Dynamic Empirical Bayes Models and Their Applications to Longitudinal Data Analysis and Prediction

Tze Leung Lai, Yong Su and Kevin Haoyu Sun

Stanford University and Numerical Methods, Inc.

Abstract: Empirical Bayes modeling has a long and celebrated history in statistical theory and applications. After a brief review of the literature, we propose a new dynamic empirical Bayes modeling approach which provides flexible and computationally efficient methods for the analysis and prediction of longitudinal data from many individuals. This dynamic empirical Bayes approach pools the cross-sectional information over individual time series to replace an inherently complicated hidden Markov model by a considerably simpler generalized linear mixed model. We apply this new approach to modeling default probabilities of firms that are jointly exposed to some unobservable dynamic risk factor, and to the well-known statistical problem of predicting baseball batting averages studied by Efron and Morris and recently by Brown.

Key words and phrases: Dynamic frailty model, empirical Bayes, generalized linear mixed models, longitudinal data, prediction, time series.

1. Introduction

The empirical Bayes methodology, introduced by Robbins (1956) and Stein (1956), considers n independent and structurally similar problems of statistical inference on unknown parameters θ_i from observed data Y_i $(i = 1, \ldots, n)$, where Y_i has probability density $f(y|\theta_i)$. Here and in the sequel, θ_i and Y_i can represent vectors. The θ_i are assumed to have a common prior distribution G that has unspecified hyperparameters. Let $d_G(y)$ denote the Bayes decision rule (with respect to some loss function and assuming known hyperparameters) when $Y_i = y$ is observed. The basic principle underlying empirical Bayes is that d_G can often be consistently estimated from $Y_1, \ldots, Y_n$, leading to the empirical Bayes rule $d_{\widehat{G}}$. Thus, the n structurally similar problems can be pooled to provide information about unspecified hyperparameters in the prior distribution, thereby yielding $\widehat{G}$ and the decision rules $d_{\widehat{G}}(Y_i)$ for the independent problems. In particular,

Robbins (1956) considered Poisson Y_i with mean θ_i, as in the case of the number of accidents by the ith driver in a sample of size n (in a given year) from a population of drivers, with distribution G for the accident-proneness parameter θ. In this case the Bayes estimate (with respect to squared error loss) of θ_i when $Y_i = y$ is observed is

$$d_g(y) = (y+1)g(y+1)/g(y),\ y = 0, 1, \ldots, \tag{1.1}$$

where $g(y) = \int_0^\infty \theta^y e^\theta dG(\theta)/(y!)$. Using $\hat{g}(k) = n^{-1}\sum_{i=1}^n I_{\{Y_i=k\}}$ to replace $g(k)$ in (1.1) yields the empirical Bayes (EB) estimate $d_{\hat{g}}(y)$. The case $Y_i \sim N(\theta_i, \sigma^2)$ with known σ, considered by Stein (1956), yields the following Bayes estimate for the prior distribution $G \sim N(\mu, \nu)$ of the θ_i:

$$d_{\mu,\nu}(y) = \mu + \{\nu/(\nu+\sigma^2)\}(y-\mu). \tag{1.2}$$

Since $\mu = E(E(Y_i|\theta_i))$ can be consistently estimated by $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $Var(Y_i) = \nu + \sigma^2$ can be consistently estimated by $s^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2/(n-1)$, replacing μ and $\nu + \sigma^2$ by these consistent estimates yields an EB estimate of the form

$$d_{\bar{Y},s^2}(y) = \bar{Y} - (1 - \sigma^2/s^2)_+(y - \bar{Y}). \tag{1.3}$$

This linear EB estimator and the subsequent variant by James and Stein (1961) have spawned a large literature covering both theory and applications.

One class of applications is in insurance. Besides estimating the accident-proneness of a driver (in a future period) for his/her automobile insurance premium, another important problem in determining insurance rates is prediction of the claim size of a policy in a future period. This is called "credibility theory" in actuarial science, to which linear EB methods have been applied to derive the premiums for insurance policies that balance the policy holder's individual risk and the class risk. The linear EB estimate (1.3) can be written in the form $\hat{d}(Y_i) = A_n Y_i + (1 - A_n)\bar{Y}$. This is called a *credibility formula* in insurance rate-making, and A_n is called a *credibility factor*. Here Y_i corresponds to the "individual premium" and $\bar{Y}$ the "collective premium". Bühlmann (1967) made use of the linear EB approach to determine the credibility factors. The monograph by Bühlmann and Gisler (2005) describes a variety of extensions of (1.3) to more general settings. A closely related class of applications is prediction of the performance of an individual in a future period using the data on the performance

in the last period of a group that includes the individual and similar subjects. A well known example, which is considered in Section 4, is prediction of batting averages of baseball players first studied by Efron and Morris (1975, 1977) and recently by Brown (2008).

