A SPATIAL MIXTURE MODEL OF INNOVATION DIFFUSION

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

Sangyoung Song
Wharton School
University of Pennsylvania
Philadelphia, PA 19104

February 15, 2004

Abstract

The diffusion of new product or technical innovation over space is here modeled as an event-based process in which the likelihood of the next adopter being in region r is influenced by two factors: (i) the potential interactions of individuals in r with current adopters in neighboring regions, and (ii) all other attributes of individuals in r that may influence their adoption propensity. The first factor is characterized by a logit model reflecting the likelihood of adoption due to spatial contacts with previous adopters, and the second by a logit model reflecting the likelihood of adoption due to other intrinsic effects. The resulting spatial diffusion process is then assumed to be driven by a probabilistic mixture of the two. A number of formal properties of this model are analyzed, including its asymptotic behavior. But the main analytical focus is on statistical estimation of parameters. Here it is shown that standard maximum-likelihood estimates require large sample sizes to achieve reasonable results. Two estimation approaches are developed which yield more sensible results for small sample sizes. These results are applied to a small data set involving the adoption of a new Internet grocery-shopping service by consumers in the Philadelphia metropolitan area.

1. Introduction

A variety of behavioral phenomena involve some form of diffusion behavior, ranging from epidemics of communicable diseases to the spread of gossip [Rogers (1995)]. The present paper is concerned specifically with how the spatial diffusion of new-product adoptions may be influenced by communications with previous adopters. In the marketing literature there is a long history of efforts to model such phenomena. This work [summarized for example in Mahajan, et al. (1990)] has focused mainly on the temporal aspects of innovation diffusion stemming from the original differential-equation approach of Bass (1969). However, our present focus is on the spatial aspects of such processes: if word-of-mouth communication is in fact a significant component of adoption behavior, then one should be able to detect this in terms of the spatial proximity of current adopters to previous adopters. Spatial models of innovation diffusion date from the work of Hägerstrand (1967), as summarized (along with its many extensions) in Morrill, et al. (1988). The most relevant extension of Hägerstrand's work for our present purposes is the model proposed by Haining (1983). This model focuses on adoptions at the individual level within a discrete-time framework, where the current probability of adoption by any individual is taken to be a logistic function of proximities to previous adopters (defined as negative exponentials of distances). A more recent model of individual adoption behavior by Strang and Tuma (1993) incorporates spatial and temporal components in a more explicit manner. Here it is postulated that at each point of (continuous) time, the instantaneous rate of adoption for individuals depends both on their "intrinsic" rate of adoption and their "infectious" rate of adoption. The first is assumed to depend linearly on relevant attributes of the individual, and the second to depend linearly on factors affecting the individual's likelihood of contacts with previous adopters.

The present model combines elements of each of these approaches. Like Haining, we characterize adoptions as a discrete sequence of events rather than points in a time continuum.[1] We also employ logit models of adoption probabilities. In this context, we develop explicit models of both the "intrinsic" factors and "contact" factors influencing adoptions, as in Strang and Tuma. But rather than focusing on the specific locations of individuals, we choose to aggregate individual locations into spatial regions (zip code areas, census tracts, etc.) that are more

[1] However Haining does introduce certain time effects by allowing the model parameters to differ from period to period. In contrast, the present model focuses solely on the sequence of adoption events, and not when they occur.

commensurate with most available data sets. In this context it is assumed that the likelihood of the next adopter being in region r is influenced by two factors: (i) the potential interactions of individuals in r with current adopters in neighboring regions, and (ii) all other attributes of individuals in r that may influence their adoption propensity. The first factor is characterized by a logit model reflecting the likelihood of adoption due to spatial contacts with previous adopters, and the second by a logit model reflecting the likelihood of adoption due to other intrinsic effects. The resulting spatial diffusion process is then assumed to be driven by a *probabilistic mixture* of the two.[2]

A number of formal properties of this model are analyzed. First it is shown that this model exhibits strong stochastic convergence to a unique steady state. Moreover, this steady state turns out to be expressible as an explicit function of the model parameters, thus providing additional useful information both for estimation purposes and for a fuller behavioral understanding of the model itself. Our subsequent analysis focuses on the statistical estimation of parameters. Here it is shown that standard maximum-likelihood estimates can be badly behaved, and require large sample sizes to achieve reasonable results. Two alternative approaches are developed. The first draws on general mixture-distribution results to construct an EM estimation algorithm that avoids many of the problems inherent in the direct maximum-likelihood approach. The second is a Bayesian approach that first smoothes the problem by postulating prior distributions for parameters, and then estimates parameter values as the mode of their joint posterior distribution given the observed data. Both approaches are shown to yield more sensible results for small sample sizes. This is particularly important in situations where interest focuses on the early stages of an adoption process.

The paper begins in section 2 below with a formal development of the model. The steady-state properties of this model are then studied in section 3, and methods for estimation are given in section 4 Finally, in section 5, these results are applied to a small data set involving the adoption of a new Internet grocery-shopping service by consumers in the Philadelphia metropolitan area.

[2] Mixture models have been applied to a wide range of phenomena, as exemplified by the many applications in McLachlan (2000). With particular reference to economic modeling, (e.g., brand choice and market segmentation), see also the many applications discussed in Wedel and Kamakura (2000, chapters 6 and 7).

2. The Basic Model

Consider the diffusion of information about an economic innovation (new product or technology) within a system of *spatial regions*, $r \in \mathbf{R} = \{1,..,R \geq 2\}$. Adoption of this innovation by individuals within the system can be modeled as realizations, $\{r_n : n = 0,1,..,N\}$, of an event-based *adoption process*, where $r_n \in \mathbf{R}$ denotes the region in which the n^{th} adoption of the innovation occurs.[3] In particular, the initial event r_0 identifies the region where adoption first occurs. If $p_n(r\,|r_0, r_1,..,r_{n-1})$ denotes probability that the n^{th} adoption occurs in region r, given the current history $(r_0, r_1,..,r_{n-1})$ of the process, then this stochastic process is completely specified by the sequence of conditional probabilities (p_n, $n = 1,..,N$), together with an initial distribution, $p_0(r)$, specifying the probability that the initial adoption occurs in region r.

This initial probability can depend on a number of relevant factors $(x_{r1},..,x_{rJ})$ which are intrinsic to the individuals in region r and serve to distinguish them from individuals in other regions. Such factors may for example include various income and age attributes of the population in r. Other relevant factors may of course include local advertising and media information about the innovation itself. If $p_0(i)$ denotes the probability that individual i in region r is the first adopter, then we postulate that this *intrinsic probability* of adoption is the same for all individuals in region r, and is proportional to an *exponential* function of these regional factors for r, i.e., is of the form

$$p_0(i) = \alpha_0 \exp\left(\sum\nolimits_{j=1}^{J} \beta_j x_{rj}\right) \quad , \quad i \in r \tag{2.1}$$

where the coefficients, $\beta_0, \beta_1,.., \beta_J$ are assumed to be common to all regions, and where α_0 is an undetermined multiplier. It then follows that the corresponding regional event probability, $p_0(r)$, is of the form

$$\begin{aligned} p_0(r) &= \alpha_0 \sum\nolimits_{i \in r} \exp\left(\sum\nolimits_{j=1}^{J} \beta_j x_{rj}\right) \\ &= \alpha_0 M_r \exp\left(\sum\nolimits_{j=1}^{J} \beta_j x_{rj}\right) \end{aligned} \tag{2.2}$$

where M_r denotes the *population size* of region r. Using the normalization condition $\sum_{r \in \mathbf{R}} p_0(r) = 1$, we may solve for α_0 and rewrite $p_0(r)$ more explicitly as a

[3] As mentioned in the introduction, this event-based approach focuses only on the location of the next adoption, and not on the time at which it occurs. A possible temporal extension is mentioned briefly in the Concluding Remarks.

weighted logit model:

$$p_0(r) = \frac{M_r \exp\left(\sum_{j=1}^{J} \beta_j x_{rj}\right)}{\sum_{s\in\mathbf{R}} M_s \exp\left(\sum_{j=1}^{J} \beta_j x_{sj}\right)} \ , \ r \in \mathbf{R} \tag{2.3}$$

Turning next to the conditional probabilities for all subsequent adoption events, $n = 1, .., N$, it is postulated that in addition to the above factors,[4] such adoptions may be due to a direct contact with previous adopters. To model such contacts, let $p_c(i|j)$ denote the probability that a contact by some adopter j is made with some individual i. If $i \in r$ and $j \in s$, then we postulate that this probability is proportional to a decreasing exponential function of the *contact cost*, c_{sr}, from region s to r, so that

$$p_c(i|j) = \alpha \exp\left(-\theta c_{sr}\right) \quad , \quad i \in r, j \in s, r, s \in R \tag{2.4}$$

for some multiplier, α, and nonnegative exponent, θ, common to all regions. Note that as in (2.1) above, one could in principle consider a linear combination of potentially relevant types of contact costs here, with coefficient vector θ reflecting the relative importance of each type of cost. But to analyze sensitivity to contact costs in the simplest possible way, we choose here to restrict θ to a single "cost-sensitivity" parameter.[5]

The regional event probability, $p_c(r|j)$, corresponding to (2.4) is then given by

$$p_c(r|j) = \sum\nolimits_{i\in r} p_c(i|j) = \alpha M_r \exp\left(-\theta c_{sr}\right) \ , \tag{2.5}$$

which together with the normalization condition, $\sum_{r\in\mathbf{R}} p_c(r|j) = 1$, allows one to solve for α and write p_c also as a weighted logit model:

$$p_c(r|j) = \frac{M_r \exp\left(-\theta c_{sr}\right)}{\sum_{v\in\mathbf{R}} M_v \exp\left(-\theta c_{sv}\right)} \quad , \quad j \in s, \ r \in \mathbf{R} \tag{2.6}$$

[4] At this point it should be noted that the regional factors relevant for the first adopter in (2.3) are here assumed to be the same for all subsequent adopters. Hence this model ignores any special features of "early adopters" versus "late adopters"

[5] This "cost sensitivity" interpretation of θ involves two implicit assumptions, namely that (i) higher contact costs tend to *impede* contacts, and (ii) contacts with previous adopters have a *positive* influence on potential adopters. These two assumptions are difficult to separate in practice. For example, if adopters tend to be dissatisfied with the product, then higher contact levels may act to discourage further adoptions. In this case, θ could be negative even when contacts levels are quite sensitive to contact costs.

Hence if the n^{th} adoption results from a contact with some previous adopter in the sequence, $(r_0, .., r_{n-1})$, then the probability, $p_c(r|r_1, .., r_{n-1})$, that this adoption will occur in region r is given by

$$\begin{aligned} p_c(r|r_1, .., r_{n-1}) &= \sum\nolimits_{s \in \mathbf{R}} p_c(r|j) p(j \in s|r_0, .., r_{n-1}) \\ &= \sum\nolimits_{s \in \mathbf{R}} \frac{M_r \exp(-\theta c_{sr})}{\sum_{v \in \mathbf{R}} M_v \exp(-\theta c_{sv})} p(j \in s|r_0, .., r_{n-1}) \quad (2.7) \end{aligned}$$

If $m_n(s)$ denotes the number of times that region s appears in the sequence, $(r_0, .., r_{n-1})$, then the last probability on the right hand side must be given by the current fraction (relative frequency) of adopters in s, i.e., by

$$p(j \in s|r_0, .., r_{n-1}) = \frac{m_n(s)}{n} \qquad (2.8)$$

Thus letting $f_n = [f_{ns} = m_n(s)/n : s \in \mathbf{R}]$ denote the corresponding relative-frequency distribution for the n^{th} adoption event, it follows that the adoption probability (2.7) may be rewritten as

$$p_c(r|f_n) = \sum\nolimits_{s \in \mathbf{R}} \frac{M_r \exp(-\theta c_{sr})}{\sum_{v \in \mathbf{R}} M_v \exp(-\theta c_{sv})} f_{ns} \ , \quad r \in \mathbf{R} \qquad (2.9)$$

where all relevant information in $(r_0, .., r_{n-1})$ is seen to be summarized by the current relative-frequency distribution of adopters, f_n.

Finally, to capture both spatial-contact effects and regional-factor effects, it is postulated that the conditional event probabilities, $p_n(r\,|r_0, .., r_{n-1})$, are in fact a *probabilistic mixture* of these two effects, i.e., that for all $n = 1, .., N$ and $(r_0, .., r_{n-1}) \in \mathbf{R}^n$,

$$p_n(r\,|r_0, .., r_{n-1}) = \lambda p_c(r|f_n) + (1 - \lambda) p_0(r) \qquad (2.10)$$

with *mixture probability*, $\lambda \in (0, 1)$. Note that as in (2.9) one may replace $(r_0, .., r_{n-1})$ on the left-hand side of (2.10) with f_n, and write simply

$$p_n(r\,|f_n) = \lambda p_c(r|f_n) + (1 - \lambda) p_0(r) \ , \quad r \in \mathbf{R} \qquad (2.11)$$

By way of summary, adoptions are thus assumed to be modeled by a *spatial-mixture process* as defined by [(2.3),(2.9),(2.11)].[6] There are several important assumptions implicit in this model, which we now discuss in turn.

[6] For the case of independent random samples, mixtures of logits have been studied extensively [as for example in Robert (1998, section 24.3.2) and McLachlan (2000, section 5.11)]. However for the present type of sequentially-dependent samples, such mixture processes appear to be less well known.

2.1. The Spatial-Mixture Assumption

A spatial-mixture process above essentially treats adoptions as a two-stage process in which (i) an "λ-weighted" coin is first flipped to determine whether an adoption is due to contact effects or other non-contact effects, and (ii) the appropriate distribution (either p_c or p_0) is then sampled to determine the region in which the new adoption occurs. This model obviously oversimplifies the actual adoption process in that adoptions may well involve both of these effects. However, in the absence of any clear hypotheses about the possible interactions between contact and non-contact effects, the present model attempts to capture their relative importance in the simplest possible way. Thus from a practical viewpoint, the mixture probability, λ, is best viewed simply as measure of the relative importance of spatial-contact effects in the adoption process.

2.2. The Constant-Population Assumption

Note also from equation (2.9) that the regional populations of potential adopters, M_r, are treated as *constant.* This ignores that fact that such populations must be reduced as adoptions occur (as recognized for example in "Bass-type" models of innovation diffusion). Hence a second key assumption in the present model is that the number of actual adopters in any region is a sufficiently small portion of the total regional population to allow these populations to be treated as constant. This is most reasonable for adoption processes involving, say, new products in competitive environments where attainable market shares are not likely to be large. More generally, this assumption is almost always appropriate for analyzing the early stages of the adoption processes – where innovation diffusion effects are most interesting. (A possible relaxation of this assumption is considered briefly in the Concluding Remarks.) Notice also from (2.9) that it is only *relative* population sizes that need to be considered here (i.e., doubling all populations has no effect on the adoption process).[7] This is even more clear in terms of the population of adopters, which at any stage n of the process need only be specified in terms of its associated *relative frequency distribution*, f_n. Indeed, the evolution of this adopter-distribution constitutes a major focus of the present analysis. In the next section we show that as n becomes large, these distributions converge to a unique *steady state distribution* that essentially characterizes the latter stages of the adoption process. In addition it is shown that this convergence is generally

[7] However, doubling populations will certainly affect adoption *rates.* Such temporal issues are discussed further in the Concluding Remarks.

quite rapid as long as λ is not too large, i.e., as long as there are significant non-contact components to adoption behavior.

3. Analysis of Steady States

The main objective of this section is to construct a deterministic "mean representation" of spatial-mixture process and show that this deterministic version always converges to a unique mean-frequency distribution that describes the steady state of the system. In section 7.1 of the Appendix it is shown (by appealing to deeper results) that the full spatial-mixture process in fact converges with probability one to the same steady-state. Hence the present simpler development is intended primarily to motivate the essential features of this stochastic convergence result.

