LEARNING FROM EXPERIENCE TO IMPROVE EARLY FORECASTS: A POSTERIOR MODE APPROACH[a]

Siddhartha R. Dalal and Yu-Yun K. Ho, Bell Communications Research

Robert P. Sherman, California Institute of Technology

ABSTRACT

This paper presents a new Bayesian method that can incorporate knowledge gained from past experience to improve early forecasting of new products and services. The method is based on maximum likelihood estimation of a pure birth process with a Bass-type adoption rate that can capture price and promotion effects. Direct maximum likelihood estimation of the model parameters is numerically infeasible. We develop an indirect estimator using a Monte-Carlo EM algorithm and Gibbs sampling. We also develop procedures for estimating standard errors of parameter estimates and forecasts. We apply the method to data on a new telecommunications service.

I. INTRODUCTION

Forecasting growth in net sales of new products and services is an important objective of business decision makers. Over the years, diffusion models like the the Bass model (Bass, 1969, 1994) and its various generalizations have been used to describe the empirical adoption curve for a wide variety of new products and services. Mahajan and Wind (1986) give a broad treatment of this subject covering a wide range of diffusion models and estimation methods. Mahajan, Muller, and Bass (1990) provide a more recent review. For more recent developments, see Young (1993), Mahajan, Sharma, and Buzzel (1993), and Weerahandi and Moitra (1995).

When an empirical adoption curve consists of many points and has passed its inflection point, there exists a number of reasonable procedures for estimating diffusion model

parameters (again, see the work of Mahajan and Wind, 1986). However, it is well known (see, for example, Sultan, Farley, and Lehmann, 1990) that estimates of diffusion model parameters, based on only a few early data points, can be highly variable. Such unstable parameter estimates can lead to unreliable forecasts, even in the short run.

Auxiliary information may exist that can shed light on how a new product or service may spread into a market. A Bayesian scheme, utilizing this information, can leverage a small number of data points to stabilize parameter estimates. Lilien, Rao, and Kalish (1981), Sultan et al. (1990), and Lenk and Rao (1990) provide examples of different Bayesian implementations illustrating this point. Lilien et al. use data on similar products or expert judgement in tandem with Bayesian regression to update parameter estimates for a new product. Sultan et al. mix linear regression estimates based on prior information from a meta-analysis of diffusion applications with estimates based on a few data points on a new product to obtain more robust posterior estimates. Lenk and Rao fuse a Hierarchical Bayes approach with a modification of the nonlinear regression technique of Srinivasan and Mason (1986) to achieve greater stability in estimation.

In this paper, we propose a new Bayesian scheme for fortifying early parameter estimates and forecasts. Our procedure is similar in spirit to that of Lilien et al. (1981) in that we bolster new product estimates with data on similar services or expert judgement. The distinguishing feature of our approach is that it is based on maximum (posterior) likelihood estimation of the parameters of a pure birth model with a Bass-type adoption rate that can capture price and promotion effects.

The pure birth model (see, for example, Karlin and Taylor, 1984) is a natural stochastic model of the underlying new product diffusion process. Estimation procedures that directly exploit this stochastic structure can do better than those that do not. For example, Dalal and Weerahandi (1995) develop a beta-binomial approximation of maximum likelihood

estimates of the parameters of a pure birth model with a Bass formulation for the adoption rate. Because of the extreme non-normality of the errors in birth-process data, in a number of simulations the beta-binomial procedure significantly outperformed the nonlinear least squares procedure of Srinivasan and Mason (1986).

The beta-binomial procedure developed by Dalal and Weerahandi provides reliable parameter estimates only when the number of eventual new product adopters is large. Nonetheless, the success of this procedure in this somewhat restricted setting suggests the potential fruitfulness of the method of maximum likelihood in the context of the pure birth model. Furthermore, it is well known (see, for example, Rao, 1973) that the method of maximum likelihood produces optimal parameter estimates under general conditions.

Direct maximum likelihood of the parameters of the pure birth model is infeasible due to the intractable nature of the likelihood function. However, by reformulating the estimation problem as a missing data problem, it is possible to proceed in an indirect manner. We blend a Monte-Carlo version the EM algorithm (Dempster, Laird, and Rubin, 1977) with the Gibbs sampling technique (e.g., Tanner, 1996) to produce a sequence of parameter estimates that, under regularity conditions, converges to maximum (posterior) likelihood estimates without any restrictions on the number of eventual adopters. Such a procedure is called a Monte-Carlo EM, or MCEM procedure. Sherman, Ho, and Dalal (1998) develop convergence theory for general MCEM sequences.

The MCEM procedure can be applied either with or without auxiliary information. When auxiliary information is incorporated, we call the procedure a BMCEM, or Bayesian MCEM procedure. Figure 1 illustrates the use of the BMCEM procedure with data on a new telecommunications service. Circles represent 38 months of cumulative net sales of the new service. We use the first 10 months of data to forecast sales for the next 28 months. As a reference, plus signs represent MCEM projections using only the first 10

data points. Solid points represent BMCEM projections leveraging these first 10 points with data on the same service rolled out earlier in a neighboring region. There is a marked improvement in accuracy. We note that the time periods for the new and auxiliary data sets overlap. Nonetheless, the example serves to illustrate the potential benefits offered by the Bayesian procedure.

The rest of this paper is devoted to explicating the BMCEM procedure. In the next section, we give a general description of the data and the associated diffusion models we consider. In addition, we present a high-level overview of the procedure and illustrate its use in the context of the telecommunications data featured in Figure 1. In Section 3, we provide a detailed development of the methodology. Section 4 develops estimates of standard errors for model parameters and forecasts. Section 5 presents simulation results. We summarize and give directions for future work in Section 6.

II. DATA AND MODELS

We observe the data matrix $\{(t_j, n_j, p(t_j), v(t_j)), j = 0, 1, \ldots, q\}$ where the t_j's are successive time points and the n_j's are the corresponding cumulative numbers of adopters of a new service, $p(t_j)$ is the price of the service at time t_j, and $v(t_j)$ is the value of promotions at time t_j. Our objective is to describe some of the dynamics of the temporal diffusion and predict net sales into the future.