For these applications, one actually has longitudinal data and it seems that combining individual and collective histories may lead to better predictions. For insurance policies, allowing the prior means to change over time has led to evolutionary credibility as an extension of traditional credibility theory (Bühlmann and Gisler, 2005). For baseball batting averages, using a player's batting average in the past season besides his batting average to date in the current season should provide considerably more information to predict his batting average for the remainder of the season than his average from the first 45 at-bats used by Efron and Morris (1975, 1977). On the other hand, there are obvious difficulties to carry this out, as some of these players may not have played or may have only played sparingly in the past season. In addition, how can one pool information from different players over different time periods to implement the EB idea? In this paper we show how these difficulties can be resolved and develop a *dynamic EB* methodology for longitudinal data. The methodology is described in Section 2 in a general framework in which $Y_{i,t}$ belongs to an exponential family of distributions for $t \in T_i$, the set of times when the ith subject is observed. The mean of $Y_{i,t}$ is related to covariates, some of which may be time-varying, via a generalized linear model with subject-specific regression parameters that have a common prior distribution across subjects. Section 2 shows how the EB principle described in the first paragraph can be extended to incorporate dynamics in the joint prior distribution over time. This results in a generalized linear mixed model (GLMM) of the type introduced by Breslow and Clayton (1993) that can be easily implemented by existing software, despite the inherent complexity of individual and collective histories.

Section 3 illustrates the usefulness of the dynamic EB methodology developed in Section 2 by considering a problem of timely relevance in the finance literature, namely modeling joint default probabilities of multiple firms. In Section 4 we use the dynamic EB approach to re-analyze Brown's (2008) data on baseball batting averages and compare the predictive performance of our approach with his EB

methods. In this connection we also introduce a more general methodology for the evaluation of predictive performance than that used by Brown (2008). Section 5 gives some concluding remarks.

2. Dynamic Empirical Bayes Models of Longitudinal Data

2.1. Cross-sectional means and dynamic linear EB models

We begin by introducing dynamic linear EB models in the context of evolutionary credibility. Bühlmann and Gisler (2005) generalized the linear EB approach to credibility theory described in Section 1 by developing *evolutionary credibility* that assumes a first-order autoregressive model for the prior means μ_t of $\theta_{i,t}$, with $E(Y_{i,t}|\theta_{i,t}) = \theta_{i,t} = \mu_t + b_i$ and

$$\mu_t = \rho\mu_{t-1} + (1-\rho)\mu + \eta_t, \tag{2.1}$$

in which η_t are i.i.d. unobservable errors with mean 0 and variance V. They use Kalman filtering to estimate μ_t. The Kalman filter involves unspecified parameters ρ, μ and V, which can be estimated by maximum likelihood or method of moments. In particular, the method of moments proceeds similarly to (1.3) and also yields for large n a consistent estimate $\bar{Y}_{t-1}$ of μ_{t-1}.

Note that replacing μ_{t-1} by $\bar{Y}_{t-1}$ in (2.1) yields

$$\mu_t = \rho\bar{Y}_{t-1} + \omega + \eta_t, \tag{2.2}$$

where $\omega = (1-\rho)\mu$. Whereas (2.1) describes the dynamics of the unobserved μ_t when the observations are $Y_{i,t}$, yielding a linear state-space model with unknown parameters ρ, μ, V, we can obtain a simpler model without hidden states by using (2.2) instead of (2.1) to model μ_t. The model thus obtained is a linear mixed model (LMM)

$$Y_{i,t} = \rho\bar{Y}_{t-1} + \omega + b_i + \epsilon_{i,t}, \tag{2.3}$$

in which $\epsilon_{i,t} = (Y_{i,t} - \theta_{i,t}) + \eta_t$. The b_i are i.i.d. random effects with $E(b_i) = 0$. Since (2.3) is in the form of a regression model, one can easily include additional covariates and lags to increase the predictive power of the model in the LMM

$$Y_{i,t} = \sum_{j=1}^{p} \rho_j \bar{Y}_{t-j} + a_i + \boldsymbol{\beta}'\mathbf{x}_{i,t} + \mathbf{b}_i'\mathbf{z}_{i,t} + \epsilon_{i,t}, \tag{2.4}$$

where a_i and $\mathbf{b}_i$ are subject-specific random effects, $\mathbf{x}_{i,t}$ represents a vector of subject-specific covariates that are available prior to time t (for predicting $Y_{i,t}$

prior to observing it at time t), and $\mathbf{z}_{i,t}$ denotes a vector of additional covariates that are associated with the random effects $\mathbf{b}_i$. Throughout the sequel, we use a_i and $\mathbf{b}_i$ to denote random effects that have zero means.

