AGGREGATION BIAS IN MAXIMUM LIKELIHOOD ESTIMATION OF SPATIAL AUTOREGRESSIVE PROCESSES

Tony E. Smith
Department of Systems Engineering
University of Pennsylvania
Philadelphia, PA 19104

October 16, 2001

Abstract

In statistical models of spatial behavior, there is often a mismatch between the scale at which data is available and the scale at which key spatial dependencies are known to occur. However, in attempting to incorporate finer grain information about spatial dependencies, certain estimation problems arise. Here it is shown that maximum likelihood procedures can produce significantly negative estimates of positive autocorrelation. This problem is analyzed in the context of a simple spatial autoregressive process, and possible correction procedure is proposed for reducing aggregation bias.

1. Introduction

In statistical models of spatial behavior, there is often a mismatch between the scale at which data is available and the scale at which key spatial dependencies are known to occur. For example, spatial dependencies in housing values may be strongest between adjacent houses or houses on the same block face, while available data may only be available at the blockgroup or even census tract level. Hence when modeling such dependencies in terms of spatial lag or spatial autocorrelation effects, the corresponding proximity weight matrices are necessarily limited to the same degree of aggregation as the given data. However, in many cases it is possible to obtain *finer grain* weight matrices in term of existing map data. In the illustration above, for example, it may be possible to determine proximity relations at a more appropriate scale, such as shared block faces rather than shared blockgroup boundaries.

In this context, the central question of interest here is how to combine such finer grain information about spatial dependencies with aggregated data in a manner which improves the overall goodness of fit. A number of initial efforts in this direction (including various attempts at lower-dimensional approximations to weight matrices for regression models with spatial autoregressive errors and/or spatial lags) all produced disappointing results. However, further investigation revealed that the poor fit of these models was partly a consequence of the *maximum likelihood estimation* procedure itself. In particular, when spatial interaction effects are at a finer scale than the basic data, the standard maximum likelihood procedure has a strong tendency to *underestimate the degree of spatial autocorrelation*. At first glance, this would appear to be consequence of the well-known fact that correlations among aggregates tend to diminish (shrink toward zero) as aggregation size increases [see for example Chapter 5 of Arbia, 1989]. However, the present situation is quite different, and appears to be more a consequence of the *variance minimizing* tendency of maximum likelihood estimation which, in the presence of aggregation, tends to favor negative autocorrelation. In many instances, a substantial portion of the estimates are thus *negative* when actual autocorrelation is *positive*. Moreover, this is not simply a 'small sample' problem. While such estimates are theoretically consistent, examples show that even for very large sample sizes these problems need not disappear. Hence they serve to illustrate some of the practical limitations of consistency itself.

For purposes of analysis, we focus in the present paper on the simplest possible case of a pure *spatial autoregressive model* involving only a single variable. While

this model is generally of interest only as one component of a larger model, in the present context it has the advantage of exhibiting all of the key difficulties above while removing many extraneous factors. This model is formalized in Section 2 below. In addition, both the small sample and large sample properties of spatial autocorrelation estimates are developed for selected examples. In Section 3 the nature of this aggregation bias problem is explored in more depth. Here it is shown by means of a small example that this bias is at least in part due to the variance minimizing tendency of maximum likelihood estimation. In addition, a possible method is proposed for reducing this bias.

2. An Aggregated Spatial Autoregressive Model

Consider the n-vector *spatial autoregressive process*,

$$y = \rho W y + \varepsilon \tag{2.1}$$

with nonnegative (nonzero) $n \times n$ *weight matrix*, $W = (w_{ij})$, satisfying $w_{ii} = 0$ for all $i = 1, .., n$, together with *influence parameter*, ρ,and $n \times 1$ *disturbance vector*, ε, normally distributed as

$$\varepsilon \sim N(0, \sigma^2 I_n) \tag{2.2}$$

In the analysis to follow, W will be assumed to be either symmetric or the row normalization of a symmetric matrix, both of which are known to possess real eigenvalues. If we let

$$B = I_n - \rho W \tag{2.3}$$

and denote the minimum and maximum eigenvalues of W by $\lambda_{\min}$ and $\lambda_{\max}$, respectively, then it is well known that the inverse matrix

$$D = B^{-1} = (I_n - \rho W)^{-1} \tag{2.4}$$

exists for all ρ in the open interval $(1/\lambda_{\min}, 1/\lambda_{\max})$, and allows y in (2.1) to be expressed in terms of ε as[1]

$$y = D\varepsilon \tag{2.5}$$

In this context, it is assumed that the vector y is *not* directly observable. Only an aggregated form of y is observable, here denoted by

$$x = Ay \tag{2.6}$$

[1]For notational simplicity we have suppressed the dependency of both B and D on ρ.

where A is a nonnegative $k \times n$ *aggregation matrix* ($1 \leq k < n$) satisfying the 'partition' condition that for each $j = 1, .., n$, $a_{ij} > 0$ for exactly one $i = 1, .., k$, *i.e.*, that each *subregion* (micro zone), j, belongs to exactly one *region* (macro zone), i. In the introductory example, if each component y_j were to represent the average housing value in block j, and a_{ij} were to represent the fraction of housing units of block group i belonging to block j, then each component, $x_i = \sum_j a_{ij} y_j$, would represent the average housing value in block group i.[2]

By combining (2.5) and (2.6) we obtain the following *aggregate model* corresponding to the *disaggregate model* in (2.5):

$$x = AD\varepsilon \tag{2.7}$$

To analyze this model, note first that by definition the rows of the aggregation matrix A are *orthogonal* (since the partition condition above implies that each column j has exactly one nonzero coefficient a_{ij}). Hence A is of full row rank, which together with the nonsingularity of D implies (from standard matrix results) that AD is of full row rank, and that $ADD'A'$ is nonsingular. It thus follows at once from (2.7) that the aggregate data x is normally distributed as $N(0, \sigma^2 ADD'A')$, with k-dimensional density:

$$\phi_k(x; \rho, \sigma^2) = (2\pi\sigma^2)^{-k/2} |ADD'A'|^{-1/2} \exp\{-\frac{1}{2\sigma^2} x'(ADD'A')^{-1}x\} \tag{2.8}$$

So even though y is not directly observable, it is still possible to obtain *maximum-likelihood estimates* of ρ and σ^2 in this aggregate model by employing the log likelihood function,

$$L_k(\rho, \sigma^2; x) = const. - \frac{k}{2} \ln\left(\sigma^2\right) - \frac{1}{2} \ln |ADD'A'| - \frac{1}{2\sigma^2} x'(ADD'A')^{-1}x \tag{2.9}$$

2.1. Estimation of ρ

Since our main concern here is the estimation of ρ, it is convenient to solve for the *mle* of σ^2 as

$$\widehat{\sigma}^2 = \frac{1}{k} x'(ADD'A')^{-1}x \tag{2.10}$$

[2] It should be noted that this interpretation is for convenience only, and is not meant to imply that a spatial autoregressive process in (2.1) should represent even a first approximation to the distribution of housing values.

and substitute (2.10) in (2.9) to obtain the *concentrated likelihood* of ρ,

$$l_k(\rho; x) = -\frac{1}{2}\ln\left(x'(ADD'A')^{-1}x\right) - \frac{1}{2k}\ln|ADD'A'| \tag{2.11}$$

where for convenience we have multiplied through by $1/k$ and supressed the constant, $-(1+\ln k)/2$. To maximize this function, we begin with the first derivative,

$$\frac{\partial}{\partial\rho}l_k = -\frac{1}{2}\frac{x'\left[(\partial/\partial\rho)\,(ADD'A')^{-1}\right]x}{x'(ADD'A')^{-1}x} - \frac{1}{2k}\frac{\partial}{\partial\rho}\ln|ADD'A'| \tag{2.12}$$

where

$$\begin{aligned}\frac{\partial}{\partial\rho}(ADD'A')^{-1} &= -(ADD'A')^{-1}A\left[\frac{\partial}{\partial\rho}DD'\right]A'(ADD'A')^{-1} \\ &= -(ADD'A')^{-1}AD\left[WD + D'W'\right]D'A'(ADD'A')^{-1}\end{aligned} \tag{2.13}$$

and

$$\begin{aligned}\frac{\partial}{\partial\rho}\ln|ADDA'| &= tr\left[(ADD'A')^{-1}A\left(\frac{\partial}{\partial\rho}DD'\right)A'\right] \\ &= 2\cdot tr\left[(ADD'A')^{-1}ADWDD'A'\right]\end{aligned} \tag{2.14}$$

By substituting (2.13) and (2.14) into (2.12) and employing the trace identity, $tr(M_1M_2) = tr(M_2M_1)$, one obtains the first-order condition:

$$\begin{aligned}\frac{\partial}{\partial\rho}l_k &= \frac{x'(ADD'A')^{-1}ADWDD'A'(ADD'A')^{-1}x}{x'(ADD'A')^{-1}x} \\ &\quad -\frac{1}{k}tr\left[D'A'(ADD'A')^{-1}ADWD\right] = 0\,.\end{aligned} \tag{2.15}$$

In the case of no aggregation ($A = I_n$) it follows it follows that (2.15) reduces to the simpler condition that

$$\frac{\partial}{\partial\rho}l_k = \frac{y'B'WBy}{y'B'By} - \frac{1}{k}tr(WD) = 0 \tag{2.16}$$

where B is given by (2.3) above. It should be clear from a comparison of (2.15) and (2.16) that the case of aggregation involves a much more complex first-order

condition. In particular, while it is readily shown [see Section 1 of the Appendix] that (2.16) always has a unique (maximal) root, this is not true of (2.15).

This is seen most easily by constructing a simple spatial example, as shown schematically in Figure 1. Here there are $n = 6$ subregions all connected by a simple $\{0, 1\}$-contiguity relation, yielding the symmetric binary weight matrix W shown in the figure.[3] These subregions are aggregated into $k = 3$ regions (where again the values, a_{ij}, of the aggregation matrix A in Figure 1 might reflect the fraction of housing units or population of region i located in subregion j). The resulting aggregate model (2.7) was then simulated using many choices of ρ, and the concentrated likelihood functions were plotted. The two examples shown in Figure 2 are both for $\rho = .5$ and $\sigma = 1$, and illustrate multiple maxima with Figure 2a showing a *positive* global maximum at $\rho = .69$ and Figure 2b showing a *negative* global maximum at $\rho = -.77$.[4]

This example shows that local search procedures (such as gradient methods) can be very misleading in estimating ρ for the present class of models. But since the maximization is only over a one-dimensional bounded interval of possible ρ values, it should be equally clear that global maximization presents no real problem in this case. Hence it will be assumed throughout that global maximization procedures are used. In this context, the real questions of interest relate to the statistical properties of these global maxima.