Let N denote the the number of individuals (e.g., households, small businesses, etc.) in the population of interest and assume that N is known and constant over time. Let π denote the unknown proportion of eventual adopters of the new service. Thus, the quantity $N\pi$ represents the number of eventual adopters.

We view n_j as a realization of a stochastic process $N(t)$ at time t_j. We assume that $N(t)$ is a pure birth Markov process with stationary transition probabilities and adoption

rate $[N\pi - N(t)]\lambda(N(t))$ (cf. Karlin and Taylor, 1984). The quantity $N\pi - N(t)$ represents the number of eventual adopters and $\lambda(N(t))$ the individual adoption rate at time t.

Bass (1969) popularized the specification $\lambda(N(t)) = \alpha + \beta N(t)$. The parameter $\alpha \geq 0$ is called the innovator coefficient and the parameter $\beta \geq 0$ is called the imitator coefficient. The rate of adoption due to innovation is greatest at the beginning of the diffusion when $N\pi - N(t)$ is largest, and decreases as time goes on. The rate of adoption due to imitation is smallest near the beginning of the diffusion when $N(t)$ is near zero and near the end of the diffusion when $N\pi - N(t)$ is near zero. The rate is greatest when $N(t) = N\pi/2$. The overall adoption rate for the Bass model is greatest when $N(t) = (N\pi\beta - \alpha)/2\beta$. The corresponding point is the inflection point of the S-shaped diffusion curve.

The Bass model is simple and has intuitive appeal in explaining how a homogeneous group of eventual adopters comes to adopt a new service. Further, the model can be modified in a straightforward manner to capture the effects of price changes and promotions.

Take price. A price increase is likely to decrease the proportion of eventual adopters while a decrease in price is likely to increase it. This notion can be captured by replacing π in the Bass model with $\pi(t) = \pi f(-\gamma p(t))$ where $f(x)$ is a known, increasing function satisfying $f(0) = 1$, γ is nonnegative, and $p(t)$ is the difference between the price of the service at time t and a reference price. In this formulation, $\pi(t)$ represents the proportion of eventual adopters at time t, and π the proportion of eventual adopters at the reference price. The function $f(x)$ can be chosen to reflect beliefs about the way price changes affect the change in the proportion of eventual adopters.

Promotions are likely to increase the strength of the innovation tendency. Increasing the value of promotions should increase the rate of adoption due to innovation. This notion can be captured by replacing the parameter α in the Bass model with $\alpha(t) = \alpha f(\delta v(t))$ where $f(x)$ is defined as before, δ is nonnegative, and $v(t)$ is the value of promotions offered

at time t. The quantity $\alpha(t)$ can be interpreted as the strength of the innovation tendency at time t. The parameter α is the strength of the innovation tendency when $v(t) = 0$.

III. THE BMCEM METHOD: AN OVERVIEW

Let $\boldsymbol{d}$ denote the observed data matrix $\{(t_j, n_j, p(t_j), v(t_j)), j = 0, 1, \ldots, q\}$ for a new service. Define the parameter vector $\theta = (\pi, \alpha, \beta, \gamma, \delta)$ and let $\theta_0 = (\pi_0, \alpha_0, \beta_0, \gamma_0, \delta_0)$ denote the value of θ that generated $\boldsymbol{d}$. We wish to develop a good estimate of θ_0 based on $\boldsymbol{d}$ and relevant auxiliary information. This estimate will be used to forecast net sales of the new service.

View θ as a random variable and write $f(\theta)$ for the prior density for θ based on the auxiliary information. Write $f(\boldsymbol{d} \mid \theta)$ for the likelihood function or conditional density of $\boldsymbol{d}$ given θ and $f(\theta \mid \boldsymbol{d})$ for the posterior density of θ given $\boldsymbol{d}$. By Bayes' rule,

$$f(\theta \mid \boldsymbol{d}) \propto f(\theta) f(\boldsymbol{d} \mid \theta)\,. \tag{1}$$

A natural estimate of θ_0 is the mode of $f(\theta \mid \boldsymbol{d})$ which can be obtained by maximizing the right-hand side of (1). However, as mentioned in the introduction, $f(\boldsymbol{d} \mid \theta)$ is numerically unstable. Dalal and Weerahandi (1995) report that for values of $N\pi \geq 25$, $f(\boldsymbol{d} \mid \theta)$ cannot be reliably maximized, apparently due to extreme nonlinearities in θ leading to numerical instablility. Therefore direct maximization is infeasible. However, we can obtain a sequence of estimates that converges to the mode of $f(\theta \mid \boldsymbol{d})$ through the indirect BMCEM method developed in detail in Section 3.

For now, we describe how we construct the prior density $f(\theta)$. This density is derived from auxiliary information such as data on the same service rolled out earlier in a neighboring region, data from a similar past service, or data developed by subject matter experts (or any combination of these sources). Auxiliary data should reflect, according to

expert opinion, how the new service is likely to spread into the market and should contain enough information to reliably estimate the model parameters.

Let $\boldsymbol{d}^* = \{(t_j^*, n_j^*, p^*(t_j), v^*(t_j)), j = 0, 1, \ldots, q^*\}$ denote the auxiliary data. We use the MCEM method developed in Section 3 to estimate θ_0 and the variances of its components from $\boldsymbol{d}^*$. Let $\theta^* = (\pi^*, \alpha^*, \beta^*, \gamma^*, \delta^*)$ denote the resulting point estimates and $(v(\pi^*), \ldots, v(\delta^*))$ the corresponding variance estimates. This information is used to construct $f(\theta)$.