2.2. Dynamic EB models in the generalized linear setting

A widely used model for longitudinal data $Y_{i,t}$ in biostatistics is the generalized linear model that assumes $Y_{i,t}$ to have a density function of the form

$$f(y;\theta_{i,t},\phi) = \exp\{[y\theta_{i,t} - g(\theta_{i,t})]/\phi + c(y,\phi)\}, \tag{2.5}$$

in which for some smooth increasing function (the link function) h and d-dimensional vector $\mathbf{x}_{i,t}$ of covariates,

$$h(\mu_{i,t}) = \boldsymbol{\beta}'\mathbf{x}_{i,t}, \text{ where } \mu_{i,t} = \frac{dg}{d\theta}(\theta_{i,t}), \tag{2.6}$$

$i = 1, \cdots, n$. In particular, for the case $h(\mu) = \theta$, or equivalently, $h = (dg/d\theta)^{-1}$, h is called the canonical link.

In the case $n = 1$, Zeger and Qaqish (1988) have extended the autoregressive time series model to the generalized linear setting in which the conditional density of Y_t given $Y_{t-1}, \cdots, Y_{t-p}$ is specified by (2.5) and (2.6) with $h(\mu_t) = \beta + \sum_{j=1}^{p} \rho_j h(Y_{t-j})$. For $n > 1$ time series $Y_{i,t}$, to extend dynamic EB models to the generalized linear setting, note that μ_s is the mean of $\mu_{i,s}$ and can be consistently estimated by $\bar{Y}_s = n^{-1}\sum_{i=1}^{n} Y_{i,s}$, which is the basic idea underlying the linear EB approach. Therefore, an EB version of the preceding model of Zeger and Qaqish (1988) for $n \geq 1$ is $h(\mu_t) = \beta + \sum_{j=1}^{p} \rho_j h(\bar{Y}_{t-j})$. As in the linear case (2.4), we can increase the predictive power of the model by including fixed and random effects and other time-varying covariates of each subject i, thereby extending the LMM (2.4) to the GLMM

$$h(\mu_{i,t}) = \sum_{j=1}^{p} \rho_j h(\bar{Y}_{t-j}) + a_i + \boldsymbol{\beta}'\mathbf{x}_{i,t} + \mathbf{b}_i'\mathbf{z}_{i,t}, \tag{2.7}$$

in which $\rho_1, \ldots, \rho_p$ and $\boldsymbol{\beta}$ are the fixed effects and a_i and $\mathbf{b}_i$ are subject-specific random effects. Note that the LMM in (2.4) is a special case of (2.7) with $h(\mu) = \mu$, as it can be written in the form $\mu_{i,t} = \sum_{j=1}^{p} \rho_j \bar{Y}_{t-j} + a_i + \boldsymbol{\beta}'\mathbf{x}_{i,t} + \mathbf{b}_i'\mathbf{z}_{i,t}$, where $\mu_{i,t}$ denotes the conditional mean of $Y_{i,t}$ given $\bar{Y}_{t-j}, \mathbf{x}_{i,t}, \mathbf{z}_{i,t}$ and $\mathbf{b}_i$. Following Breslow and Clayton (1993), we assume a_i and $\mathbf{b}_i$ to be independent normal

with zero means. For notational simplicity, we can augment $\mathbf{b}_i$ to include a_i so that (2.7) can be written as $h(\mu_{i,t}) = \sum_{j=1}^{p} \rho_j h(\bar{Y}_{t-j}) + \mathbf{x}'_{i,t}\boldsymbol{\beta} + (1, \mathbf{z}_{i,t})\mathbf{b}_i$ such that $\mathbf{b}_i$ has covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\alpha})$. Lai and Shih (2003a, b) have shown by asymptotic theory and simulations that the choice of a normal distribution, with unspecified parameters, for the random effects $\mathbf{b}_i$ in GLMM is innocuous; heuristically, this is due to very low resolution in estimating the actual distribution of the $\mathbf{b}_i$ nonparametrically in mixture models. The Appendix gives details on the implementation, such as computation of the likelihood function, and refinements of the GLMM (2.7).

2.3. Prediction and variable selection

An important application of the dynamic EB model (2.5) and (2.7) is to estimate some future function ψ_{t+1} of the unobserved $\mathbf{b}_i$, e.g., predicting the response of subject i at the next period entails estimating $\mu_{i,t+1} = h^{-1}(\sum_{j=1}^{p} \rho_j h(\bar{Y}_{t+1-j}) + \mathbf{x}'_{i,t+1}\boldsymbol{\beta} + (1, \mathbf{z}_{i,t+1})\mathbf{b}_i)$, in which $\mathbf{x}_{i,t+1}$ and $\mathbf{z}_{i,t+1}$ are assumed to be known at time t. When the parameters $\phi, \boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)'$ in (2.5) and (2.7) are known, $\psi_{t+1}(\mathbf{b}_i)$ can be estimated by the conditional expectation of $\psi_{t+1}(\mathbf{b}_i)$ given the data of the ith subject up to time t. Without assuming these parameters of the GLMM (2.7) to be known, we can estimate them by maximum likelihood using all the observations up to time t. Letting $\hat{\phi}_t$, $\hat{\boldsymbol{\alpha}}_t$, $\hat{\boldsymbol{\beta}}_t$ and $\hat{\boldsymbol{\theta}}_t$ be the corresponding MLEs, we can estimate the future value $\psi_{t+1}(\mathbf{b}_i)$ by