To analyze the asymptotic behavior of spatial-mixture processes as N becomes large, it is convenient to rewrite the system in vector notation. For any fixed *data matrix*, $X = (x_{rj} : r = 1,..,R, j = 1,..,J)$, *contact-cost matrix*, $C = (c_{rs} : r, s = 1,..,R)$, and *parameter values*, λ, θ, and $\beta = (\beta_1,..,\beta_J)'$, let the *contact-probability matrix*, $P_c = [P_c(r,s) : r,s = 1,..,R]$, be defined for all $r,s = 1,..,R$ by

$$P_c(r,s) = \frac{M_r \exp(-\theta c_{sr})}{\sum_{v\in\mathbf{R}} M_v \exp(-\theta c_{sv})} \tag{3.1}$$

and let the *intrinsic-probability vector*, $p_0 = (p_0(r) : r = 1,..,R)'$, be defined by (2.3).

3.1. The State-Probability Mapping

Next observe that if the *unit simplex* in $\Re^R$ (equivalently, the set of probability distributions on R) is denoted by

$$\Delta = \{x = (x_1,..,x_R) \in \Re^R_+ : \sum_{r=1}^{R} x_r = 1\} \tag{3.2}$$

then by definition, each relative frequency vector $f_n = (f_{nr} : r \in \mathbf{R})$ in (2.11) above must be an element of Δ. Hence if the *state-probability mapping*, $p : \Delta \to \Delta$, is now defined for all $f \in \Delta$ by

$$p(f) = \lambda P_c f + (1-\lambda)p_0 \tag{3.3}$$

then from (2.11) it is seen that this mapping essentially defines the entire process in the sense that for each n, the value $p(f_n)$ yields the n^{th} state probability distribution for the process, i.e.,

$$p_r(f_n) = p_n(r|f_n)\ ,\ n = 1, .., N \tag{3.4}$$

In particular it is should be clear that the asymptotic properties of each spatial-mixture process are governed by its state-probability mapping.

Before analyzing the properties of this mapping, it is useful to interpret $p(f_n)$ as the conditional distribution of an appropriately defined random vector. If the r^{th} column of the R-square identity matrix, I_R, is denoted by $e_r = (0, .., 1, .., 0)'$ then the region in which the n^{th} adoption occurs can be treated as the realization of a random *regional-outcome vector*, Y_n, where $\Pr(Y_n = e_r|f_n)$ denotes the conditional probability that the n^{th} adoption occurs in region r given the current distribution, f_n , of previous adopters. With this definition it follows at once from (2.11) and (3.4) that for all n:

$$p_r(f_n) = \Pr(Y_n = e_r|f_n)\ ,\ \ r \in \mathbf{R} \tag{3.5}$$

Even more important for our present purposes is the fact that p can also be interpreted as the *conditional expectation* of these regional-outcome vectors, i.e.,

$$\begin{aligned} E(Y_n|f_n) &= \sum\nolimits_{r=1}^{R} e_r \Pr(Y_n = e_r|f_n) = \sum\nolimits_{r=1}^{R} e_r p_r(f_n) \\ &= I_R p(f_n) = p(f_n). \end{aligned} \tag{3.6}$$

This view of p allows one to construct a useful deterministic approximation to spatial-mixture processes.

3.2. Mean Representations of Spatial-Mixture Processes

If we now designate the sequence of random frequency vectors $(f_n : n = 1, .., N)$ as the *adoption-frequency sequence* for the spatial-mixture process, then this sequence is seen to be generated by the sequence of regional-outcome vectors $(Y_n : n = 0, 1, .., N)$ according to the following recursive relation

$$f_{n+1} = \frac{n}{n+1} f_n + \frac{1}{n+1} Y_n\ ,\ \ n = 1, .., N-1 \tag{3.7}$$

with initial condition, $f_1 = Y_0$, defined by the regional-outcome vector for the first adopter. In other words, the adoption-frequency vector f_{n+1} for event $n + 1$ is constructed by adding the realization of the regional-outcome vector, X_n, to the sum, $nf_n = \sum_{m=0}^{n-1} Y_m$, of all previous regional outcomes and rescaling by $1/(n+1)$. Hence by taking conditional expectations, $E(\cdot|f_n)$, of both sides, we obtain the corresponding mean-value identity:

$$\begin{aligned} E\left(f_{n+1}|f_n\right) &= \frac{n}{n+1}f_n + \frac{1}{n+1}E(Y_n|f_n) \\ &= \frac{n}{n+1}f_n + \frac{1}{n+1}p(f_n)\ ,\ \ n = 1,..,N-1 \end{aligned} \tag{3.8}$$

Finally, by treating these conditional expectations as representative *mean frequencies*, $\overline{f}_1 = E(f_1)\ ,\ \overline{f}_2 = E(f_2|\overline{f}_1)\ ,..,\ \overline{f}_{n+1} = E\left(f_{n+1}|\overline{f}_n\right)$, we obtain a deterministic difference equation

$$\overline{f}_{n+1} = \frac{n}{n+1}\overline{f}_n + \frac{1}{n+1}p(\overline{f}_n)\ ,\ \ n = 1,..,N-1 \tag{3.9}$$

that can be viewed as the *mean representation* of the spatial-mixture process. While this deterministic representation in no way captures the full stochastic behavior of the process, it is reasonable to expect (from the Law of Large Numbers) that as $N \to \infty$, the asymptotic behavior of this deterministic version should accurately reflect that of the stochastic process.

3.3. Convergence of Mean-Representations

Hence the main objective of this section is to show that this sequence exhibits strong global convergence properties to a unique relative-frequency distribution that in fact describes the steady state of the system. In section 7.1 of the Appendix (Theorem 2) it is shown that the stochastic sequence in (3.7) *converges with probability one* to the same steady-state. We begin by rewriting (3.9) as follows:

$$\frac{\overline{f}_{n+1} - \overline{f}_n}{1/(n+1)} = p(\overline{f}_n) - \overline{f}_n \tag{3.10}$$

In this form expression (3.10) looks roughly like a differential equation, where the left-hand side is an approximate derivative. This can be made precise by first noting that the integer sequence ($n = 1, 2, ...$) conveys only event-ordering information that represents no explicit points of time. Hence we are free to associate these events with any correspondingly ordered sequence of time points. A particularly useful choice is given by the following recursive definition. Let $t_0 = 0$, and for each $n \geq 1$ let

$$t_n = t_{n-1} + \frac{1}{n} \tag{3.11}$$

If we then associate the n^{th} adoption event with time t_n and replace $\overline{f}_n$ with $\overline{f}_{t_n}$then (3.10) takes the form:

$$\frac{\overline{f}_{t_{n+1}} - \overline{f}_{t_n}}{t_{n+1} - t_n} = p(\overline{f}_{t_n}) - \overline{f}_{t_n} \tag{3.12}$$

which is now seen to be a genuine discrete approximation to the (autonomous) differential equation:

$$\dot{f}_t = p(f_t) - f_t \ , \ \ t \geq 0 \tag{3.13}$$

Two additional points should be made here. First observe from (3.9) that since (3.12) can also be written as

$$\overline{f}_{t_{n+1}} = \frac{n}{n+1}\overline{f}_{t_n} + \frac{1}{n+1}p(\overline{f}_{t_n}) \tag{3.14}$$

and since the unit simplex, Δ, is closed under convex combinations we see that $\overline{f}_{t_n} \in \Delta \Rightarrow p(\overline{f}_{t_n}) \in \Delta \Rightarrow \overline{f}_{t_{n+1}}$. Hence for any starting point $\overline{f}_{t_1} \in \Delta$, (3.12) generates a well-defined sequence $(\overline{f}_{t_n} : n = 1, 2, ..)$ in Δ, so that these discrete sequences are seen to approximate the behavior of the differential equation (3.13) *on the unit simplex*. Even more important is the fact that for this sequence of time values:

$$\begin{aligned} \lim_{n\to\infty} t_n &= \lim_{n\to\infty} \sum\nolimits_{m=1}^{n} \frac{1}{m} \\ &= \sum\nolimits_{m=1}^{\infty} \frac{1}{m} = \infty \end{aligned} \tag{3.15}$$

Hence, one the one hand, the time intervals, $t_n - t_{n-1}$, in the denominator of (3.12) get smaller, and thereby yield better and better approximations to the derivative on the left side of (3.13). But on the other hand, these time points slowly diverge to infinity, thus implying that the *asymptotic* behavior of (3.12) as $n \to \infty$ should be reflected by that of (3.13) as $t \to \infty$. This is the essential feature of the time sequence chosen in (3.11), and it continues to play a major role in the stochastic version developed in the Appendix. Given these observations, it suffices to say at this point that the asymptotic behavior of (3.12) [and hence of (3.9)] can be studied in terms of the asymptotic behavior of (3.13). But since (3.13) is essentially a differential equation, one can draw on large body of knowledge to determine its asymptotic properties.

To begin with it is clear that the *steady states* for solutions to this equation must be precisely the set of points $f \in \Delta$ where there is no further change, i.e., where $\dot{f} = 0$. By (3.13) this in turn is turn equivalent to the condition that $f = p(f)$, so that the steady states of solutions to (3.13) are precisely the *fixed points* of the state-probability mapping, p. Hence (as asserted above) the properties of this mapping are indeed central to the behavior of spatial-mixture processes. In the present case, by expanding this map, we see that fixed points of p must satisfy the linear equation:

$$\begin{aligned} f &= p(f) = \lambda P_c f + (1-\lambda)p_0 \\ &\Rightarrow (I_R - \lambda P_c)f = (1-\lambda)p_0 \end{aligned} \tag{3.16}$$

Moreover, as shown in section 7.1 of the Appendix (Theorem 1), the matrix $(I_R - \lambda P_c)$ is always nonsingular, so that we can solve for the *unique fixed point* of p as

$$f^* = (1-\lambda)(I_R - \lambda P_c)^{-1} p_0 \tag{3.17}$$

Given this result, the key question is whether or not solutions to (3.13) actually converge to this steady state. Here we can appeal to standard properties of linear differential equations to show that this is the case. In particular, if we rewrite (3.13) more explicitly as

$$\begin{aligned} \dot{f}_t &= \lambda P_c f_t + (1-\lambda)p_0 - f_t \\ &= (\lambda P_c - I_R)f_t + (1-\lambda)p_0 \end{aligned} \tag{3.18}$$

then (as shown in Theorem 1 of the Appendix, section 7.1) global convergence of solutions $(f_t : t \geq 0)$ to (3.18) is assured if the real parts of all eigenvalues of the matrix $\lambda P_c - I_R$ are negative. Here it can be shown (Lemma 2 of section 7.1 in the Appendix) that the real part, α, of each eigenvalue of $\lambda P_c - I_R$ satisfies the inequality

$$\alpha \leq \lambda - 1 \tag{3.19}$$

and hence must be negative. This implies (as in Theorem 1 of section 7.1 in the Appendix) that for every solution path $(f_t : t \geq 0)$ to (3.13) we must have

$$\lim_{t\to\infty} f_t = f^* \tag{3.20}$$

So the same must be true for the approximating mean-frequency sequences in (3.9) [and (3.12)]i.e., we must also have

$$\lim_{n\to\infty} \overline{f}_n = f^* \tag{3.21}$$

More generally, it can be shown (Theorem 2 of section 7.1 in the Appendix) that if this deterministic sequence is replaced by the adoption-frequency sequence $(f_n : n \geq 1)$, then this stochastic sequence converges to f^* with probability one, i.e.,

$$\Pr(\lim_{n\to\infty} f_n = f^*) = 1 \tag{3.22}$$

3.4. Rates of Convergence

The results above also help to clarify the important role of λ in determining the *rate of convergence* to f^*. As is well known [from the eigenvalue bound in (3.19)], the rate of convergence in (3.20) is of the same order of magnitude as the convergence of $\exp[(\lambda - 1)t]$ to zero as $t \to \infty$. In terms of the present discrete approximation, this implies that for all n,

$$\left|\overline{f}_n - f^*\right| = O\left[\exp\{(\lambda - 1)t_n\right] \tag{3.23}$$

where [by (3.11)],

$$t_n = \sum_{k=1}^{n} \frac{1}{k} \tag{3.24}$$

Hence smaller values of λ ensure faster rates of convergence. In the present case, observe that if $\lambda = 0$, then there no contact effects, so that all adoptions are independent random samples from the same fixed distribution, p_0, and convergence to a steady state is simply the Law of Large Numbers applied to distribution p_0. In the present model, this constitutes the maximum possible rate of convergence for the innovation diffusion process.

As spatial contacts become more important, i.e., as λ increases toward 1, the fixed distribution p_0 plays an ever-diminishing role. The resulting process then tends to behave in a "sticky" manner, depending on initial conditions. For example, if λ is close to one, and the first adoption occurs in a relatively populous region, then there is a good chance that subsequent contacts will continue to remain inside that region for some time. In the limiting case where all effects are due to contacts (i.e., a *pure-contact process* with $\lambda = 1$) there must still be a unique steady state, which is seen from (3.3) to be the stationary distribution for the Markov chain with transition matrix P_c. But while this steady state must eventually be achieved, the underlying contact process tends to exhibit long-range dependencies, and can be very slow to converge.[8]

4. Estimation of Parameters

Since the model itself is essentially a likelihood function, the obvious method to employ for parameter estimation is maximum-likelihood estimation. Unfortunately, this method turns out to be rather badly behaved for mixture models. For without further constraints, there is no guarantee that the key mixture parameter, λ, will lie between zero and one. Of course this would in principle be no problem if one had sufficiently many samples. However, simulations show that "sufficiently many" in the present case can be rather large (see section 7.2.3 of the Appendix). Two alternative estimation procedures are developed below that avoid these difficulties. We begin with a brief look at the maximum-likelihood approach, and then consider these alternative approaches.

4.1. Maximum-Likelihood Estimation

If for convenience we now denote the observed *regional-outcome data* by $y = (y_n : n = 0, 1, .., N)$ [with $y_n (= r_n)$ denoting the region in which the n^{th} adoption occurs], and let $f_n = f(y_0, .., y_{n-1})$ denote the adoption frequency distribution derivable from $y_0, .., y_{n-1}$ at stage n, then the *joint probability function* for this data can be written explicitly as

$$p(y; \beta, \lambda, \theta) = p(y_0; \beta) \prod\nolimits_{n=1}^{N} p(y_n | y_{n-1}, .., y_0; \beta, \lambda, \theta) \qquad (4.1)$$

where semicolons are used to separate variables from parameters. By expressions (2.3),(2.9) and (2.10), the probabilities on the right-hand side of (4.1) are seen to

[8] This pure-contact process is formally an instance of an *interactive Markov process* (as for example in Brumelle and Gerchak, 1980) and can in fact be characterized as a Markov chain on the *denumerable* state space of rational-valued distributions in the simplex Δ. However, while this representation is useful for establishing the existence of steady states, it offers little help in analyzing rates of convergence. Indeed, the "sticky" behavior of this process is due precisely to the fact that the positive supports of the associated transition-probability distributions (each with only R positive mass points) are very localized in this large state space. Simulated examples show that even for small regional systems, the realized adoption-frequency distributions, f_n, can remain far from the steady state for values of n in the tens of thousands.

have the respective forms

$$p(y_0;\beta) = \frac{M_{y_0}\exp\left(\sum_{j=1}^{J}\beta_j x_{y_0 j}\right)}{\sum_{s\in\mathbf{R}} M_s \exp\left(\sum_{j=1}^{J}\beta_j x_{sj}\right)} \tag{4.2}$$

and

$$\begin{aligned} p(y_n|y_{n-1},..,y_0;\beta,\lambda,\theta) &= \lambda p_c\left[y_n|f(y_0,..,y_{n-1})\right] + (1-\lambda)p_0(y_n) \\ &= \lambda\sum_{s\in\mathbf{R}} \frac{M_{y_n}\exp\left(-\theta c_{sy_n}\right)}{\sum_{v\in\mathbf{R}} M_v \exp\left(-\theta c_{sv}\right)} f_{ns} + (1-\lambda)\frac{M_{y_n}\exp\left(\sum_{j=1}^{J}\beta_j x_{y_n j}\right)}{\sum_{s\in\mathbf{R}} M_s\exp\left(\sum_{j=1}^{J}\beta_j x_{sj}\right)} \end{aligned} \tag{4.3}$$

Hence the standard *log likelihood function* for parameters (β,λ,θ) is given by[9]

$$L(\beta,\lambda,\theta|y) = \log p(y_0;\beta) + \sum_{n=1}^{N} \log p(y_n|y_{n-1},..,y_0;\beta,\lambda,\theta) \tag{4.4}$$

Our primary interest focuses on the behavior of this function with respect to the mixture parameter λ. To clarify this behavior, recall from (2.10) that each term of the summation on the right hand side can be written as

$$\begin{aligned} \log p(y_n|y_{n-1},..,y_0;\beta,\lambda,\theta) &= \log\left\{\lambda p_\theta(y_n|f_n) + (1-\lambda)p_\beta(y_n)\right\} \\ &= \log\left\{\lambda\left[p_\theta(y_n|f_n) - p_\beta(y_n)\right] + p_\beta(y_n)\right\} \end{aligned} \tag{4.5}$$

where the *contact* probability, $p_\theta(y_n|f_n)$, and *intrinsic* probability, $p_\beta(y_n)$, are now subscripted by their relevant parameters, θ and β, respectively. Hence each term in the summation of (4.4) is seen to be a concave increasing [decreasing] function of λ whenever $p_\theta(y_n|f_n) > p_\beta(y_n)$ [$p_\theta(y_n|f_n) < p_\beta(y_n)$]. It should also be clear that L is well defined for negative values of λ as long as bracketed values in (4.5) are positive for all $n = 1,..,N$. In fact, simulations (below) show that such cases are quite possible.