For example, consider the proportion π which takes values in the interval $[0, 1]$. We use a beta prior for π having mean π^* and variance $v(\pi^*)$. The parameters α, β, γ, and δ take on nonnegative values and so we use gamma priors for these parameters with means and variances matching the corresponding point and variance estimates. These variance estimates can be adjusted up or down to reflect the preceived reliability of the auxiliary data for modeling the diffusion of the new product. Finally, we assume independence of components so that $f(\theta)$ is the product of the component priors.

IV. A TELECOMMUNICATIONS APPLICATION

In this section, we briefly describe data on a new telecommunications service and present results of applying the BMCEM scheme to the subset of this data featured in Figure 1. For proprietary reasons, only the barest description of the data can be given at this time.

We have monthly cumulative net sales data on a new telecommunications service covering a period of about 5 1/2 years beginning in late 1989 for small businesses in different regions of the United States. Roll-out of the service was staggered, that is, the service was introduced in some regions earlier than in others. Prices in all regions remained essentially unchanged over this period so it was not possible to estimate price effects from this data. Two types of promotions were offered in the regions we analyze, one a service order charge

waiver, the other, a get-acquainted offer waiving the fee for the first month of service. We code no promotions in a given time period as 0, and either of the two promotion types as 1.

Figure 2 depicts cumulative net sales and net sales of the new service for the region whose diffusion curve is represented by circles in Figure 1. We shall call this region R. The plotted points (0's and 1's) indicate the timing of promotions in region R for the period shown. Note the jumps in net sales in periods when promotions are offered.

We now describe how we produced the output in Figure 1. To produce the curve represented by the plus signs, we first applied the MCEM procedure to estimate the parameters of the diffusion curve for region R using only the first 10 points from this region. From statistical abstracts, we determined the number of small businesses in region R to be approximately 104000. The parameter vector θ_0 is $(\pi_0, \alpha_0, \beta_0, \delta_0)$ and the corresponding MCEM estimates are $(.03, .05, 0, .3)$. Based on these estimates, we projected the forecast out 28 months. The projected, or fitted values, which are estimates of the mean of the process $N(t)$ at the given time points, are obtained by numerically solving a set of approximate stochastic differential equations relating the mean and variance of $N(t)$ to their derivatives. These equations are provided in Section 4. It is interesting to note that the MCEM forecasts fairly accurately track the observed diffusion curve in the short term, for months 11 through 19. However, they seriously understate the curve beyond this period. Note that these estimates suggest, contrary to expectation, that there is no "word of mouth" effect for this diffusion.

To produce the curve represented by solid points in Figure 1, we first applied the MCEM procedure to estimate the parameters of the diffusion curve for a neighboring region, call it R^*. The new telecommunications service was introduced in region R^* prior to its introduction in region R. Economic forces influencing the decision to purchase the

new service were deemed similar in both of these regions. We found the value of N for region R^* to be approximately 123000. Plots of this data (0's and 1's indicate promotion type) along with fitted values (solid points) based on the point estimates are given in Figure 3. The fit is tight, with the solid points accurately tracking the jumps in net sales due to promotions. Note that although the diffusion in region R^* has apparently not reached its inflection point, there are still enough points to yield stable parameter estimates: the estimate θ^* is $(.3, .0009, .0000008, 1.1)$ and the corresponding vector of standard error estimates is $(.09, .0002, .0000003, .06)$. These values are used to construct the prior density $f(\theta)$ as prescribed in the previous subsection. The BMCEM procedure takes $f(\theta)$ and the data matrix $\boldsymbol{d}$ based on the first 10 points from region R and produces an estimate of the mode of $f(\theta \mid \boldsymbol{d})$. This estimate is $(.36, .0014, .0000003, 1.09)$. Note the positive "word of mouth" effect, as one would expect. The sales projections represented by the solid points in Figure 1 are based on these parameter estimates.

We note that the monthly net sales data on the new service in region R^* used to construct $f(\theta)$ covers the time period November, 1989 to March, 1995. The projections in Figure 1 are for the time period May, 1992 to August, 1994. Because of this overlap, we would not have been able to carry out this analysis in May of 1992 without supplementing the data from R^* with other data to ensure stable estimates of the parameters needed to construct $f(\theta)$. Nonetheless, the example illustrates the use of the procedure as well as its potential benefits.

As explained in the next section, the BMCEM procedure is iterative. For the diffusion applications we consider, convergence seems to occur after only a few iterations. For example, the parameter estimates produced above are based on 10 iterations of the BMCEM procedure. Running the entire BMCEM procedure for this application on a Sparc 10 workstation required approximately 10 minutes.

V. THE BMCEM PROCEDURE

This section provides a detailed development of the BMCEM methodology.

Direct maximum likelihood estimation in incomplete, or missing, data problems can be numerically intractable. The EM algorithm (Dempster, Laird, and Rubin, 1977) can circumvent this problem by replacing one difficult maximization with a sequence of simpler maximizations that yield the desired maximum likelihood estimates in the limit. The E step of the algorithm requires computation of a conditional expectation of the complete, or augmented, data likelihood function. When this computation is infeasible, recourse can be made to direct Monte-Carlo integration or, more commonly, to a blending of Monte-Carlo integration with Markov chain sampling techniques like the Gibbs sampler, the Metropolis algorithm, or the Hastings algorithm. In any case, the resulting scheme is called a Monte-Carlo EM, or MCEM algorithm. Wei and Tanner (1990) were the first to propose such a scheme. Guo and Thompson (1991) and Chan and Ledolter (1995) present substantial applications. We now develop notation for a BMCEM, or Bayesian MCEM procedure.

Let $\boldsymbol{D} = (D_1, \ldots, D_n)$ denote the observed data matrix where n is the sample size. Let $\mathcal{D}$ denote the support of $\boldsymbol{D}$ and $\boldsymbol{d}$ an element of $\mathcal{D}$. Write $f(\boldsymbol{d} \mid \theta)$ for the observed data likelihood where θ is an element of the parameter space, Θ. Let θ_0 denote the parameter value that generated the observed data. Our objective is to estimate θ_0 with the mode of the posterior density $f(\theta \mid \boldsymbol{d}) \propto f(\theta) f(\boldsymbol{d} \mid \theta)$, where $f(\theta)$ is the prior density for θ.

