LOGISTIC REGRESSSION FOR SPATIAL PAIR-POTENTIAL MODELS

Merlise Clyde and David Strauss

Department of Statistics
University of California
Riverside, California 92521

ABSTRACT

The spatial models considered in this paper are Gibbs processes with pairwise interaction potentials, which provide a rich framework for models where the likelihood of a particular configuration of points depends on attraction or repulsion between neighboring pairs of points. However, standard statistical estimation techniques, such as maximum likelihood estimation, have been extremely difficult or impossible to use because of an awkward normalizing constant in the probability density function.

We develop an estimation method based on an idea of Besag's (1975), who outlined a straightforward estimation procedure that places a fine grid over the realization of a point process and uses a pseudolikelihood method to estimate the parameters of the resulting lattice process.

We show that Besag's pseudolikelihood procedure is equivalent to maximum likelihood estimation of a certain logistic regression model, and we prove convergence of the sequence of the pseudolikelihood (parameter) estimates as the mesh of the grid becomes fine, and consistency as the domain's size increases.

We compare the pseudolikelihood method to a graphical maximum likelihood method with a simulation study. In addition we illustrate the procedure using logistic regression to fit several models to Strauss' (1975) redwood seedling data.

1. Introduction

Spatial point patterns in a Euclidian region arise in several disciplines, such as forestry, genetics, ecology, epidemiology, and physics. Although many of the spatial point processes of interest will occur in a two-dimensional or three-dimensional region, we may also consider one-dimensional point processes such as the location of plants along a line transect or the traditional point process in time. Given a sample from a spatial point process, one often begins an analysis by looking at departures from a homogeneous Poisson process (Diggle, 1983, Chapter 2). If we reject the Poisson process hypothesis we may wish to fit a model to the data, to make inferences about the mechanisms generating the point process or to give a concise description of the data. In many cases one is interested in spatial point processes that result from local pairwise-interactions between individuals, such as attractions or repulsions.

The spatial models considered in this paper are Gibbs processes with pairwise interaction potentials (see Cox and Isham, 1980, pp 155-159; Diggle, 1983, pp 63-66). These processes provide a rich framework for models where the likelihood of a particular configuration of points depends on attraction or repulsion between neighboring pairs of points. However, standard statistical estimation techniques, such as maximum likelihood estimation, have been extremely difficult or impossible to use because of an awkward normalizing constant in the probability density function. Various methods have been proposed (Strauss, 1975; Ripley, 1977; Ogata and Tanemura, 1989) to approximate the normalizing constant so as to permit maximum likelihood estimation. In general, these methods can be quite difficult to implement in practice. In order to make these spatial models more accessible, we would like to have an estimation method that can be easily applied with the use of standard statistical packages.

Besag (1975) outlined a straightforward estimation procedure that places a fine grid over the realization of a point process and uses a pseudolikelihood method to estimate the parameters of the resulting lattice process. Besag et al. (1982) justify this procedure by proving a limit theorem that shows that general pairwise-interaction point processes can be obtained as the limit of a suitable sequence of auto-logistic lattice processes. Convergence of the lattice process to the point process as the cell size of the grid decreases does not, however, necessarily imply convergence of the sequence of lattice parameter estimates or consistency of the estimator.

In this paper, we show that Besag's pseudolikelihood procedure is equivalent to maximum likelihood estimation of a certain logistic regression model; thus any statistical package that will perform logistic regression can be used to estimate the parameters of many Gibbsian processes. We show convergence of the sequence of the pseudolikelihood (parameter) estimates as the mesh of the grid becomes fine, and consistency as the domain's size increases. We compare the pseudolikelihood method to a graphical maximum likelihood method with a simulation study. In addition we illustrate the procedure using logistic regression to fit several models to Strauss' (1975) redwood seedling data.

2. Background

2.1 Models

We begin by describing the spatial processes of interest in this paper. The treatment is rather informal; for a more mathematical account see Besag et al. (1982). We then

proceed to give some examples of the spatial processes, and indicate the pseudolikelihood approach to parameter estimation.

The spatial processes are taken on D, a bounded subset of two dimensional Euclidean space. The treatment for dimensionality other than two is entirely analogous. Until further notice D will be taken to have unit area. Events in D will be denoted by ξ_i. For $n > 0$ let x_n denote an ordered set of n such elements. Thus x_n is an element of D^n, and we write x for a generic element of the set

$$\Omega = \bigcup_{n=0}^{\infty} \{D^n\}.$$

The spatial point process defines a distribution on the subsets x of Ω. Let μ denote the the distribution corresponding to the Poisson process on D with unit intensity, such that n is a Poisson random variable with mean 1 and the distribution of x given n is uniform on D^n. We can then define the distribution for a new spatial process in terms of a function $g(x)$ that specifies how likely the realization x is relative to the unit Poisson process.

The processes we will consider in this paper are Gibbs distributions. For such processes we can write the probability density function with respect to the unit Poisson process as

$$f(x) = (1/Z)\exp[U(x)] \tag{2.1}$$

where the normalizing constant $Z = \int e^{U(x)} d\mu$ is finite. In particular, we will consider pair-potential models where the potential function U is of the form

$$U(x) = \alpha n(x) + \sum_{\xi_i \neq \xi_j} \sum u(\|\xi_i - \xi_j\|). \tag{2.2}$$

In (2.2), $n(x)$ denotes the number of events in x and the pair-potential function $u(r)$ indicates the interaction between events ξ_i and ξ_j. Positive values of $u(r)$ correspond to attraction between events at separation r, and negative values to repulsion. If we take $u(r) \equiv 0$ we obtain a new Poisson process with intensity e^{α}.

In applications, u is taken to be of some fairly simple form characterized by a small number of parameters. Some potential models that have been used to model biological phenomena include the square-well and the Lennard-Jones (Fig. 1). For the square-well potential, $u(r)$ is given by

$$\begin{array}{rcll} & = & -\infty & 0 < r \leq c_o \\ u(r) & = & \beta & c_o < r \leq c_1 \\ & = & 0 & c_1 < r \end{array} \tag{2.3}$$

which has the Strauss process and the hard-core model as special cases. The Strauss process is obtained when $c_0 = 0$ and β is non-positive; in the hard-core model β is zero. The square-well potential can be generalized to a step function potential with

$$u(r) = \beta_i \quad \text{if } c_{i-1} < r \leq c_i \quad \text{for } i = 1, \ldots, k \tag{2.4}$$

where $c_o = 0$, $c_k = \infty$. In order for 2.4 to describe a valid potential, β_1 must be non-positive and $\beta_k = 0$. For k large enough, we should be able to approximate many

continuous potentials. However, we may wish to fit a continuous model with fewer parameters such as the Lennard-Jones potential, where $u(r)$ is given by

$$u(r) = \beta r^{-\kappa} + \gamma r^{-\ell} \qquad \beta < 0. \tag{2.5}$$

Here κ and ℓ are suitable integers with $\kappa > \ell$. (For the three-dimensional models in physics, the usual choice is $\kappa = 12$, $\ell = 6$.) Other potential models are described in Ogata and Tanemura (1989).

