \sigma) \propto g(\theta, \sigma). \tag{13}$$

When prior distributions are unknown, then a noninformative or diffuse prior may be defined as:

$$g(\theta, \sigma) \propto \kappa \quad \text{(Laplace, 1812)}$$
$$\text{or} \tag{14}$$
$$g(\log(\theta, \sigma)) \propto \kappa \text{ (Jeffreys, 1939; Box and Tiao, 1973)}$$

for κ = constant. Thus, the final form of the posterior distribution of $\{\theta, \sigma\}$ is:

$$p(\theta, \sigma|\mathbf{Y}) \propto g(\theta, \sigma)\cdot\mathcal{L}(\theta|\mathbf{Y}) \tag{15}$$

On combining Jeffreys noninformative prior (14), with the likelihood function (12), the posterior probability density function becomes:

$$p(\theta, \sigma|\mathbf{Y}) = (\text{const})\cdot\{\sigma^{-(n+1)}\}\exp\{1/2\textstyle\sum_{i=1}(Y_i - \mu_i)^2/(\sigma^2)\}. \tag{16}$$

Estimation and inference for the Bayesian model is obtained through integration of the posterior distribution. This may be accomplished by either analytical or numerical methods. In the analytical case, integrating out σ in (16) results in $\theta_j \sim$ Univariate $t_{(n-p)}$ and joint $\theta \sim$ Multivariate t. Conversely, integrating out θ in (16) gives $\sigma \sim$ inverse gamma (Zellner, 1971). However, for nonlinear models, analytical integration is rarely feasible (Seber and Wild, 1989), and thus, numerical techniques must be used. Numerical integration of posterior

distributions can be carried out with Monte Carlo integration methods (Press, et. al., 1995). This would result in parameter probability distributions from which moments and probability intervals may then be achieved directly.

III. EMPIRICAL RESULTS

The data used in this study concern the germination of onion seed. Eight replications of 100 seeds were incubated at a constant temperature of 25 C and cumulative germination recorded daily for 20 days at various water potential levels. Only the control treatment is presented here. For a full analysis of the data see Shafii, et. al., 1991.

Computations were performed using Linux 2.0.25 and GCC 2.7.2 (Public domain). Least squares solutions and graphics were carried out using SAS 6.12 (SAS, 1991). Program codes are available from the authors upon request.

Least squares estimates are presented in Table 1a. Solution convergence was quick, however it was noted that the estimation procedure was sensitive to the starting values for parameter L and careful consideration of its value and the Levenberg-Marquardt iteration method were necessary to obtain convergence. The underlying residuals were of acceptable magnitudes and showed no excessive patterns or trends. Correlations among parameter estimates were nominal. Maximum germination was estimated at 83% with a lag time to initial germination of approximately 2 days. 95% confidence bounds based on asymptotic normality were reasonably narrow and did not encompass zero. The resulting predicted Weibull curve appeared to follow the data well (Figure 3a).

For the genetic algorithm, the floating point method of Michalewicz (1992) was used with settings: population size = 500, probability of cross-over = 0.8 and probability of mutation = 0.18. The resulting parameter estimates are given in Table 1b. The estimate values were similar to those of least squares with a maximum germination of 86% and lag time of 2 days. The rate parameter K and shape parameter C were also similar in value to the least squares estimates. Residual analysis indicated no problematic patterns or trends, so no action was deemed warranted. Estimation of parameter variability was based on a bootstrap simulation using 1500 samples of size 60. The 95% percentile intervals were wider than the corresponding confidence bounds of least squares, but still tightly bounded the estimates and did not cover zero. Wider bounds are not unexpected since the bootstrap method assumes less information is known about the system, i.e. no distributional assumptions, and thus is more conservative. The predicted model from the bootstrap genetic algorithm also followed the data well (Figure 3b).