$$\hat{\psi}_{t+1,i} = E_{\hat{\phi}_t, \hat{\boldsymbol{\alpha}}_t, \hat{\boldsymbol{\beta}}_t, \hat{\boldsymbol{\theta}}_t}[\psi_{t+1}(\mathbf{b}_i) | \text{data of the } i\text{th subject up to time } t], \tag{2.8}$$

which can be computed by the hybrid method described in the Appendix.

In the preceding section we have assumed that the observations $(Y_{i,t}, \mathbf{x}_{i,t}, \mathbf{z}_{i,t})$ are available at every $1 \le t \le T$, for all $1 \le i \le n$. In longitudinal data in biostatistics, however, there is often between-subject variations in the observation times. Lai, Sun and Wong (2010) recently addressed this difficulty by using a prediction approach that customizes the predictive model for an individual by choosing predictors that are available at the individual's observation times. By making use of similar ideas, we can extend the dynamic EB approach of the preceding section to the setting where there is between-subject variations in the observation times, and also address the more basic problem concerning selection of variables for prediction of the individual's future response. Specifically, we propose to divide the subjects into K structurally similar subgroups. In many

applications, subjects belonging to the same subgroup have similar observation times because of their structural similarity. For example, patients who have more serious ailments are monitored more frequently than others in a study cohort, causing the irregularity of observation times over different subgroups. We assume the cross-sectional dynamics (2.7) separately for each subgroup, i.e., with μ_t and $\bar{Y}_{t-j}$ replaced by $\mu_t^{(k)}$ and $\bar{Y}_{t-j}^{(k)}$ for the kth subgroup, in which $\bar{Y}_s^{(k)}$ is the sample average from all subjects (from the subgroup) who are observed at time s. Moreover, for $\mu_{i,t}$ in (2.7) with i belonging to the kth subgroup, we only choose predictors $\mathbf{x}_{i,t}$ and $\mathbf{z}_{i,t}$ that are common to all subjects in the kth group, by using the BIC for the GLMM associated with that subgroup. The Appendix gives the definition and computational details of the BIC in GLMM.

3. Dynamic EB Models of Joint Default Intensities of Multiple Firms

In the wake of the 2007-08 financial crisis, it was widely recognized that models used previously to price credit derivatives such as CDOs (collateralized debt obligations) for a portfolio of firms had neglected the "frailty" traits of latent macroeconomic variables and the "contagion" effects of a firm's default on other firms in the portfolio. To account for the frailty effects, Duffie et al. (2009) introduced a dynamic frailty model for the default intensities $\lambda_i(t)$ of firms in the portfolio at time t, assuming an unobserved frailty process F_t in

$$\lambda_i(t) = \exp(\beta_0 + \boldsymbol{\beta}_1' \mathbf{X}_{i,t} + \boldsymbol{\beta}_2' \mathbf{U}_t + F_t) \tag{3.1}$$

to capture the cumulative effect of various unobserved fundamental common shocks to the default intensities of the n firms. The latent frailty process F_t is assumed to be an Ornstein-Uhlenbeck (OU) process

$$dF_t = \kappa(\mu - F_t)dt + \sigma dB_t, \quad F_0 = 0, \tag{3.2}$$

where B_t is a standard Brownian motion which volatility parameter is fixed to be 1, and $\kappa \geq 0$ is the mean-reversion rate of F_t. Because F_t is not observable, (3.1)-(3.2) is a hidden Markov model (HMM). In Section 3.1 we apply the dynamic EB approach to come up with a considerably simpler alternative to a common latent frailty process F_t, and show that its performance in predicting future default probabilities is comparable to that of the HMM even when the defaults are actually generated by (3.1)-(3.2). Section 3.2 describes further background and

applications of the dynamic EB approach to joint default modeling of corporate bonds and bank loans.