To gain further insight here, consider the following situation. Suppose that true value of λ is close to zero, and that intrinsic probability model, $p_\beta(\cdot)$, fits

[9] At this point it should be noted that for estimation purposes it is of course assumed that the number of regions (R) exceeds the number of parameters to be estimated (J) [so that intrinsic probabilities are not overparameterized]. In addition it is implicitly assumed that the number of adoptions (N) is larger that the number of regions [so that it is at least possible for the data to include adoptions in all regions].

the observed data quite well for an appropriate choice of β, say β^*. Suppose moreover that these intrinsic probabilities are fairly uniform, so that none of the probabilities $p_{\beta^*}(y_n)$ is extremely small. Finally, suppose that the true value of θ is very large, so that contact probabilities are small when contact costs are large. It is then reasonable to expect that the location, y_n, of each new adopter will tend to be closest to (i.e., have lowest contact costs with) those regions containing the largest number of previous adopters. This means that a large *negative* value of θ will tend to concentrate contact-probability mass on those regions with the fewest previous adopters, thus making these realized contacts look very improbable. In this type of situation is quite possible that for sufficiently negative values, θ^*, the parameter combination (β^*, θ^*) will yield contact probabilities that are uniformly smaller than the associated intrinsic probabilities, i.e., $p_{\theta^*}(y_n|f_n) < p_{\beta^*}(y_n)$ for all n. In such a case, (4.4) becomes a *globally decreasing function of* λ which is maximized at $\lambda^* = -\infty$.[10]

One explicit example is produced in the first column of Table 4.1 below. This example is taken from a set of 1000 simulations run on a model involving a sequence of 100 adoptions from a system of 18 regions ($N = 100, R = 18$).[11] Values for two intrinsic variables were generated randomly, and the model was simulated with parameter values ($\beta_1 = 1, \beta_2 = -2, \lambda = 0.3, \theta = 10$)[12]. The maximum-

	Max Lik		**EM**		**MAP**	
	Estimate	P-value	Estimate	P-value	Estimate	P-value
β_1	1.00038	0.00002	1.00038	0.00002	0.99939	0.00002
β_2	-2.17349	0.00000	-2.17342	0.00000	-2.17165	0.00000
λ	-1805646	0.00000	0.00013	0.99962	0.00001	0.92034
θ	-2838.631	0.99999	260.007	1.00000	153.963	0.99999

Table 4.1: Example of Negative Lambda

[10] The possible nonexistence of maximum-likelihood estimates for mixture distributions is well known, and was evidently first noted for mixtures of normal densities by Lehmann (1983). See also the discussions in Robert (1998) and McLachlan (2000, section 2.5).

[11] For sake of convenience, the 18 regions chosen consist of a tightly grouped set of contiguous counties in (an ArcView file of) North Carolina. The relevant contact costs were then taken to be centroid distances between counties.

[12] It should be noted here that interregional distances were normalized to lie between zero and one (with the maximum pairwise distance equaling one). At this scale, a negative exponential with $\theta = 10$ falls almost to zero at a about half the maximum distance, and essentially excludes interaction effects beyond this range.

likelihood estimates of parameters in the first column were obtained from one of the simulated "bad" cases as described above. (A detailed development of this estimation procedure is given in section 7.2.1 of the Appendix.) Notice that the β estimates for the intrinsic component of the model are right on target, while the estimates for λ and θ exhibit the type of pathological behavior described above.[13] Here the inequalities $p_\theta(y_n|f_n) < p_\beta(y_n)$ do indeed hold for all $n = 1, .., 100$, so that the algorithm would continue to produce greater negative λ values if allowed to run indefinitely. Note also that while we have included the standard (asymptotic) P-value diagnostics, these values are not really meaningful for either λ or θ, since the partial derivatives of the likelihood function with respect to these parameters are not quite zero (as is required for asymptotic normality of maximum-likelihood estimates).

Such examples show quite dramatically that in this type of mixture model, additional constraints are needed in order to ensure reasonable behavior of the estimates.[14] Here we consider two approaches. The first involves the well-known EM algorithm, and has the theoretical advantage of building in a natural constraint on λ with essentially no additional modeling assumptions. The second "Bayesian" approach turns out to be somewhat more attractive from a practical viewpoint, but does require additional modeling assumptions.

4.2. An EM Algorithm based on Data Augmentation

The estimation problem for mixture models discussed above is well known [see for example the discussion in Hastie, Tibshirani, and Friedman (2001, section 8.5)]. A standard way to avoid this problem is to employ the two-stage interpretation of mixture distributions mentioned in section 2.1 above, and to treat the outcome of the first "coin-flipping" stage as an unobserved dichotomous variable, δ, where in the present case, $\delta = 1$ denotes a *contact* adoption and $\delta = 0$ denotes an *intrinsic* (non-contact) adoption. The data is then "augmented" to a larger data set $(y, \delta) =$

[13] In the present case the true value, $\lambda = .3$, is *not* close to zero, so that the situation is not exactly as described above. But even for this value of λ, approximately 3% of the simulated λ-estimates were negative. This percentage of course eventually vanishes if N is allowed to increase without bound.

[14] On the positive side, however, it is worth noting that the functional differences between the intrinsic probabilities, $p_\beta(y_n)$, and contact probabilities, $p_\theta(y_n|f_n)$, ensure that all parameters of the model are generally *identifiable*. Hence the types of constraints often imposed to avoid "aliasing" (or "label-switching") problems in more symmetric mixture models are not necessary in the present case [see for example the discussion in Stephens, 2000].

$[(y_n,\delta_n) : n = 1,..,N]$ where the random vector, $\delta = (\delta_n : n = 1,..,N)$, is now treated as "missing data". Finally, an EM algorithm is constructed to estimate the parameters (β,λ,θ) in the presence of this missing data. This algorithm was first formalized by Dempster, Laird, and Rubin (1977), who showed that a wide range of estimation problems can be reformulated in terms of missing data.[15] For sake of completeness, a detailed development of this procedure for the present case is given in section 7.2.2 of the Appendix. Hence for our present purposes it is enough to sketch the main ideas.

Observe first that in this two-stage approach, λ becomes the sole parameter of the (Bernoulli) distribution of each random variate, δ_n. Hence the joint probability distribution in (4.1) is now replaced by a higher-dimensional joint probability distributions of the form:

$$p(y,\delta;\beta,\lambda,\theta) = p(y_0;\beta)\prod\nolimits_{n=1}^{N} p\left(y_n|\ \delta_n, y_0,..,y_{n-1};\beta,\theta\right) p(\delta_n;\lambda) \tag{4.6}$$

with corresponding log likelihood,

$$\begin{aligned}\log p(y,\delta;\beta,\lambda,\theta) = \sum\nolimits_{n=1}^{N} \log p(\delta_n;\lambda) + \\ \left[\log p(y_0;\beta) + \sum\nolimits_{n=1}^{N} \log p\left(y_n|\ \delta_n, y_0,..,y_{n-1};\beta,\theta\right)\right]\end{aligned} \tag{4.7}$$

Observe that if one were to average out (i.e., take expectations with respect to) δ in the density (4.6) then the original likelihood function would obviously be recovered. What is far less obvious is that if one averages out δ in the *log* likelihood (4.7) then the resulting function continues to be *monotone* in the original likelihood function, and hence has the same (local) maxima. This monotonicity property suggests a natural algorithm: Given any current parameter values $(\beta_k,\lambda_k,\theta_k)$, (i) first take the expectation of (4.7) with respect to δ, and then (ii) obtain $(\beta_{k+1},\lambda_{k+1},\theta_{k+1})$ by maximizing this function in (β,λ,θ). The expectation in (i) is called the *E-step*, and the maximization in (ii) is called the *M-step*. One can see immediately that this approach offers computational advantages since λ appears only in the first term of (4.7), and can be maximized separately. In the present case, it turns out that if in the E-step we let π_n^k denote the expectation

[15] With respect to mixture distributions in particular, the basic idea of this algorithm is evidently much older, as discussed in section 4.3 of Dempster, Laird, and Rubin (1977). A detailed treatment of EM algorithms for mixture distributions is given in McLachlan (2000, section 2.8).

of δ_n given current values $(\beta_k, \lambda_k, \theta_k)$, then this expectation has the form

$$\pi_n^k = E(\delta_n; \beta_k, \lambda_k, \theta_k) = \frac{\lambda_k p_k(y)}{\lambda_k p_k(y) + (1-\lambda_k) q_k(y)} \tag{4.8}$$

where p_k and q_k are positive probabilities. Moreover, in the M-step it turns out that the value of λ which maximizes the expectation of (4.7) has the simple closed form

$$\lambda_{k+1} = \frac{1}{N} \sum\nolimits_{n=1}^{N} \pi_n^k \tag{4.9}$$

Hence it is clear from (4.8) and (4.9) that if $0 < \lambda_k < 1$, then it must *always* be true that $0 < \lambda_{k+1} < 1$. In other words, if we start with some initial guess $0 < \lambda_1 < 1$, then all subsequent estimates will necessarily satisfy this constraint condition. But since this procedure must always increase the likelihood function in (4.4) one is led to ask how this is possible given the above behavior of this function. The answer is that this procedure essentially builds in a natural *parameter constraint*, $0 \leq \lambda \leq 1$, and then maximizes (4.4) subject to this constraint.

This is made clear by applying the EM algorithm to the above example. The estimation results are shown in the second set of columns. Here it is clear that the algorithm has converged to the zero-boundary value of λ. Moreover, a zero estimate for λ is not altogether unreasonable given this realization of the data. Note also that the sign of θ is now much more reasonable. The apparent insignificance of both λ and θ in terms of asymptotic P-values again has little meaning since the maximum occurs on the boundary of the constraint space, which violates the classical conditions for asymptotic normality.

Finally, we note that in addition to this particular illustration, the behavior of the EM algorithm was simulated for a range of sample sizes, N, in the 18-region example described above. Since the results of these simulations are qualitatively identical to those of the next procedure, we choose not to present them separately.

4.3. A Bayesian Estimation Approach

While the EM algorithm has many attractive features (including its simplicity of calculation), it suffers from several practical shortcomings. First, since in the present case it can be viewed formally an algorithm for likelihood maximization subject to the constraint $0 \leq \lambda \leq 1$, it will generally converge to zero values of λ whenever the likelihood function is maximized at negative λ values. Hence, as shown in the left histogram of Figure 4.2 below, the sampling distributions of such estimates tend to exhibit a clumping at zero, even though the true value of λ in this example was chosen to be 0.3. In addition, this algorithm is notoriously slow to converge. There are several well-known methods for speeding up this procedure: most notably to truncate the computationally intensive M-step by taking only a few gradient-steps (usually one) in each iteration.[16] This generalization, known as the GEM algorithm, was employed here but continued to be very slow to converge. Even in cases where the maximum-likelihood estimates were quite close to the true values, it generally proved to be most efficient to simply terminate the GEM algorithm after several hundred iterations.

With these limitations in mind, we are led to consider an alternative approach that amounts to smoothing the 0-1 constraint on λ. The most natural way to achieve smoothing is to adopt a Bayesian viewpoint, and treat parameters as random variables with prior distributions. The advantage of this approach in the present case is that even for small sample sizes, one can not only guarantee that λ lies in the unit interval, but can also impose sufficient smoothing to avoid concentrations at the end points.

4.3.1. A Prior Distribution for λ

The most natural prior distribution for λ is of course the *Beta Distribution*, which acts as the conjugate prior for simple binomial likelihoods. While prior skewness can be incorporated by employing the general two-parameter family of Beta distributions, $B(a,b)$, we choose to adopt the simple "symmetry" assumption that (without further information) small values of λ are no more likely than large values, i.e., that diffusion with very few spatial contacts is no more likely than diffusion dominated by spatial contacts. Hence, in addition to the above model assumptions, we now impose the Bayesian hypothesis that λ has a prior Beta distribution with $a = b$, so that (up to a constant factor) its prior density has the form:

$$\pi(\lambda) \propto \lambda^{a-1}(1-\lambda)^{a-1} \ , \quad a > 1 \tag{4.10}$$

While it is possible to treat a as an unknown parameter to be estimated (by imposing a "hyperprior distribution" on a) we choose to treat a as a simple smoothing parameter to be specified. The shape of this prior density is illustrated in Figure 4.1 below for two values: $a = 1.01$ and $a = 2.00$.

The first density is seen to be very flat, and is here employed as a Bayesian approximation to the constraint, $0 \leq \lambda \leq 1$, above. This will serve as the default

[16] See also the methods discussed in McLachlan (2000, section 2.17).

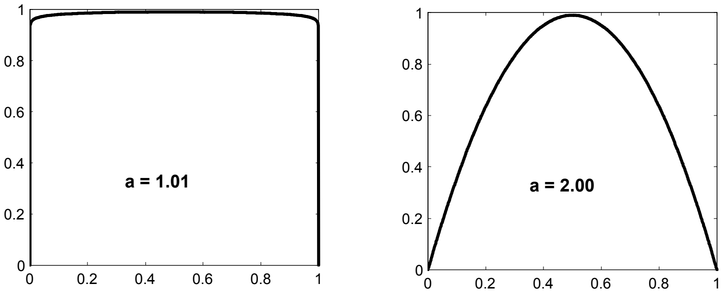

Figure 4.1: Prior Distributions for Lambda

value for most of the estimations discussed below. The second density is seen to be more concentrated about the middle value, $\lambda = 0.5$, and is here taken to represent the hypothesis that neither contact and nor non-contact elements are dominant in the adoption process. As illustrated below, this type of prior tends to yield more reasonable results in small-sample situations.

The other parameters are here taken to have flat (noninformative) priors with densities of the form:

$$\pi(\beta_j) \propto 1 \ \ , \ \ j = 1, .., J \tag{4.11}$$

and

$$\pi(\theta) \propto 1 \tag{4.12}$$

Observe that (4.12) allows both positive and negative values for θ. While it may be reasonable to postulate that $\theta > 0$ [as is implicit in (2.4)], this not only requires the introduction of an additional parameterized family of densities (such as a very flat Gamma density), but also precludes certain relevant types of contact behavior (as discussed in footnote 5 above). Hence we choose here to treat θ and β in a parallel manner.