If direct maximization of $f(\theta \mid \boldsymbol{d})$ is impractical, it may be possible to obtain the mode indirectly through the sequential EM procedure. Let $\boldsymbol{T} = (T_1, \ldots, T_p)$ denote the missing data vector, $\mathcal{T}$ the support of $\boldsymbol{T}$, and $\boldsymbol{\tau}$ an element of $\mathcal{T}$. For example, in the diffusion modeling application, the individual adoption times may be viewed as missing data. Write $f(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta)$ for the complete data likelihood. Let ϕ denote an arbitrary element of Θ. Write $f(\cdot \mid \boldsymbol{d}, \phi)$ for the conditional density of $\boldsymbol{T}$ given $\boldsymbol{D} = \boldsymbol{d}$ and ϕ and $l(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta)$

for $\log[f(\theta)f(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta)]$. For each θ and ϕ in Θ, let

$$Q(\theta \mid \phi) = \int_{\mathcal{T}} l(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta)\, f(\boldsymbol{\tau} \mid \boldsymbol{d}, \phi)\, d\boldsymbol{\tau}$$

and define

$$\theta(\phi) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, Q(\theta \mid \phi)\,.$$

In other words, for a given value of ϕ, $\theta(\phi)$ is that value of θ in Θ that maximizes $Q(\theta \mid \phi)$. The EM algorithm generates a sequence $\theta^1, \theta^2, \ldots$ where $\theta^i = \theta(\theta^{i-1})$, $i = 1, 2, \ldots$ with θ^0 an arbitrary starting value. We assume that $f(\theta \mid \boldsymbol{d})$ is unimodal and that θ^i converges to this mode as $i \to \infty$. Wu (1983) shows that if $f(\theta \mid \boldsymbol{d})$ is unimodal, then under mild regularity conditions, θ^i converges to this mode as i tends to infinity.

It may not be possible to evaluate $Q(\theta \mid \phi)$ either directly or through numerical integration. However, it may be possible to estimate $Q(\theta \mid \phi)$ through Monte-Carlo integration facilitated with a Markov Chain sampling technique like the Gibbs sampler. We now develop notation for a Gibbs implementation.

For each $k \geq 1$, write $f_k(\cdot \mid \boldsymbol{d}, \phi)$ for the conditional density of $\boldsymbol{T}$ given $\boldsymbol{D} = \boldsymbol{d}$ and ϕ after k Gibbs iterations. Let $\boldsymbol{T}_i^{(k)}$, $i = 1, \ldots, m$, denote a sample of independent observations from $f_k(\cdot \mid \boldsymbol{d}, \phi)$ and write P_k^m for the empirical measure that puts mass $\frac{1}{m}$ on each $\boldsymbol{T}_i^{(k)}$. Notice that $\boldsymbol{T}_i^{(k)}$ depends on θ_0 (through $\boldsymbol{d}$) and ϕ. For each θ and ϕ in Θ, let

$$Q_k^m(\theta \mid \phi) = P_k^m l(\cdot, \boldsymbol{d} \mid \theta)$$

and define

$$\hat{\theta}(\phi) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, Q_k^m(\theta \mid \phi)\,.$$

The BMCEM algorithm generates a sequence of iterates $\hat{\theta}^1, \hat{\theta}^2 \ldots$ where $\hat{\theta}^i = \hat{\theta}(\hat{\theta}^{i-1})$, $i =$

$1, 2, \ldots$ with θ^0 an arbitrary starting value. We stop iterating when a sensible convergence criterion is satisfied and define the last iterate, $\hat{\theta}^i$, as the BMCEM estimate of the mode of $f(\theta \mid \boldsymbol{d})$.

Notice that if the prior density $f(\theta)$ is constant, or noninformative, then the BMCEM procedure reduces to the MCEM procedure, producing estimates of the mode of $f(\boldsymbol{d} \mid \theta)$, the observed data likelihood function.

We now explain the rationale for a Monte-Carlo procedure for estimating the parameters of the diffusion model described in Section 2.1. For notational simplicity we will present our argument in the context of the simple Bass model where $\theta = (\pi, \alpha, \beta)$. The generalization accommodating price and promotion is immediate.

We observe the data matrix $\boldsymbol{d} = \{(t_j, n_j), j = 0, 1, \ldots, q\}$ where n_j is a realization of a process $N(t)$ at time t_j. For simplicity, take $(t_0, n_0) = (0, 0)$. We assume that $N(t)$ is a pure birth Markov process with stationary transition probabilities and population adoption rate $[N\pi - N(t)][\alpha + \beta N(t)]$. Let S_i denote the ith "sojourn time", or time between the ith and $(i+1)$st adoptions, $i = 0, 1, \ldots, n_q - 1$. The Markovian and stationarity assumptions guarantee that the S_i's are independent and exponentially distributed (cf. Ross, 1983, p.142). The rate assumption implies that S_i has rate $(N\pi - i)(\alpha + \beta i)$. These facts form the basis for writing down the likelihood function for the observed data.