2.2 Maximum Pseudolikelihood Estimator

As noted earlier, maximum likelihood estimation of (2.1) for non-Poisson processes is generally infeasible because the normalizing constant Z is intractable. One way around this, suggested by Besag (1975, 1977), begins by placing a grid over D. Let A_i denote the ith cell of the grid and let n_i be the number of events in A_i. For a sufficiently fine grid n_i is either 0 or 1. The pseudolikelihood of the $\{n_i\}$ is defined to be

$$\prod_i P(n_i | \{n_j : j \neq i\}), \tag{2.6}$$

and a maximum pseudolikelihood estimator (abbreviated MPE) is a parameter value that maximizes (2.6). In particular, since (2.6) does not involve Z, such estimators are not difficult to compute. In the next section we examine their properties more closely.

3. Some Aspects of the MPE

3.1 Logistic Regression Estimation

Let A_i be a cell of size a in the unit domain D, and write x_i for the intersection of x and $D \backslash A_i$. Let η_i be the centroid of A_i and let ξ be a generic event in A_i. Then the odds-ratio of the conditional probabilities in (2.6) can be written as

$$\frac{P(n_i = 1|x_i)}{P(n_i = 0|x_i)} = \frac{\int_{\xi \in A_i} f(\xi, x_i) d\xi}{f(x_i)} = a f(\eta_i, x_i)/f(x_i) + o(a) \tag{3.1}$$

where $o(a)$ denotes a quantity ε_a such that ε_a / a tends to zero as a tends to zero for every x. Thus the log-odds or logit is

$$\text{Logit}\{P(n_i = 1|x_i)\} = \ln(a) + \Delta U_i(x) + o(a) \tag{3.2}$$

where ΔU_i is the change in potential resulting from the addition of an event at η_i to the configuration x_i. This is a logistic regression estimation model, with binary dependent variables n_i. Although the n_i are, of course, dependent, maximization of (2.6) is equivalent to formal maximum likelihood estimation for the logistic regression (3.2), and can thus be implemented by an iteratively reweighted least squares procedure on standard computer packages such as BMDP (Jennrich and Moore, 1975). For the potentials discussed in section 2, the logits are linear functions of the unknown parameters, and thus any statistical package that will perform logistic regression can be used.

As an example, consider the step function potential (2.4). Neglecting the $o(a)$ term, (3.2) becomes

$$\text{Logit}\{P(n_i = 1|x_i)\} = \ln(a) + \alpha + \sum_{j=1}^{k} \beta_j z_{ji} \tag{3.3}$$

where n_i is an indicator variable of presence/absence of an event in A_i and z_{ji} is the number of events $\xi \in x_i$ where $c_{j-1} < ||\eta_i - \xi|| < c_j$. Similarly for the Lennard-Jones potential (2.5),

$$\text{Logit}\{P(n_i = 1|x_i)\} = \ln(a) + \alpha + \beta_1 r_{\kappa i} + \beta_2 r_{\ell i} \tag{3.4}$$

where $r_{\kappa_i} = \Sigma||\eta_i - \xi_j||^{-\kappa}, R_{\ell i} = \Sigma||\eta_i - \xi_j||^{-\ell}$, and $\beta_1 < 0$. The intercept from most programs will estimate $[\ln(a) + \alpha]$; it may be adjusted to give the estimate of α, since a is known.

The logistic regression procedure has a number of practical advantages. Step function models can be used to suggest other continuous potential models. Different models can be easily fit to the data and compared informally through their pseudolikelihood ratios. Other diagnostics provided by some of the programs, including the predicted probabilities and maximum number of correct predictions using an optimal cut-off value, may be useful for model comparisons.

3.2 Convergence of the MPE as a Tends to Zero

The logistic regression procedure involves division of the domain D into $N = 1/a$ cells. The limiting behavior of the logistic regression estimator as a tends to zero deserves some attention. For a given realization x with n events, the number of cells with $n_i = 1$ stays fixed at n while the number of cells with $n_i = 0$ namely, $(N - n)$ tends to infinity as a tends to zero. For simplicity, suppose we use all N cells in the regression. Let θ be the vector of model parameters and denote by θ_a the MPE corresponding to a grid with N cells each of size a. We now show that the sequence of MPE's θ_a converge to a nontrivial limit θ^* as a tends to zero. This limit will, of course, generally differ from the true θ, as it depends on the single realization x.

Number the cells so that the first n are those with $n_i = 1$. This will be possible if a is taken to be sufficiently small. Let Δ_i denote the increment $U(\eta_i, x_i) - U(x_i)$, resulting from the inclusion of an event at the centroid of cell i. Neglecting terms that are $o(a)$, we have

$$\begin{aligned} P(n_i = 1|x_i) &= a \exp(\Delta_i) \\ P(n_i = 0|x_i) &= 1 - a \exp(\Delta_i) \end{aligned} \tag{3.5}$$

Then the logarithm of the pseudolikelihood (2.6) is

$$LPL = n \ln(a) + \sum_1^n \Delta_i + \sum_{n+1}^N \ln(1 - a \exp(\Delta_i)),$$

which reduces to

$$LPL = \sum_1^n \Delta_i - a \sum_{n+1}^N \exp(\Delta_i) \tag{3.6}$$

apart from a constant. We can write this as

$$LPL^* = \sum_1^n \Delta_i - \langle \exp(\Delta_i) \rangle + O(a) \tag{3.7}$$

where $\langle \ \rangle$ denotes the operation of averaging over the N cells, and the $O(a)$ term arises from the inclusion of the first n items in that average. Now as a tends to zero

the average tends to the spatial integral of $\exp(\Delta_i)$ over D. Hence, apart from the $O(a)$ term, (3.7) is independent of a. A simple continuity argument then shows that if θ_a denotes a parameter value that maximizes (3.7), then the sequence $\{\theta_a\}$ tends to a limit as a tends to zero.