4.3.2. Maximum a posteriori estimation

In this Bayesian framework, the probability function in (4.1) is now interpreted as a *conditional density*:[17]

$$p(y|\beta, \lambda, \theta) = p(y_0|\beta) \prod\nolimits_{n=1}^{N} p(y_n|y_{n-1}, .., y_0, \beta, \lambda, \theta) \tag{4.13}$$

[17] Following standard Bayesian convention, we here refer to all probability distributions as "densities" defined with respect to appropriate continuous or discrete reference measures.

in which semicolons are replaced by conditioning notation, so that $p(y|\beta,\lambda,\theta)$ now denotes the conditional density of y given (β,λ,θ). The *posterior density* of the parameters (β,λ,θ) given y is then obtainable from (4.10) through (4.13) by the standard identity

$$p(\beta,\lambda,\theta|y)p(y) = p(\beta,\lambda,\theta,y) = p(y|\beta,\lambda,\theta)p(\beta,\lambda,\theta)$$

$$\begin{aligned}
\Rightarrow \quad p(\beta,\lambda,\theta|y) &\propto p(y|\beta,\lambda,\theta)p(\beta,\lambda,\theta) \\
&= p(y|\beta,\lambda,\theta)\pi(\beta)\pi(\theta)\pi(\lambda) \\
&\propto p(y|\beta,\lambda,\theta)\lambda^{a-1}(1-\lambda)^{a-1} \\
&= p(y_0|\beta)\prod\nolimits_{n=1}^{N} p(y_n|f_n,\beta,\lambda,\theta)\lambda^{a-1}(1-\lambda)^{a-1} \qquad (4.14)
\end{aligned}$$

Given this joint posterior density, the most standard Bayesian approach is to derive the conditional posterior densities for each parameter, and then employ Gibbs sampling techniques to simulate the marginal posterior distributions of each parameter given y. This provides not only estimates of the posterior mean (and median) of each parameter, but also estimates of their standard deviations which can be used for testing purposes. This full approach will be developed in a subsequent paper.[18]

For the present, we choose simply to employ the natural Bayesian generalization of maximum-likelihood estimation: namely to find the most likely *posterior* values of the parameters given the data y. This procedure, known as *maximum a posteriori* (*MAP*) *estimation*,[19] is seen [from a comparison of (4.4) and (4.14)] to reduce to maximum-likelihood estimation when all priors are flat. In the present case, the log posterior density of (β,λ,θ) has the form:

$$\begin{aligned}
\log p(\beta,\lambda,\theta|y) &= \log\left[p(y_0|\beta)\prod\nolimits_{n=1}^{N} p(y_n|f_n,\beta,\lambda,\theta)\right] + \log\left[\lambda^{a-1}(1-\lambda)^{a-1}\right] \\
&= L(\beta,\lambda,\theta|y) + (a-1)\left[\log\lambda + \log(1-\lambda)\right] \qquad (4.15)
\end{aligned}$$

But since the posterior mode of (4.14) is by definition the point $(\beta^*,\lambda^*,\theta^*)$ that maximizes (4.14) [and hence (4.15)], it is clear that this estimation problem is

[18] This Gibbs sampling approach requires that the joint posterior density yield a *proper distribution* (i.e., with finite probability mass). Hence it is worth noting in passing that even though the above priors are improper with respect to β and θ, it can be shown [by employing the results of Speckman, Lee and Sun (2001)] that the joint posterior density given by (4.4) together with (4.2) and (4.3) is proper.

[19] For discussions of MAP estimation in the context of mixture distributions see McLachlan (2000, section 2.10), and Hastie, et. al. (2001, section 8.5.1).

formally equivalent to a *penalized* version of maximum-likelihood with penalty function, $(a-1)\,[\log\lambda + \log(1-\lambda)]$, approaching $-\infty$ as λ approaches zero or one. Thus the practical effect of this MAP estimation approach is to replace the sharp 0-1 constraint implicit in the EM algorithm above with a smoothed penalty-function version. Here it is evident that the parameter $a \in (1,\infty)$ governs the degree of smoothing, with a close to one yielding almost no smoothing of the 0-1 constraint (as can also be seen from Figure 4.1). Hence, as with all Bayesian estimation, the addition of this parameter serves to add flexibility to the constrained maximum-likelihood approach above.

A final issue that should be mentioned here is the possibility of *multiple maxima* for the objective function in (4.15). In fact the very nature of mixture distributions often yields multiple modes corresponding to the separate statistical populations in the mixture.[20] Hence, when sample sizes are relatively small, it may be difficult to distinguish a single dominant mixture of intrinsic and contact events. In such cases it is important try alternative starting points when maximizing (4.15) [as discussed further in section 7.2.3 of the Appendix].

4.3.3. Simulated Estimation Results

MAP estimation (with $a = 1.01$) was applied to 1000 simulations of the 18-region example above for the range of sample sizes shown in the first column of Table 4.2. The mean, median, and standard deviation of the sampling distributions for each parameter estimate are shown in the rest of the table. The last column for each variate shows the fraction of "bad" cases, defined for $\lambda(= 0.3)$ to be a value less than .01, and for all other parameters $[\beta_1(= 1), \beta_2(= -2), \theta(= 10)]$ to be a value with the wrong sign. As mentioned above, the parameter choice, $a = 1.01$, is taken to approximate an unsmoothed 0-1 constraint on λ. Before discussing the results of this approach in detail, it should be reiterated that the results obtained for the EM algorithm above are almost exactly the same. However, the EM algorithm takes on average about *ten times* as long to converge. Hence, while there are many known methods for marginally improving the efficiency of the EM algorithm, MAP estimation seems win hands down in the present application.[21]

Turning now to the estimation results themselves, notice first that the estimates for β_1 and β_2 seem quite reasonable even for the smallest sample size tested.

[20] See for example the illustrations given in Robert (1998) and McLachlan (2000).

[21] It is worth noting however that the improvement procedure (M-step) in each iteration of the EM algorithm is computationally somewhat simpler that the general gradient algorithm for MAP estimation. See section 7.2.2 for further discussion.

Beta 1					**Beta 2**			
Size	Mean	Median	St Dev	% < 0	Mean	Median	St Dev	% < 0
100	1.1727	1.057	0.615	0	-2.199	-2.046	0.675	0
200	1.102	1.029	0.413	0	-2.109	-2.039	0.465	0
500	1.077	1.017	0.372	0	-2.104	-2.035	0.411	0
1000	1.012	1.002	0.159	0	-2.014	-2.003	0.176	0
2000	1.008	0.999	0.098	0	-2.007	-2.001	0.124	0
Lambda					**Theta**			
Size	Mean	Median	St Dev	% < .01	Mean	Median	St Dev	% < 0
100	0.264	0.266	0.130	.029	16.378	9.268	169.19	.041
200	0.259	0.261	0.106	.005	11.114	9.126	123.69	.020
500	0.269	0.263	0.096	.001	5.105	9.073	130.35	.002
1000	0.274	0.274	0.071	0	9.311	9.415	2.207	0
2000	0.280	0.275	0.061	0	9.406	9.477	1.777	0

Table 4.2: Simulation Results for a = 1.01

In particular, there were no estimates with the wrong sign.[22] However, the results for λ and θ are far less satisfactory. Turning first to the mixture parameter, λ, it is clear that for sample sizes as small as 100 the estimates are quite unreliable. This is seen most clearly by the fact that almost 3% of the samples are clumped at zero. This means that about 3 in every 100 samples can be expected to yield a *negative* maximum-likelihood estimate of λ, even though the true value of λ is substantially larger than zero. This is shown in more detail in the corresponding histogram for λ estimates on the left in Figure 4.2 below. For comparison, the estimation results are also given for a smoothing value of $a = 2.0$. Notice that here the negativity problem has disappeared altogether, and that the values are now more concentrated around the true value, $\lambda = .3$.

However, it is important to emphasize that smoothing by itself is not guaranteed to improve the situation. For example, if λ were in fact close to zero, it is clear that the prior distribution for $a = 2$ in Figure 4.1 above would tend to overstate the significance of contact effects. More generally, it is well known that prior distributions tend to "pull" estimated modal values toward the prior mean. Hence a small maximum-likelihood value would be pulled upward toward $\lambda = .5$. The same is true for values of λ close to one, where smoothing tends to understate

[22] The corresponding ranges of asymptotic P-values were all quite significant, and are not shown.

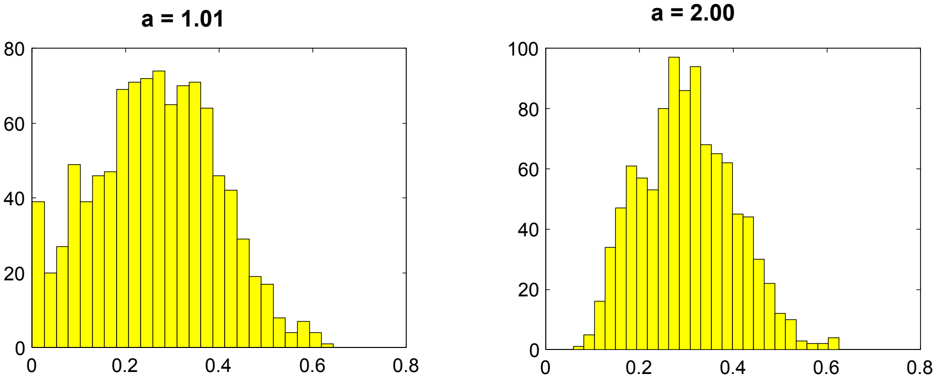

Figure 4.2: Lambda Estimates for $N = 100$

the significance of contact effect. All that can be said here is that when samples are relatively small (say less than 200 in the present case), a smoothing value of around $a = 2.0$ tends to produce better results than an unsmoothed estimate as long as the true value of λ is not too extreme.

4.4. Steady-State Consequences for Estimation

As stated in section 3.3, the present model is not simply a mixture model. In fact it is a discrete dynamical model with adoption frequencies (f_n) that converge stochastically to a unique steady given by the fixed point, f^*, of the state-probability mapping in (3.17) Hence if this innovation diffusion process has progressed sufficiently far, then it is reasonable to suppose that the most recently observed adoption frequencies, f_N, are fairly close to f^*. This has several consequences for parameter estimation. First, since the fixed point in (3.17) is seen to be an explicit function of the parameters, this provides an additional piece of information that is potentially useful for estimation purposes. However, if these frequencies are "too close" to the steady state, f^*, then there is the possibility of an identification problem in estimation. We now discuss each of these issues in turn.

4.4.1. Steady-State Regression

In view of the computational difficulties with MAP estimation discussed in section 7.2.3 of the Appendix, it is important to obtain a good starting point for the maximization algorithm. With this in mind, our present objective is to use steady state information to obtain reasonable initial parameter values for the estimation procedure. To do so we begin by observing from (3.16) that if f^* were known, then we would have the additional parametric relation:

$$\begin{aligned} (I_R - \lambda P_\theta) f^* &= (1-\lambda) p_\beta \\ &\Rightarrow \left[\frac{1}{1-\lambda}(I_R - \lambda P_\theta) \right] f^* = p_\beta \end{aligned} \tag{4.16}$$

where [as in (4.5)] we now index the contact-probability matrix, P_θ, and intrinsic-probability vector, p_β, by their associated parameters. Next let

$$\alpha = \left[\sum_{s \in \mathbf{R}} M_s \exp\left(\sum\nolimits_{j=1}^{J} \beta_j x_{sj} \right) \right]^{-1} \tag{4.17}$$

so that p_β has the form

$$\begin{aligned} p_\beta(r) &= \alpha M_r \exp(x_r' \beta) \\ &\Rightarrow \log\ p_\beta(r) = \log \alpha + \log M_r + x_r' \beta \\ &\Rightarrow \log\ p_\beta = (\log \alpha)\, u + \log M + X\beta \end{aligned} \tag{4.18}$$

where $M = (M_r : r = 1, .., R)'$ is the vector of regional populations, and again $u = (1, .., 1)'$ denotes the unit vector in $\Re^R$. In this context, if f_N is assumed to approximate f^*,[23] then we may write

$$\log \left[\frac{1}{1-\lambda}(I_R - \lambda P_\theta) \right] f_N - \log M \approx (\log \alpha)\, u + X\beta \tag{4.19}$$

Finally, by letting $v(\lambda, \theta)$ denote the left-hand side of (4.19) we obtain the approximate linear relation

$$v(\lambda, \theta) = [u, X] \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} \tag{4.20}$$

where $\beta_o (= \log \alpha)$ can now be viewed as an unknown intercept term. Hence for each given (λ, θ) pair, we can estimate β using ordinary least squares. While

[23] If N is large, then it may be more reasonable to average frequencies over say the last ten percent of adoptions, $\overline{f} = \frac{1}{m} \Sigma_{n=N-m}^{N} f_n$ (with $m/N \approx .10$), and use $\overline{f}$ rather than f_N.

neither λ or θ are known, they are *scalar* parameters. Moreover, λ is bounded by the unit interval, and θ is assumed to be positive. So in view of the speed of linear regression, it is a simple matter to try a selected range of values of these variables, compute the associated β estimate, and choose the triple (β, λ, θ) with maximum log likelihood, $L(\beta, \lambda, \theta|y)$, to be the initial set of parameter values $(\beta_0, \lambda_0, \theta_0)$ for the MAP estimation procedure. In the present model it turns out that β is not overly sensitive to the choice of $\theta > 0$, so we simply set $\theta = 1$ and focused on values of λ.[24] In this context, one additional simplification is possible since (4.16) requires that $(I_R - \lambda P_\theta)f_N \geq 0$. This is easily seen to restrict feasible values of λ to an interval $(0, \overline{\lambda})$ with $\overline{\lambda}$ defined by the condition that

$$\min_r \left[f_N(r) - \overline{\lambda} P_\theta(r, \cdot) f_N\right] = 0 \tag{4.21}$$

Even for the bad case involving only $N = 100$ adoptions, this procedure produced a reasonable starting point $[\beta_0 = (0.98073, -1.7987),\ \lambda_0 = 0.00899,\text{and } \theta_0 \equiv 1]$, for obtaining the final estimates in the MAP column of Table 4.1.

4.4.2. Parameter Identifiability

To motivate the second consequence of stochastic convergence, suppose that the diffusion process is not observed from its inception, but rather that data is only available for adoptions $n = m, .., N$. Then the joint density in (4.1) now has the form:

$$p(y^m; \beta, \lambda, \theta) = \prod\nolimits_{n=m}^{N} p(y_n | f_n, \beta, \lambda, \theta) \tag{4.22}$$

where $y^m = (y_m, .., y_N)$. In addition, suppose that m is sufficiently large to ensure by stochastic convergence that $f_n \approx f^*$ for all $n \geq m$, and hence that (4.22) has the approximate form:

$$p(y^m; \beta, \lambda, \theta) \approx \prod\nolimits_{n=m}^{N} p(y_n | f^*, \beta, \lambda, \theta) \tag{4.23}$$

This is now seen to be the likelihood of $N - m$ independent random samples from the distribution f^*, so parameter estimation is essentially classical maximum-likelihood estimation for independent random sampling. By solving for f^* in (4.16) as,

$$f^*(\beta, \lambda, \theta) = (1 - \lambda)(I_R - \lambda P_\theta)^{-1} p_\beta \tag{4.24}$$

[24] In the context of footnote 12 above, a negative exponential with $\theta = 1$ is fairly disbursed, with weights falling from one at zero to about 0.40 at the maximum interregional distance.

we can rewrite (4.23) in classical form as

$$p(y^m;\beta,\lambda,\theta) \approx \prod\nolimits_{r=1}^{R} f_r^*(\beta,\lambda,\theta)^{Nf_r^m} \tag{4.25}$$

Now suppose also that $N-m$ is sufficiently large to ensure (by the Law of Large Numbers) that the relative frequencies, f^m, corresponding to y^m are almost the same as f^*, i.e., $f^m \approx f^*$. Then since it is well known that (4.25) is maximized when each f_r^* equals f_r^m, it follows that any parameter values (β,λ,θ) satisfying the identity $f^m = f^*(\beta,\lambda,\theta)$ should be approximately maximum-likelihood estimates. To see that this yields a nonidentifiability problem, observe first from (4.24) that this identity can be written more explicitly as follows:

$$\begin{aligned} f^m &= (1-\lambda)(I_R - \lambda P_\theta)^{-1} p_\beta \\ &\Rightarrow \;\; (I_R - \lambda P_\theta) f^m = (1-\lambda) p_\beta \end{aligned} \tag{4.26}$$

Hence if we invoke the natural assumption that within-region contact costs, c_{rr}, are always smaller that between-region contact costs, c_{rs}, and allow θ to become large, then by (3.1) it follows that $\lim_{\theta\to\infty} P_\theta = I_R$, so that for large θ expression (4.26) reduces to

$$(1-\lambda) f^m \approx (1-\lambda) p_\beta \tag{4.27}$$

Thus for $\theta = \infty$ and $\lambda = 1$, it is clear from (4.27) that β is not identifiable. Of course $\lambda = 1$ is not possible in MAP estimation. So what this means from a practical viewpoint is that when m and $N-m$ are both large, one can expect to find cases yielding MAP estimates with θ large, λ close to one, and β estimates highly erratic. Simulations show that this is exactly what happens. In particular, the simulations shown in the "1000" row of Table 4.2 above are actually constructed as the first 1000 observations in simulations with $N = 2000$. The second thousand observations were then used to construct estimation examples as above with $m = 1000$ and $N - m = 1000$. The MAP estimation results for these samples are shown in the columns marked "Second 1000 Samples" in Table 4.3 below, and those in the "First 1000 Samples" are repeated from Table 4.2 below. The single most important comparison here is in terms of standard deviations. Notice in particular that those for the beta estimates have increased by a full order of magnitude. Notice also that λ both and θ are now significantly skewed toward larger values (mean > median). One specific example is illustrated in the last two columns. Notice that in the second 1000 samples, the estimates of both λ and θ are considerably larger than for the first 1000 samples, and that the beta

	First 1000 Samples			**Second 1000 Samples**			**Example**	
Par	Mean	Median	St Dev	Mean	Median	St Dev	First	Second
β_1	1.012	1.002	0.159	1.1071	1.007	2.783	1.314	-10.303
β_2	-2.014	-2.003	0.176	-2.1851	-2.009	3.910	-2.369	-38.599
λ	0.274	0.274	0.071	0.3083	0.2855	0.123	0.325	0.954
θ	9.311	9.415	2.207	10.572	9.679	12.70	8.596	40.605

Table 4.3: Comparison of Early and Late Estimates

estimates are now way off base. These are exactly the results predicted above, and show that identification problems do indeed tend to emerge as the given samples segment moves toward the steady state. The inference here is simply that diffusion parameters become harder to identify in later stages of the process, where adoption behavior is itself more diffuse.