3.1. A logistic mixed model for dynamic frailty

Partitioning the time interval $(0, T^*]$ of default events in the study into disjoint intervals $I_0 = (0, t_1], \cdots, I_K = (t_K, t_{K+1}]$ with $t_{K+1} = T^*$, let $\pi_{i,k}$ denote the conditional probability of default of firm i in the time interval I_k given that it has not defaulted up to time t_k. Let $Y_{i,k}$ be the binary variable taking the value 0 or 1 for the event of the ith firm surviving or defaulting in the time interval I_k. Note that the value 1 (default) for $Y_{i,k}$ is an absorbing state. Let $\mathcal{H}_k$ denote the set of firms in the study that have not defaulted up to time k, and let $\bar{Y}_k = \sum_{i\in\mathcal{H}_k} Y_{i,k}/|\mathcal{H}_k|$. The dynamic EB approach in Section 2.2 amounts to the logistic mixed model $Y_{i,k+1}|Y_{i,k} = 0 \sim \text{Bernoulli}(\pi_{i,k})$, where

$$\text{logit}(\pi_{i,k}) = \beta + b_i + \rho\text{logit}(\bar{Y}_k) + \boldsymbol{\beta}_1'\mathbf{X}_{i,t_k} + \boldsymbol{\beta}_2'\mathbf{U}_{t_k}, \tag{3.3}$$

in which the b_i are the random effects and $\text{logit}(p) = \log(p/(1-p))$ is the canonical link of the Bernoulli distribution. Note that we use the coarser binary data $Y_{i,k}$ (instead of the default times up to T^*) to fit the logistic mixed model (3.3) (instead of the HMM (3.1)). The rationale behind this will be explained in Section 3.2.

To see the relationship between (3.3) and (3.1), we first assume that defaults only occur at integer times $t \geq 1$ and consider the discrete-time analog of (3.1), in which F_t is the discrete-time analog of the OU process (3.2), namely an AR(1) process of the form $F_t = \gamma F_{t-1} + \omega + \xi_t$, in which ξ_t are i.i.d. unobservable errors with mean 0 and variance V. Taking $t_k = k$ in $I_k = (t_k, t_{k+1}]$ of the preceding paragraph, there is no loss of information in using $Y_{i,k}$ for $k = 1, \cdots, T^* - 1$, because the actual default times up to T^* are integers. The default intensity $\lambda_{i,k}$ in the discrete-time HMM (3.1) is the conditional probability $P(Y_{i,k+1} = 1|Y_{i,k} = 0)$ and is given by $\exp(\beta_0 + \boldsymbol{\beta}_1'\mathbf{X}_{i,k} + \boldsymbol{\beta}_2'\mathbf{U}_k + F_{k+1})$, in which F_{k+1} is the unobserved common frailty of the firms at time $k+1$. Since $F_{k+1} = \gamma F_k + \omega + \xi_{k+1}$, the HMM can be written in the form

$$\log(\lambda_{i,k}) = \beta_0 + \omega + \gamma F_k + \xi_{k+1} + \boldsymbol{\beta}_1'\mathbf{X}_{i,k} + \boldsymbol{\beta}_2'\mathbf{U}_k. \tag{3.4}$$

Let $\beta = \beta_0 + \omega$ and compare (3.4) with (3.3), in which $t_k = k$. Instead of using a latent state F_t, (3.3) attempts to capture the effect of the common frailty of the

firms via the cross-sectional average default rate $\bar{Y}_t$. Since $\pi_{i,k}$ is typically small, $\text{logit}(\pi_{i,k}) \approx \log(\pi_{i,k})$, hence (3.3) essentially replaces γF_k in (3.4) by $\rho\text{logit}(\bar{Y}_k)$, and the normally distributed random disturbance ξ_{k+1} in the AR(1) model for F_{k+1} by subject-specific random effect b_i. Note that $\bar{Y}_k$ lies between 0 and 1 but F_k is normally distributed, which shows the importance of the link function $\text{logit}(\cdot)$ in using $\rho\text{logit}(\bar{Y}_k)$ as a surrogate for γF_k. We can alternatively use $\log(\cdot)$ as the link function h in (2.6) instead of the canonical link $\text{logit}(\cdot)$. Since F_k is an unobserved state and β, β_0, ρ and γ are unknown parameters that have to be estimated from the data, $\hat{\beta} + \hat{\rho}\text{logit}(\bar{Y}_k)$ and $\hat{\beta}_0 + \hat{F}_{k+1}$ may perform similarly as estimates of $\beta_0 + E(F_{k+1}|F_k)$ when the defaults are actually generated by the HMM. This is illustrated in the following simulation study which compares the performance of the 1-year ahead predictor, based on the logistic mixed model (3.3), of a firm's default probability with that based on the adaptive particle filter for the HMM (3.4).