2.2. Sampling Distributions of ρ

We begin with 1000 simulations of global maximum estimates for simple example above with $\rho = .5$ and $\sigma = 1$. Figure 3a shows the results for the standard disaggregate model (2.5) with no aggregation. Here we see that the distribution is somewhat negatively skewed with sample mean, $\overline{\rho} = .335$, indicating a tendency to underestimate ρ in the standard model. Such biases have been well documented for a wide range of maximum likelihood estimates involving small sample sizes, such as the present case of $n = 6$. However, for the results of the aggregate model (2.7) the situation is decidedly worse, as seen in Figure 3b. Here the sample mean, $\overline{\rho} = .02$, shows a pronounced bias. Much more important however is the erratic multimodal nature of the histogram. The strong mode near $\rho = -1$ is particularly

[3] The symmetric normalization, $diag(W'u)^{-1/2}[W]diag(W'u)^{-1/2}$, of W [with unit vector, $u' = (1, .., 1)$] was employed.in these examples. However, the standard row normalization, $diag(W'u)^{-1}[W]$, of W yields essentially the same results.

[4] The respective y-vectors for Figures 2a and 2b are $(.0043, -0.318, 1.095, -1.874, 0.428, 0.896)$ and $(-0.399, 0.690, 0.816, 0.712, 1.290, 0.669)$.

troublesome. But since there are effectively only three samples here, namely the values of $x = Ay$ for the three aggregate regions, it can be argued that this may be simply an artifact of sample size.

To show that this is not the case, a number of larger models were constructed and simulated as well. The first is a square 10×10 grid of subregions aggregated into 16 regions, as shown in Figure 4a. Here the weight matrix W is again based on simple contiguities (with row normalized values w_{ij} now interpretable as the fraction of shared boundary for subregion i which is contiguous with subregion j). The aggregation matrix, A, in this case gives equal weight to the members of each region i (so that if n_i denotes the number of subregions in i then $a_{ij} = 1/n_i$ for all $j = 1, .., n_i$). The results of 1000 simulations for the disaggregate model (2.5) with $\rho = .4$ and $\sigma = 1$ are shown for this case in Figure 5. Here it is seen that estimates for this model (with $n = 100$) are now behaving very well with a bell-shaped histogram centered at $\overline{\rho} = .39$. However, the results for the aggregate model (2.7) continue to exhibit the same difficulties: a low sample mean, $\overline{\rho} = .097$, and a multimodal histogram with a strong mode near $\rho = -1$.

A final example involving both a larger number of samples and a more realistic regional scheme is the set of blocks and block groups from West Philadephia shown in Figure 4b. In this case W is again based on boundary shares as above, and A is based on the fraction of housing units in each block of a block group. There are 312 blocks within the 43 block groups shown. The results for 1000 simulations of the disaggregate model (2.5) with $\rho = .4$ and $\sigma = 1$ are shown for this case in Figure 6a. Here again that estimates for this model (with $n = 312$) are right on target, with a sample mean $\overline{\rho} = .395$. While the results for the aggregate model (2.7) are somewhat better than the example above, with a mean of $\overline{\rho} = .252$, there continues to be a secondary mode near $\rho = -1$. Hence while there is some clear improvement in this case, there is still a significant fraction of *negative* estimates $\widehat{\rho}$ [more than 22%] even though the degree of positive spatial autocorrelation ($\rho = .4$) is considerable. Moreover, while the effective sample size ($k = 43$) is not overwhelming, it should certainly be adequate to obtain estimates in such a simple two-parameter model.

2.3. Asymptotic Properties of ρ Estimates

Before proceeding to a more detailed investigation of the possible causes of this undesirable behavior, it is of interest to ask whether such behavior persists in the limit. The examples above suggest that this might simply be a case where

relatively large samples are required in order to achieve reasonable sampling distributions of the maximum likelihood estimates. However, the following extension of the example in Figure 1 provides an informative counterexample.

While asymptotic results for spatial models are somewhat more tenuous than for temporal models, it is nontheless possible to imagine that a given spatial pattern can be extended to the infinite plane by some form of expansion scheme [designated as 'increasing-domain asymptotics' by Cressie (1993)]. In this context, Mardia and Marshall (1984) developed a set of 'growth, convergence, and continuity' conditions [based on the more general results of Sweeting (1980)] for both consistency and asymptotic normality of maximum likelihood estimates. In principle, this result can be applied to establish a comparable consistency result for the present aggregated case, under appropriate conditions.[5]. However it is far too general to allow any conclusions to be drawn about the finite-sample behavior of such estimators.[6]

But there is one case in which such analysis is possible. In particular, if we simply *replicate* a given system of regions an arbitrarily large number of times, and treat each replicate as an 'island' then the (block diagonal) covariance structure of this composite system not only satisfies all conditions for convergence, but actually allows the limiting behavior of estimates to be studied. In this case, a set of N replicates can be viewed as a sequence of N independent random sample from the k-dimensional aggregate model in (2.7), so that all classical results for maximum-likelihood estimation in the independence case can be applied. In particular, it follows from the results of Wald (1949) that for a random vector x distributed with density $f(x;\theta)$, if $\widehat{\theta}_N$ denotes any choice of a global maximum of the associated likelihood function for N independent random samples, then the sequence $\left(\widehat{\theta}_N\right)$ converges in probability (and in fact converges almost surely) to the true value of θ. Hence by applying this result to the sequence of N replicates with parameter $\theta = (\rho, \sigma^2)$, we could establish consistency of such estimators for this replicated case.

However, for our present purposes it is much more insightful to employ the 'ex-

[5]Such conditions must for example include the requirement that the average number of subregions per region converge [*i.e.*, that k_n/n have a positive limit in (0,1)], and that the sequence of aggregation matrices (A_n) as well as the weight matrices (W_n) exhibit appropriate 'growth, covergence and continuity' properties.

[6]It is worth noting in particular that the general result of Sweeting (1980) shows only that there is *a* consistent root of the likelihood equations. Hence in cases of multiple local maxima (such as those illustrated below) such general arguments offer no help in picking a consistent root. [See footnote 7 below for further discussion of this point.]

tremum estimator' approach of Amemiya (1985). In particular, Amemiya shows (Theorem 4.1.1) that in our case if any positive monotone transformation of the concentrated likelihood function can be constructed which converges (uniformly in probability) to a nonstochastic function for which the true ρ value yields the unique global maximum, then the sequence of *maximal-root estimators*, $\widehat{\rho}_N$, will converge in probability to ρ. The advantage of this approach is that if the limiting nonstochastic objective function can be computed, then the qualitative nature of this convergence can be examined in some detail.[7] To do so, we begin by observing that if the true values of ρ and σ^2 are denoted respectively by ρ_0 and σ_0^2, and if $D_0 = (I_n - \rho_0 W)^{-1}$, then by (2.7) it follows that $x = AD_0\varepsilon$. Hence for a single replication, the concentrated likelihood, $l_k(\cdot; x)$, in (2.11) is a random function of the form

$$\begin{aligned} l_k(\cdot, \varepsilon) &= -\frac{1}{2}\ln\left(\varepsilon' D_0' A'(ADD'A')^{-1}AD_0\varepsilon\right) - \frac{1}{2k}\ln|ADD'A'| \\ &= -\frac{1}{2}\ln\left[tr\left(D_0'A'(ADD'A')^{-1}AD_0\varepsilon\varepsilon'\right)\right] - \frac{1}{2k}\ln|ADD'A'| \quad (2.17) \end{aligned}$$

Next let us replicate this system N times, so that the $n \times n$ weight matrix W is replaced by the $Nn \times Nn$ block diagonal weight matrix $\langle W\rangle_N$ with diagonal blocks all equal to W (indicating in particular that there are no spatial connections between replicates). The $k \times n$ aggregation matrix A also becomes an $Nk \times Nn$ block diagonal matrix $\langle A\rangle_N$. It can then be easily verified that matrices $D_0'A'(ADD'A')^{-1}AD_0$ and $ADD'A'$ in (2.17) are both replaced by their block diagonal counterparts $\langle D_0'A'(ADD'A')^{-1}AD_0\rangle_N$ and $\langle ADD'A'\rangle_N$. If for N independent samples $\varepsilon_1, .., \varepsilon_N$ from $N(0, \sigma_0^2 I_n)$ we denote the $Nn \times 1$ stacked vector by $\varepsilon_{(N)} = (\varepsilon_1', .., \varepsilon_N')'$, then the resulting concentrated likelihood function takes the form

$$\begin{aligned} l_{Nk}(\cdot, \varepsilon_{(N)}) &= -\frac{1}{2}\ln\left[tr\left(\langle D_0'A'(ADD'A')^{-1}AD_0\rangle_N \varepsilon_{(N)}\varepsilon_{(N)}'\right)\right] - \frac{1}{2k}\ln|\langle ADD'A'\rangle_N| \\ &= -\frac{1}{2}\ln\left[\sum_{i=1}^N tr\left(D_0'A'(ADD'A')^{-1}AD_0\varepsilon_i\varepsilon_i'\right)\right] - \frac{1}{2Nk}N\ln|ADD'A'| \\ &= -\frac{1}{2}\ln\left[tr\left(D_0'A'(ADD'A')^{-1}AD_0\left[\sum_{i=1}^N \varepsilon_i\varepsilon_i'\right]\right)\right] - \frac{1}{2k}\ln|ADD'A'| \end{aligned} \quad (2.18)$$

[7] An additional advantage of this approach is that it permits consistent estimation in the case of multiple local maxima. For as long as the global maximum is unique (generally a reasonable assumption), one can be assured that a global search of the parameter space will yield the *consistent* estimator.

Hence if we define the new function, $z_N(\cdot, \varepsilon_{(N)})$ by

$$\begin{aligned} z_N(\rho, \varepsilon_{(N)}) &= 2 \cdot l_{Nk}(\rho, \varepsilon_{(N)}) + \frac{1}{2} \ln \left(\frac{N}{\sigma_0^2} \right) \\ &= -\ln \left[\frac{1}{\sigma_0^2} tr \left(D_0' A' (ADD'A')^{-1} AD_0 \left[\frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \varepsilon_i' \right] \right) \right] \\ &\quad -\frac{1}{k} \ln |ADD'A'| \end{aligned} \tag{2.19}$$

then it follows from the first line of (2.19) that $z_N(\cdot, \varepsilon_{(N)})$ is a positive monotone transformation of $l_{Nk}(\cdot, \varepsilon_{(N)})$. Moreover, since it is well known that

$$\frac{1}{N} \sum_{i=1}^{N} \varepsilon_i \varepsilon_i' \underset{prob}{\rightarrow} \sigma_0^2 I_n \tag{2.20}$$

it also follows that $z_N(\cdot, \varepsilon_{(N)})$ converges in probability to the nonstochastic function, $z(\cdot)$, defined by

$$z(\rho) = -\ln \left[tr \left(D_0' A' (ADD'A')^{-1} AD_0 \right) \right] - \frac{1}{k} \ln |ADD'A'| \tag{2.21}$$

As is shown in Section 2 of the Appendix, $z(\cdot)$ is the desired limiting function.