Bootstrap frequency distributions and least squares approximations for each parameter are given in Figure 4. Here differences from least squares approximations become evident. The bootstrap distributions for parameters M, L, and C appear skewed. The distribution for L also shows a sharp truncation at 2 days. These are indications of parameter-effect nonlinearity and demonstrate the potential inadequacy of the least squares method. Only parameter K has a good correspondence with the least squares distribution. Joint distributions between parameter sets may also be examined. For example, the joint distribution between

parameters M and C is given in Figure 5. Although this plot does not necessarily have a biological significance, it effectively illustrates the parameter-effect nonlinearity. The ellipse represents the least squares 95% joint confidence region as defined in (4), while the dots indicate the 1500 bootstrap solutions. Clearly the least squares approximation does a poor job of representing the joint region. Joint inferences based on linear approximations would therefore be misleading without adjustments to the significance level.

Before Bayesian analysis can be carried out a prior distribution for θ must be chosen. Jefferys noninformative prior (14) does not provide a good approximation with high parameter-effects nonlinearity (Seber and Wild, 1989). Furthermore, controversy surrounds the use of improper priors in multiparameter situations (Stone, 1976). In this case, it was possible to construct an informative prior due to previous experience with germination data and known theoretical considerations. Bounds on M and L were defined from practical limits, i.e. U[0,100] and U[0, 20], respectively. Earlier use of the Weibull model in germination found the maximum values for K and C to be positive and practically limited to 2 and 3, respectively, leading to U[0,2] and U[0,3] (Shafii, 1991). Given the range of the response, 0 to 100, the largest possible value for σ is 70.71 producing a prior of U[0, 70.71]. Although both the informative and noninformative priors were investigated, only the former is reported here.

Most probable values based on numeric integration of the resulting posterior distribution are presented in Table 1c along with 95% probability intervals. The estimates corresponded closely to those of the previous two methods and provided a Weibull curve which fit the data well (Figure 3c). Maximum germination is predicted at 83% while lag time is 2 days. Rate and shape parameters are 1.08 and 0.36 , respectively. Bayesian analysis also provides an estimate of σ of 4.69. The 95% intervals are much narrower than either least squares or the genetic algorithm. This is evident in the spread of the marginal distributions shown in Figure 6. The Bayesian analysis allows for less variability in parameter values than traditional normal approximations. The skewness in distributions for some parameters is less noticeable than was observed with the genetic algorithm, but is still present. Parameter L also continues to show a sharp truncation at approximately 2 days. Only the distribution of σ agrees closely with that of least squares.

Unlike the genetic algorithm, joint confidence regions were not feasible due to computational restrictions of p-dimensional integration. Further refinement of the numerical integration algorithm or program codes may reduce computational requirements, however, the large number of iterative integrations needed for this problem presented overwhelming time and storage constraints.

The alternative estimation procedures described above produced essentially identical fitted germination response curves as that of the least squares method (Figure 7). Although this would seem to negate any necessity in using alternative methods, it actually demonstrates their value in cases where standard nonlinear estimation algorithms fail. This is particularly true in multiparameter situations where problems such as local minima, parameter-effects nonlinearity and lack of convergence can arise with Gauss-Newton, Levenberg-Marquardt and other iterative optimization routines. Furthermore, these alternatives are a means of achieving

simultaneous estimation in problems which are composed of multiple nonlinear relationships or systems which cannot be expressed in closed form.

IV. CONCLUSIONS

Nonlinear growth models provide a flexible framework for describing biological processes. While ordinary nonlinear least squares techniques are appropriate for many applications, they may encounter difficulties in estimation or inference phases of the analysis. The bootstrap genetic algorithm provides a nonparametric approach which yields analogous results to that of least squares with fewer restrictions and assumptions. The Bayesian method is an attractive alternative which leads directly to parameter probability distributions and achieves the required moments. Some inferential aspects of the Bayesian approach warrants further investigation.

REFERENCES

Bates, D. M. and D. G. Watts. 1988. Nonlinear Regression and its Applications. John Wiley and Sons, NY.

Box G. E. P. and G. C. Tiao. 1973. Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading, MA.

Brown, R.F. and D. G. Mayer. 1988. Representing cumulative germination. 2. The use of Weibull and other empirically derived curves. Ann. Bot. 61: 127-38.

Davis, L. 1991. Handbook of Genetic Algorithms. Van Nostrand Reinhold, NY.

Efron, B. and R. J. Tibshironi. 1993. An Introduction to the Bootstrap. Chapman Hall, NY.