5. An Application to Internet Grocery Shopping

The following application involves the adoption of a new Internet grocery shopping service, Netgrocer.com, by consumers in the Philadelphia metropolitan area. The spatial units (regions) are here taken to be zipcode areas, and the time span is from the introduction of this web site in May 1997 through January 2001. The zipcode-level demographic data used was provided by CACI Marketing Systems. Philadelphia was chosen as a study area from a much larger data set including all zipcode areas throughout the country.[25]

5.1. Spatial Data and Intrinsic Variables

The Philadelphia data set consisted of $N = 1288$ adoptions over the given time period, and involved a total of $R = 46$ zipcode areas.[26] As in the simulation study above, contact costs were taken to be linear in distance between centroids of zipcode areas. Populations for each zipcode area were based on 1999 levels, and were assumed to be essentially the same for all other years. The demographic

[25] The authors are grateful to David Bell for supplying the Netgrocer.com data used for the empirical application in this paper.

[26] There are 48 zipcode areas in Philadelphia. Two (peripheral) areas were excluded because of missing information.

data used for the intrinsic variables at the zipcode level are described in Table 5.1 below.

Variable	Description
SUPMAS	number of supermarkets per person
NOVEHICLE	% of housing units with no vehicle available
APER	% of Asians
BPER	% of Blacks
HPER	% of Hispanics
ELDERLY	% of population over 65 years old
COLDEG	% of over 25 year-olds with college degrees
SOLO	% of single-member households
FAMLARG	% of households with more than five members

Table 5.1: Description of Intrinsic Variables

The choice of these variables was based on a number of considerations that can be roughly classified as follows:

Access to off-line retail stores: Adoption of online grocery shopping would appear to be more attractive to consumers who do not own cars and/or live in zip-code areas that have fewer grocery stores. To reflect accessibility to off-line groceries, the availability of vehicles (NOVEHICLE) and the number of supermarkets divided by the population in the zipcode area (SUPMAS) were included.

Access to the Internet: According to a study by U.S. Department of Commerce[27], households of different ethnic and racial backgrounds have disparate rates of Internet access. Black and Hispanic households show lower Internet penetration rates (23.5% and 23.6%, respectively) than White and Asian households (46.1% and 56.8%, respectively). Hence we also expect that the percent of Black (BPER), Hispanic (HPER), and Asian (APER) households should influence the probability of adoption. Specifically, zipcode areas with higher percent of Black and Hispanic households are expected to be less likely to adopt the innovation. In addition, the elderly tend to be less willing to accept new technologies, partly because their learning costs are higher than for the young. Therefore we expect that adoptions

[27] *Falling through the net: Toward digital inclusion*, A Report on Americans' Access to Technology Tools (October 2000), U.S. Department of Commerce, Economics and Statistics Administration, National Telecommunications and Information Administration. This report can be obtained at the website, *http://digitaldivide.gov/reports.htm*

will be less likely in zipcode areas with higher percentages of elderly residents (ELDERLY).

Attractiveness of on-line shopping: We expect that more educated consumers (COLDEG) will tend to value time more highly, and hence find online grocery shopping more attractive . We also expect that the size of households will influence the attractiveness of on-line shopping. To capture such effects, we have included both the percentage of single-member households (SOLO) and the percentage of households with more than five members (FAMLARG).

5.2. Estimation Results

Estimates were obtained using both the MAP estimation procedure and the EM algorithm. The results, shown in Table 5.2 below, are seen to be essentially the same for both procedures.[28] In addition to the initial parameter values obtained by steady-state regression (discussed above), several alternative starting values for the parameters were tried and no secondary maxima were found.

	MAP Estimates		**EM Estimates**	
Variable	Estimate	P-value	Estimate	P-value
SUPMAS	-2197.695	< 0.0000	-2197.459	< 0.0000
NOVEHICLE	7.254	< 0.0000	7.253	< 0.0000
APER	5.410	< 0.0000	5.410	< 0.0000
BPER	-0.899	0.0064	-0.899	0.0064
HPER	-1.655	0.3075	-1.656	0.3073
ELDERLY	-6.830	0.0005	-6.830	0.0005
COLDEG	4.318	0.0021	4.318	0.0021
SOLO	-4.230	0.0998	-4.228	0.1000
FAMLARG	-12.550	0.0008	-12.547	0.0008
LAMBDA	0.582	< 0.0000	0.582	< 0.0000
THETA	-1212.148	0.9995	-1085.000	0.9985

Table 5.2: Comparison of MAP and EM Estimates

The key mixture parameter λ is statistically very significant, and suggests that word-of-mouth contacts may indeed be an important component of new-adoption

[28] However, the MAP gradient procedure converged in 12 iterations, while the EM algorithm required 636 iterations to converge.

behavior. This is also supported by the estimated value for θ, which is seen to be very negative. Plots show that the log posterior density approaches its maximum asymptotically as $\theta \to -\infty$, and hence that it is not possible to ascertain a meaningful asymptotic P-value here (as discussed above). However, it does appear that contacts are very sensitive to distance. To see that these findings are in fact consistent with the data, consider the spatial distribution of cumulative adoptions shown in Figure 5.1 below. Here it is clear that adoptions in the Philadelphia area are highly concentrated in the two zipcode areas (19104,19103) near the middle of the map. The biggest of these (19104) contains both the University of Pennsylvania and Drexel University, thus suggesting that many adoptions were by word of mouth between students, faculty and others associated with this academic complex. However, it should also be emphasized that since the Internet is used extensively by this particular population, many adoptions may have resulted from direct exposure to Internet advertising. In any case, the large negative value of θ suggests that most contacts are occurring *inside* these zipcode areas. So for this adoption process it appears, unfortunately, that zipcode areas are too large to capture diffusion effects in much detail.

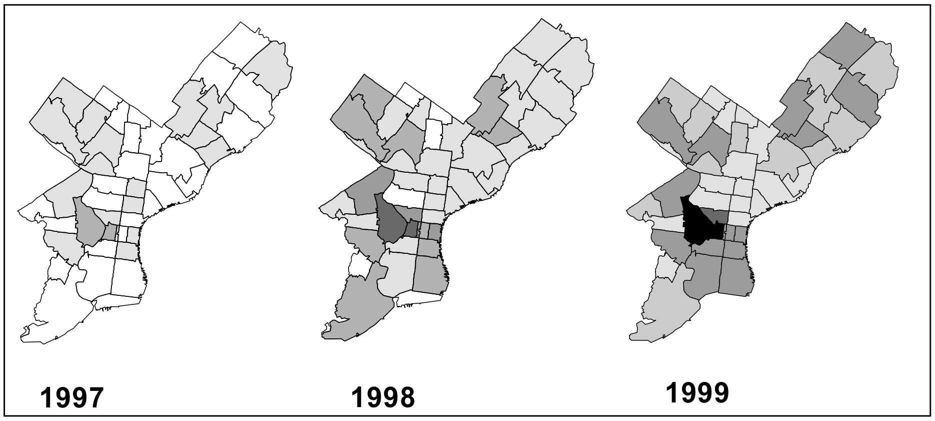

Figure 5.1: Cumulative Adoption Frequencies

Turning finally to the β estimates, most of our hypotheses were confirmed. First the significant positive sign on NOVEHICLE and negative sign on SUPMAS[29] both suggest that accessibility to off-line shopping does decrease the like-

[29] The parameter estimate for SUPMAS is large because SUPMAS is the only variable that is not normalized between zero and one.

lihood of adoption. As for ethnic differences in Internet usage, the signs of APER and BPER are very significant and consistent with our hypothesis. (The sign of HPER is also consistent but not significant). The effects of age are also evident in the significant negative sign on ELDERLY. Finally, the significant positive sign on COLDEG is consistent with our hypothesis about the attractiveness of online shopping to those with high values for time.

But perhaps what is most important to note here is that most of these findings can again be explained largely in terms of the academic populations within which this innovation process is concentrated. This is particularly clear with respect to the strong negative significance of ELDERLY and FAMLARG, both of which are consistent with young student populations. The positive significance of APER is also consistent with the current composition of these populations. The only real surprise here is the insignificance of SOLO, which is clearly an attribute of student populations. While SOLO does exhibit some collinearities with other variables, this is not the main source of the problem. Rather, the key difficulty appears to be the fact that less than 17% of households in the university zipcode (19104) are actually single-member households. In particular, this large zipcode also contains a sizable residential population in West Philadelphia not related to the university.[30] This again suggests that in the present application, zipcode areas are too large to capture some of the significant population heterogeneities that exist.

6. Concluding Remarks

In this paper we have developed a probabilistic mixture model for analyzing new-product adoption processes in space. The primary intent of the model is to allow the relevance of spatial contact effects to be estimated in a simple manner by means of a single mixture parameter. But while the model is appealing in terms of its simplicity, it is clearly limited in scope. Hence it is appropriate here to consider two possible extensions of the model that would widen its applicability.

One simple extension would be to relax the *constant-population assumption*, and allow the populations of past adopters and potential future adopters in each region to change as new adoptions occur. The primary advantage of the constant-population assumption employed here is to allow a simple formulation and analysis of the steady-state properties of the model – assuming relatively large regional

[30] Most off-campus students now live in the Center City area of Philadelphia, as reflected by tract 19103 which contains almost 50% single households.

populations of potential adopters. But in cases where regions are quite small (such as the extreme case of single individuals, when such data is available), or more generally, cases where the numbers of adopters are likely to grow to a significant fraction of total regional populations, this notion of a steady state is not tenable. Here it would be more appropriate to replace fixed regional populations, M_r, with variable populations. More precisely, if for any given adoption data, $y = (y_0, y_1, .., y_N)$, we define indicator functions by $\delta_r^n(y) = 1 \Leftrightarrow y_n = r$, then for any set of initial regional populations, $(M_r : r \in \mathbf{R})$, one may consider *state-dependent populations*, $(M_r^n : r \in \mathbf{R}, n = 1, .., N)$ defined by

$$M_r^n(y) = M_r - \sum\nolimits_{i=0}^{n} \delta_r^n(y) \tag{6.1}$$

The present model can easily be reformulated in terms of these variable populations.[31]

A more important limitation of the present model relates to *time*. In particular, this event-based process ignores any real-time considerations. However, if one is interested for example in the number of new customers expected during the first year following the introduction of a product, or the time required for adoptions to reach say 20% of "market saturation", then it is clear that time-based rates of adoption are crucial. With this in mind, it is important to observe that event-based models such as the present one can always be viewed formally as "jump processes" embedded within continuous-time stochastic processes, where the only missing ingredient is the relevant sequence of "sojourn times" between jumps. By choosing the right state space, one can in fact regard the present model as the simplest type of a jump process, namely a *Markov chain*. In particular, if regional populations M_r are taken as fixed integers, and partitioned into *adopter subpopulations*, M_r^a, and *non-adopter subpopulations*, $\overline{M}_r^a = M_r - M_r^a$, then one can easily reformulate the present model as a Markov chain on the finite state space, $\mathbf{M}$, consisting of all possible population profiles,

$$\mathbf{M} = \{[(M_r^a, M_r - M_r^a), r \in \mathbf{R}] : 0 \leq M_r^a \leq M_r, r \in \mathbf{R}\} \tag{6.2}$$

(Note also that this state space automatically incorporates variable populations, as discussed above.) Such a Markov chain can easily be extended to a *continuous-time Markov process* by the introduction of appropriately defined exponential sojourn times [as detailed for example in Kulkarni (1995,Chapter 6)]. Of course

[31] It should be noted, however that (like most Bass-type models) these state-dependencies still ignore other relevant population changes, such as in- and out-migration from each region r.

it should be emphasized that the state space in (6.2) is generally enormous, and difficult to analyze directly. But it nonetheless provides a basic framework within which more succinct modeling representations can be developed. Such possibilities will be explored in subsequent work.

7. Appendix

The following appendix includes stochastic steady-state analysis as well as the details of the maximum-likelihood, EM, and MAP estimation procedures discussed in the text.

7.1. Steady State Analysis

The purpose of this section of the Appendix is to establish the stochastic convergence property in expression (3.22) of the text. The approach used here draws heavily on the work of Kushner and Clark (1978), as summarized in Kushner and Kim (1997), now designated as [KK]. From a conceptual viewpoint, this approach closely parallels the development of sections 3.2 and 3.3 , except that we now consider approximations to the autonomous differential equation (3.13) on Δ by the *stochastic* relation in (3.7), rather than the deterministic relation in (3.9). We begin by rewriting (3.7) in a more general form as:

$$\dot{f}_t = \Psi(f_t) \ , \ \ f_0 \in \Delta \tag{7.1}$$

with (affine) kernel function, $\Psi : \Re^R \to \Re^R$, defined by

$$\begin{aligned} \Psi(f) &= p(f) - f \\ &= [\lambda P_c f + (1-\lambda) p_0] - f \end{aligned} \tag{7.2}$$

To analyze this differential equation, we begin by showing that the restriction of starting points, f_0, in (7.1) to the unit simplex Δ ensures that solutions will be entirely contained in Δ:

Lemma 1. *The unit simplex,* Δ, *is an invariant set for the differential equation with kernel* (7.2), *i.e., for any solution* $(f_t : t \geq 0)$ *of* (7.1) *it must be true that*

$$f_t \in \Delta \ , \ t \geq 0 \tag{7.3}$$

Proof: To see that a solution path $(f_t : t \geq 0)$ with $f_0 \in \Delta$ can never leave Δ, note first that if $u = (1,..,1)' \in \Re^R$, then one may define a corresponding differential equation for the *sum*, $s = u'f$, by

$$\begin{aligned} \dot{s}_t &= u'\dot{f}_t = u'\left[\lambda P_c f_t + (1-\lambda)p_0 - f_t\right] \\ &= \lambda\left(u'P_c f_t\right) + (1-\lambda)u'p_0 - u'f_t \end{aligned} \tag{7.4}$$

But $u'p_0 = 1$ and $u'f = 1$ for all $f \in \Delta$. Moreover, since P_c is column stochastic by definition, it also follows that $u'P_c = u'$, and hence that for all solutions $(f_t : t \geq 0)$ with $f_0 \in \Delta$ we must have,

$$\dot{s}_t = \lambda(1) + (1-\lambda)(1) - 1 = 0\ , \quad t \geq 0 \tag{7.5}$$

This in turn implies that these solutions must stay in the *unit flat*, $F = \{f \in \Re^R : u'f = 1\}$ containing Δ. Thus it follows from the continuity of solution paths for (7.1) that any path leaving Δ must cross the boundary $\partial(\Delta)$ of Δ in F at least once. Hence suppose t_0 is the first exit time for this path, so that by continuity, $f_{t_0} \in \partial(\Delta)$. Then since Δ is the intersection of F with the nonnegative quadrant in $\Re^R$, it follows that (for $R \geq 2$) we must have $f_{t_0 r} = 0$ for some component of f_{t_0}. To see that this is not possible, let $P_c(r,\cdot)$ denote the r^{th} row of P_c and let p_{0r} denote the r^{th} component of p_0. Then by the positivity of both $P_c(r,\cdot)f_{t_0}$ and p_{0r}, we see that

$$f_{t_0 r} = 0 \;\Rightarrow\; \dot{f}_{t_0 r} = \Psi_r(f_t) = \lambda P_c(r,\cdot)f_{t_0} + (1-\lambda)p_{0r} > 0$$

By the continuity of Ψ it then follows that $\dot{f}_{tr} > 0$ for all t in some neighborhood $(t_0 - \varepsilon, t_0 + \varepsilon)$ of t_0, and thus that f_{tr} can only increase in this neighborhood. But since t_0 is the first exit time, f_t must then lie in the relative interior of Δ for all $t \in (t_0 - \varepsilon, t_0)$. Hence we must have $f_{tr} > 0$ for all $t \in (t_0 - \varepsilon, t_0)$, and it follows that f_{tr} can only reach zero by decreasing. Thus we obtain a contradiction, and may conclude that no first exit time, t_0, can exist. ■

In analysis to follow, we may thus treat (7.1) as general differential equation on $\Re^R$ subject only to the initial condition that $f_0 \in \Delta$. To characterize the stability properties of this differential equation, we begin with two useful eigenvalue results. If the modulus of each complex number $\omega = a + ib$ is denoted by $|\omega| = \sqrt{a^2 + b^2}$, then we have the following eigenvalue property of contact-probability matrices, P_c:

Lemma 2. If ω is an eigenvalue of P_c, then $|\omega| \leq 1$.