For $s, t \geq 0$, write $P_{n,m}(t, \theta)$ for $I\!P(N(s+t) = n \mid N(s) = m)$. Stationarity implies that $P_{n,m}(t, \theta)$ does not depend on s. In particular, $P_{n,m}(t, \theta) = I\!P(N(t) = n \mid N(0) = m)$. Apply this fact and the Markovian property to write

$$f(\boldsymbol{d} \mid \theta) = \prod_{j=1}^{q} P_{n_j, n_{j-1}}(t_j - t_{j-1}, \theta)\,.$$

Write $\Lambda_i(\theta)$ for $(N\pi - i)(\alpha + \beta i)$, $i = 1, 2, \ldots, n_q$. Standard results on birth processes

(cf. Bartlett, 1978, p.59) show that

$$P_{n,m}(t,\theta) = \sum_{i=m}^{n} \omega_i(\theta) \exp(-\Lambda_i(\theta)\, t)$$

where

$$\omega_i(\theta) = \frac{\prod_{k=m}^{n-1} \Lambda_k(\theta)}{\prod_{k=m,k\neq i}^{n} [\Lambda_k(\theta) - \Lambda_i(\theta)]}\,.$$

Recall that θ_0 denotes the true value of the parameter vector. We would like to estimate θ_0 by maximizing $f(\theta \mid \boldsymbol{d}) \propto f(\theta) f(\boldsymbol{d} \mid \theta)$. However, as mentioned previously, except for very small values of $N\pi$, neither $f(\boldsymbol{d} \mid \theta)$ nor its logarithm can be reliably maximized. Apparently, this is because $f(\boldsymbol{d} \mid \theta)$ is a product of sums of ratios of products, making it highly nonlinear and therefore subject to numerical instability. However, by viewing this estimation problem as a missing data problem, where the vector of individual adoption times are the missing data, we may apply BMCEM.

Write T_i for $\sum_{j=0}^{i-1} S_j$, the ith adoption time, and $\boldsymbol{T}$ for $(T_1, T_2, \ldots, T_{n_q})$. Let $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_{n_q})$ denote a realization of $\boldsymbol{T}$ and write $f(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta)$ for the complete-data likelihood function. Notice that $f(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta)$ is equal to the joint density function of $\boldsymbol{T}$ and $N(t_q)$ given θ. Next, Let Θ denote the parameter space for θ_0. We assume that Θ is such that $\Lambda_i(\theta) > 0$ for all θ in Θ and for all i. For the Bass model, $\Lambda_i(\theta) > 0$ provided both α and β are nonnegative, either α or β is positive, and $\pi \geq n_q/N$. Deduce that for all θ in Θ,

$$f(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta) = \left[\prod_{i=0}^{n_q-1} \Lambda_i(\theta) \exp(-\Lambda_i(\theta)[\tau_{i+1} - \tau_i]) \right] \exp(-\Lambda_{n_q}(\theta)[t_q - \tau_{n_q}])$$

where $\tau_0 = 0$, $\tau_i \leq \tau_j$ for $i < j$, and $\tau_{n_j} \leq t_j < \tau_{n_j+1}$, $j = 1, 2, \ldots, q$. The final factor is the probability of not observing the $(n_q + 1)$st adoption in the interval $[\tau_{n_q}, t_q]$. This function has a simple form compared to the incomplete data likelihood.

Because $f(\cdot \mid \boldsymbol{d}, \phi)$ involves the intractable incomplete data likelihood, it is infeasible to either evaluate or estimate $Q(\theta \mid \phi)$ through direct Monte-Carlo integration. Therefore, we turn to the Gibbs sampling technique to facilitate Monte-Carlo integration.

The Gibbs sampler is a cyclic, iterative technique for generating observations from the joint distribution of a random vector when it is possible to sample from all full univariate conditional distributions. See Gelfand and Smith (1990) or Tanner (1996) for a basic introduction. We now explain how to implement the Gibbs sampler in our problem.

We wish to generate m observations from $f_k(\cdot \mid \boldsymbol{d}, \phi)$ to construct the objective function $Q_k^m(\theta \mid \phi)$. To generate one such observation we must be able to sample from the conditional distribution of each T_i given all the other adoption times, $\boldsymbol{D} = \boldsymbol{d}$, and ϕ.

Rearrange terms in the complete-data likelihood to get, for each ϕ in Θ,

$$f(\boldsymbol{\tau}, \boldsymbol{d} \mid \phi) = \left[\prod_{i=1}^{n_q} \Lambda_{i-1}(\phi) \exp(-[\Lambda_{i-1}(\phi) - \Lambda_i(\phi)]\tau_i)\right] \exp(-\Lambda_{n_q}(\phi) t_q) \tag{2}$$

where $\tau_i \leq \tau_j$ for $i < j$ and $\tau_{n_j} \leq t_j < \tau_{n_j+1}$, $j = 0, 1, \ldots, q$. Let $\boldsymbol{T}_{-i}$ denote the vector of adoption times excluding T_i. Write $\boldsymbol{\tau}_{-i}$ for a realization of $\boldsymbol{T}_{-i}$ and $f(\tau_i \mid \boldsymbol{\tau}_{-i}, \boldsymbol{d}, \phi)$ for the density of T_i given $\boldsymbol{T}_{-i} = \boldsymbol{\tau}_{-i}$, $\boldsymbol{D} = \boldsymbol{d}$ and ϕ. Write $f(\boldsymbol{\tau}_{-i}, \boldsymbol{d} \mid \phi)$ for the joint density of $\boldsymbol{T}_{-i}$ and $\boldsymbol{D}$ given ϕ. Note that $f(\tau_i \mid \boldsymbol{\tau}_{-i}, \boldsymbol{d}, \phi)$ is the ratio of $f(\boldsymbol{\tau}, \boldsymbol{d} \mid \phi)$ to $f(\boldsymbol{\tau}_{-i}, \boldsymbol{d} \mid \phi)$ on the interval $[l_i, u_i]$ where

$$\begin{aligned} l_i &= \tau_{i-1} + [t_j - \tau_{i-1}] \{i - 1 = n_j\}, \\ u_i &= \tau_{i+1} + [t_j - \tau_{i+1}] \{i = n_j\}. \end{aligned}$$

Moreover, $f(\boldsymbol{\tau}_{-i}, \boldsymbol{d} \mid \phi)$ does not involve τ_i and only the ith factor of $f(\boldsymbol{\tau}, \boldsymbol{d} \mid \phi)$ involves