Jansen, J. G. M. 1973. A method of recording germination curves. Ann. Bot. 37: 705-8.

Jeffreys, H. 1939. Theory of Probability. Oxford Univ. Press.

Laplace, P. S. 1812. Theorie analytique des probabilities. Editions Culture et Civilisation, Brussels.

Linux 2.0.25 Linus Torvalds. 1996. Public domain software.

Michalewicz, Z. 1992. Genetic Algorithms + Data Structure = Evolution Programs, Second ed. Springer-Verlag, Berlin.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1995. Numerical Recipes in C, Second Edition. Cambridge Univ. Press.

SAS Institute Inc. 1991. SAS/STAT User's Guide, Vol. 2, Ver. 6. SAS Institute Inc, Cary, NC.

Seber, G. A. F. and C. J. Wild. 1989. Nonlinear Regression. John Wiley and Sons, NY.

Shafii, B., W. J. Price, J. B. Swenson, and G. A. Murray. 1991. Nonlinear estimation of growth curve models for germination data analysis. Proceedings of the 1991 Kansas State University Conference on Applied Statistics in Agriculture, pp. 19-36.

Stone, M. 1976. Strong inconsistency from uniform priors. J. Am. Stat. Assoc., 71, 114-116.

Tipton, J. L. 1984. Evaluation of three growth curve models for germination data analysis. J. Amer.. Soc. Hort. Sci. 4: 51-54.

Torres, M. and G. Frutes. 1990. Logistic function analysis of germination behavior of aged fennel seeds. Env. Exp. Bot. 30: 383-90.

Zellner, A. 1971. An Introduction to Bayesian Inference in Econometrics. John Wiley and Sons, NY.

Table 1. Parameter estimates and upper and lower 95% bounds for a) Least Squares, b) Genetic Algorithm, and c) Bayesian Estimation.

a) Least Squares		Asymptotic 95% Confidence Bounds	
Parameter	Estimate	Lower	Upper
M	83.36	80.54	86.17
K	1.10	0.86	1.33
L	1.95	1.89	2.00
C	0.56	0.41	0.71
b) Genetic Algorithm		**95% Percentile Intervals**	
Parameter	Estimate	Lower	Upper
M	86.57	81.58	94.12
K	1.09	0.84	1.38
L	2.05	1.67	2.49
C	0.49	0.27	0.93
c) Bayesian Estimation		**95% Probability Intervals**	
Parameter	Estimate	Lower	Upper
M	83.98	82.26	85.12
K	1.10	0.95	1.19
L	1.99	1.98	2.00
C	0.36	0.35	0.37
σ	4.69	3.49	6.20

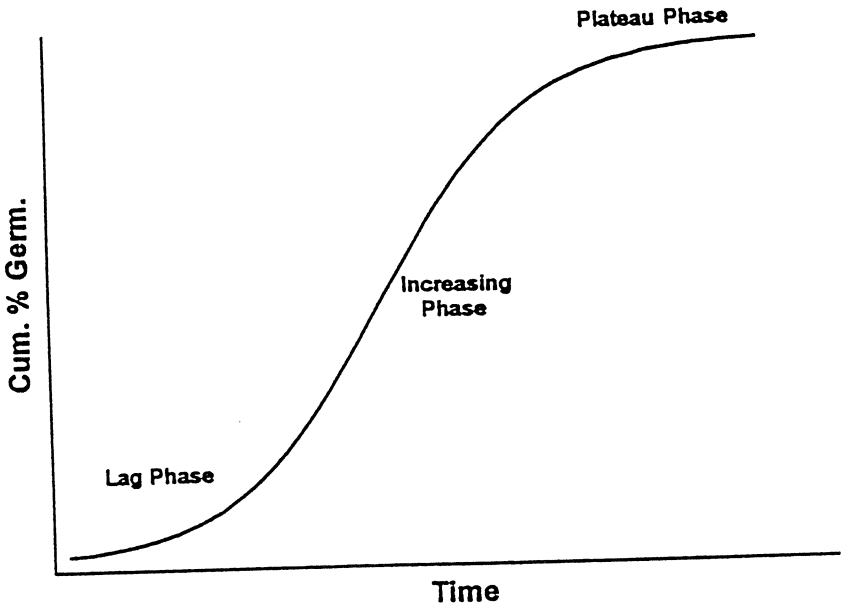


Figure 1. The three phases of cumulative germination over time.