Example 1. Consider $n = 500$ firms over a 30-year period. For simplicity, we choose univariate $X_{i,t}$ and U_t, which are the distance to default (Crosbie and Bohn, 2002; Duffie et al., 2009) for firm i and the three-month Treasury bill rate, respectively. Duffie, Saita and Wang (2007) have fitted AR(1) models to these covariates:

$$X_{i,t} = X_{i,t-1} + 0.04(\mu_i - X_{i,t-1}) + 0.3\eta_{i,t}, \quad U_t = 0.9U_{t-1} + 0.6 + 1.8\epsilon_t, \quad (3.5)$$

which our simulation study uses to generate the covariates, with $\mu_i \sim N(2, 0.5^2)$, $\eta_{i,t} \sim N(0,1), X_{i,1} \sim N(\mu_i, 0.3^2)$, $\epsilon_t \sim N(0,1)$ and $U_1 \sim N(6, 1.8^2)$. We also generate the AR(1) model $F_{t+1} = \gamma F_t + \omega + \xi_{t+1}$ with $\gamma = 0.5$, $\omega = 0.5$ and $\xi_{t+1} \sim N(0, 0.5^2)$. The discrete-time default intensity $\lambda_{i,t}$ is given by (3.4) with $(\beta_0, \beta_1, \beta_2) = (-2, -1, -0.3)$; we choose these regression parameters to match roughly the empirical results in Duffie, Saita and Wang (2007). Since the conditional probability of firm i defaulting at time $t+1$ given that it has not defaulted up to time t is $\pi_{i,t} = e^{-\lambda_i(t)}$, we generate $Y_{i,t+1} \sim \text{Bernoulli}(\pi_{i,t})$.

We fit the logistic mixed model (3.3) with $t_k = k$ to the simulated data for the 500 firms over a period of $T^* = 30$ years and compare the estimated $\hat{\pi}_{i,t}^{(1)}$ to the actual $\pi_{i,t}$, for $t = 16, \cdots, 30$; the comparison is only for firms that still survive at time t. We also compare $\hat{\pi}_{i,t}^{(1)}$ with the estimate $\hat{\pi}_{i,t}^{(2)}$ that uses the adaptive particle filter (Lai and Bukkapatanam, 2013) for the HMM to estimate the posterior

distribution of F_{t+1} and therefore also of $\lambda_{i,t}$. Both estimates use training data up to time t. Figure 1 gives the result for a simulated firm that survives throughout the entire 30-year period. Figure 2 plots the estimates $\hat{\beta}+\hat{\rho}\text{logit}(\bar{Y}_k)$ and $\hat{\beta}_0+\hat{F}_{k+1}$ based on data up to time k, and compares them with $\beta_0+\omega+\gamma F_k$ in a simulated set of 500 firms used in Figure 1. Note that although $\hat{\beta}+\hat{\rho}\text{logit}(\bar{Y}_k)$ differs substantially from $\beta_0+E(F_{k+1}|F_k)$, $\hat{\beta}_0+\hat{F}_{k+1}$ is also not close to $\beta_0+E(F_{k+1}|F_k)$

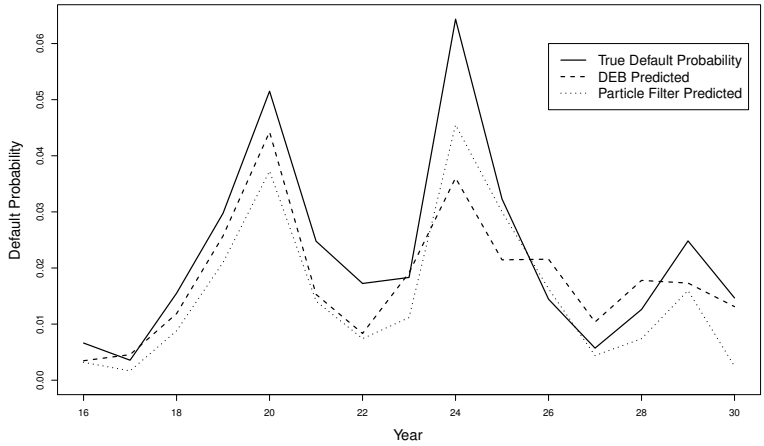

Figure 1. Predicted default probabilities for year t based on data up to $t-1$

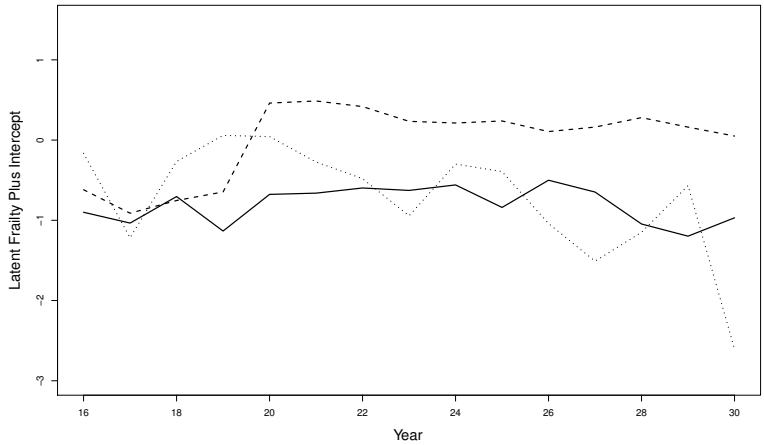

Figure 2. Comparison of the solid curve $\beta_0 + E(F_t|F_{t-1})$ with the dash curve $\hat{\beta} + \hat{\rho}\ \text{logit}(\bar{Y}_{t-1})$ and the dotted curve $\hat{\beta}_0 + \hat{F}_t$