Proof: First observe that since P_c must have the same eigenvalues as its transpose, P_c', we may shift attention to P_c'. Observe next from the column stochasticity of P_c that $P_c'u = u$, and hence that 1 is an eigenvalue of P_c' with eigenvector u. Hence the result follows from the nonnegativity of P_c', which together with the Perron-Frobenius Theorem implies that this positive eigenvalue must have maximum modulus for P_c'. ■

As a direct consequence of this property for P_c we next show that:

Lemma 3. If $\omega = a + ib$ is any eigenvalue of the matrix $\lambda P_c - I_R$, then

$$-(1+\lambda) \leq a \leq \lambda - 1 \tag{7.6}$$

Proof: If ω is an eigenvalue of $\lambda P_c - I_R$ then by definition there is some (possibly complex-valued) vector x with $(\lambda P_c - I_R)x = \omega x$, so that

$$\lambda P_c x - x = \omega x \;\Rightarrow\; P_c x = \left(\frac{1+\omega}{\lambda}\right) x \tag{7.7}$$

Hence $(1+\omega)/\lambda$ must be an eigenvalue of P_c and it follows from Lemma 2 that

$$1 \geq \left|\frac{1+\omega}{\lambda}\right| = \left|\frac{1+a}{\lambda} + i\frac{b}{\lambda}\right| = \sqrt{\left(\frac{1+a}{\lambda}\right)^2 + \left(\frac{b}{\lambda}\right)^2} \tag{7.8}$$

and thus that

$$\begin{aligned} 1 \geq \left|\frac{1+a}{\lambda}\right| \quad &\Rightarrow \quad -1 \leq \frac{1+a}{\lambda} \leq 1 \\ &\Rightarrow \quad -\lambda \leq 1 + a \leq \lambda \\ &\Rightarrow \quad -(1+\lambda) \leq a \leq \lambda - 1 \;\; . \;\; ■ \end{aligned} \tag{7.9}$$

Using these eigenvalue properties we can now establish the fundamental convergence property of (7.1), namely that [as in (3.17)] the fixed point for p,

$$f^* = (1-\lambda)(I_R - \lambda P_c)^{-1} p_0 \tag{7.10}$$

is the (unique) globally asymptotically stable stationary point of (7.1)

Theorem 1.[Deterministic Convergence] *If* $(f_t : t \geq 0)$ *is any solution to* (7.1) *then*

$$\lim\nolimits_{t\to\infty} f_t = f^* \tag{7.11}$$

Proof: It must first be verified that f^* is well defined, i.e., that the matrix $(I_R - \lambda P_c)$ in (7.10) is nonsingular. By rewriting this matrix as

$$(I_R - \lambda P_c) = -\lambda \left(P_c - \frac{1}{\lambda} I_R \right) \tag{7.12}$$

it is clear that nonsingularity of $(I_R - \lambda P_c)$ is equivalent to that of $\left(P_c - \frac{1}{\lambda} I_R\right)$. But since the determinantal equation

$$|P_c - \omega I_R| = 0 \tag{7.13}$$

is precisely the *characteristic equation* of the contact-probability matrix, P_c, it then follows that $\left(P_c - \frac{1}{\lambda} I_R\right)$ is nonsingular (i.e., has a nonzero determinant) iff $\frac{1}{\lambda}$ is *not* an eigenvalue of P_c. Finally since every eigenvalue of P_c must have modulus less than or equal to one by Lemma 2, and since $\lambda \in (0,1) \Rightarrow 1/\lambda > 1$, it follows that $1/\lambda$ cannot be an eigenvalue of P_c. Hence the unique fixed point, f^*, of p in (7.10) is well defined, and is seen from (7.2) to be the unique stationary point of (7.1).

It thus remains to be shown that this stationary point is globally asymptotically stable, i.e., that (7.11) holds. To do so, it is convenient to convert the affine differential equation (7.1) to a linear differential equation by employing the change of variables,

$$z_t = f_t - f^* \ , \quad t \geq 0 \tag{7.14}$$

which, by the fixed-point property of f^*, yields

$$\begin{aligned} \dot{z}_t \ &= \ \dot{f}_t \ = \ [\lambda P_c f_t + (1-\lambda)p_0] - f_t \\ &= \ [\lambda P_c (z_t + f^*) + (1-\lambda)p_0] - (z_t + f^*) \\ &= \ [\lambda P_c z_t - z_t] + [\lambda P_c f^* + (1-\lambda)p_0 - f^*] \\ &= \ \lambda P_c z_t - z_t \ = \ (\lambda P_c - I_R) z_t \end{aligned} \tag{7.15}$$

and thus yields the linear differential equation

$$\dot{z}_t = A z_t \ , \quad t \geq 0 \tag{7.16}$$

with matrix kernel

$$A = \lambda P_c - I_R \tag{7.17}$$

To establish (7.11) it then suffices from (7.14) to show that for every solution $(z_t : t \geq 0)$ of (7.16) we must have

$$\lim_{t\to\infty} z_t = 0 \tag{7.18}$$

But for linear differential equations it is well known [for example, Theorem 6.5.2 in Hirsch and Smale (1974, p.136)] that this is equivalent to showing that all eigenvalues of A have negative real parts. Hence the desired result follows at once from Lemma 3. ■

Given these deterministic convergence results for (7.1), we are ready to state our central result on stochastic convergence to (7.1). If the positive integers are denoted by $\mathbf{N} = \{1, 2, ..\}$, then we now consider stochastic *adoption-frequency sequences* $(f_n : n \in \mathbf{N})$ in Δ generated as in (3.7) by

$$f_{n+1} = \frac{n}{n+1} f_n + \frac{1}{n+1} Y_n \ , \quad n \in \mathbf{N} \tag{7.19}$$

with initial condition, $f_1 = Y_0$, where $(Y_n : n = 0, 1, 2, ..)$ is the sequence of *regional-outcome vectors* for a spatial-mixture process. Next observe that this sequence can be rewritten as:

$$f_{n+1} = f_n + \epsilon_n Z_n \ , \quad n \in \mathbf{N} \tag{7.20}$$

where the stochastic process $(Z_n : n \in \mathbf{N})$ is defined for all n by

$$Z_n = Y_n - f_n \tag{7.21}$$

and where the *step-size sequence* $(\epsilon_n : n \in \mathbf{N})$ is defined by,

$$\epsilon_n = 1/(n+1) \ , \quad n \in \mathbf{N} \tag{7.22}$$

In this form, (7.20) is seen to be an instance of expression (5.1.3) in [KK](with $\theta_n \equiv f_n$ and $Y_n \equiv Z_n$). Notice in particular that this step-size sequence satisfies the critical "slow divergence" condition that $\sum_{n=1}^{\infty} \epsilon_n = \infty$ with $\epsilon_n \to 0$, as in expression (5.1.1) of [KK]. Notice also that since (f_n) always lies in Δ, there is no need for a projection operator, Π_Δ, constraining this sequence to Δ. With this reformulation of (7.19), we are now ready to state our main result:

Theorem 2.[Stochastic Convergence] *If* $(f_n : n \in \mathbf{N})$ *is the adoption-frequency sequence for a spatial-mixture process, and* f^* *is the unique fixed point of the associated state-probability mapping as in* (3.17), *then*

$$\Pr\left(\lim_{n\to\infty} f_n = f^*\right) = 1 \tag{7.23}$$

Proof: Observe first that since (f_n) always lies in the (compact) domain of attraction, Δ, for the asymptotically stable point, f^*, and since the step-size sequence (ϵ_n) has already been observed to satisfy the slow-divergence condition (5.1.1) in [KK], it follows from Theorem 5.2.1 in [KK] that it is enough to verify conditions (A2.1) through (A2.5) in [KK, p.94], which can equivalently be stated in our terms as follows: Condition (A2.1) requires that

$$\sup_n E\left(\|Z_n\|^2\right) < \infty \tag{7.24}$$

and conditions (A2.2) and (A2.3) together state that there must exist a continuous function Φ and sequence of random vectors (φ_n) such that for all $n \in \mathbf{N}$,

$$E\left(Z_{n+1}|f_1, Z_1, .., Z_n\right) = \Phi(f_{n+1}) + \varphi_{n+1} \tag{7.25}$$

Finally, conditions (A2.4) and (A2.5) require respectively that the step-size sequence (ϵ_n) in (7.20) and random vector sequence (φ_n) in (7.25) must satisfy the additional "shrinkage" conditions that

$$\sum_{n=1}^{\infty} \epsilon_n^2 < \infty \tag{7.26}$$

and

$$\Pr\left(\sum_{n=1}^{\infty} \epsilon_n \left|\varphi_n\right| < \infty\right) = 1 \tag{7.27}$$

To verify these conditions, we first recall from (3.6) together with the recursive construction of (f_n) in (7.19) that for all $n \in \mathbf{N}$,

$$E\left(Y_{n+1}|f_1, Y_1, .., Y_n\right) = E(Y_{n+1}|f_{n+1}) = P(f_{n+1}) \tag{7.28}$$

Hence, observing from (7.19) that f_{n+1} is totally determined by $(f_1, Y_1, .., Y_n)$, and that Y_{n+1} depends on the past only through f_{n+1}, it follows from (7.21) together with the definition of Ψ that

$$\begin{aligned} E\left(Z_{n+1}|f_1, Z_1, .., Z_n\right) &= E(Y_{n+1} - f_{n+1}|f_1, Y_1 - f_1, .., Y_n - f_n) \\ &= E(Y_{n+1}|f_1, Y_1, .., Y_n) - E(f_{n+1}|f_1, Y_1, .., Y_n) \\ &= E(Y_{n+1}|f_{n+1}) - f_{n+1} \\ &= P(f_{n+1}) - f_{n+1} \\ &= \Psi(f_{n+1}) \end{aligned} \tag{7.29}$$

Hence by setting $\Phi = \Psi$ and $\varphi_n \equiv 0$, it follows at once from the continuity of Ψ that (7.25) holds for these choices. Moreover, from the well-known convergence result

$$\sum\nolimits_{n=1}^{\infty} \frac{1}{(n+1)^2} < \infty \tag{7.30}$$

we also see that conditions (7.26) and (7.27) hold for our choices of (ϵ_n) and (φ_n). Finally to verify (7.24), observe if we set $\Delta_{\text{sup}} = \sup\{\|f\| : f \in \Delta\}$ then by the boundedness of Δ, $\Delta_{\text{sup}} < \infty$. But since the sequences (Y_n) and (f_n) in (7.21) both lie in Δ, we may then conclude that

$$\|Z_n\| = \|Y_n - f_n\| \leq \|Y_n\| + \|f_n\| \leq 2\Delta_{\text{sup}} \tag{7.31}$$

and hence that

$$\sup\nolimits_n E\left(\|Z_n\|^2\right) \leq 4\left(\Delta_{\text{sup}}\right)^2 < \infty \tag{7.32}$$

Thus all conditions are satisfied, and the result is established. ■

7.2. Estimation Details

In this final section we give explicit developments of the maximum-likelihood, EM, and MAP estimation procedures discussed the text. First we consider the maximum-likelihood estimates discussed in section 4.1 above.

7.2.1. Maximum-Likelihood Estimation

For convenience, we restate a number of relevant definitions in the text. First let the row vector of intrinsic attributes for region r be denoted by $x_r = (x_{rj} : j = 1, .., J)$, and let the realized adoption data again be denoted by y_0 for the first adopter, and $y = (y_n : n = 1, .., N)$ for all subsequent adopters. Next let the *intrinsic probability* of r be denoted by

$$p_\beta(r) = \frac{M_r \exp(\beta' x_r)}{\sum_{s=1}^{R} M_s \exp(\beta' x_s)} \tag{7.33}$$

and let the *contact probability* of r given relative frequency vector $f = (f_s : s = 1, .., R)$ be denoted by

$$\begin{aligned} p_\theta(r|f) &= \sum\nolimits_{s\in\mathbf{R}} \frac{M_r \exp(-\theta c_{sr})}{\sum_{v\in\mathbf{R}} M_v \exp(-\theta c_{sv})} f_s \\ &= \sum\nolimits_{s\in\mathbf{R}} P_\theta(r,s) f_s \end{aligned} \tag{7.34}$$

where the contact probability matrix, $P_\theta = [P_\theta(r,s)]$, in (3.1) is now indexed by its parameter θ. The *mixture probability* of r given f together with the parameter vector, $\phi = (\beta, \lambda, \theta)$ is then given [as in (2.11)] by

$$p_\phi(r|f) = \lambda p_\theta(r|f) + (1-\lambda)p_\beta(r) \tag{7.35}$$

In these terms, the log likelihood function in (4.4) has the form

$$\begin{aligned} L(\phi|y) &= L(\beta,\lambda,\theta|y) \;=\; \log p_\beta(y_0) + \textstyle\sum_{n=1}^N \log p_\phi(r|f)) \\ &= \log p_\beta(y_0) + \textstyle\sum_{n=1}^N \log\left[\lambda p_\theta(r|f) + (1-\lambda)p_\beta(r)\right] \end{aligned} \tag{7.36}$$

If the relative relative frequency vector generated by events $(y_0, y_1, .., y_n)$ in (2.8) is again denoted by f_n, then the gradient of L is given by $\nabla_\phi L = (\nabla_\beta L', \nabla_\lambda L, \nabla_\theta L)'$, where

$$\nabla_\beta L = (x'_{y_0} - \textstyle\sum_{s\in\mathbf{R}} p_\beta(s)x'_s) + \textstyle\sum_{n=1}^N \dfrac{(1-\lambda)p_\beta(y_n)}{p_\phi(y_n|f_n)}(x'_{y_n} - \textstyle\sum_{s\in\mathbf{R}} p_\beta(s)x'_s) \tag{7.37}$$

$$\nabla_\lambda L = \textstyle\sum_{n=1}^N \dfrac{p_\theta(y_n|f_n) - p_\beta(y_n)}{p_\phi(y_n|f_n)} \tag{7.38}$$

$$\nabla_\theta L = \textstyle\sum_{n=1}^N \dfrac{\lambda}{p_\phi(y_n|f_n)} \textstyle\sum_{s\in\mathbf{R}} P_\theta(y_n,s)f_n(s)\left[\textstyle\sum_{v\in\mathbf{R}} P_\theta(v,s)c_{sv} - c_{sy_n}\right] \tag{7.39}$$