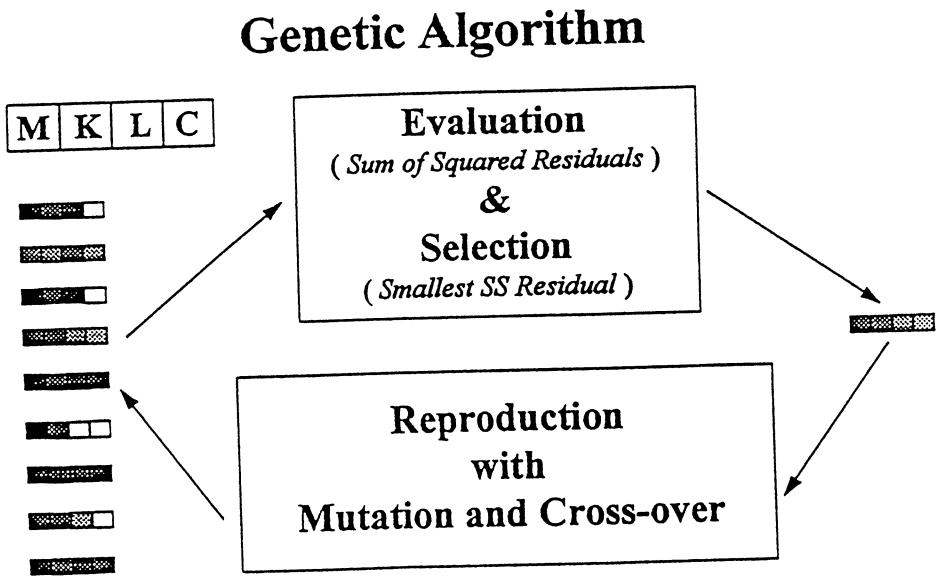
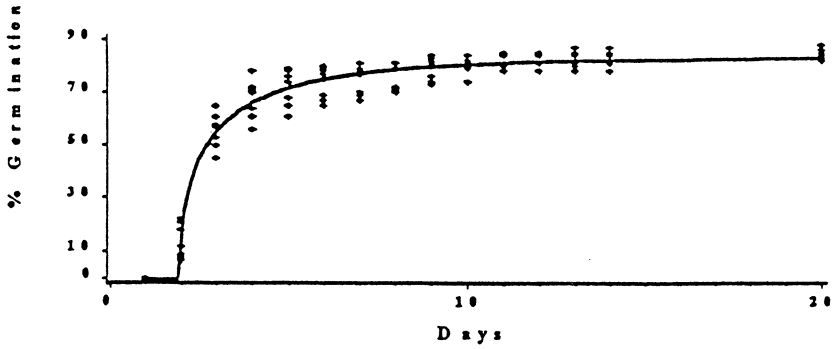
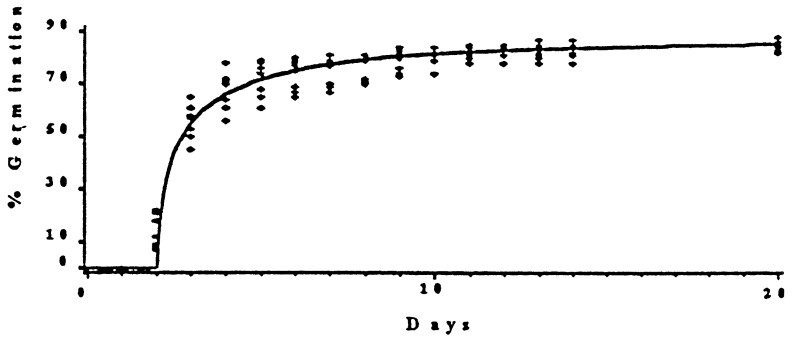


Figure 2. Flow chart for the genetic algorithm.

a)



b)



c)