since it is an adaptive filter that predicts F_{k+1} from the observations $Y_{i,s}$, $s \leq k, 1 \leq i \leq n$, rather than from the unobserved state F_k. Thus, $\hat{\pi}_{i,t}^{(1)}$ and $\hat{\pi}_{i,t}^{(2)}$ have similar performance as estimates of the conditional probability $\pi_{i,t}$.

We have generated 100 simulated data sets in this way and computed $\pi_{i,t}$, $\hat{\pi}_{i,t}^{(1)}$ and $\hat{\pi}_{i,t}^{(2)}$ for each data set. Table 1 gives the mean and 5-number summaries of the absolute prediction errors $\sum_{i\in\mathcal{H}_t} |\pi_{i,t} - \hat{\pi}_{i,t}^{(j)}|/|\mathcal{H}_t|$ for $j = 1$ (dynamic EB via logistic mixed model, denoted DEB) and $j = 2$ (adaptive particle filter, denoted APF), $t = 16, 18, 20, 25, 30$. It shows that the dynamic EB approach performs favorably in comparison with the adaptive particle filter.

Table 1. Five-number summaries (minimum, 1st quartile Q_1, median Q2, 3rd quartile Q_3, and maximum) and mean of absolute prediction errors, all multiplied by 100

	t	Min	Q_1	Q_2	Q_3	Max	Mean
DEB	16	0.0117	0.295	0.812	1.84	8.44	1.49
APF	16	0.0185	0.261	0.555	1.75	95.6	2.99
DEB	18	0.0531	0.309	0.635	1.38	14.2	1.50
APF	18	0.0342	0.337	0.666	1.46	98.1	3.61
DEB	20	0.0648	0.244	0.476	1.01	18.5	0.980
APF	20	0.0206	0.276	0.569	1.30	99.4	3.96
DEB	25	0.0392	0.221	0.453	1.32	13.2	1.18
APF	25	0.0232	0.303	0.819	2.41	98.8	12.3
DEB	30	0.0233	0.254	0.470	1.54	7.22	1.09
APF	30	0.0132	0.457	1.15	6.33	99.0	14.1

We now consider the case of continuous default times with default intensities (3.1) for the n firms. We use the life-table approach described in the first paragraph of this section. Although using the default indicator $Y_{i,k}$ in the time interval loses some information contained in the observed default times, the loss is relatively minor, as illustrated in the following example that shows the logistic mixed model (3.3) to have comparable performance in predicting default probabilities as the HMM (3.1) that actually generates the default events.

Example 2. In this example, suppose the latent frailty F_t follows a continuous-time O-U process (3.2) with $\kappa = 0.125, \mu = 1$ and $\sigma = 0.5$, instead of the discrete-time AR(1) model, and still assume $n = 500$ firms over a period of $T = 30$ years, with $e_i = 0$ for all i. We use (3.1) and (3.2) to generate the firms' default times by using the "thinning algorithm" for non-homogeneous Poisson

processes (Ross, 2013). We can use the adaptive particle filter to estimate the posterior distribution of F_t and thereby compute the APF estimate $\hat{\lambda}_{i,t}^{(2)}$ of $\lambda_i(t)$. Details of the APF, which basically involves a set of $N = 1000$ atoms and their associated weights to represent the posterior distribution of the parameter vector $\boldsymbol{\theta} = (\kappa, \mu, \sigma, \beta_0, \beta_1, \beta_2)$, $K = 1000$ MCMC iterations to choose the atoms sequentially, and $M = 5000$ trajectories ("particles") of the latent process, are given in Lai and Bukkapatanam (2013). We also use the coarser binary data $Y_{j,s}$ $(s \le t, j = 1, \cdots, n)$ to fit the considerably simpler logistic mixed model (3.3) and thereby compute the dynamic EB estimate $\hat{\pi}_{i,t}^{(1)}$ of the default probability $\pi_{i,t} = P(t < \tau_i \le t+1 | \tau_i \ge t) = 1 - \exp\left(-\int_t^{t+1} \lambda_i(s) ds\right)$. Figure 3 plots the actual default intensities of a simulated firm that survives throughout the entire 30-year period and the estimated intensities $\hat{\lambda}_{i,t}^{(2)}$ and $\hat{\lambda}_{i,t}^{(1)} = -\log(1 - \hat{\pi}_{i,t}^{(1)})$ at $t = 15, \cdots, 29$.