It thus remains to examine the properties of this limiting function in specific cases. Here it is instructive to consider once again the simple example in Figure 1. In this case, the limiting function $z(\cdot)$ has the bimodal form shown in Figure 7a. More importantly, this example shows that while the global maximum is indeed at the true value $\rho_0 = .5$, the secondary mode is *strongly negative*, $\rho_{00} = -.71$, and has a z-value very close to the maximum value [$z(\rho_{00}) = -1.0954 \approx -1.0782 = z(\rho_0)$]. Hence it should be clear in this case that even for large numbers of replications, N, the maximal-root estimate, $\widehat{\rho}_N$, may well be close to ρ_{00} rather than ρ_0. This is illustrated by the histogram of Figure 7b, which shows the estimation results for 1000 simulated samples of the replicated process with $N = 100$. As expected from the general consistency result above, at sample sizes this large ($k = 300$) the majority of maximal-root estimates $\widehat{\rho}_N$ cluster around ρ_0. However, a substantial fraction (more that 12%) still cluster around the secondary mode ρ_{00}. Hence, even at these large sample sizes, the values $z(\rho_0)$ and $z(\rho_{00})$ are sufficiently close to ensure that the global maximum of the concentrated likelihood still has a significant chance of occurring near ρ_{00} rather than ρ_0.[8]

[8] Simulations of larger replication numbers show that by $N = 1000$ ($k = 3000$) the second

It should also be noted that an additional consequence of this bimodal behavior is the failure of standard significance tests based on asymptotic normality. So if one is unlucky enough to come up with strongly negative estimates of ρ in such situations where positive autocorrelation is expected, then standard significance tests will only serve to reinforce these negative findings.[9]

3. Analysis of Aggregation Bias

The negative results above raise the obvious questions: What is going on? What can we do about it? While there appear to be no definitive answers to these questions, some insight can be gained by studying the behavior of the log-likelihood function in (2.9). Since our main focus is on ρ, it is of interest to consider the asymptotic behavior of (2.9) with respect to ρ as the variance parameter, σ^2, becomes large. To do so, observe that if for any aggregate data value, x, and distinct ρ-values, ρ_1 and ρ_2, we let $D_i = (I_n - \rho_i W)^{-1}$, $i = 1, 2$, then

$$\begin{aligned} L_k(\rho_1, \sigma^2; x) - L_k(\rho_2, \sigma^2; x) &= -\frac{1}{2}\ln|AD_1D_1'A'| + \frac{1}{2}\ln|AD_2D_2'A'| \\ &\quad -\frac{1}{2\sigma^2}x'\left[(AD_1D_1'A')^{-1} - (AD_2D_2'A')^{-1}\right]x \end{aligned}$$

so that as σ^2 becomes large,

$$\lim_{\sigma^2\to\infty}\left[L_k(\rho_1, \sigma^2; x) - L_k(\rho_2, \sigma^2; x)\right] = -\frac{1}{2}\ln|AD_1D_1'A'| + \frac{1}{2}\ln|AD_2D_2'A'|$$

Hence we may conclude that the limiting form of L_k is a positive monotone function of the negative log determinant,

$$L(\rho) = -\ln|ADD'A'| \tag{3.1}$$

mode has finally disappeared, so that classical consistency and asymptotic normality properties are in full force. However, it should be emphasized that this case was chosen mainly for its simplicity. A local search for 'worst cases' in the neighborhood of this example produced a case in which $|z(\rho_0) - z(\rho_{00})|$ was so small that even for $N = 1000$ the fraction of 1000 samples in the ρ_{00}-cluster exceeded 40%.

[9] This was verified in the present case by calculating the asymptotic covariance matrix and constructing (Wald) significance tests for the simulated estimates. Not surprisingly, for the given sample size of $k = 300$ all negative estimates near ρ_{00} were shown to be highly significant. These results are not reported here.

which may be viewed as the *asymptotic log-likelihood of ρ under infinite dispersion*. Notice also that this asymptotic function is *nonstochastic* and hence can be analyzed independently of any x-data. For later use, we also note that the concentrated likelihood function in (2.11) can be written in terms of L as

$$l_k(\rho; x) = -\frac{1}{2} \ln \left(x'(ADD'A')^{-1}x \right) + \frac{1}{2k} L(\rho) \tag{3.2}$$

Given these observations, we next ask: how should this as asymptotic function behave? An examination of (2.1) suggests that as the dispersion of ε becomes arbitrarily large, any autocorrelation effects among the y value should eventually be overwhelmed. This conjecture is confirmed by analyzing L for the disaggregate model with $A = I_n$. In this case L reduces to

$$L(\rho) = -\ln |DD'| = -\ln \left(|D|^2 \right) = const. - \ln |D| \tag{3.3}$$

which is well known to be strictly concave.[10] Hence the unique maximum of L is given by the first-order condition in (2.14) [with $A = I_n$] as

$$0 = \frac{\partial L}{\partial \rho} = -2 \cdot tr \left[(DD')^{-1} DWDD' \right] = -2 \cdot tr(WD) \tag{3.4}$$

But at $\rho = 0$ we see that $D = I_n \;\Rightarrow\; tr(WD) = tr(W) = 0$,[11] and hence may conclude that the unique solution to (3.4) is given by $\rho = 0$. Thus for the disaggregate model we obtain the intuitively satisfying result that as dispersion of ε becomes arbitrarily large, the most likely value of ρ converges to zero.

However, for the aggregate model this is not the case. While a full analysis of this problem appears to be quite difficult, some insight can be gained by examining the derivative of L at $\rho = 0$ for the dissagregate model. Here again we see from (2.14) [with $D = I_n$] that

$$\left. \frac{\partial L}{\partial \rho} \right|_{\rho=0} = -2 \cdot tr \left[(AA')^{-1} AWA' \right] \tag{3.5}$$

But since A is nonnegative with orthogonal rows, it follows that AA' is a positive diagonal matrix, and hence that $(AA')^{-1}$ is positive diagonal. This together

[10] In section (3.1) of the Appendix it is shown that $\partial^2 L / \partial \rho^2 = -k \cdot \sum_i \omega_i^2 < 0$, where the ω_i's are the (real) eigenvalues of WD.

[11] Recall that $w_{ii} \equiv 0$ by assumption.

with the nonnegativity of W implies that $tr\,[(AA')^{-1}AWA'] \geq 0$, and hence that $(\partial L/\partial\rho)_{\rho=0} \leq 0$. Moreover (barring exceptional cases) this derivative is *strictly negative.* Hence there must always be at least a local maximum of L at *negative* values of ρ. While sharper results here are somewhat elusive, it can be shown [Section 3 of the Appendix] that if W is *symmetric* then L is strictly concave, and hence that the unique global maximizer of L is *always negative.* For *row normalizations* of symmetric weight matrices, it also appears that all maxima are achieved at negative ρ values.[12]

What this means from a practical viewpoint is that at least for highly dispersed aggregate models, the asympotically most likely values of ρ tend to be *negative.* This leads naturally to the next question of why this should be true. Here again there seems to be no completely satisfactory answer. However, some insight can be gained by looking at the simplest possible case.

3.1. The Case of One Region and Two Subregions

At first glance the case of single region ($k = 1$) divided into two subregions ($n = 2$) would appear to be degenerate from a spatial viewpoint. With only one region in the aggregate model, there can be no *observable* spatial interaction. However, the model is still influenced by the *unobservable* interaction between subregions. Hence this case serves to emphasize the effects of this unobservable spatial component. One additional advantage here is that the regional covariance matrix reduces to simple variance, which is more readily interpretable.

To formalize this case, let the weight matrix W be given by

$$W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{3.6}$$

so that B and D have the respective forms

$$B = \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}, \qquad D = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{3.7}$$

Next let the aggregation matrix be an unspecified positive linear combination $A = a' = (a_1, a_2)$, so that for normally distributed $\varepsilon = (\varepsilon_1, \varepsilon_2)'$ in (2.2), the single

[12] Extensive studies of numerical examples for such matrices show that L need not be concave, but that the derivative $\partial L/\partial\rho$ is always negative for $\rho > 0$. It thus seems reasonable to conjecture in this case that all maxima continue to occur at negative values of ρ.

aggregate regional variate x has the form

$$x = a'D\varepsilon = \frac{1}{1-\rho^2}\left[(a_1 + \rho a_2)\varepsilon_1 + (a_1\rho + a_2)\varepsilon_2\right] \tag{3.8}$$

with corresponding variance given by

$$\begin{aligned} var(x) &= \sigma^2 a'DD'a \\ &= \sigma^2 \frac{(a_1 + \rho a_2)^2 + (a_1\rho + a_2)^2}{(1-\rho^2)^2} \end{aligned} \tag{3.9}$$

Note that influence of ρ on the variance of x is summarized by the value, $a'DD'a$, which may here be designated as the *relative variance* of x (since it defines variance up to a scalar multiple). With this terminology it is clear that the asymptotic log-likelihood $L(\rho)$ in (3.1) depends entirely on this relative variance:

$$\begin{aligned} L(\rho) &= -\ln\left[a'DD'a\right] \\ &= const. - \ln\left[\frac{1}{(1-\rho^2)^2}\right] - \ln\left[(a_1 + \rho a_2)^2 + (a_1\rho + a_2)^2\right] \end{aligned} \tag{3.10}$$

Hence the value of ρ which is asymptotically most likely (as dispersion of ε becomes large) is precisely that value that *minimizes the relative variance of* x. More generally, if the determinant $|ADD'A'|$ in (2.8) is designated as the *relative generalized variance* of x,[13] then this same interpretation holds for (3.1) as well.

In this context, it should be clear from (3.9) that determining the minimizing value of ρ is complex even in this simple case. But here one can gain qualitative insight by observing from (3.3) that for the *disaggregate model*, the asymptotic log-likelihood function L becomes

$$\begin{aligned} L(\rho) &= const. - \ln|D| \\ &= const. - \ln\left[\frac{1}{(1-\rho^2)^2}\right] \end{aligned} \tag{3.11}$$

Hence this disaggregate likelihood is seen to be essentially the first term of (3.10), which corresponds to the denominator of relative variance in (3.9). It is thus reasonable in this case to interpret the denominator of relative variance (first

[13] This terminology follows the *generalized variance* interpretation of covariance matrix determinants first introduced by Wilks (1932).

term of L) as the 'disaggregate' effect of ρ on relative variance, arising from interaction between the individual subregions. The numerator of relative variance (second term of L) then represents the additional 'aggregate' effect of ρ arising from regional aggregation. Moreover, since the terms $(a_1 + \rho a_2)^2$ and $(a_1\rho + a_2)^2$ in (3.9) are essentially the contributions to relative variance of the two aggregate variates $(a_1 + \rho a_2)\varepsilon_1$ and $(a_1\rho + a_2)\varepsilon_2$ in (3.8), it is appropriate to designate this numerator as *aggregation variance*,

$$v(\rho) = (a_1 + \rho a_2)^2 + (a_1\rho + a_2)^2 \tag{3.12}$$

By observing that the denominator of (3.9) achieves its maximum at $\rho = 0$ [yielding minimum relative variance for the disaggregate case], it is clear that the key effect of aggregation on relative variance is in terms of $v(\rho)$. Moreover by differentiating $v(\rho)$, this aggregation variance is seen to be minimized at

$$\rho_v = -\frac{2a_1a_2}{a_1^2 + a_2^2} < 0 \tag{3.13}$$

This implies that the value of ρ minimizing relative variance in (3.9) must lie between the extremes, $\rho = \rho_v$ and $\rho = 0$, and hence must be *negative*.