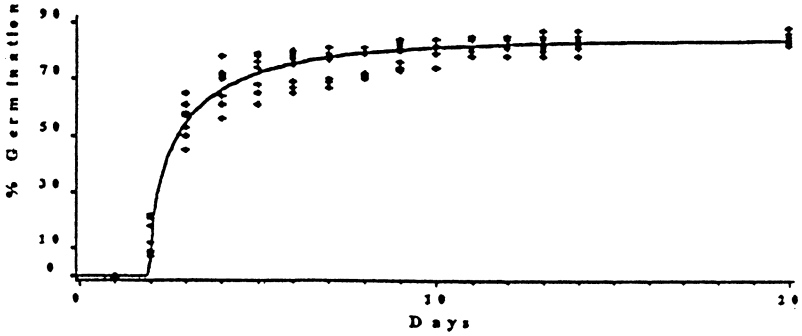


Figure 3. Fitted Weibull curves for a) Least Squares, b) Genetic algorithm, and c) Bayesian estimation.

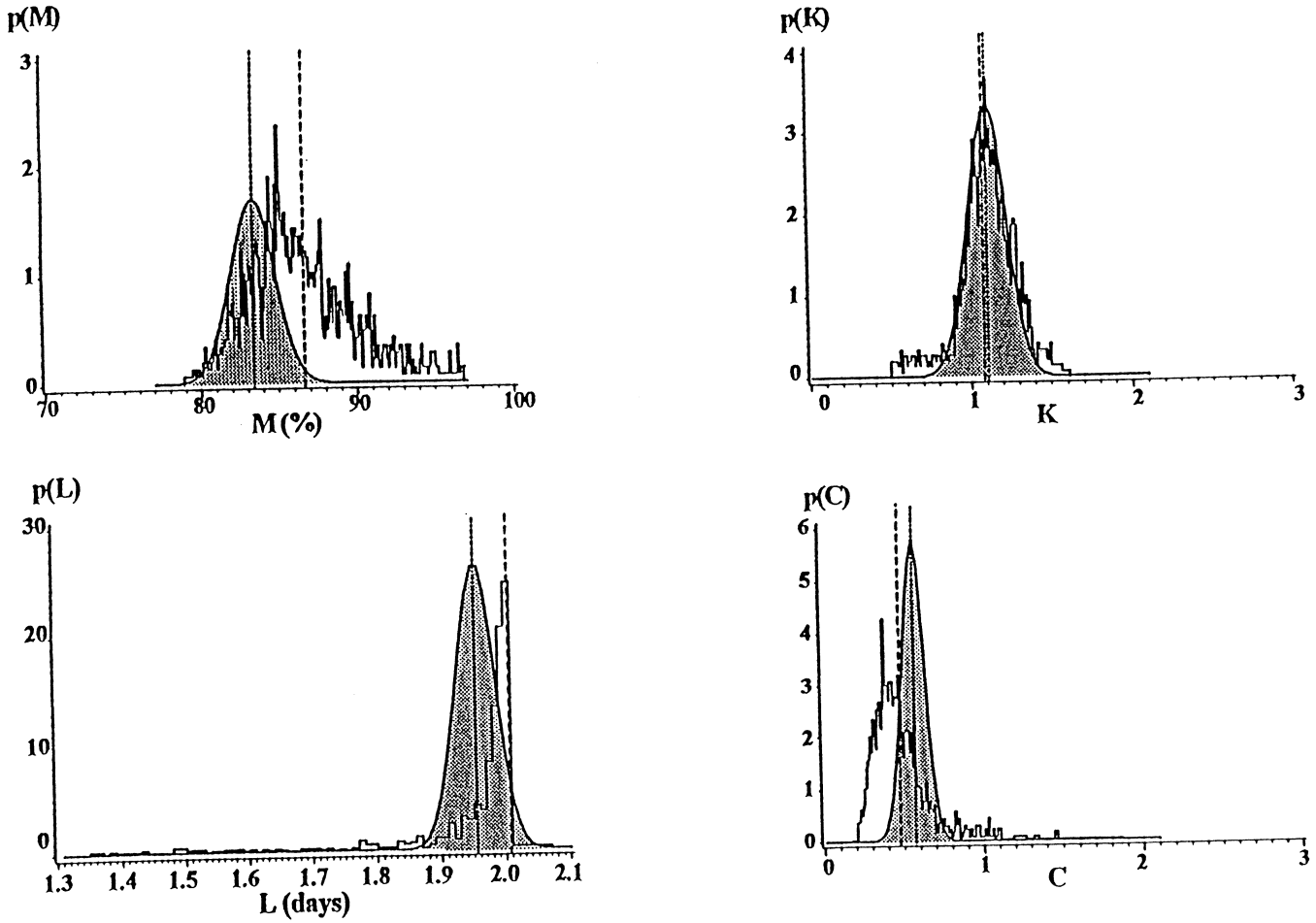


Figure 4. Bootstrap parameter frequency distributions for the Weibull model fitted with a genetic algorithm. Shaded