3.3 Consistency

A minimal requirement of an estimator is that it is consistent as the domain D becomes large. To discuss this we need to relax the requirement that D is of unit size and instead consider a sequence of domains D_m, with sizes d_m that tend to infinity as m tends to infinity. Associated with these will be a sequence of functions $\{f_m(x)\}$ specifying the density of the point process with respect to a unit Poisson process on D_m, where $f_m(x)$ has potential function

$$U_m(x) = \alpha n_m(x) + \sum_{\xi_i \neq \xi_j} \sum u(\|\xi_i - \xi_j\|), \tag{3.8}$$

and α and the parameters β characterizing u are independent of m. Let $\theta = (\alpha, \beta)$, and denote by $\theta_{a,m}$ the MPE corresponding to a grid with cell sizes all equal to a. Then the consistency result is

> As m tends to infinity and a tends to zero, the sequence of MPE's converges almost surely to θ.

We offer an informal justification. For fixed m let $f_m^*(\underline{n}, \theta)$ be the density for a Markov random field on the cells of D_m; f_m^* is of form (2.1), with potential function U specified by θ. Corresponding to f_m^* is a point process obtained by choosing an event from a uniform distribution on A_i for each i where $n_i = 1$. Besag et. al. (1982) show that the resulting density functions tend to $f_m(\theta)$ for (almost) all x as a tends to zero. On the other hand, Geman and Graffigne (1987) have shown that, for a sequence of Markov random fields $\{f_m^*\}$ with fixed a and common parameter θ, the MPE sequence θ_m converges almost surely to θ. Hence, taking both limits, we obtain the result.

We note that it is not necessary to take the product (2.6) over all cells A_i for consistency to obtain. It follows from the arguments of Geman and Graffigne (1987) that it is sufficient that the number of cells used in the pseudolikelihood tends to infinity with m. It is, of course, required that the choice of cells to use is made independently of the realization x. Two obvious procedures are

(i) sample the requisite number of cells at random

(ii) select the cells as far apart as possible, in a regularly spaced pattern.

We would recommend the latter, as it would be expected to yield the more efficient estimator and is easier to carry out. Evidently a compromise must be made in the choice of a; a small value will require a large number of cells to result in an acceptable number with $n_i = 1$, whereas a large value will produce a biased estimator. We offer some comments on this in the next section based on a simulation study.

4. Simulation Study

We carried out a simulation study to examine the behavior of the logistic regression procedure for a one-dimensional spatial point process on the unit interval. We used a square-well potential (2.3) with a hard-core radius of 0.0001 and interactive range $0.0001 < r < 0.0005$. The intensity parameter, a, was chosen to be 5.2 so that there would be approximately 100 (= exp(5.2)) points in the interval over the range of β's studied from -1.0 to 0.0. Realizations of the process were generated using a spatial birth and death process (Ripley 1977).

In general maximum likelihood estimation using (2.1) is infeasible because of the normalizing constant Z. However, we can estimate β using a maximum likelihood procedure based on graphical methods. Let Y be the number of interacting pairs; that is,

$$Y = \#(\xi_i, \xi_j) \text{ such that } 0.0001 < ||n_i - n_j|| < 0.0005.$$

This is the sufficient statistic for β if α is known, so that the MLE for β satisfies the likelihood equation

$$Y = E_\beta(Y) = Z'(\beta)/Z(\beta) = g(\beta).$$

Although g is unknown, we can approximate it by plotting simulation based sample averages of Y against β. The curve in Figure 2 represents the smoothed means of Y for 500 simulations each for 50 β's between -1.0 and 0.0.

Using Figure 2, we can obtain the MLE of β and compare this to the MPE, either with α known or with both α and β jointly estimated from the data. For the MPE, the cell size a was taken as the hard core radius, 0.0001, and the complete grid was used. Figure 3 compares the three methods in terms of mean-square error and bias for 11 different β's. The MPE with α known is actually somewhat better than the MLE in terms of bias and MSE. The MPE with both α and β estimated from the data is only slightly worse than the MLE in terms of MSE. Since the MLE requires α to be known, the former comparison seems to be more appropriate.

We also examined how changes in a affected the MPE by estimating α and β with a ranging from 0.001 to 0.00005. Figures 4 and 5 suggest that there is very little change in the parameters for a's smaller than twice the hardcore radius. In general, if one has the computing facilities available decreasing values of a should be used until the parameter estimates stabilize.

5. Example

We applied the logistic regression procedure to Strauss's (1975) redwood data, using the region previously analyzed by Ripley (1977) and Diggle (1983) so that the results will be comparable (Figure 6). We considered a series of models based on the step function potential $u(r)$ defined below, with an intensity parameter α.

We define models 1 through 9 as follows:

Model 1 is the Poisson Process with intensity α.

Model 2 is a square-well with intensity α and interaction parameter β_2.

For $2 < i \leq 9$, Model i is a step function with intensity α and interaction parameters $\beta_2, \ldots, \beta_i$.

$$u(r) = \begin{array}{lllll} \beta_1 = -\infty & 0 & < r & \leq 0.016 & \text{(the hard core)} \\ \beta_2 & 0.016 & < r & \leq 0.050 & \\ \beta_3 & 0.050 & < r & \leq 0.075 & \\ \beta_4 & 0.075 & < r & \leq 0.100 & \\ \beta_5 & 0.100 & < r & \leq 0.125 & \\ \beta_6 & 0.125 & < r & \leq 0.150 & \\ \beta_7 & 0.150 & < r & \leq 0.200 & \\ \beta_8 & 0.200 & < r & \leq 0.250 & \\ \beta_9 & 0.250 & < r & \leq 0.350 & \\ 0 & & r & > 0.350 & \end{array}$$

The hard core radius was set equal the minimum inter-point distance. We used SAS's CATMOD procedure to perform the logistic regressions for Models 1 - 9. Figure 7 shows the parameter estimates based on Model 9 (full model). The parameter estimates appeared to be fairly stable over the different models. Figure 8 compares the log pseudolikelihoods for the various models and the log pseudolikelihood ratio of Model i to Model $i+1$ for the contribution of β_{i+1} in the model. Although we cannot do any formal significance testing, it appears that three parameters (i.e. Model 3) are sufficient in this context. With this model the range of interactions corresponds to about 6 feet in the original units, and had been previously suggested as the likely range of interaction (Strauss 1975). The negative estimate for β_7 may not be significantly different from zero, or may correspond to repulsion among the parent trees (stumps) in the area from which the seedlings possibly sprouted.

We could also continue to subdivide the interval and add more steps to determine if a continuous potential with a decaying interaction function is appropriate. Although the step functions require estimation of more parameters they do provide a way to explore pair-wise interactions over different ranges, so that one may then select possible continuous potentials to fit to the data.