Thus in this simple case, the effect of aggregation on minimum relative variance is clear: the positivity of a_1 and a_2 imply that any positive correlation effect, ρ, in (3.12) must necessarily *increase* aggregation variance, and hence that relative variance can only be minimized at *negative* ρ values. This is of course completely analogous to standard 'variance reduction' techniques in sample design, where one reduces the variance of sums by creating negatively correlated samples. In the present case, the relevant 'sums' are simply the aggregated data. While this interpretation is less clear in the general case, where variance must be replaced by 'generalized variance' (as in foonote 13 above), it nonetheless appears that this type of variance reduction effect is at the root of the aggregation-bias problem in the present case. The maximum-likelihood values of ρ tend to be those which yield smaller relative (generalized) variances, and in the presence of aggregation, these values tend to be negative.

3.2. A Possible Bias-Correction Procedure

While there appears to be no quick fix for this type of aggregation bias, the above observations suggest at least one possible approach. For if one focuses on the asymptotic case when the dispersion of ε becomes large, then it is reasonable

to require that (as in the bias-free disaggregate case) the aggregate maximum-likelihood estimate of ρ approach zero. As observed above, this is equivalent to requiring that relative generalized variance, $|ADD'A'|$, achieve its minimum at $\rho = 0$. There are many ways to implement such a modification. For example, by simply subtracting the derivative, $(\partial L/\partial\rho)_{\rho=0}$, of the asymptotic log-likelihood at zero [expression (3.5)] from $L_k(\rho, \sigma^2; x)$, one produces a modified log-likelihood function

$$\begin{aligned} L_k^0(\rho, \sigma^2; x) &= -\frac{k}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}x'(ADD'A')^{-1}x \\ &\quad -\frac{1}{2}\ln|ADD'A'| + tr\left[(AA')^{-1}AWA'\right] \end{aligned} \tag{3.14}$$

for which the associated asymptotic log-likelihood must have zero derivative at $\rho = 0$.

However, after experimenting with numerous variations on this theme, the most successful approach found is simply to rescale ρ in the asymptotic log-likelihood, $L(\rho)$, in a manner which achieves the same effect. In particular, if the boundary values of ρ are denoted by $\rho_{\min} = 1/\lambda_{\min}$ and $\rho_{\max} = 1/\lambda_{\max}$, and if ρ^* denotes the point in $(\rho_{\min}, \rho_{\max})$ where L achieves its maximum, then the appropriate rescaling of ρ amounts to a piecewise linear transformation, $\rho \to \widetilde{\rho}$ which moves ρ^* to the origin while leaving the boundaries fixed, *i.e.*,

$$\widetilde{\rho} = \begin{cases} \rho + (1 - \frac{\rho}{\rho_{\min}}) \cdot \rho^* & , \rho_{\min} < \rho \leq 0 \\ \rho + (1 - \frac{\rho}{\rho_{\max}}) \cdot \rho^* & , 0 < \rho < \rho_{\max} \end{cases} \tag{3.15}$$

If we denote the resulting transformation of L by $\widetilde{L}(\rho) = L(\widetilde{\rho})$, then the corresponding modification of the concentrated likelihood function in (3.2) now given by

$$\widetilde{l}_k(\rho; x) = -\frac{1}{2}\ln\left(x'(ADD'A')^{-1}x\right) + \frac{1}{2k}\widetilde{L}(\rho) \tag{3.16}$$

In the present case, the relevant values ($\rho_{\min} = -1.16, \rho^* = -.226$, $\rho_{\max} = 1$) yield the tranformed asymptotic log-likelihood, $\widetilde{L}$, shown in Figure 8a, where the dotted curve represents the corrected values (achieving a maximum at $\rho = 0$).

The limiting likelihood function, $\widetilde{z}$, under this rescaling is shown in Figure 8b. Notice first that while there are still two modes, the secondary mode in the negative range of ρ has now diminished markedly. As a consequence, the histogram of 1000 simulated estimates of ρ shown in Figure 9a no longer exhibits

a second mode, yielding a dramatic improvement over Figure 7a. Figure 9b shows a similar result for the example in Figure 5 above.

However, one should hasten to add that this correction is meaningful only for 'small' samples, and is surely not consistent. In the present case, the primary mode is at $\rho_1 = .52 > .5 = \rho_0$, so that estimates will eventually exhibit a small upward bias. The is already seen in Figure 9a, where the sample mean is .513. Moreover, while this bias is small in the present case, this need not be true in general. In the present case, the table below shows the ρ_1 values calculated for a selection of nonnegative ρ_0 values. As ρ_0 approaches zero, bias clearly increases (and is even worse for negative ρ_0).

ρ_0	0	.1	.2	.3	.4	.5	.6	.7	.8
ρ_1	.278	.299	.331	.378	.440	.517	.605	.700	.800

Moreover, while the higher values look promising, the secondary mode in these cases is more severe (indicating simply that corrections of the likelihood function are relatively slight for large ρ_0). So if ρ_0 is close to zero in the present example (with $\sigma = 1$), then this correction procedure will reject the null hypothesis much too often.

Hence a better correction procedure (assuming $\rho_0 \geq 0$) might involve some compromise between l_k and $\widetilde{l}_k$. In particular, for any convex combination of these functions, $\alpha_k\ \widetilde{l}_k + (1 - \alpha_k) l_k$, with $\lim_{k\to\infty} \alpha_k = 0$, it is clear that the maximum-root estimate of ρ must necessarily be consistent. The task here is to find a choice of α_k's which avoids the serious bimodality properties of l_k while at the same time minimizing the positive bias introduced by $\widetilde{l}_k$. Such possibilities will be considered in subsequent work.

4. Concluding Remarks

The central task of this paper has been to illuminate the difficulties of estimating parameters in the presence of 'misalligned' data. Here we have focused on the simple case of a pure spatial autoregressive process in which the observable data is at a higher level of aggregation than the actual process itself, as represented by the weight-matrix data. In this setting, the resulting sampling distributions of maximum likelihood estimates for selected examples were shown to exhibit serious bias (even for large sample sizes), and a possible correction procedure was

proposed. This procedure is admittedly *ad hoc*, and clearly represents only a first pass at the problem.

However, it can be argued that in the simple context of spatial autoregressive processes, there is perhaps no need for such correction procedures at all. Given model [(2.1),(2.2),(2.6)] for a specific weight matrix, W, and aggregation matrix, A, one can in principle simulate the sampling distribution of ρ estimates for a selected range of ρ_0 values and simply observe how they behave. For example, in the case illustrated by Figure 1 above, the sampling distribution (for 1000 simulations) under null hypothesis, $\rho_0 = 0$, is shown in Figure 10. Hence it is possible to test the null hypothesis of 'no spatial interaction' directly in terms of this sampling distribution. The fact that the sample mean is significantly negative (about $-.10$ in this case) is of no consequence. One can still test whether a given estimate is 'significantly big' with respect to this distribution, and hence test for the presence of spatial autocorrelation even though the estimates are themselves very biased.[14]

But in the more important cases of multiple regression models with spatial autoregressive errors or spatial lags, the situation is far more complex. Here the estimate of ρ is mainly of interest as an intermediate step in estimating and testing the significance of the key β parameters. Hence the *value* of the ρ estimate is crucial for estimating the primary β parameters. An even more fundamental issue in these models is whether consistent estimation of parameters is even possible. In the case of spatial lag models for example, even if explanatory variables are observed at the same level of aggregation as the dependent variable, there exists no simple reduced form such as (2.7) above. Hence the possibility of consistent estimation in such cases is questionable. In these more complex models it may thus only be meaningful to introduce micro spatial lags in combination with other micro data that allow the possibility of consistent estimation [such as the 'auxilliary variable' techniques of Holt, Steel, Tranmer and Wrigley (1996) and others].[15]

[14] It is also of interest to note in the present case that the behavior of these sampling distributions is well predicted by the corresponding asymptotic form of the concentrated likelihood function (2.21), as shown in Figure 7. This is also seen in Figure 10 where the associated asymptotic function is plotted above the histogram. Notice in particular, that at $\rho_0 = 0$, the secondary mode has now merged with the primary mode, to produce a single 'flat' mode extending significantly into the negative range of ρ. Again this form is roughly mirrored by the sampling distribution below. Hence one can (very quickly) plot these asymptotic functions for a selected range of ρ_0 values, and see how the estimates behave.

[15] Some initial results along these lines using Best Linear Unbiased Prediction methods have been obtained by James LeSage (personal communication).

Alternatively, one might consider the general Bayesian approach to misaligned data recently proposed by Zhu, Gelfand, and Carlin (2000) in which micro spatial lags are treated simply as another type of data misalignment. Such possibilities will be explored in subsequent work.

References

[1] Amemiya, T. (1985) *Advanced Econometrics*, Cambridge, Massachusetts: Harvard University Press.

[2] Arbia, G. (1989) *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*, Kluwer, Dordrecht.

[3] Cressie, N.A. (1993) *Statistics for Spatial Data*, New York: Wiley

[4] Holt, D., D.G. Steel, M. Tranmer and N. Wrigley (1996) Aggregation and ecological effects in geographically based data, *Geographical Analysis*, 28:244-262.

[5] Mardia, K.V. and R.J. Marshall (1984) Maximum likelihood estimation of models for residual covariance in spatial regression, *Biometrika*, 71:135-146.

[6] Ripley, B. (1988) *Statistical Inference for Spatial Processes*, Cambridge: Cambridge University Press.

[7] Wald, A., (1949) Note on the consistency of the maximum likelihood estimate, *Annals of Mathematical Statistics*, 20:595-601.

[8] Wilks, S.S., (1932) Certain generalizations in the analysis of variance, *Biometrika*, 24:471-494.

[9] Zhu, L., Gelfand, A.E., and Carlin B.P. (2000) Hierarchical regression with misaligned spatial data: Relating ambient ozone and pediatric asthma ER visits in Atlanta." Research Report 2001-010, Division of Biostatistics, University of Minnesota, 2000. Submitted to Biometrics.

5. Appendix

In this Appendix, a number of the results in the text are proved.