τ_i. Deduce that

$$f(\tau_i \mid \boldsymbol{\tau}_{-i}, \boldsymbol{d}, \phi) = c_i(\phi) \exp(-\delta_i(\phi)\tau)\{l_i \leq \tau_i \leq u_i\} \tag{3}$$

where $\delta_i(\phi) = \Lambda_{i-1}(\phi) - \Lambda_i(\phi)$. Since $f(\tau_i \mid \boldsymbol{\tau}_{-i}, \boldsymbol{d}, \phi)$ must integrate to unity we see that

$$c_i(\phi) = \delta_i(\phi)/[\exp(-\delta_i(\phi)l_i) - \exp(-\delta_i(\phi)u_i)]\,.$$

The distribution function corresponding to $f(\tau_i \mid \boldsymbol{\tau}_{-i}, \boldsymbol{d}, \phi)$ is a linear function of $\exp(-\delta_i(\phi)\tau_i)$ and so is easily inverted. Therefore, it is trivial to sample from $f(\tau_i \mid \boldsymbol{\tau}_{-i}, \boldsymbol{d}, \phi)$ using the probability integral transformation technique.

Here are the details of the Gibbs scheme. Let $\boldsymbol{T}^{(j)} = (T_1^{(j)}, T_2^{(j)}, \ldots, T_{n_q}^{(j)})$ denote the jth Gibbs iterate, $j = 0, 1, \ldots, k$. The case $j = 0$ specifies a set of starting values for the adoption times that satisfy $T_i^{(0)} \leq T_j^{(0)}$ for $i < j$ and $T_{n_j}^{(0)} \leq t_j < T_{n_j+1}^{(0)}$, $j = 0, 1, \ldots, k$, but are otherwise arbitrary. Define $T_0^{(j)} = 0$, $j = 1, 2, \ldots, k$. Starting with $j = 1$, draw $T_i^{(j)}$, $i = 1, 2, \ldots, n_q$, from the distribution with density

$$\frac{\delta_i(\phi)\exp(-\delta_i(\phi)\tau_i)}{\exp(-\delta_i(\phi)l_i^{(j)}) - \exp(-\delta_i(\phi)u_i^{(j)})}\{l_i^{(j)} \leq \tau_i \leq u_i^{(j)}\}$$

where

$$\begin{aligned} l_i^{(j)} &= T_{i-1}^{(j)} + \left[t - T_{i-1}^{(j)}\right]\{i - 1 = n\}\,, \\ u_i^{(j)} &= T_{i+1}^{(j-1)} + \left[t - T_{i+1}^{(j-1)}\right]\{i = n\}\,, \end{aligned}$$

and (t, n) is a component of $\boldsymbol{d}$. Repeat this last step for $j = 2, 3, \ldots, k$. The final Gibbs iterate, $\boldsymbol{T}^{(k)}$, is an observation from $f_k(\cdot \mid \boldsymbol{d}, \phi)$. Independently generate m such

observations, then construct $Q_k^m(\theta \mid \phi)$.

VI. ESTIMATING VARIANCES

In this section, we describe how to obtain variance estimates for $\hat{\theta}^i$ and for $N(t)$. In the course of doing so, we will indicate how to estimate the mean of the diffusion process $N(t)$ at any time t. The latter estimates can be used to fit an observed diffusion curve or to make projections into the future.

To estimate the variance of $\hat{\theta}^i$ we use a result due to Louis (1982) which states that observed information is equal to the difference between complete information and missing information. That is,

$$-\nabla_{\theta\theta} \log f(\theta \mid \boldsymbol{d}) = -I\!E[\nabla_{\theta\theta} l(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta)] - V[\nabla_\theta l(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta)]$$

where $\nabla_\theta = \frac{\partial}{\partial \theta}$, $\nabla_{\theta\theta} = \frac{\partial^2}{\partial \theta^2}$, V denotes the variance operator, and expectations are taken with respect to $\boldsymbol{\tau}$ given $\boldsymbol{d}$ and θ. Notice that the expression on the right is a sum of functionals of the tractable complete data likelihood. The inverse of this expression, evaluated at $\hat{\theta}^i$, is an estimate of the variance of $\hat{\theta}^i$. In our problem, we cannot evaluate the expression on the right directly. Instead we develop a Monte-Carlo version of Louis' method. Specifically, we use numerical derivatives to estimate the bracketed partial derivatives at $\hat{\theta}^i$ and then form sample averages corresponding to the expectations using the m final Gibbs iterates of $\boldsymbol{\tau}$.

Next, turn to the problem of estimating the variance of $N(t)$. Note that

$$V[N(t)] = I\!EV[N(t) \mid \theta] + VI\!E[N(t) \mid \theta]$$

where the outer expectations are taken with respect to θ. Write $M(t, \theta)$ for $I\!E[N(t) \mid \theta]$

and $V(t,\theta)$ for $V[N(t) \mid \theta]$. Taylor expand $V(t,\theta)$ about $\hat{\theta}^i$ and take expectations with respect to θ to see that $I\!\!E V(t,\theta)$ can be approximated by $V(t,\hat{\theta}^i)$. Handle the second term similarly to see that

$$V[N(t)] \approx V(t,\hat{\theta}^i) + \left[\nabla_\theta M(t,\hat{\theta}^i)\right]' \hat{V}(\hat{\theta}^i) \left[\nabla_\theta M(t,\hat{\theta}^i)\right]$$

where $\hat{V}(\hat{\theta}^i)$ is the estimate of the variance of $\hat{\theta}^i$ from the Monte-Carlo version of Louis' method. All that remains is to develop estimates of $V(t,\hat{\theta}^i)$ and $\nabla_\theta M(t,\hat{\theta}^i)$ to substitute into the last expression. To do this, we simultaneously solve a set of two approximate differential equations for $V(t,\hat{\theta}^i)$ and $M(t,\hat{\theta}^i)$. We use numerical derivatives in tandem with the differential equations to develop an estimate of $\nabla_\theta M(t,\hat{\theta}^i)$. Write ∇_t for $\frac{\partial}{\partial t}$. For each t and θ,

$$\begin{aligned}
\nabla_t M(t,\theta) &\approx \lambda(M(t,\theta))[N\pi - M(t,\theta)] - \lambda'(M(t,\theta))V(t,\theta) \\
\nabla_t V(t,\theta) &\approx \nabla_t M(t,\theta) + 2V(t,\theta)[[N\pi - M(t,\theta)]\lambda'(M(t,\theta)) - \lambda(M(t,\theta))]
\end{aligned}$$

where $\lambda'(x)$ denotes the derivative of the individual adoption rate $\lambda(x)$ with respect to x. See Dalal and Weerahandi (1992) for a derivation of these approximations.