A final cautionary note should be sounded about the use of potential models that are attractive in part of their range. For sufficiently strong attraction (depending on the range and number of points) the probability is very large that the realization will consist mostly of a single cluster of points. For any particular case the issue can In principle be resolved by long simulation runs of the model; but in practice the results often depend critically on the initial realization.

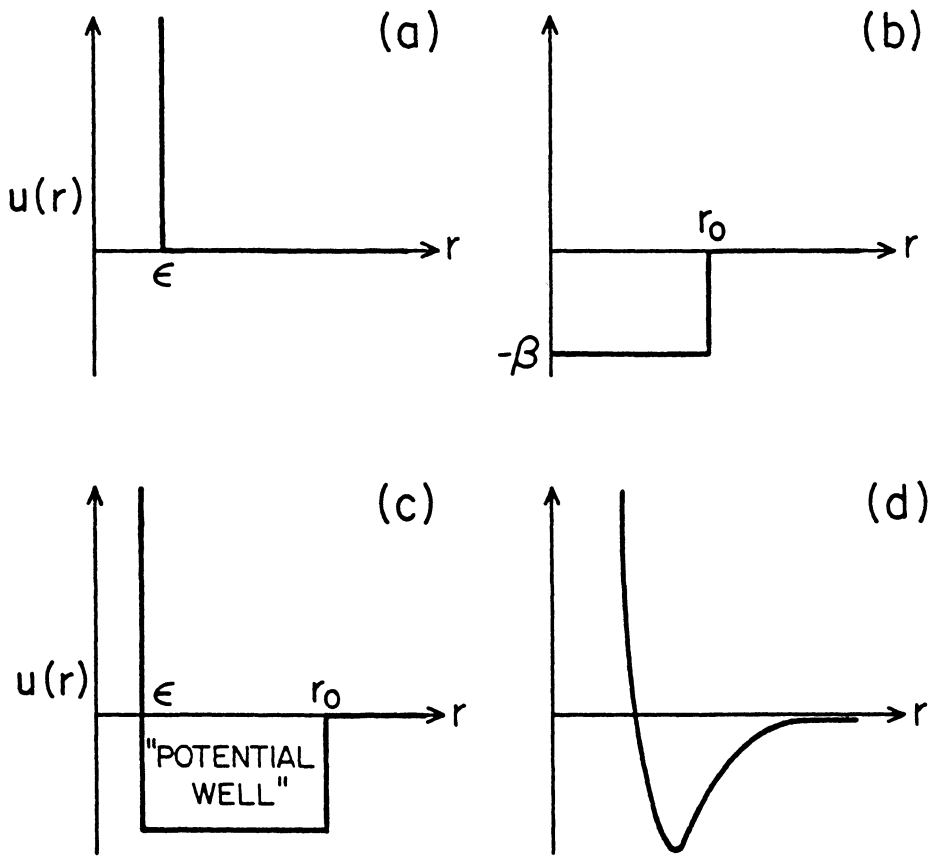


Figure 1. Pair potentials: (a) hard core; (b) Strauss; (c) square-well; (d) Lennard–Jones.

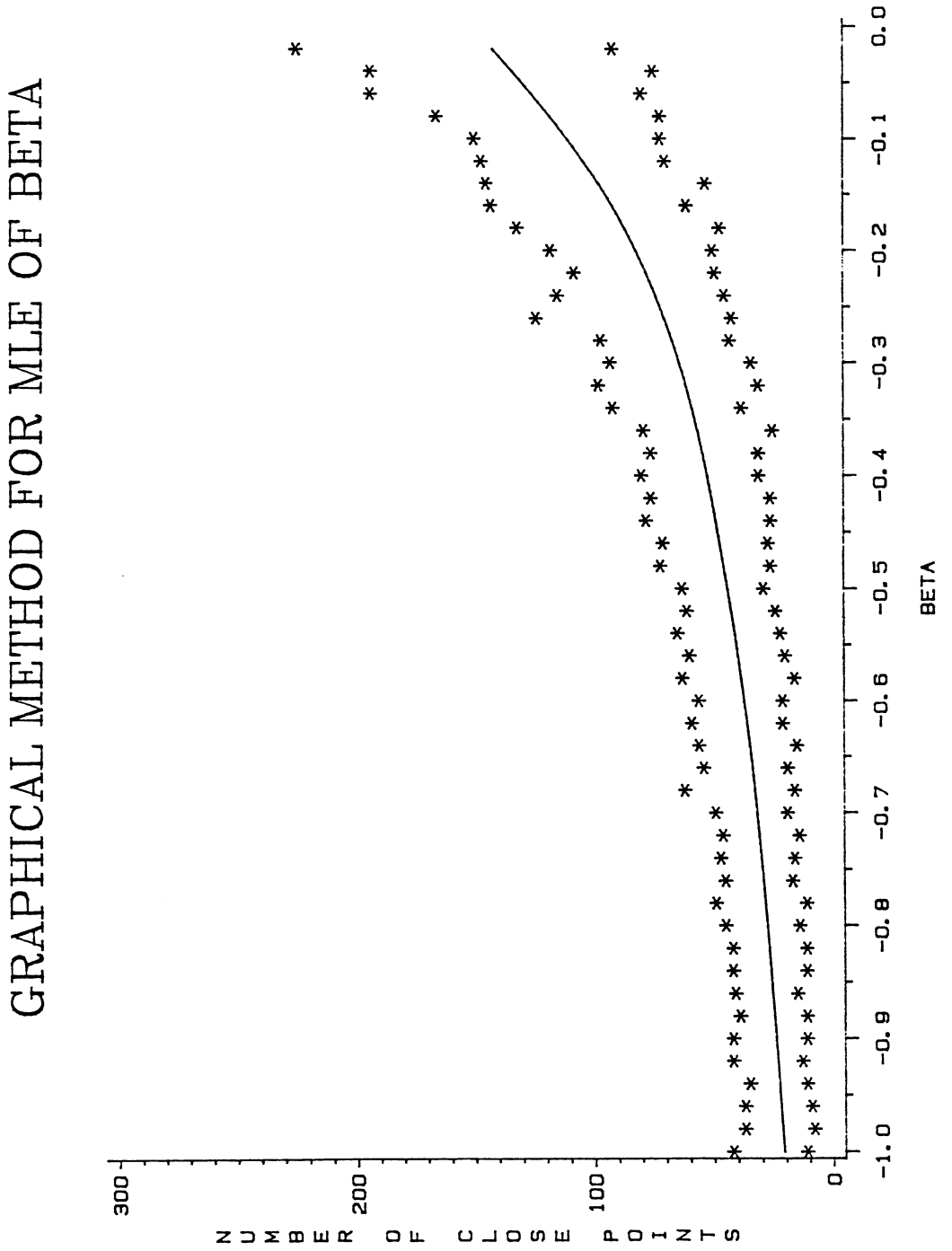
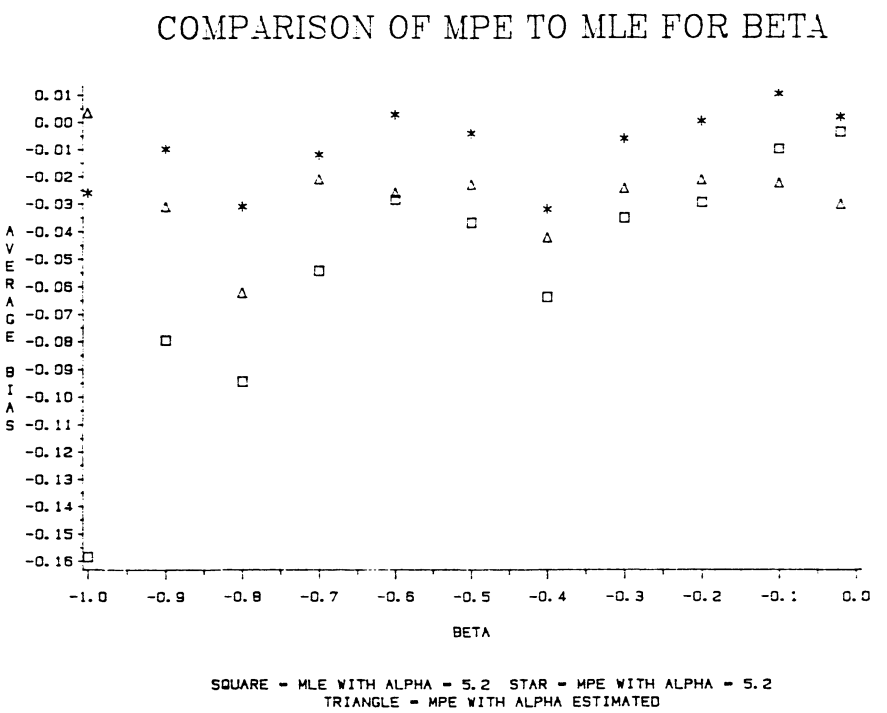