5.1. Uniqueness of MLEs for the Disaggregate Case

In view of the central role plays by multiple roots of the likelihood function in aggregate case, it is appropriate to document uniqueness of roots for the standard disaggregate case [see also Ripley (1988, section 2.2)]. To do so, it suffices to show that the second derivative of l_k with respect to ρ is always negative whenever the first derivative is zero. We begin by differentiating (2.16) to obtain

$$\begin{aligned} \frac{\partial^2}{\partial \rho^2} l_k &= 2\left(\frac{y'B'Wy}{y'B'By}\right)^2 - \frac{y'W'Wy}{y'B'By} - \frac{1}{k}tr(WDWD) \\ &= [A_1] + [A_2] \end{aligned} \tag{A.1}$$

where

$$A_1 = \left(\frac{y'B'Wy}{y'B'By}\right)^2 - \frac{y'W'Wy}{y'B'By} \tag{A.2}$$

and

$$A_2 = \left(\frac{y'B'Wy}{y'B'By}\right)^2 - \frac{1}{k}tr(WDWD) \tag{A.3}$$

Turning first to A_1, observe that if $x = By$ and $w = Wy$ then

$$\begin{aligned} A_1 &= \frac{(x'w)^2}{\|x\|^4} - \frac{\|w\|^2}{\|x\|^2} \\ &= \frac{\|w\|^2}{\|x\|^2}\left(\frac{(x'w)^2}{\|x\|^2\|w\|^2} - 1\right) \leq 0 \end{aligned} \tag{A.4}$$

by the Cauchy-Schwartz Inequality. Moreover, if $\partial l_k/\partial\rho = 0$, then by (2.16) it follows that

$$\frac{y'B'Wy}{y'B'By} = \frac{1}{k}tr(WD) \tag{A.5}$$

and hence from (A.3) that

$$\begin{aligned} A_2 &= \left(\frac{1}{k}tr(WD)\right)^2 - \frac{1}{k}tr(WDWD) \\ &= \frac{1}{k^2}\left[tr(WD)^2 - k \cdot tr(WDWD)\right] \end{aligned} \tag{A.6}$$

But since the spectrum, $\sigma(W) = (\lambda_1, .., \lambda_n)$, of W is real, and since the corresponding spectrum of WD, say $\sigma(WD) = (\omega_1, .., \omega_n)$, is easily seen to be given

by

$$\omega_i = \frac{\lambda_i}{1 - \rho\lambda_i}\,,\ i = 1, .., n$$

it follows that $\lambda(WD)$ is also real. Hence, observing by definition that $\lambda(WDWD) = (\omega_1^2, .., \omega_n^2)$, and recalling the trace of a matrix is the sum of its spectrum, it follows (from an application of Hölder's Inequality) that

$$A_2 = \frac{1}{k^2}\left[\left(\sum\nolimits_i \omega_i\right)^2 - k \cdot \sum\nolimits_i \omega_i^2\right] \leq 0 \tag{A.7}$$

with strict inquality holding unless $\omega_1 = \cdots = \omega_n$. To see that the latter case is not possible, observe that since our condition, $\rho \in (1/\lambda_{\min}, 1/\lambda_{\max})$, implies that $1 - \rho\lambda_i > 0$ for all $i = 1, .., n$, it follows that each ω_i has the same sign as λ_i. Finally since $\lambda_{\max} > 0$ for every nonnegative (nonzero) matrix and since $w_{ii} \equiv 0$ implies that $0 = tr(W) = \sum_i \lambda_i$, we must also have $\lambda_{\min} < 0$, and may conclude that $\omega_{\min} < 0 < \omega_{\max}$.

5.2. Limiting Objective Function

To show verify that $z(\cdot)$ is the desired limit function, it must be shown that (i) the sequence of random functions $\left(z_N(\rho, \varepsilon_{(N)})\right)$ in (2.19) exhibit appropriate uniform probabilistic convergence to $z(\cdot)$, and that (ii) the unique global maximum of this function is at $\rho = \rho_0$. To establish (i), we begin by observing that while $z(\cdot)$ diverges at the boundaries of the parameter space $(1/\lambda_{\min}, 1/\lambda_{\max})$, it is easily seen to be uniformly continuous on every closed interval $\Gamma \subset (1/\lambda_{\min}, 1/\lambda_{\max})$. Hence it is enough to establish uniform probabilistic convergence on every closed interval $\Gamma \subset (1/\lambda_{\min}, 1/\lambda_{\max})$ containing ρ_0. To do so, let the random function $h(\cdot, \varepsilon)$ be defined on each Γ by

$$h(\rho, \varepsilon) = tr\left[D_0' A'(ADD'A')^{-1} AD_0 \varepsilon\varepsilon'\right] \tag{A.8}$$

and observe that

$$\begin{aligned} E\left[h(\rho, \varepsilon)\right] &= tr\left[D_0' A'(ADD'A')^{-1} AD_0 E(\varepsilon\varepsilon')\right] \\ &= tr\left[D_0' A'(ADD'A')^{-1} AD_0 \left(\sigma_0^2 I_n\right)\right] \\ &= \sigma_0^2 \cdot tr\left[D_0' A'(ADD'A')^{-1} AD_0\right] \end{aligned} \tag{A.9}$$

Hence letting the nonstochastic mean-value function $h(\cdot)$ be defined by the right hand side, *i.e.*, by

$$h(\rho) = \sigma_0^2 \cdot tr\left[D_0' A'(ADD'A')^{-1} AD_0\right] \tag{A.10}$$

it follows easily that the difference function

$$g(\rho, \varepsilon) = h(\rho, \varepsilon) - h(\rho) \tag{A.11}$$

has zero mean for all $\rho \in \Gamma$, and is continuous in both ρ and ε. In addition if we write the matrix $D_0' A'(ADD'A')^{-1} AD_0$ as $M(\rho) = [m_{ij}(\rho)]$, then

$$\begin{aligned} h(\rho, \varepsilon) &= tr[M(\rho)\varepsilon\varepsilon'] = tr[\varepsilon' M(\rho)\varepsilon] = \textstyle\sum_{ij} \varepsilon_i m_{ij}(\rho)\varepsilon_j \\ &\Rightarrow \quad |h(\rho, \varepsilon)| \leq \left|\textstyle\sum_{ij} \varepsilon_i m_{ij}(\rho)\varepsilon_j\right| \leq \textstyle\sum_{ij} |\varepsilon_i \varepsilon_j| \cdot |m_{ij}(\rho)| \\ &\Rightarrow \quad \sup_{\rho\in\Gamma} |h(\rho, \varepsilon)| \leq \textstyle\sum_{ij} |\varepsilon_i \varepsilon_j| \cdot \sup_{\rho\in\Gamma} |m_{ij}(\rho)| = \textstyle\sum_{ij} |\varepsilon_i \varepsilon_j| \cdot m_{ij} \end{aligned} \tag{A.12}$$

where $m_{ij} = \sup_{\rho\in\Gamma} |m_{ij}(\rho)| < \infty$ for all $i, j = 1, .., n$. Hence the finiteness of $E(|\varepsilon_i \varepsilon_j|)$ for all $i, j = 1, .., n$ implies that

$$E\left(\sup_{\rho\in\Gamma} |h(\rho, \varepsilon)|\right) = \textstyle\sum_{ij} m_{ij} E(|\varepsilon_i \varepsilon_j|) < \infty \tag{A.13}$$

and the function g is seen to satisfy all conditions of Theorem 4.2.1 in Amemiya (1985) for the *iid* sequence of random vectors (ε_i). Thus $\frac{1}{N}\sum_{i=1}^N g(\cdot, \varepsilon_i)$ converges uniformly in probability to zero on Γ, implying from (A.11) that $\frac{1}{N}\sum_{i=1}^N h(\cdot, \varepsilon_i)$ converges uniformly in probability to $h(\cdot)$ on Γ. Finally since (2.19) shows that $z_N(\cdot, \varepsilon_{(N)})$ can written as

$$z_N(\cdot, \varepsilon_{(N)}) = -\ln\left[\frac{1}{\sigma_0^2}\left\{\frac{1}{N}\textstyle\sum_{i=1}^N h(\cdot, \varepsilon_i)\right\}\right] - \frac{1}{k}\ln|ADD'A'| \tag{A.14}$$

and hence is uniformly continuous in both ρ and $\frac{1}{N}\sum_{i=1}^N h(\cdot, \varepsilon_i)$, it follows that uniform probabilistic convergence of $\frac{1}{N}\sum_{i=1}^N h(\cdot, \varepsilon_i)$ to $h(\cdot)$ implies uniform probabilistic convergence of $z_N(\cdot, \varepsilon_{(N)})$ to the function

$$z(\cdot) = -\ln\left[\frac{1}{\sigma_0^2} h(\cdot)\right] - \frac{1}{k}\ln|ADD'A'| \tag{A.15}$$

which is precisely (2.21).

Establishing (ii) is somewhat more delicate [as should be clear from Figure 7a, which shows that other local maxima not only exist, but can also be very close in value to the global maximum]. The following argument starts with the full log-likelihood function in (2.9) and makes use of the fundamental 'information inequality' which asserts that for any parameter pair (ρ, σ^2) distinct from the true values (ρ_0, σ_0^2), if the functions $L_k(\rho, \sigma^2; \cdot)$ and $L_k(\rho_0, \sigma_0^2; \cdot)$ differ on a set of positive probability measure, then

$$E\left[L_k\left(\rho, \sigma^2; x\right)\right] < E\left[L_k\left(\rho_0, \sigma_0^2; x\right)\right] \tag{A.16}$$

where expectation is taken with respect to x under the true parameter values (ρ_0, σ_0^2).[16] In the present case, it is enough to require that for all $(\rho, \sigma^2) \neq (\rho_0, \sigma_0^2)$, the covariance matrices, $\sigma^2 ADD'A'$ and $\sigma_0^2 AD_0 D_0' A'$ be distinct.[17] To evaluate the left hand side of (A.16) we employ the arguments in (2.17) and (A.9) to obtain,

$$\begin{aligned} E\left[L_k(\rho, \sigma^2; x)\right] &= E\left[L_k(\rho, \sigma^2; AD_0\varepsilon)\right] \\ &= -\frac{k}{2}\ln\left(\sigma^2\right) - \frac{1}{2}\ln|ADD'A'| - \frac{1}{2\sigma^2} tr\left[D_0'A'(ADD'A')^{-1}AD_0 E(\varepsilon\varepsilon')\right] \\ &= -\frac{k}{2}\ln\left(\sigma^2\right) - \frac{1}{2}\ln|ADD'A'| - \frac{\sigma_0^2}{2\sigma^2} tr\left[D_0'A'(ADD'A')^{-1}AD_0\right] \end{aligned} \tag{A.17}$$

where the constant term has been dropped for convenience. Next we maximize (A.17) with respect to σ^2 for each fixed value of ρ to obtain the unique value

$$\sigma_\rho^2 = \frac{\sigma_0^2}{k} tr\left[D_0'A'(ADD'A')^{-1}AD_0\right] \tag{A.18}$$

Since (A.16) holds for all $(\rho, \sigma^2) \neq (\rho_0, \sigma_0^2)$, it follows in particular that for each $\rho \neq \rho_0$,

$$E\left[L_k(\rho, \sigma_\rho^2; AD_0\varepsilon\right] < E\left[L_k(\rho_0, \sigma_0^2; AD_0\varepsilon\right] \tag{A.19}$$

[16] It is worth noting that one of the earliest applications of this basic inequality was by Wald (1949) in his proof of consistency of maximal roots for the independent sampling case. Hence its relevance here is not surprising.