regions represent least squares asymptotic normal distributions.

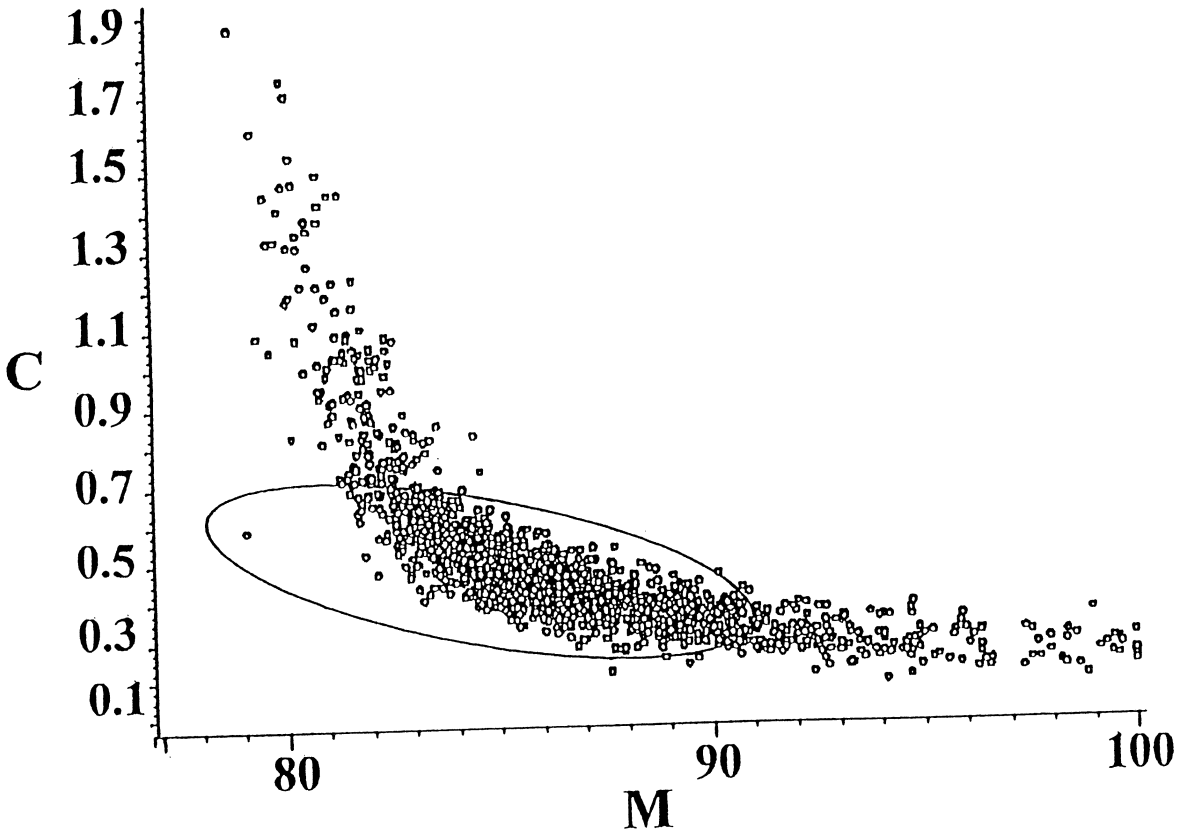


Figure 5. Joint distribution for Weibull parameters C and M. Dots represent result of 1500 bootstrap genetic algorithm simulations, ellipse represents least squares linear approximation.