Figure 2. Curve for approximate MLE of β. The $*$'s represent the maximum and minimum number of interacting pairs from the simulations.



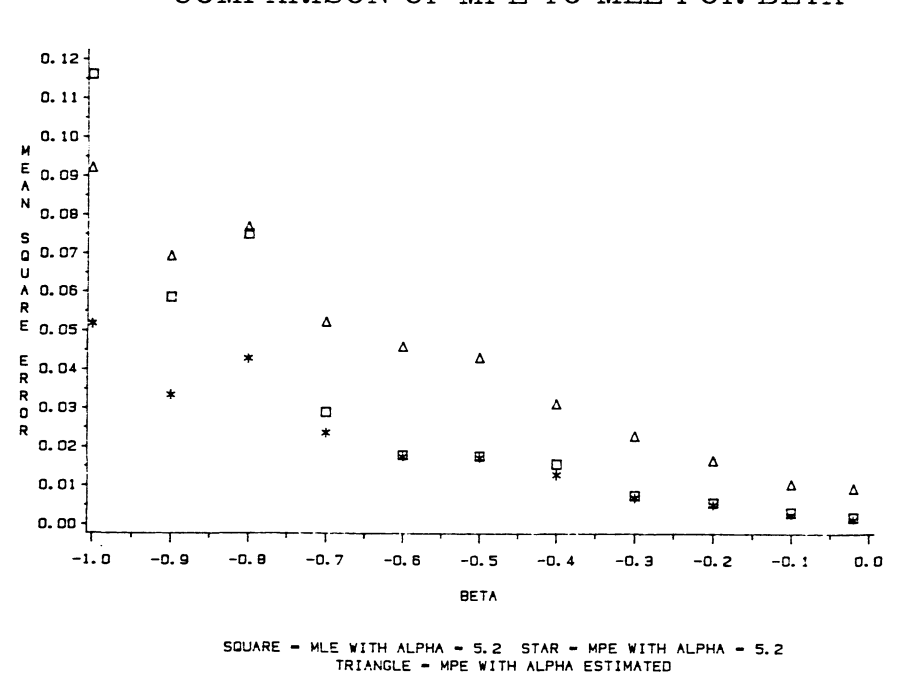


Figure 3. Comparison of MPE to the approximate MLE for β in terms of bias and mean-square error.