[17] This requires that the (highly overdetermined) system of $k(k+1)/2$ distinct polynomial equations in the two unknowns (ρ, σ^2) [implied by the matrix equality $\sigma^2 ADD'A' = \sigma_0^2 AD_0 D_0' A'$] have no solutions in $(1/\lambda_{\min}, 1/\lambda_{\max}) \times R_+$ other than (ρ_0, σ_0^2).

To evaluate the right hand side of (A.19) we next observe that $(\rho, \sigma^2) = (\rho_0, \sigma_0^2)$ implies $D = D_0$, so that

$$\begin{aligned} tr\left[D_0'A'(AD_0D_0'A')^{-1}AD_0\right] &= tr\left[(AD_0D_0'A')^{-1}AD_0D_0'A'\right] \\ &= tr(I_k) = k \end{aligned} \tag{A.20}$$

Hence for these true values, we have

$$E\left[L_k(\rho_0, \sigma_0^2; AD_0\varepsilon\right] = -\frac{k}{2}\ln\left(\sigma_0^2\right) - \frac{1}{2}\ln|AD_0D_0'A'| - \frac{k}{2} \tag{A.21}$$

and may use (A.17),(A.18),(A.20), and (A.21) to rewrite (A.19) as follows

$$\begin{aligned} &-\frac{k}{2}\ln\left(\frac{\sigma_0^2}{k}tr\left[D_0'A'(ADD'A')^{-1}AD_0\right]\right) - \frac{1}{2}\ln|ADD'A'| - \frac{k}{2} \\ &\quad < -\frac{k}{2}\ln\left(\sigma_0^2\right) - \frac{1}{2}\ln|AD_0D_0'A'| - \frac{k}{2} \\ \\ \Rightarrow\quad &-\frac{k}{2}\ln\left(\sigma_0^2\right) - \frac{k}{2}\ln\left(\frac{1}{k}tr\left[D_0'A'(ADD'A')^{-1}AD_0\right]\right) - \frac{1}{2}\ln|ADD'A'| \\ &\quad < -\frac{k}{2}\ln\left(\sigma_0^2\right) - \frac{1}{2}\ln|AD_0D_0'A'| \\ \\ \Rightarrow\quad &\frac{k}{2}\ln(k) - \frac{k}{2}\ln\left(tr\left[D_0'A'(ADD'A')^{-1}AD_0\right]\right) - \frac{1}{2}\ln|ADD'A'| \\ &\quad < -\frac{1}{2}\ln|AD_0D_0'A'| \\ \\ \Rightarrow\quad &-\frac{k}{2}\ln\left(tr\left[D_0'A'(ADD'A')^{-1}AD_0\right]\right) - \frac{1}{2}\ln|ADD'A'| \\ &\quad < -\frac{k}{2}\ln(k) - \frac{1}{2}\ln|AD_0D_0'A'| \\ \\ \Rightarrow\quad &-\frac{1}{2}\ln\left(tr\left[D_0'A'(ADD'A')^{-1}AD_0\right]\right) - \frac{1}{2k}\ln|ADD'A'| \\ &\quad < -\frac{1}{2}\ln(tr\left[D_0'A'(AD_0D_0'A')^{-1}AD_0\right] - \frac{1}{2k}\ln|AD_0D_0'A'| \\ \\ \Rightarrow\quad &z(\rho) < z(\rho_0) \end{aligned} \tag{A.22}$$

Hence $z(\rho)$ achieves its unique global maximum at $\rho = \rho_0$.

5.3. Concavity of L for Symmetric W

To establish concavity of L for the case of symmetric weight matrices, W, we begin by observing that if the orthogonal projection onto $span(D'A')$ is denoted by

$$\mathbf{P} = D'A'(ADD'A')^{-1}AD \tag{A.23}$$

then it follows at once from (2.14) and (3.1) together with the identity $tr(M_1M_2) = tr(M_2M_1)$ that

$$\begin{aligned} \frac{\partial L}{\partial \rho} &= -2 \cdot tr\left[D'A'(ADD'A')^{-1}ADWD\right] \\ &= -2 \cdot tr\left[\mathbf{P}WD\right] \end{aligned} \tag{A.24}$$

Moreover, by using (2.13) we see that

$$\begin{aligned} \frac{\partial}{\partial \rho}\mathbf{P} &= \left(\frac{\partial}{\partial \rho}D'\right)A'(ADD'A')^{-1}AD \\ &\quad + D'A'\left[\frac{\partial}{\partial \rho}(ADD'A')^{-1}\right]AD \\ &\quad + D'A'(ADD'A')^{-1}A\left(\frac{\partial}{\partial \rho}D\right) \\ &= D'W'\mathbf{P} - \mathbf{P}(WD + D'W')\mathbf{P} + \mathbf{P}WD \\ &= (I_n - \mathbf{P})D'W'\mathbf{P} + \mathbf{P}WD(I_n - \mathbf{P}) \end{aligned} \tag{A.25}$$

Hence the second derivative of L is given by

$$\frac{\partial^2 L}{\partial \rho^2} = -2 \cdot tr\left[\left(\frac{\partial}{\partial \rho}\mathbf{P}\right)WD + \mathbf{P}W\left(\frac{\partial}{\partial \rho}D\right)\right]$$

$$= -2 \cdot tr\left[(I_n - \mathbf{P})D'W'\mathbf{P}WD + \mathbf{P}WD(I_n - \mathbf{P})WD + \mathbf{P}WDWD\right]$$

$$= -2 \cdot \{tr\left[(I_n - \mathbf{P})D'W'\mathbf{P}WD\right] + tr\left[\mathbf{P}WD(I_n - \mathbf{P})WD\right] + tr\left[WD\mathbf{P}WD\right]\} \tag{A.26}$$

To show that this expression is negative for symmetric W, we next observe from (2.3) and (2.4) that by definition, $W = W' \Rightarrow B = B' \Rightarrow D = D'$. Moreover,

since

$$\begin{aligned} WB &= W - \rho W^2 = (I_n - \rho W)W = BW \\ &\Rightarrow DW = WD \end{aligned} \tag{A.27}$$

we see that $(WD)' = D'W' = DW = WD$, and hence that WD is also symmetric. Finally, since $\mathbf{P}$ and $I_n - \mathbf{P}$ are each orthogonal projections, they are symmetric, idempotent and positive semidefinite. Hence by letting $M_1 = WD(I_n - \mathbf{P})$, the first term in the brackets of (A.26) is seen to be of the form

$$\begin{aligned} tr\left[(I_n - \mathbf{P})D'W'\mathbf{P}WD\right] &= tr\left[(I_n - \mathbf{P})^2 D'W'\mathbf{P}WD\right] \\ &= tr\left[(I_n - \mathbf{P})D'W'\mathbf{P}WD(I_n - \mathbf{P})\right] \\ &= tr\left[M_1'\mathbf{P}M_1\right] \end{aligned} \tag{A.28}$$

But since the symmetric positive semidefiniteness of $\mathbf{P}$ implies that $M_1'\mathbf{P}M_1$ is also symmetric positive semidefinite, it then follows that all eigenvalues are nonnegative, and hence that $tr\left[M_1'\mathbf{P}M_1\right] \geq 0$. Similarly, by setting $M_2 = WD\mathbf{P}$, the second term in the brackets of (A.26) is of the form

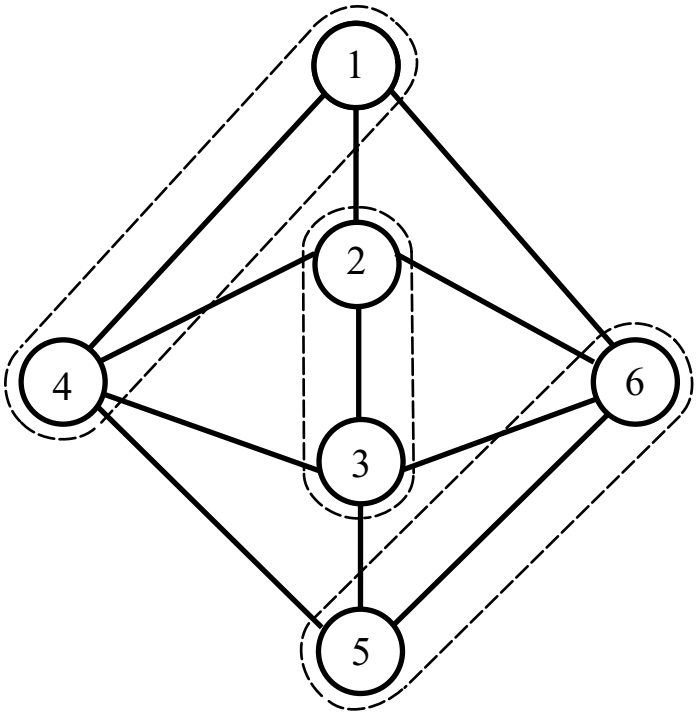
$$\begin{aligned} tr\left[\mathbf{P}WD(I_n - \mathbf{P})WD\right] &= tr\left[\mathbf{P}^2 WD(I_n - \mathbf{P})WD\right] \\ &= tr\left[\mathbf{P}WD(I_n - \mathbf{P})WD\mathbf{P}\right] \\ &= tr\left[M_2'(I_n - \mathbf{P})M_2\right] \end{aligned} \tag{A.29}$$

and hence is nonnegative by the same arguments. Finally, since the third term the brackets of (A.26) is also of this same form with $M_3 = WD$, it follows that all terms are nonnegative.