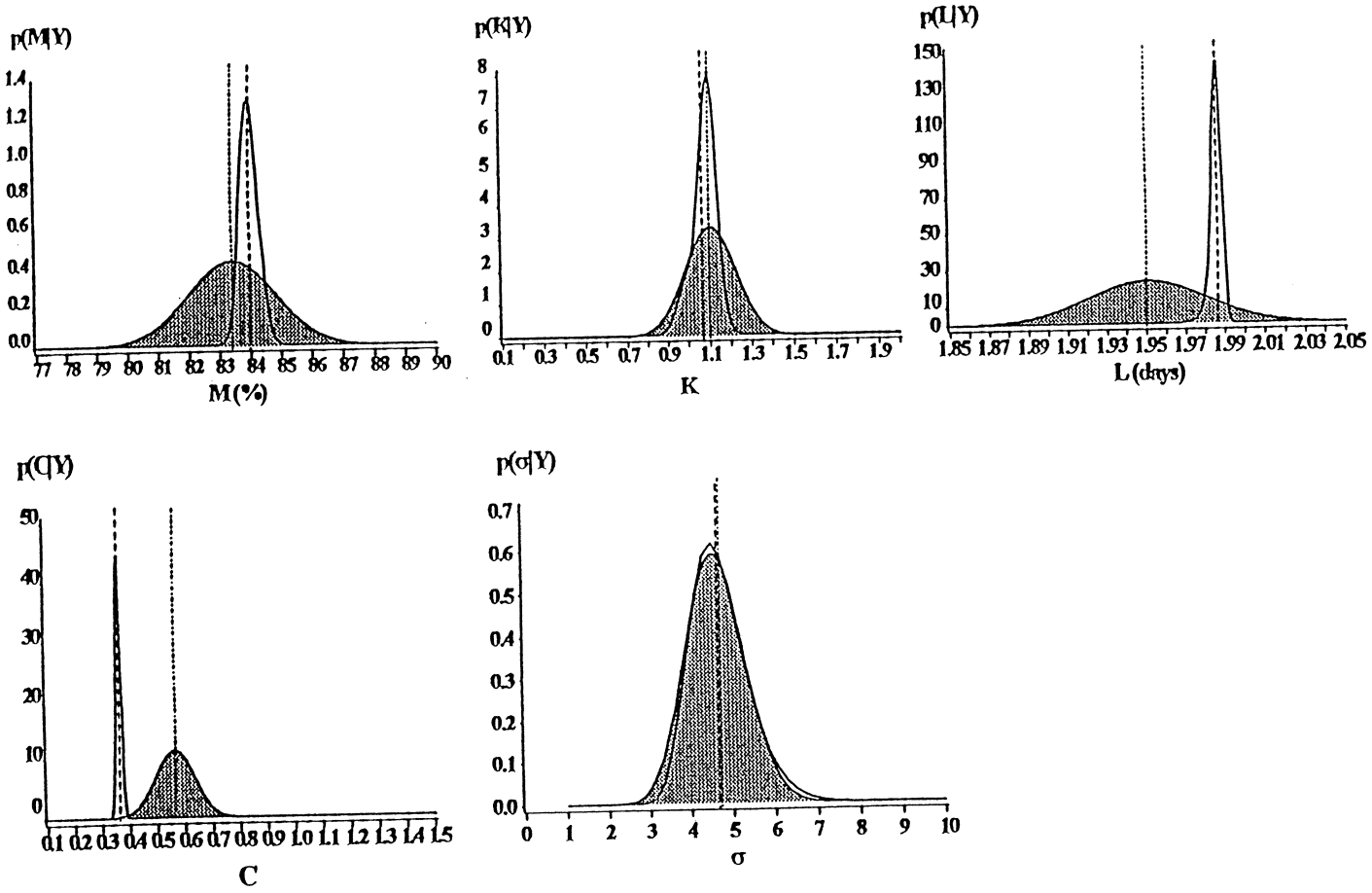


Figure 6. Bayesian parameter probability distributions for the Weibull model. Shaded regions represent least squares asymptotic normal distributions.