The terms of the Hessian matrix

$$H_\phi = \begin{pmatrix} \nabla_{\beta\beta}L & \nabla_{\lambda\beta}L' & \nabla_{\theta\beta}L' \\ \nabla_{\lambda\beta}L & \nabla_{\lambda\lambda}L & \nabla_{\lambda\theta}L \\ \nabla_{\theta\beta}L & \nabla_{\lambda\theta}L & \nabla_{\theta\theta}L \end{pmatrix} \tag{7.40}$$

are somewhat more complex. We begin with the important diagonal components, and start with the simplest of these, namely $\nabla_{\lambda\lambda}L$. Partial differentiation of (7.38) yields

$$\nabla_{\lambda\lambda}L = -\textstyle\sum_{n=1}^N \left(\dfrac{p_\theta(y_n|f_n) - p_\beta(y_n)}{p_\phi(y_n|f_n)}\right)^2 < 0 \tag{7.41}$$

and shows that L is *strictly concave* in λ [as already observed following (4.5) in the text]. Next, if for each n we let $\alpha_n = (1-\lambda)p_\beta(y_n)/p_\phi(y_n|f_n) \in (0,1)$, then partial differentiation of (7.37) with respect to β yields (after some reduction):

$$\nabla_{\beta\beta}L = A_1 - A_2 \tag{7.42}$$

where

$$\begin{aligned} A_1 &= \sum\nolimits_{n=1}^{N} [\alpha_n(1-\alpha_n)] \left(x'_{y_n} - \sum\nolimits_{s\in\mathbf{R}} p_\beta(s)x'_s\right)\left(x_{y_n} - \sum\nolimits_{s\in\mathbf{R}} p_\beta(s)x_s\right) \\ A_2 &= \left(1+\sum\nolimits_{n=1}^{N}\alpha_n\right)\cdot \\ &\quad \left[\sum\nolimits_{s\in\mathbf{R}} p_\beta(s)x'_s x_s - \left(\sum\nolimits_{s\in\mathbf{R}} p_\beta(s)x'_s\right)\left(\sum\nolimits_{s\in\mathbf{R}} p_\beta(s)x_s\right)\right] \end{aligned}$$

Notice first that A_1 is a nonnegative weighted sum of positive semidefinite (psd) matrices, and hence is psd. Moreover, observe that square-bracketed term in A_2 is the (psd) covariance matrix of a random vector with outcomes (x_s) and associated probabilities $p_\beta(s)$. Hence B is psd by the same argument. From (7.42) we then see that $\nabla_{\beta\beta}L$ is the difference of two psd matrices, and thus is generally *not* negative semidefinite (nsd). Moreover, the first-order condition, $0 = \nabla_\beta L$, is seen from (7.37) to offer little additional insight into the negative semidefiniteness of $\nabla_{\beta\beta}L$ at singular points. In fact, examples show that L can be multimodal in β (see section 7.2.3 below). Turning finally to $\nabla_{\theta\theta}L$, partial differentiation of (7.39) yields with respect to θ yields

$$\nabla_{\theta\theta}L = B_1 - (B_2 + B_3) \tag{7.43}$$

where the three matrices:

$$\begin{aligned} B_1 &= \sum\nolimits_{n=1}^{N} \frac{\lambda}{p_\phi(y_n|f_n)} \sum\nolimits_{s\in\mathbf{R}} P_\theta(y_n,s)f_n(s)\left(\sum\nolimits_{v\in\mathbf{R}} P_\theta(v,s)c_{sv} - c_{sy_n}\right)^2 \\ B_2 &= \sum\nolimits_{n=1}^{N} \frac{\lambda}{p_\phi(y_n|f_n)} \sum\nolimits_{s\in\mathbf{R}} P_\theta(y_n,s)f_n(s)\left[\sum\nolimits_{v\in\mathbf{R}} P_\theta(v,s)c_{sv}^2\right. \\ &\qquad \left. - \left(\sum\nolimits_{v\in\mathbf{R}} P_\theta(v,s)c_{sv}\right)^2\right] \\ B_3 &= \sum\nolimits_{n=1}^{N} \left(\frac{\lambda}{p_\phi(y_n|f_n)}\right)^2 \left(\sum\nolimits_{s\in\mathbf{R}} P_\theta(y_n,s)f_n(s)\left[\sum\nolimits_{v\in\mathbf{R}} P_\theta(v,s)c_{sv} - c_{sy_n}\right]^2\right) \end{aligned}$$

are all seen to be nonnegative. Hence it should be clear that the sign of this (scalar) second derivative is completely undetermined. As with β, examples show that the L can be multimodal in θ (see section 7.2.3 below). To determine the off-diagonal terms, observe first that the partial of (7.37) with respect to λ yields (after some reduction)

$$\nabla_{\lambda\beta}L = -\sum\nolimits_{n=1}^{N} \frac{p_\beta(y_n)p_\theta(y_n|f_n)}{p_\phi(y_n|f_n)^2}(x'_{y_n} - \sum\nolimits_{s\in\mathbf{R}} p_\beta(s)x'_s) \tag{7.44}$$

and similarly, the partial of (7.37) with respect to θ yields

$$\begin{aligned}\nabla_{\theta\beta}L &= -\lambda(1-\lambda)\sum\nolimits_{n=1}^{N}\left[\frac{p_\beta(y_n)}{p_\phi(y_n|f_n)^2}(x'_{y_n}-\sum\nolimits_{v\in\mathbf{R}}p_\beta(v)x'_v)\cdot\right.\\ &\left.\sum\nolimits_{s\in\mathbf{R}}P_\theta(y_n,s)f_n(s)\left(\sum\nolimits_{v\in\mathbf{R}}P_\theta(v,s)c_{sv}-c_{sy_n}\right)\right] \quad (7.45)\end{aligned}$$

Finally the partial of (7.38) with respect to λ yields.

$$\begin{aligned}\nabla_{\lambda\theta}L &= \sum\nolimits_{n=1}^{N}\frac{p_\beta(y_n)}{p_\phi(y_n|f_n)^2}\sum\nolimits_{s\in\mathbf{R}}\left[P_\theta(y_n,s)f_n(s)\cdot\right.\\ &\left.\left(\sum\nolimits_{v\in\mathbf{R}}P_\theta(v,s)c_{sv}-c_{sy_n}\right)\right] \quad (7.46)\end{aligned}$$

In regions where H_ϕ is negative definite (i.e., has all negative eigenvalues), one can employ these results to construct Newton-Raphson increments, $-H_\theta^{-1}\nabla_\phi L$. Otherwise, the simple gradient, $\nabla_\phi L$, can be used.

However, there are several additional complications in this type of gradient procedure that should be mentioned. First, step sizes must of course be restricted sufficiently to ensure that only logs of positive quantities are involved in evaluations of the objective function L. A special problem related to θ is that the objective function can often be extremely flat in the θ direction, thus slowing down convergence. Here is was found to be efficient to do fairly broad line-search maximization with respect to θ on each step.

7.2.2. EM Algorithm

Turning next to the EM algorithm in section 4.2 above, it is here postulated that the outcome of the first stage for the n^{th} adoption is a Bernoulli random variable, δ_n, with $\Pr(\delta_n = 1) = \lambda$. The explicit form of the joint density in (4.6) is

$$\begin{aligned}p(y,\delta;\beta,\lambda,\theta) &= p(y_0;\beta)\prod\nolimits_{n=1}^{N}p\left(y_n|\ \delta_n,y_0,..,y_{n-1};\beta,\theta\right)p(\delta_n;\lambda)\\ &= p_\beta(y_0)\prod\nolimits_{n=1}^{N}p_\theta\left(y_n|\ f_n\right)^{\delta_n}p_\beta(y_n)^{1-\delta_n}\left[\lambda^{\delta_n}(1-\lambda)^{1-\delta_n}\right] \quad (7.47)\end{aligned}$$

with associated log likelihood

$$\begin{aligned} L(\beta, \lambda, \theta | y, \delta) &= \log p_\beta(y_0) + \sum\nolimits_{n=1}^{N} \left[\delta_n \log p_\theta\left(y_n |\ f_n\right) + (1-\delta_n) \log p_{\beta(y_n)}\right] \\ &\quad + \sum\nolimits_{n=1}^{N} \left[\delta_n \log \lambda + (1-\delta_n) \log(1-\lambda)\right] \\ &= \log p_\beta(y_0) + \sum\nolimits_{n=1}^{N} \left\{\delta_n \left[\log p_\theta\left(y_n |\ f_n\right) - \log p_{\beta(y_n)}\right] + \log p_{\beta(y_n)}\right\} \\ &\quad + \sum\nolimits_{n=1}^{N} \left\{\delta_n \left[\log \lambda - \log(1-\lambda)\right] + \log(1-\lambda)\right\} \qquad (7.48) \end{aligned}$$

Hence, for any given set of parameter estimates, $\phi_k = (\beta_k, \lambda_k, \theta_k)$, the *E-step* for this procedure consists of calculating the expectation of (7.48) with respect to δ_n under the parameter values ϕ_k. Since y is taken to be given, we must first determine the conditional distribution, $p_{\phi_k}(\delta_n | y)$, of δ_n given y under ϕ_k. Note that the only relevant parts of y for δ_n are the adoption outcome, y_n, and the current relative frequency distribution, f_n, of previous adopters. Hence $p_{\phi_k}(\delta_n | y) = p_{\phi_k}(\delta_n | y_n, f_n)$, and it follows by definition that for $\delta_n = 1$,

$$\begin{aligned} p_{\phi_k}(\delta_n = 1 | y_n, f_n) p_{\phi_k}(y_n | f_n) &= p_{\phi_k}(\delta_n = 1, y_n | f_n) \\ &= p_{\phi_k}(y_n | \delta_n = 1, f_n) p_{\phi_k}(\delta_n = 1 | f_n) \\ &= p_{\theta_k}(y_n | f_n) \lambda \qquad (7.49) \end{aligned}$$

Similarly, for $\delta_n = 0$,

$$\begin{aligned} p_{\phi_k}(\delta_n = 0 | y_n, f_n) p_{\phi_k}(y_n | f_n) &= p_{\phi_k}(\delta_n = 0, y_n | f_n) \\ &= p_{\phi_k}(y_n | \delta_n = 0, f_n) p_{\phi_k}(\delta_n = 0 | f_n) \\ &= p_{\beta_k}(y_n)(1-\lambda) \qquad (7.50) \end{aligned}$$

and we see by taking ratios of both sides of (7.49) and (7.50) that

$$\frac{p_{\phi_k}(\delta_n = 1 | y_n, f_n)}{p_{\phi_k}(\delta_n = 0 | y_n, f_n)} = \frac{p_{\theta_k}(y_n | f_n) \lambda}{p_{\beta_k}(y_n)(1-\lambda)} \qquad (7.51)$$

This together with the identity $1 = p_{\phi_k}(\delta_n = 1 | y_n, f_n) + p_{\phi_k}(\delta_n = 0 | y_n, f_n)$, then yields

$$p_{\phi_k}(\delta_n = 1 | y_n, f_n) = \frac{\lambda p_{\theta_k}(y_n | f_n)}{\lambda p_{\theta_k}(y_n | f_n) + (1-\lambda) p_{\beta_k}(y_n)}$$

Thus, if we denote the conditional expectation of δ_n given (y_n, f_n) under ϕ_k by π_n^k, then (as with all Bernoulli variates)

$$\begin{aligned}\pi_n^k = E_{\phi_k}(\delta_n|y_n, f_n) &= p_{\phi_k}(\delta_n = 1|y_n, f_n) \\ &= \frac{\lambda p_{\theta_k}(y_n|f_n)}{\lambda p_{\theta_k}(y_n|f_n) + (1-\lambda)p_{\beta_k}(y_n)}\end{aligned}$$

Finally, to obtain the desired expectation, we need only observe that the right hand side of (7.48) is *linear* in δ_n, so that

$$\begin{aligned}&E_{\phi_k}\left[L(\beta,\lambda,\theta|y,\delta)\right] = \log p_\beta(y_0) \\ &\quad + \sum\nolimits_{n=1}^N E_{\phi_k}(\delta_n|y_n, f_n)\left[\log p_\theta\left(y_n|\ f_n\right) - \log p_\beta(y_n)\right] + \log p_\beta(y_n) \\ &\quad + \sum\nolimits_{n=1}^N \left\{E_{\phi_k}(\delta_n|y_n, f_n)\left[\log\lambda - \log(1-\lambda)\right] + \log(1-\lambda)\right\} \\ &\quad = \log p_\beta(y_0) + \sum\nolimits_{n=1}^N \left\{\pi_n^k\left[\log p_\theta\left(y_n|\ f_n\right) - \log p_\beta(y_n)\right] + \log p_\beta(y_n)\right\} \\ &\quad + \sum\nolimits_{n=1}^N \left\{\pi_n^k\left[\log\lambda - \log(1-\lambda)\right] + \log(1-\lambda)\right\} \qquad (7.52)\end{aligned}$$

Turning next to the *M-Step* of the procedure, the expectation in (7.52) constitutes the relevant objective function for this step. Hence if we now let

$$Z_1^k(\beta) = \log p_\beta(y_0) + \sum\nolimits_{n=1}^N (1-\pi_n^k)\log p_\beta(y_n) \qquad (7.53)$$

$$Z_2^k(\theta) = \sum\nolimits_{n=1}^N \pi_n^k \log p_\theta\left(y_n|\ f_n\right) \qquad (7.54)$$

$$\begin{aligned}Z_3^k(\lambda) &= \sum\nolimits_{n=1}^N \left\{\pi_n^k\left[\log\lambda - \log(1-\lambda)\right] + \log(1-\lambda)\right\} \\ &= (\log\lambda)\sum\nolimits_{n=1}^N \pi_n^k + \log(1-\lambda)\left(N - \sum\nolimits_{n=1}^N \pi_n^k\right) \qquad (7.55)\end{aligned}$$

then it follows that this objective function can be written as

$$Z^k(\beta,\lambda,\theta) = Z_1^k(\beta) + Z_2^k(\theta) + Z_3^k(\lambda) \qquad (7.56)$$

and hence is a *separable additive* in β, λ, and θ. We start by maximizing Z_3^k with respect to λ, which turns out to be expressible in closed form as follows:

$$\nabla_\lambda Z_3^k = \frac{1}{\lambda}\sum\nolimits_{n=1}^N \pi - \frac{1}{1-\lambda}\left(N - \sum\nolimits_{n=1}^N \pi_n^k\right) = 0$$

$$\Rightarrow \quad \lambda^* = \frac{1}{N}\sum\nolimits_{n=1}^{N} \pi_n^k \tag{7.57}$$

Moreover, one may readily verify that the second derivative is negative, and hence that λ^* uniquely maximizes this concave function. To maximize Z_1^k with respect to β, observe that by substituting (7.33) into (7.53) and taking partial derivatives:

$$\begin{aligned} \nabla_\beta Z_1^k &= x'_{y_0} + \sum\nolimits_{n=1}^{N} \left(1-\pi_n^k\right) x'_{y_n} \\ &\quad - \left(N+1-\sum\nolimits_{n=1}^{N} \pi_n^k\right)\left(\sum\nolimits_{s\in\mathbf{R}} p_\beta(s) x'_s\right) \end{aligned} \tag{7.58}$$

Similarly, by substituting (7.34) into (7.54) and differentiating, we obtain (after some manipulation):

$$\nabla_\theta Z_2^k = \sum\nolimits_{n=1}^{N} \pi_n^k \sum\nolimits_{s\in\mathbf{R}} \frac{P_\theta(y_n,s) f_n(s)}{\sum_{v\in\mathbf{R}} P_\theta(y_n,v) f_n(v)} \left[\sum\nolimits_{v\in\mathbf{R}} P_\theta(v,s) c_{sv} - c_{sy_n}\right] \tag{7.59}$$

Before proceeding to the second-order conditions, it is of interest to consider the behavior of $\nabla_\theta Z_2^k$ in this M-step. Recall that while the EM algorithm embodies a natural constraint on the admissible range of λ, no such restriction is placed on θ. Hence a natural way to investigate the sign of θ here is to consider the sign of $\nabla_\theta Z_2^k$ evaluated at $\theta = 0$. If this sign is *always* positive then this must guarantee that the optimal value of θ is positive. To see what this means in the present case, observe from (3.1) that $P_0(r,s) = M_r / \sum_{s\in\mathbf{R}} M_s$, and hence that

$$\sum\nolimits_{v\in\mathbf{R}} P_0(v,s) c_{sv} = \frac{\sum_{v\in\mathbf{R}} M_v c_{sv}}{\sum_{v\in\mathbf{R}} M_v} = \bar{c}_s \tag{7.60}$$

is simply the *average contact cost* from region s to all individuals in the system. In these terms we have

$$\begin{aligned} \nabla_\theta Z_2^k\big|_{\theta=0} &= \sum\nolimits_{n=1}^{N} \pi_n^k \sum\nolimits_{s\in\mathbf{R}} \frac{M_{y_n} f_n(s)}{\sum_{v\in\mathbf{R}} M_{y_n} f_n(v)} \left[\bar{c}_s - c_{sy_n}\right] \\ &= \sum\nolimits_{n=1}^{N} \pi_n^k \left\{\sum\nolimits_{s\in\mathbf{R}} f_n(s)\bar{c}_s - \sum\nolimits_{s\in\mathbf{R}} f_n(s) c_{sy_n}\right\} \end{aligned} \tag{7.61}$$

The first term in brackets is seen to be the average contact cost from current adopters to all individuals in the system, and the second term is the average contact cost from current adopters to individuals the region where the n^{th} adoption occurs. Hence positivity of all these terms guarantees that $\nabla_\theta Z_2^k\big|_{\theta=0}$ will be positive. In other words, if average contact costs to each successive adopter are "lower than expected" when $\theta = 0$, then the most likely value of θ should be positive (indicating sensitivity to contact costs). More generally if these contact costs *tend* to be lower than expected, then positivity should still hold.