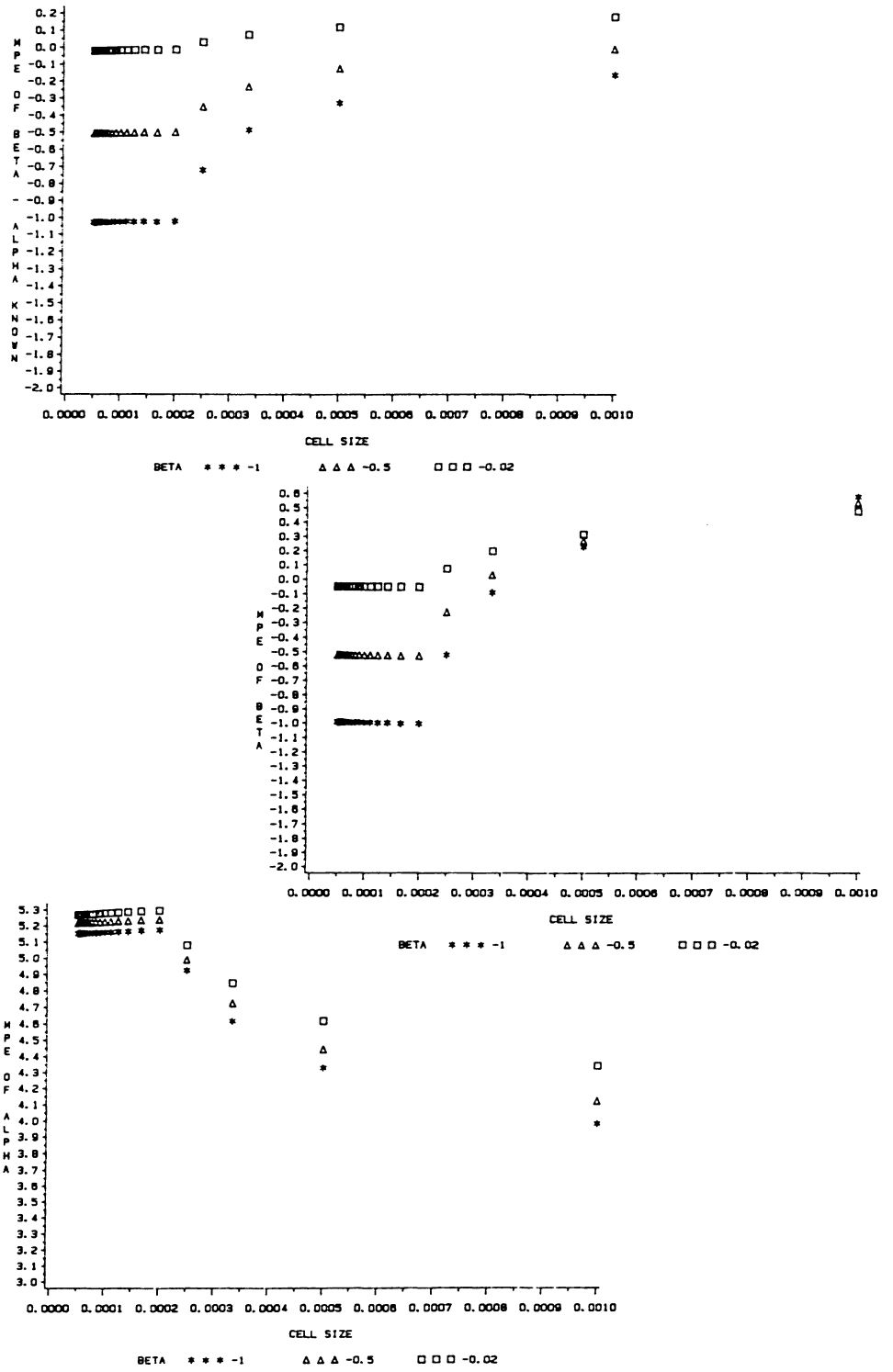


Figure 4. Convergence of MPE as the cell size decreases for β with α known and for β and α jointly estimated.

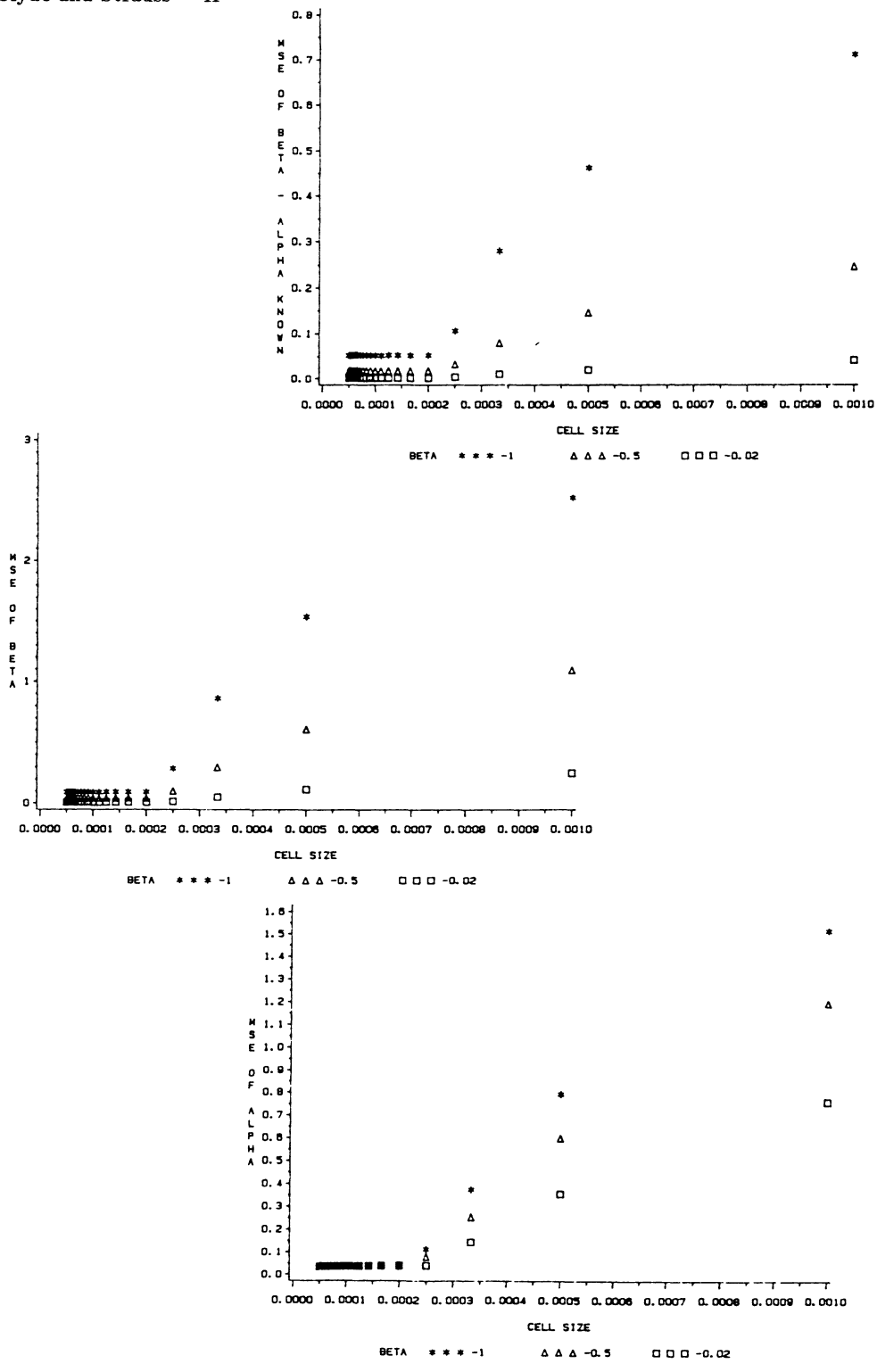


Figure 5. Mean-square error of MPE as cell size decreases for β with α known and for β and α jointly estimated.