Notice that an important by-product of solving the differential equations above is an estimate of $M(t,\theta_0)$, the mean of the diffusion curve $N(t)$ at time t. Thus, by solving this set of equations we can obtain fitted or projected values of the diffusion process at any arbitrary time point together with a corresponding estimate of the variance of the process.

VII. SIMULATIONS

In this section, we present results of applying the MCEM and BMCEM procedures to

simulated data where N, the target population size, is 2000. Even though this population size is small relative to the population size for the telecommunications application, the MCEM and BMCEM procedures still perform well, as we will show. In each simulation, we run the procedures with 10 iterations ($i = 10$), 30 samples per iteration ($m = 30$), and 50 Gibbs steps per sample ($k = 50$). At this level of precision, the procedures take approximately 3 minutes to run on a Sparc 10 workstation.

Figure 4 presents the results of applying the MCEM procedure to data from a realization of a process $N(t)$ satisfying the Bass model with parameter vector $\theta_0 = (\pi_0, \alpha_0, \beta_0) = (.5, .0296, .0004)$ and $N = 2000$. Thus, $N\pi_0 = 1000$. There are 13 observations on the process $N(t)$, namely, $N_0, N_1, \ldots, N_{12}$ at time points $t = 0, 1 \ldots, 12$. These 13 observations are plotted as circles in Figure 4. The solid points are the MCEM estimates of the mean of $N(t)$ at times $0, 1, \ldots, 12$. The estimated points fit the observations very well. The MCEM estimates of θ_0 are $(.49, .032, .00042)$. The corresponding vector of standard errors is $(.024, .0044, .000032)$. Each component is well within 2 standard errors of the corresponding true values. Further, the estimates of the mean of $N(t)$ at the time points $t = 0, 1, \ldots, 12$ are all within 2 standard errors of the corresponding observed values as indicated by the dashed lines in the plot.

Figure 5 shows what can happen when the MCEM procedure is applied to a path that has yet to reach its infection point. We fit the Bass model to the first 3 points of a realization from the same Bass process as the one used to generate the data for Figure 4. The 13 circles represent a mature realization of this process. Note that the first 3 points fall short of the inflection point of the curve. The plus signs indicate the MCEM estimates of the diffusion curve at time points $4, 5, \ldots, 12$. The predicted values seriously understate the observed values. In fact, the MCEM estimate of θ_0 is $(.12, .15, .003)$. The estimated components are far from the true values. The MCEM procedure cannot be expected to

provide accurate long-range predictions based on only a small number of data points from an incipient diffusion curve.

Figure 5 also shows the results of applying the BMCEM procedure. Again, we only use the first 3 observed data points, but we prime the pump with data from a mature, 12 point path generated from the true model. The results are excellent, with the predicted values for time points 4 through 12 hugging the corresponding observed values of $N(t)$. The estimate of θ_0 is $(.497, .0306, .00039)$.

VIII. SUMMARY AND DIRECTIONS FOR FUTURE WORK

This paper develops a new Bayesian procedure that allows decision makers to learn from past experience to improve the accuracy of early forecasts. Past experience can take the form of data on similar services, expert opinion, or a combination of these auxiliary sources of information. The procedure is based on maximum likelihood estimation of the parameters of a pure birth process with a generalized Bass adoption rate that can capture the effects of price changes and promotions.

Despite optimality under general conditions, the method of maximum likelihood has been avoided in the pure birth context due to the intractable nature of the likelihood function. We circumvent this difficulty by developing an indirect iterative procedure that yields maximum likelihood estimates in the limit. This procedure blends a Monte-Carlo version of the EM algorithm with the Gibbs sampling technique, a so-called Monte-Carlo EM or MCEM procedure. When auxiliary information is incorporated into the estimation scheme, the procedure is called a Bayesian MCEM, or BMCEM procedure. We illustrate the use of the technique through simulations and a telecommunications application.

The MCEM procedure yields accurate fitted and predicted values when there are many observations on a innovation or when the diffusion curve of the innovation has passed its

inflection point. The BMCEM procedure can provide good fits and forecasts even when there are only a few data points on an innovation provided auxiliary information exists that reflects the way the innovation will spread into the marketplace. As more information is collected on the innovation, the BMCEM procedure will adapt to the unique properties of the innovation.

The MCEM and BMCEM procedures are iterative, producing a sequence of parameter estimates that are guaranteed, under regularity conditions (see Sherman et al., 1998), to converge to the mode of the posterior density as the number of iterations (i), the number of Monte-Carlo samples (m), and the number of Gibbs steps per sample (k) tend to infinity. The development of optimal stopping rules for i, m, and k requires further research. Our experience in the diffusion context, both with simulations and applications, suggests that convergence is reached after a small number of iterations. In all simulations and the telecommunications application, we chose the settings $i = 10$, $m = 30$, and $k = 50$. In monitoring the sequence of iterates, we found that after only a few iterations, the parameter values appeared to fluctuate about a constant value. In the simulations, this value was very close to θ_0, the value of the parameter vector that generated the data. Increasing m and k diminishes the variance of the fluctuations. The settings $i = 10$, $m = 30$, and $k = 50$ seemed to provide a reasonable balance between precision and computational speed. On a Sparc 10 workstation, a single run of the MCEM procedure (parameter estimates, forecasts, and variance estimates) at the above settings when the number of observed points was about 7000 required about 7 minutes.