It thus remains only to show that at least one term is positive. But for the third term, we see from the orthogonal projection properties of $\mathbf{P}$ (including the equality $AD\mathbf{P} = AD$) together with the identity $WD = DW$ and the nonsingularity of D that

$$\begin{aligned} tr\left[(WD)'\mathbf{P}WD\right] = 0 &\Rightarrow (WD)'\mathbf{P}WD = \mathbf{O} \\ &\Rightarrow \mathbf{P}WD = \mathbf{O} \Rightarrow AD\mathbf{P}WD = \mathbf{O} \\ &\Rightarrow ADWD = \mathbf{O} \Rightarrow ADW = \mathbf{O} \\ &\Rightarrow AWD = \mathbf{O} \Rightarrow AW = \mathbf{O}\,, \end{aligned} \tag{A.30}$$

which is not possible given our definitions of A and W. Hence $tr\left[(WD)'\mathbf{P}WD\right] > 0$ and the result is established.



$$W = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix}$$

$$A = \begin{bmatrix} 0 & .35 & .65 & 0 & 0 & 0 \\ 0 & 0 & 0 & .15 & .85 & 0 \\ .80 & 0 & 0 & 0 & 0 & .20 \end{bmatrix}$$

Figure 1. A Simple Spatial Aggregation Example

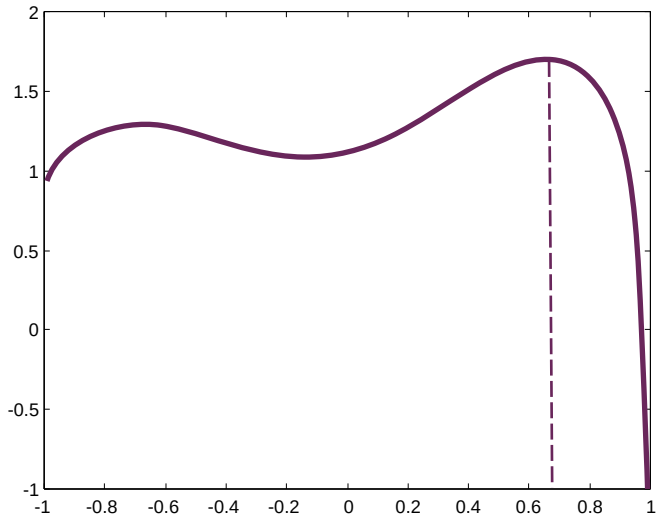


Figure 2a. Bimodal with Positive Global Maximum

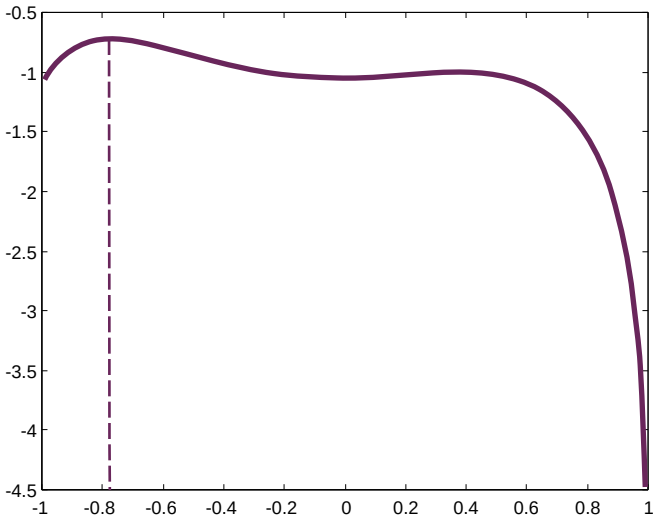


Figure 2b. Bimodal with Negative Global Maximum

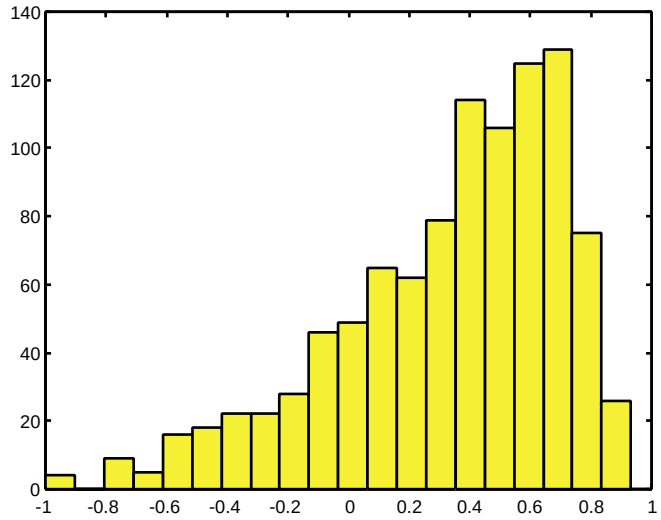


Figure 3a. Disaggregate Model estimates of Rho (3x6)

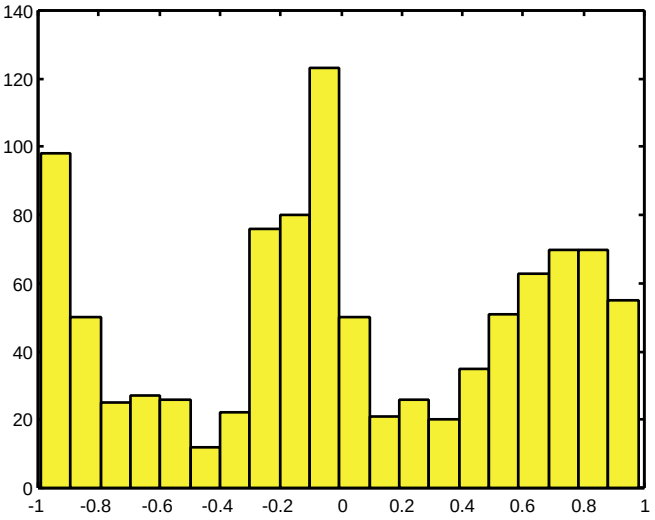


Figure 3b. Aggregate Model estimates of Rho (3x6)

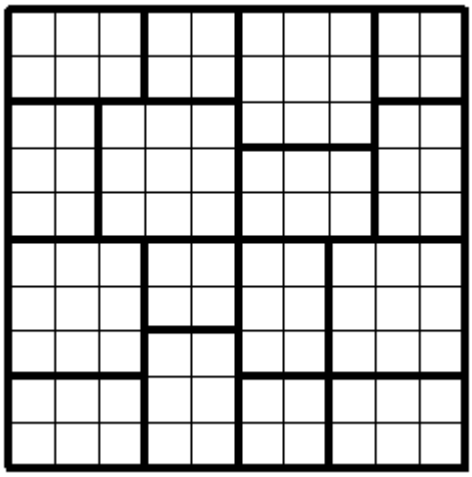


Figure 4a. Regions and Subregions (16x100)

Figure 4a. Philadelphia Block Groups (43x312)

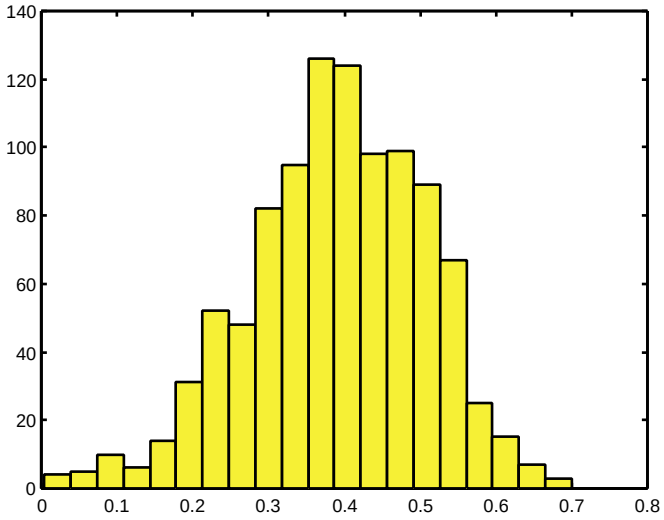


Figure 5a. Disaggregate Model estimates of Rho (16x100)

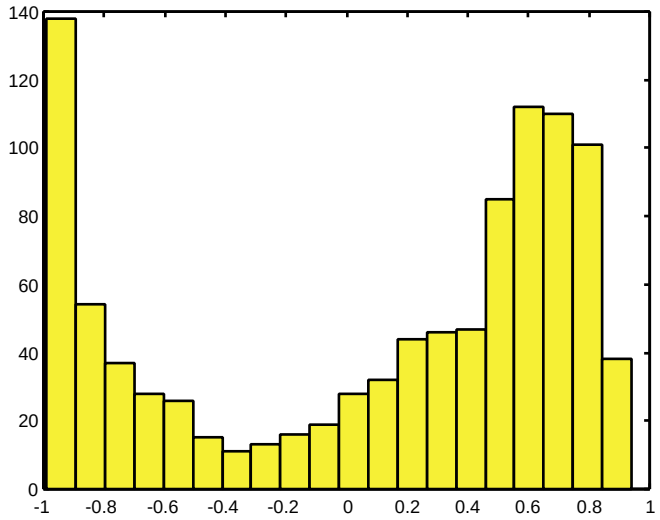


Figure 5b. Aggregate Model estimates of Rho (16x100)

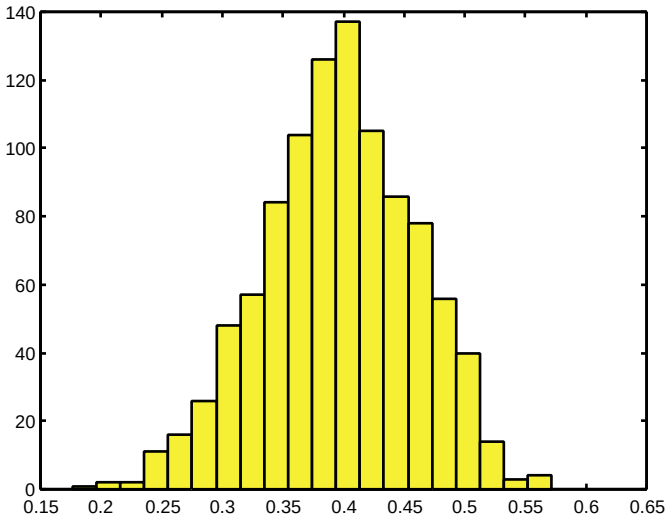


Figure 6a. Dissaggregate Model estimates of Rho (43x312)

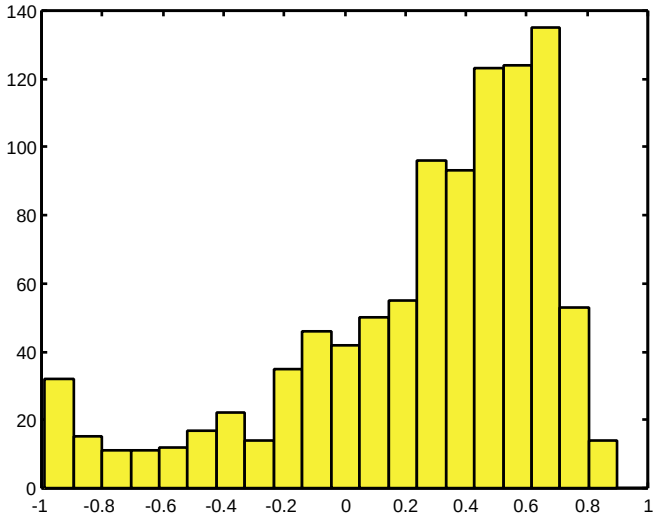


Figure 6b. Aggregate Model estimates of Rho (43x312)

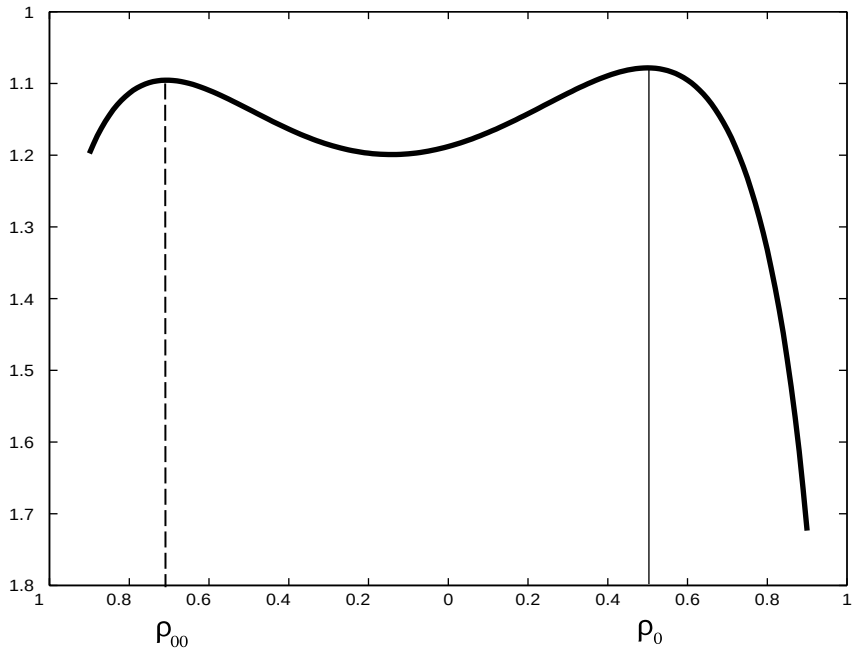


Figure 7a. Limiting Likelihood function (3x6)