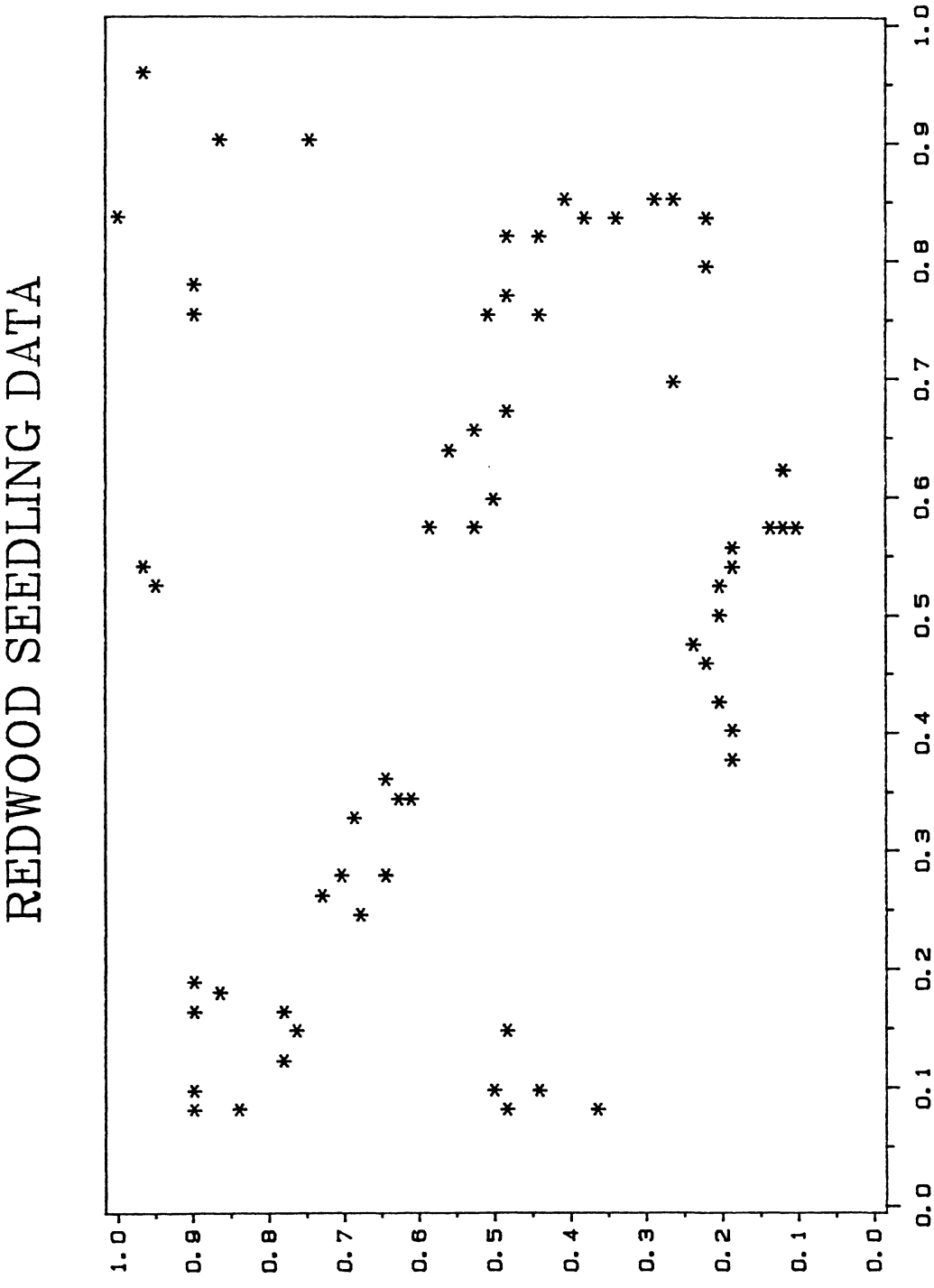


Figure 6. Locations of 62 redwood seedlings.