It is also worth noticing here that for any given parameters $\phi_k = (\beta_k, \lambda_k, \theta_k)$, the values of the gradients in (7.58) and (7.59) are exactly the same as those for the likelihood function in (7.37) and (7.39) above. For example, (7.39) can be rewritten [by adding $p_\theta(y_n|f_n)$ to both numerators and denominators] as

$$\nabla_\theta L \;=\; \sum\nolimits_{n=1}^{N} \frac{\lambda p_\theta(y_n|f_n)}{p_\phi(y_n|f_n)} \sum\nolimits_{s\in\mathbf{R}} \frac{P_\theta(y_n,s)f_n(s)}{p_\theta(y_n|f_n)} \left[\sum\nolimits_{v\in\mathbf{R}} P_\theta(v,s)c_{sv} - c_{sy_n}\right] \tag{7.62}$$

But since

$$\frac{\lambda p_{\theta_k}(y_n|f_n)}{p_{\phi_k}(y_n|f_n)} = \frac{\lambda p_{\theta_k}(y_n|f_n)}{\lambda p_{\theta_k}(y_n|f_n) + (1-\lambda)p_{\theta_k}(y_n)} = \pi_n^k \tag{7.63}$$

and since by (7.34),

$$p_\theta(y_n|f_n) = \sum\nolimits_{v\in\mathbf{R}} P_\theta(y_n,v)f_n(v) \tag{7.64}$$

it follows that $\nabla_\theta L|_{\phi=\phi_k}$ is precisely the same as $\nabla_\theta Z_2^k$. A similar argument shows that $\nabla_\beta L|_{\phi=\phi_k} = \nabla_\beta Z_1^k$. Hence from a gradient viewpoint, this M-step is seen to be essentially the same as partial likelihood maximization with respect to (β,θ). However, since π_n^k is *fixed* in the gradient for each M-step, while $\lambda p_\theta(y_n|f_n)/p_\phi(y_n|f_n)$ is *variable* in the gradient function $\nabla_\theta L$, it should also be clear that the second-order properties of these two maximization problems are generally quite different, and in particular that those for the M-step should be simpler. In particular we see here that

$$\begin{aligned}\nabla_{\beta\beta} Z_1^k \;=\; & -\left(N + 1 - \sum\nolimits_{n=1}^{N} \pi_n^k\right)\cdot \\ & \left\{\sum\nolimits_{s\in\mathbf{R}} p_\beta(s)x_s' x_s - \left(\sum\nolimits_{s\in\mathbf{R}} p_\beta(s)x_s\right)\left(\sum\nolimits_{s\in\mathbf{R}} p_\beta(s)x_s'\right)\right\}\end{aligned} \tag{7.65}$$

and hence that Z_1^k is now *negative semidefinite* in β [by the same argument as that for A_2 in (7.42)]. But the second partial with respect to θ is unfortunately not much simpler than that for maximum-likelihood. Here, if we let

$$\alpha_\theta(s,r) = \frac{P_\theta(r,s)f_n(s)}{\sum_{v\in\mathbf{R}} P_\theta(r,v)f_n(v)} \tag{7.66}$$

and

$$z_\theta(s,r) = \sum\nolimits_{v\in\mathbf{R}} P_\theta(v,s)c_{sv} - c_{sr} \tag{7.67}$$

then in terms of this notation,

$$\nabla_{\theta\theta} Z_1^k = C_1 - (C_2 + C_3) \tag{7.68}$$

where the three terms

$$\begin{aligned}
C_1 &= \sum\nolimits_{n=1}^{N} \pi_n^k \left\{ \sum\nolimits_{s\in\mathbf{R}} \alpha_\theta(s,y_n) z_\theta(s,y_n)^2 - \left(\sum\nolimits_{s\in\mathbf{R}} \alpha_\theta(s,y_n) z_\theta(s,y_n) \right)^2 \right\} \\
C_2 &= \sum\nolimits_{n=1}^{N} \pi_n^k \sum\nolimits_{s\in\mathbf{R}} \alpha_\theta(s,y_n) \left[\sum\nolimits_{v\in\mathbf{R}} P_\theta(v,s)c_{sv}^2 - \left(\sum\nolimits_{v\in\mathbf{R}} P_\theta(v,s)c_{sv} \right)^2 \right] \\
C_3 &= \sum\nolimits_{n=1}^{N} \pi_n^k \left[\sum\nolimits_{s\in\mathbf{R}} \alpha_\theta(s,y_n) c_{sy_n} \right]
\end{aligned}$$

are all nonnegative [since the bracketed terms in C_1 and C_2 are again seen to be the variances of appropriately defined random variables]. Hence the sign of $\nabla_{\theta\theta} Z_1^k$ continues to be indeterminate.

Separable additivity implies that maximization with respect to β and θ can be done separately. As with the gradient procedure for maximum likelihood, it generally proved to be more efficient to carry out a fairly broad line-search maximization for θ than to rely on gradient steps (which tend to be rather small). However even with Newton-Raphson steps for β and line searches for θ, the EM algorithm is very slow to converge. Hence the following more efficient MAP estimation procedure is recommended for the present model.

7.2.3. MAP Estimation

For the Bayesian approach discussed in section 4.3 above, the mixture parameter, λ, was postulated to be a random variable with symmetric prior Beta distribution, $\pi(\lambda) \propto \lambda^{a-1}(1-\lambda)^{a-1}$, where the parameter, $a > 1$, is treated as a smoothing parameter. In addition it was assumed that β and θ were distributed with flat priors, yielding the joint posterior density in (4.14) with log posterior given [as in (4.15)] by:

$$\Phi(\beta,\lambda,\theta|y) = L(\beta,\lambda,\theta|y) + (a-1)\left[\log\lambda + \log(1-\lambda)\right] \tag{7.69}$$

Maximization of (7.69) to estimate the posterior mode, i.e., *MAP estimation*, is thus equivalent to penalized maximization of (7.36) with λ-penalty function given by the second term in (7.69). Hence the gradient and Hessian for (7.69) are obtainable almost directly from those of (7.36). In particular, it follows at once that

$$\nabla_\beta \Phi = \nabla_\beta L \tag{7.70}$$

$$\nabla_\theta \Phi = \nabla_\theta L \tag{7.71}$$

$$\nabla_\lambda \Phi = \nabla_\lambda L + (a-1)\left[\frac{1}{\lambda} - \frac{1}{1-\lambda}\right] \tag{7.72}$$

where the gradients of L with respect to β, λ, and θ are given respectively by (7.37), (7.38), and (7.39). Similarly all terms of the Hessian matrix for Φ are exactly those of (7.40) except for the second partial of Φ with respect to λ, which now takes the form:

$$\nabla_{\lambda\lambda} \Phi = \nabla_{\lambda\lambda} L - (a-1)\left[\frac{1}{\lambda^2} + \frac{1}{(1-\lambda)^2}\right] \tag{7.73}$$

But since $\nabla_{\lambda\lambda} L < 0$ by (7.41) and since the second term in (7.73) is negative, it again follows that Φ is strictly concave in λ. Moreover, since Φ approaches $-\infty$ as λ approaches either 0 or 1, it then follows that for any values of (β, θ), Φ always achieves a unique maximum in λ on $(0,1)$.

Note also that for a close to one [$a = 1.01$ in Figure 4.1] and for $\lambda \in (0,1)$ [but not too close to 0 or 1], both gradient and Newton-Raphson steps are almost the same in both maximization procedures. The key distinctions involve cases where L achieves it maximum in λ outside $[0,1]$. Here it should be clear that the MAP estimate of λ will be close to 0 or 1 (with degree of closeness depending on the smoothing parameter a). To gain further insight into these cases, it is useful to look more closely at L, which can be written in terms of (4.5) as

$$L(\beta, \lambda, \theta | y) = \log p_\beta(y_0) + \sum\nolimits_{n=1}^{N} \log\left\{\lambda p_\theta(y_n | f_n) + (1-\lambda) p_\beta(y_n)\right\} \tag{7.74}$$

Notice here that as λ approaches zero, all terms involving θ vanish. Thus it is not surprising that estimates of θ become extremely unstable when λ is close to zero [as for example in the case depicted in Table 4.1 of the text]. Similarly, when λ approaches one, all terms involving β except the first term vanish. So when λ is close to one, only the initial observation, y_0, provides much information about β, and estimates of β can be expected to be unstable [as reflected by the example shown in Table 4.3 of the text].

As mentioned in the text, an additional difficulty that can arise is the possibility of multiple maxima for the objective function Φ. For the simulations summarized in Table 4.2 of the text, multiple maxima occurred with some frequency for sample sizes $N \leq 200$. In all such cases there were two local maxima. These usually produced similar values of the β coefficients, but radically different pairs of (λ, θ) values. One example (with $N = 200$) is illustrated in Table 5.1 below. Notice in particular the radical difference between the two values of λ.

	Global Max		**Local Max**	
Par	Estimate	P-value	Estimate	P-value
β_1	-5.271	< 0.0000	-7.449	0.921
β_2	3.939	0.0001	-27.541	0.825
λ	0.031	< 0.0000	0.911	< 0.000
θ	-5.407	0.00002	-11.676	< 0.000

Table 7.1: Case of Multiple Maxima

This illustrates a case where the two modes do indeed correspond to "best local fits" of the intrinsic model and contact model, respectively. Moreover, while the intrinsic mode is somewhat more likely, both modes are seen in Figure 7.1 below to be quite prominent.[32] So, especially for small sample sizes, it is imperative to try a number of starting points in the MAP estimation procedure to identify possible multiple modes.

Finally, as asymptotic theory predicts, when the number of adoptions, N, is large (and the true value of the mixture parameter, λ, is not too close to 0 or 1), the maximum-likelihood estimates of (β, λ, θ) are generally well behaved, and there is little need for alternative estimation procedures.[33] In the simulations of Table 4.2 in the text, this turned out to be the case for $N \geq 1000$. But since these simple simulations involved only *four* unknown parameters, this serves to underscore the limitations of direct maximum-likelihood estimation in the present context.

[32] This plot shows the log-likelihood values along a line in the parameter space through the two modes. If the parameter vectors for the two modes are denoted by ϕ^1 and ϕ^2, then the values plotted are $L\left[\alpha\phi^1 + (1-\alpha)\phi^2|y\right]$ for $\alpha \in [-0.1, 1.1]$. Hence the two modes correspond respectively to the points $\alpha = 0$ and $\alpha = 1$.

[33] As discussed in MacLachlan(2000, section 2.5), as long as such estimates are uniquely identifiable, they are guaranteed to be strongly consistent under very general conditions.

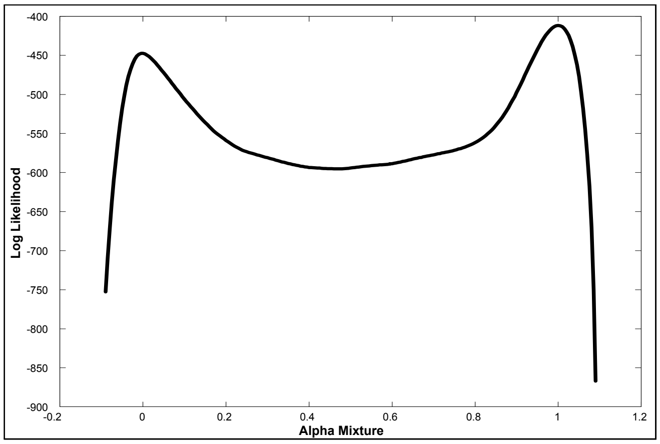

Figure 7.1: Plot of Multiple Maxima

References

[1] Bass, F.M. (1969) “A new product growth model for consumer durables”, *Marketing Science*, 15: 215-227.

[2] Brumelle, S.L. and Y. Gerchak (1980) “A stochastic model allowing interaction among individuals and its behavior for large populations”, *Journal of Mathematical Sociology*, 7: 73-90.

[3] Dempster, A., N. Laird, and D. Rubin (1977) “Maximum likelihood from incomplete data via the EM algorithm”, *Journal of the Royal Statistical Society B*, 39:1-38.

[4] Hägerstrand, T. (1967), *Innovation Diffusion as a Spatial Process*, University of Chicago Press (A. Pred, trans.).

[5] Haining, R. (1983) “Spatial and spatial-temporal interaction models and the analysis of patterns of diffusion”, *Transactions of the Institute of British Geographers*, 8:158-186.

[6] Hastie, T., R. Tibshirani, and J. Friedman (2001), *The Elements of Statistical Learning*, Springer-Verlag: New York.

[7] Hirsch, M.W. and S. Smale (1974) *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press: New York.

[8] Kulkarni, V.G. (1995), *Modeling and Analysis of Stochastic Systems*, Chapman & Hall: New York.

[9] Kushner, H.J. and Clark, D. (1978), *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag: Berlin.

[10] Kushner, H.J. and G.G. Kim (1997), *Stochastic Approximation Algorithms and Applications*, Springer-Verlag: Berlin.

[11] McLachlan, G. and D. Peel (2000), *Finite Mixture Models*, Wiley: New York.

[12] Mahajan, V., E. Muller, and F.M. Bass (1990), "New product diffusion models in marketing: A review and directions for research", *Journal of Marketing*, 54: 1-26.

[13] Morrill, R.,G.L. Gaile, and G.I. Thrall (1988), *Spatial Diffusion*, Sage Publications: Newbury Park, CA.

[14] Lehmann, E.L. (1983) *Theory of Point Estimation*, Wiley: New York

[15] Robert, C.H. (1998) "Mixtures of distributions: inference and estimation" in *Markov Chain Monte Carlo in Practice*, edited by W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, Chapman & Hall: Boca Raton, Fla.

[16] Rogers, E. M. (1995) *Diffusion of Innovations*, 4^{th} Edition, Free Press: New York.

[17] Speckman, P.L., J. Lee, and D. Sun (2001), "Existence of the MLE and propriety of posteriors for a general multinomial choice model" (1999), *http://www.stat.missouri.edu/~dsun/submit.html.*

[18] Stephens, M. (2000), "Dealing with label-switching in mixture models", *Journal of the Royal Statistical Society, Series B*, 62: 795–809

[19] Strang, David and Nancy B. Tuma (1993) "Spatial and Temporal Heterogeneity in Diffusion", *American Journal of Sociology*, 99 (Novemeber), 614-639.

[20] Wedel, M. and W.A. Kamakura (2000), *Market Segmentation: Conceptual and Methodological Foundations*, *2^{nd} ed.*, Kluwer Academic Press: Dordrecht, The Netherlands.