In order to apply the methods proposed in this paper, the innovation diffusion process must satisfy a pure birth assumption. This means that once an individual adopts the innovation, that individual never drops it. We expect that this assumption need only hold approximately for the procedures to yield reasonable predictions, and we think that

this was a reasonable assumption to make for the telecommunications data we analyzed. However, there may be other innovations for which the effect of droppers is more than a second order effect. In the future, we plan to extend the estimation procedure for the Bass model and its generalizations to cover birth and death processes.

As alluded to above, methodological and theoretical advances are still needed to develop guidelines on how best to choose the parameters of the iterative procedure. Also, more work is needed on how best to specify the function $f(x)$ associated with the price and promotion specification and the priors for the Bayesian analysis. We hope to pursue these lines of research in the future.

(a) We thank a conscientious referee for comments and suggestions that led to the improvement of this paper.

REFERENCES

Bartlett, M. S., An Introduction to Stochastic Processes (Cambridge, University Press, 3rd Edition, 1978).

Bass, F., "A New Product Growth Model for Consumer Durables", Management Science, 15, 1969, pp. 215-227.

Bass, F., "Why the Bass Model Fits Without Decision Variables", *Marketing Science*, 13, 1994, pp. 203-223.

Chan, K. S. and Ledolter, J., "Monte-Carlo EM Estimation for Time Series Models Involving Counts", Journal of the American Statistical Association, 90, 1995, pp. 242-252.

Dalal, S. and Weerahandi, S., "Some Approximations for the Moments of a Process Used in Diffusion of New Products", Statistics and Probability Letters, 15, 1992, pp. 181-189.

Dalal, S. and Weerahandi, S., "Estimation of Innovation Diffusion Models with Application to a Consumer Durable", Marketing Letters, 6, 1995, pp. 123-136.

Dempster, A., Laird, N. and Rubin, D., "Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion)", Journal of the Royal Statistical Society, Series B, 39, 1977, pp. 1-38.

Gelfand, A., and Smith, A., "Sampling-Based Approaches to Calculating Marginal Densities", Journal of the American Statistical Association, 85, 1990, pp. 398-409.

Guo, S. W., and Thompson E. A., "Monte-Carlo Estimation of Variance Component Models for Large Complex Pedigrees", IMA Journal of Mathematics Applied in Medicine and Biology, 8, 1991, pp. 171-189.

Heide, J., and Weiss. A., "Vendor Consideration and Switching Behavior for Buyers in High Technology Markets", Journal of Marketing, 59, 1995, pp. 30-43.

Karlin, S., and Taylor H., An Introduction to Stochastic Modeling (New York, Academic Press Inc., 1984).

Lilien, G., Rao, A., and Kalish, S., "Bayesian Estimation and Control of Detailing Effort in a Repeat Purchase Diffusion Environment", Management Science, 27, 1981, pp. 493-506.

Louis, T., "Finding Observed Information Using the EM Algorithm", Journal of the Royal Statistical Society, Series B, 44, 1982, pp. 98-130.

Mahajan, V. and Wind, Y., Innovation Diffusion Models of New Product Acceptance, (Cambridge, Massachusetts, Ballinger, 1986).

Mahajan, V., Muller, E., and Bass, F., "New Product Diffusion Models in Marketing: A Review and Directions for Research", Journal of Marketing, 54, 1990, pp. 1-26.

Mahajan, V., Sharma, S., and Buzzel, R., "Assessing the Impact of Competitive Entry on Market Expansion and Incumbent Sales", Journal of Marketing, 57, 1993, pp. 39-52.

Rao, C. R., Linear Statistical Inference and its Applications (New York, John Wiley and Sons, 1973).

Ross, S., Stochastic Processes, (New York, John Wiley and Sons, 1983).

Sherman, R., Ho, Y., and Dalal, S., "Convergence Conditions for Monte-Carlo EM Sequences", Journal of Econometrics, 1998, under review.

Srinivasan, V., and Mason, C., "Nonlinear Least Squares Estimation of New Product Diffusion Models", Marketing Science, 5, 1986, pp. 169-178.

Sultan, F., Farley, J., and Lehmann, D., "A Meta-Analysis of Applications of Diffusion Models", Journal of Marketing Research, 27, 1990, pp. 70-77.

Tanner, M., Tools for Statistical Inference (New York, Springer-Verlag, 1996).

Weerahandi, S. and Moitra, S., "Using Survey Data to Predict Adoption and Switching for Services", Journal of Marketing Research, 32, 1995, pp. 85-97.

Wu, J., "On the Convergence Properties of the EM Algorithm", Annals of Statistics, 11, 1983, pp. 95-103.

Young, P., "Technological Growth Curves: A Competition of Forecasting Models", Technological Forecasting and Social Change, 44, 1993, pp. 375-389.

Figure 1: An application of BMCEM methodology to data on a new telecommunications service. Circles represent the observed data, plus signs represent forecasts based on the MCEM procedure using only the first 10 observed points, and solid points represent forecasts based on the BMCEM procedure using the first 10 observed points and auxiliary data.

Figure 2: Cumulative net sales and net sales of a new telecommunications service in region R. A zero or one indicates the type of promotion offered during the previous month.

Figure 3: Cumulative net sales and net sales of a new telecommunications service in region R^*. Zeros and ones indicate promotions. Solid points represent an MCEM fit to the data in region R^*.

Figure 4: MCEM fit to a mature synthetic data path together with pointwise 95% confidence bands. Circles represent synthetic data. Solid points represent MCEM fitted values.

Figure 5: Early forecasting of a synthetic data path based on the BMCEM methodology. Circles represent the synthetic data, plus signs represent forecasts based on the MCEM procedure using only the first 3 observed points, and solid points represent forecasts based on the BMCEM procedure using the first 3 observed points and auxiliary data. 95% pointwise confidence bands for the BMCEM forecasts are displayed.