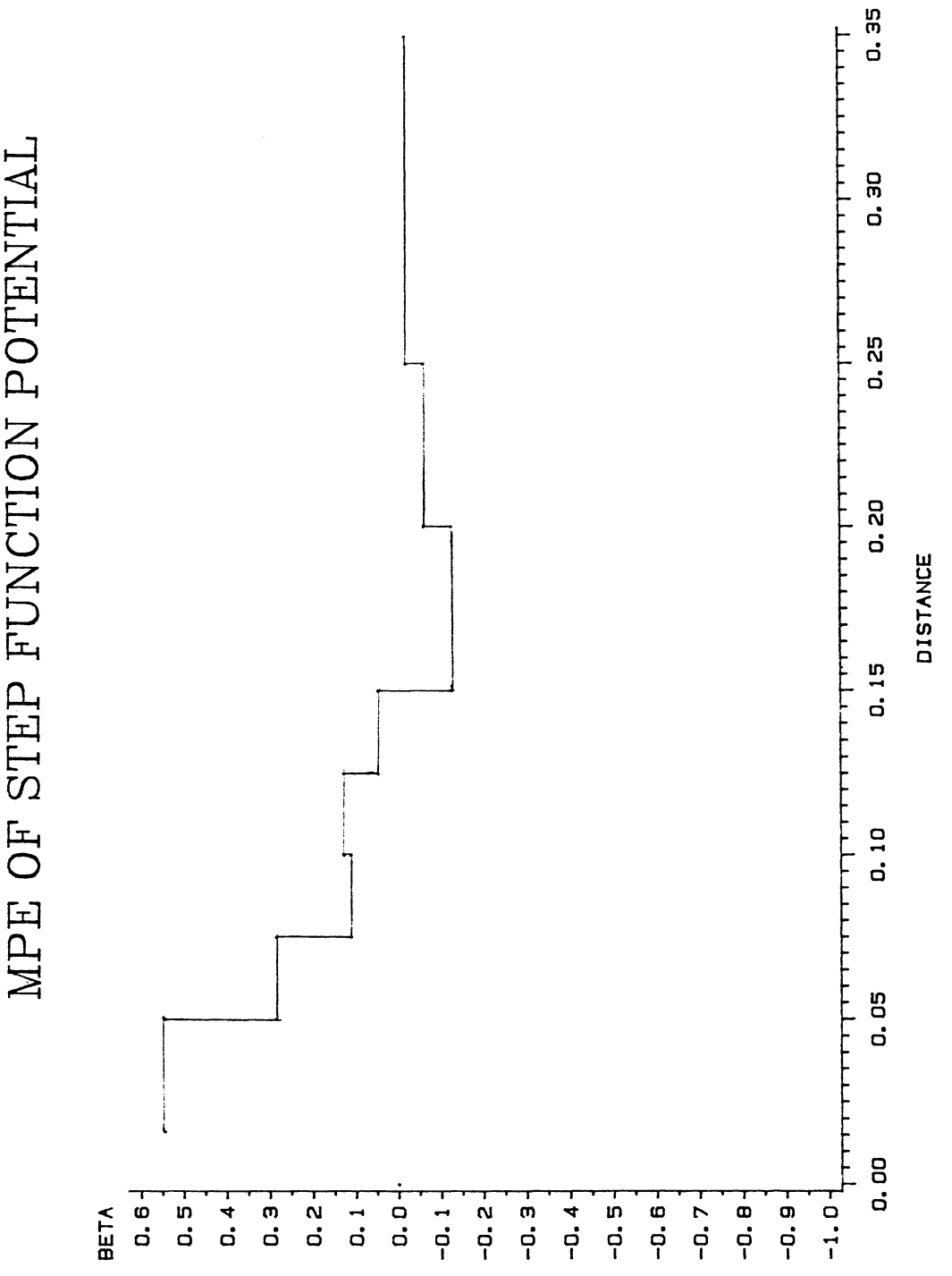
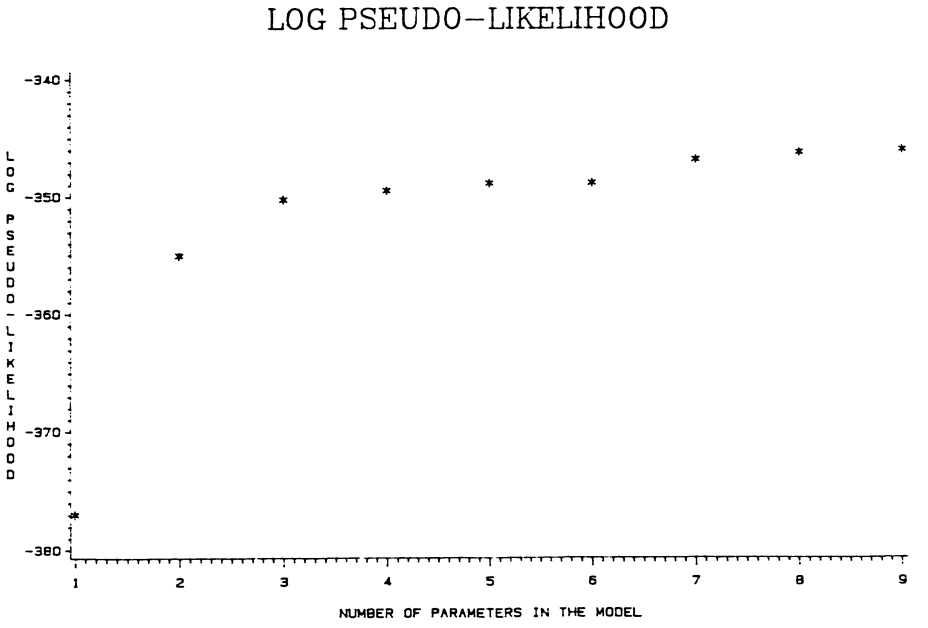


Figure 7. MPE of step function potential based on Model 9.



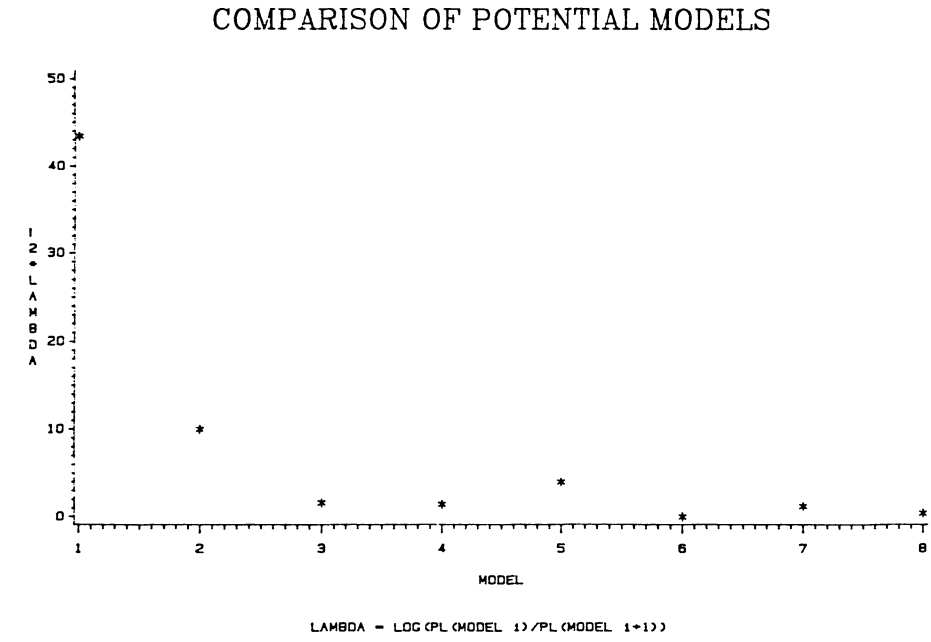


Figure 8. Comparison of models based on log pseudolikelihood and log pseudolikelihood ratio.

6. Acknowledgments

We would like to thank Paul Lowe from the Academic Computing Center at UCR for providing computer time on the Cray at the San Diego Supercomputer Center for the simulation study.

7. Literature Cited

Besag, J. E. (1975). Statistical analysis of non-lattice data. *The Statistician* **24**, 179-195.

Besag, J. E. (1977). Some methods of statistical analysis of spatial data. *Bulletin of the International Statistical Association* **47**, 77-92.

Besag, J. E., Milne R., & Zachary, S. (1982). Point process limits of lattice processes. *Journal of Applied Probability* **19**, 210-216.

Cox, D. R., & Isham, V. (1980). *Point Processes.* Chapman and Hall.

Diggle, P. J. (1983). *Statistical Analysis of Spatial Point Patterns.* Academic Press.

Geman. S., & Graffigne, C. (1987). Markov random field image models and their application of computer vision. *Proceedings of the 1986 International Congress of Mathematics.* American Mathematical Society.

Jennrich, R. I., & Moore, R. H. (1975). Maximum likelihood estimation by means of nonlinear least squares. *Proceedings of the Statistical Computing Section*, American Statistical Association, 57-65.

Ogata, Y., & Tanemura, M. (1984). Likelihood estimation of soft-core interaction potentials for Gibbsian point patterns. *Annals of the Institute of Statistical Mathematics* **41**, 583–600.

Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society* (B) **39**, 172-212.

Strauss, D. (1975). A model for clustering. *Biometrika* **62**, 467-475.