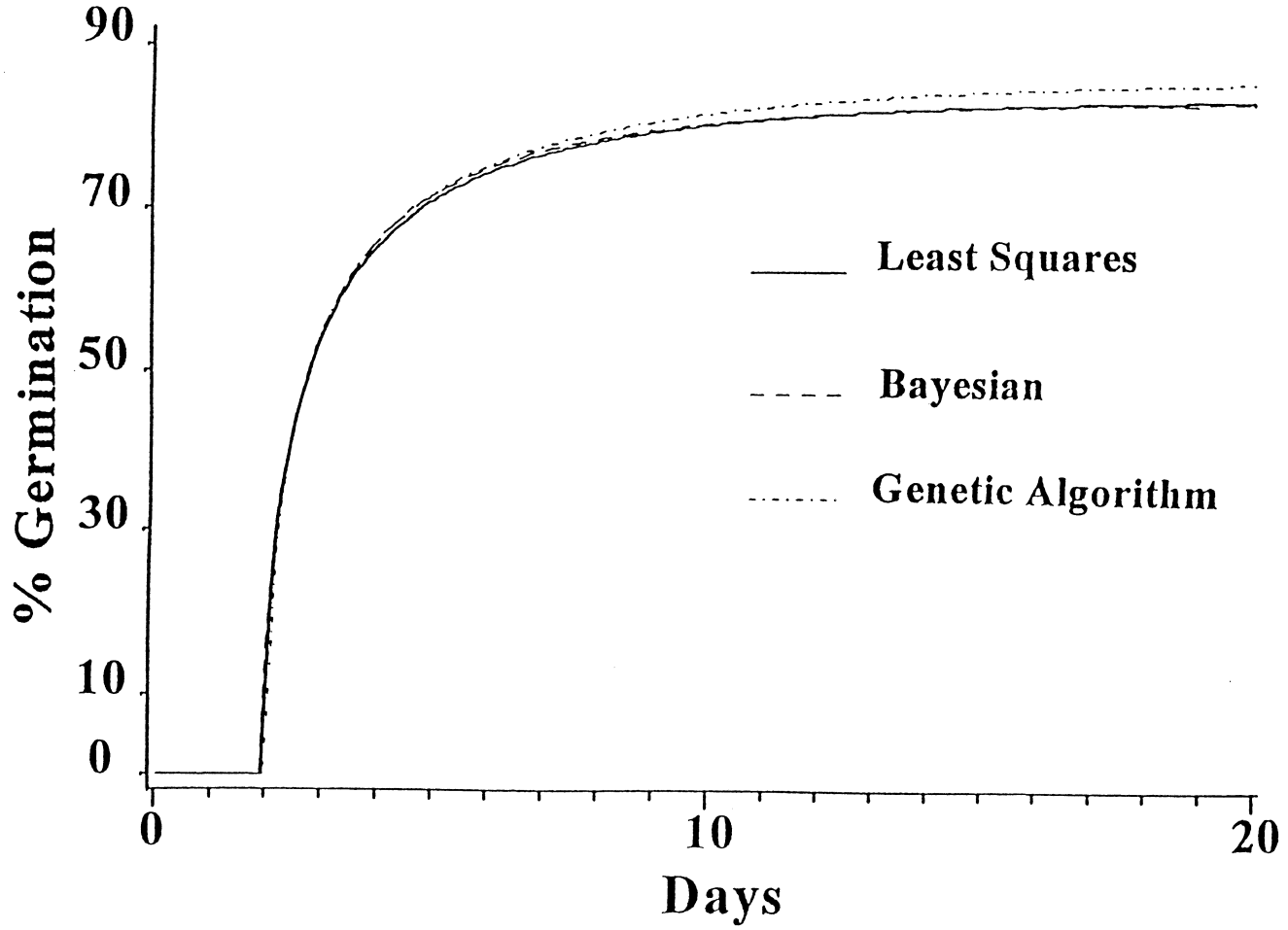


Figure 7. Comparison of fitted Weibull curves for Least Squares, Genetic algorithm and Bayesian estimation.