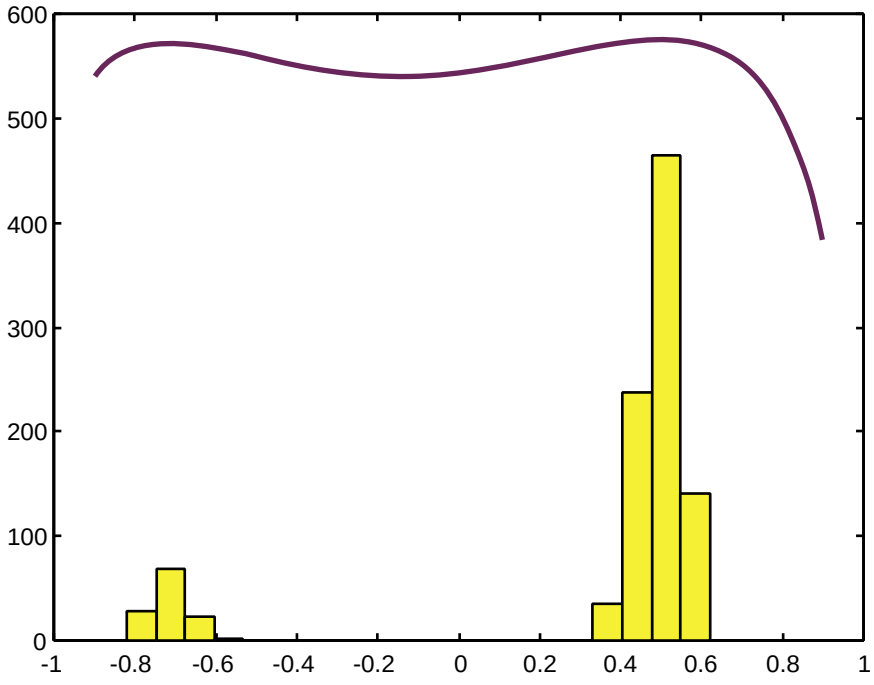


Figure 7b. Histogram for 100 Replications (3x6)

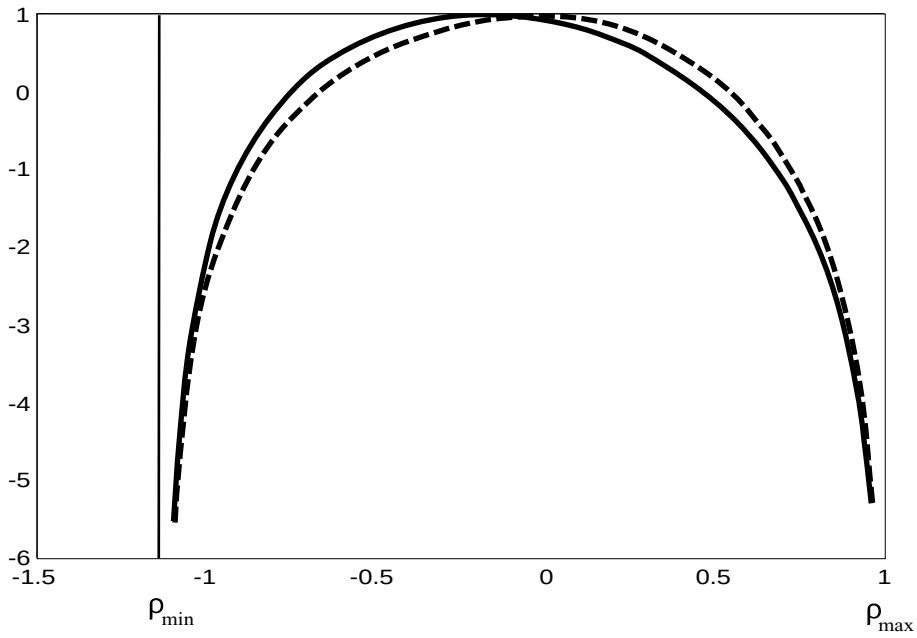


Figure 8a. Corrected Asymptotic Log-Likelihood (100 reps of 3x6)

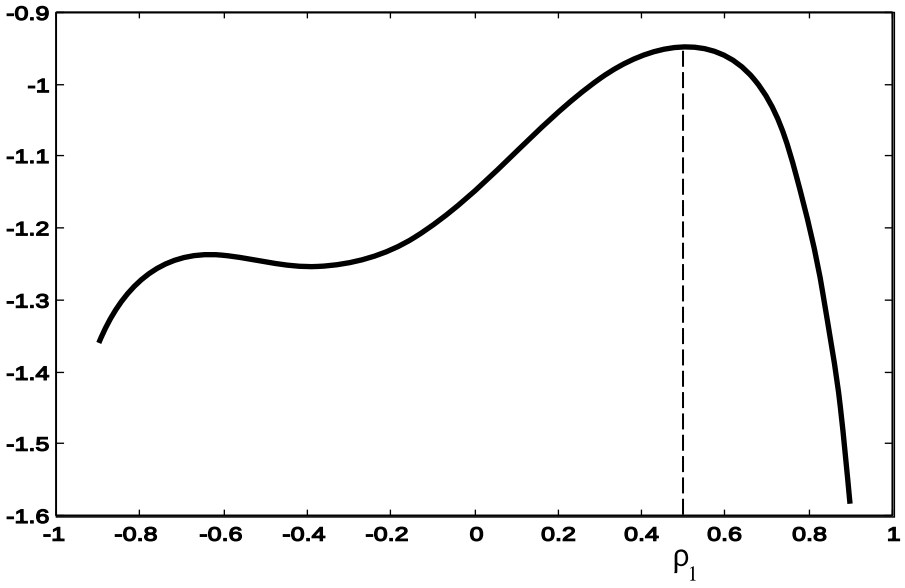


Figure 8b. Corrected Limit Likelihood (100 reps of 3x6)

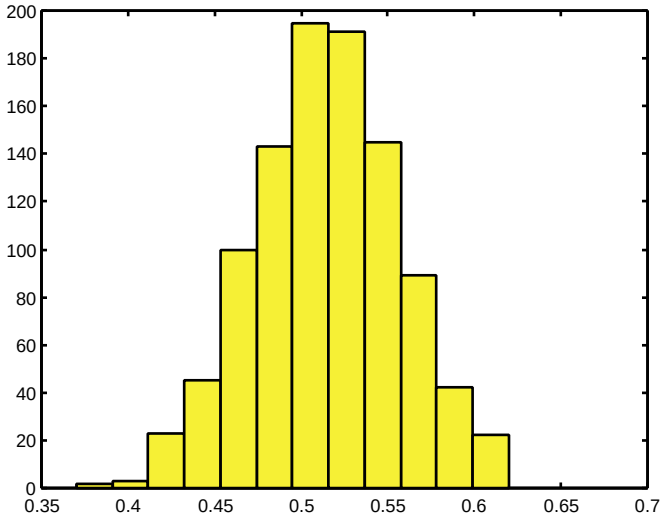


Figure 9a. Histogram of Corrected Estimates (3x6, *N* = 100)

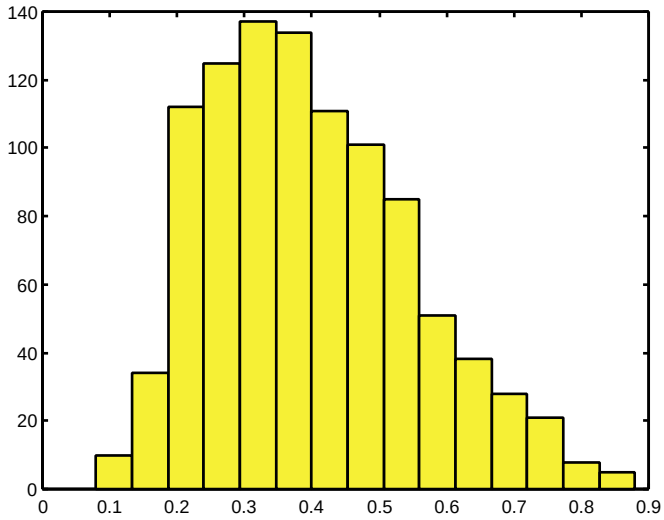


Figure 9b. Histogram of Corrected Estimates (16x100)

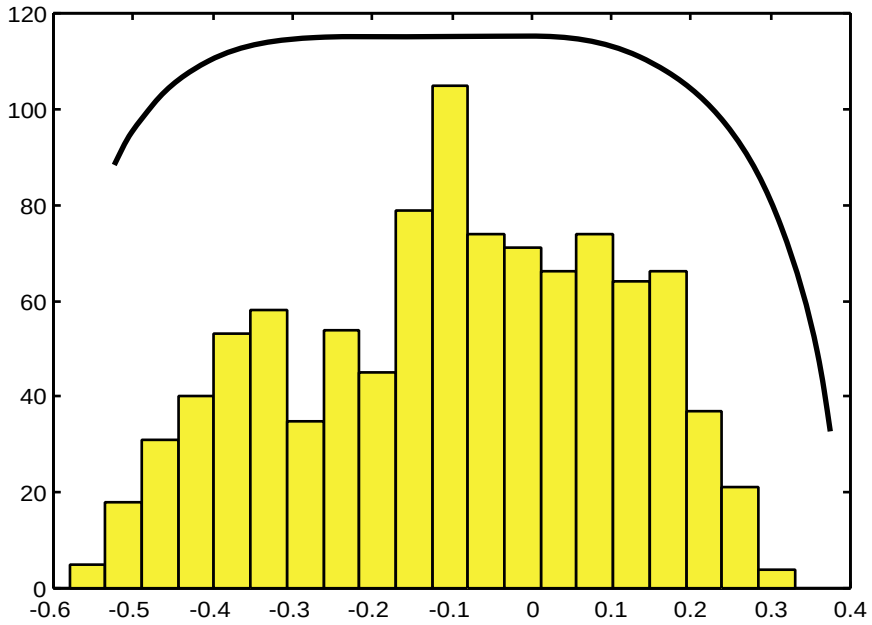


Figure 10. Histogram for $\rho_0 = 0$ (100 replications of 3x6)