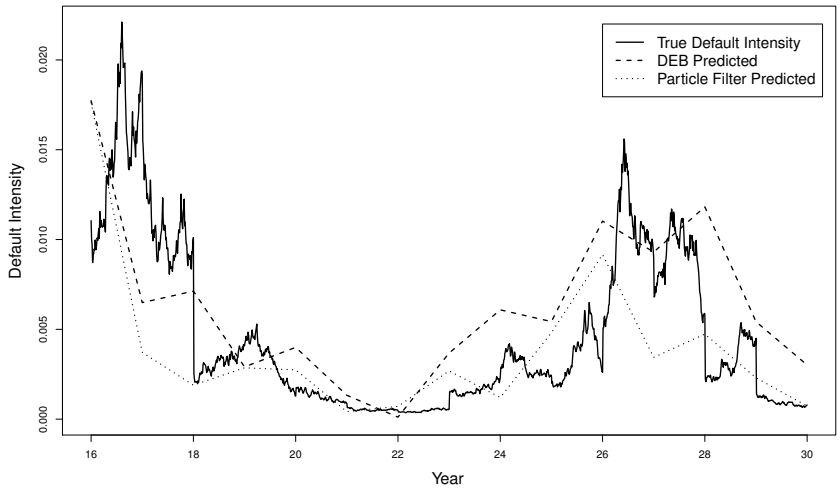

Figure 3. Estimated default intensities for year t based on data up to $t-1$

We have generated 100 simulated data sets in this way and computed $\pi_{i,t}, \hat{\pi}_{i,t}^{(1)}$ and $\hat{\pi}_{i,t}^{(2)}$ for each data set. The computation of $\pi_{i,t}$ and $\hat{\pi}_{i,t}^{(2)}$ each involves 1000 additional Monte Carlo simulations to generate the conditional distribution of $\{F_s, t < s \le t+1\}$. Table 2 gives the mean and 5-number summaries of the absolute prediction errors (as defined in the paragraph following Figure 2) for DEB and APF. The pattern is similar to that in Table 1, showing that DEB

compares favorably with APF which tends to give somewhat smaller absolute errors below the median but larger ones beyond the third quartile. One possible explanation is that even though the data are generated by the assumed HMM, the complexity of the HMM seems to result in MCMC estimates of $\boldsymbol{\theta}$ that are not accurate enough for the particle filter, for a certain fraction of the sample paths. We have increased the number K of MCMC iterations for these sample paths, but it only leads to slight improvements of the results for APF in Table 2.

Table 2. Five-number summaries (minimum, 1st quartile Q_1, median Q2, 3rd quartile Q_3, and maximum) and mean of absolute prediction errors, all multiplied by 100

	t	Min	Q_1	Q_2	Q_3	Max	Mean
DEB	16	0.00803	0.355	0.880	1.62	23.5	1.90
APF	16	0.00408	0.446	0.903	2.50	43.4	2.43
DEB	18	0.0405	0.321	0.686	1.55	19.0	1.63
APF	18	0.0193	0.296	0.765	2.08	24.2	2.18
DEB	20	0.0321	0.265	0.622	1.49	8.91	1.30
APF	20	0.0349	0.304	0.699	1.91	17.1	1.65
DEB	25	0.0349	0.310	0.640	1.32	6.00	1.11
APF	25	0.0426	0.331	0.777	1.52	14.0	1.51
DEB	30	0.0222	0.264	0.538	1.67	11.6	1.39
APF	30	0.0168	0.257	0.663	1.83	24.3	19.7

3.2. Extension to competing risks and loan portfolios

Unlike the discrete-time default indicator variables $Y_{i,t}$ in Section 3.1, Duffie et al. (2009) actually use censored survival data to fit the continuous-time HMM (3.1)–(3.2). The $\mathbf{X}_{i,t}$ in (3.1) is a firm-specific covariate vector containing the firm's distance to default and its trailing 1-year stock return, and $\mathbf{U}_t$ is a macroeconomic vector containing the 3-month Treasury bill rate and the trailing 1-year return on the S&P 500 index. Duffie et al. (2009) fit the HMM (3.1)–(3.2) to a set of 402,434 firm-months of data between January 1979 and March 2004. The data at time t can be represented by the vector $\mathbf{Y}_t = ((T_i \wedge (t - e_i)^+, \delta_{i,t}, \mathbf{X}_{i,t}, \mathbf{U}_t), i = 1, \cdots, n)$, where $T_i = \tau_i \wedge c_i$, τ_i is the default time of the ith firm (measured from the firm's entry time e_i into the empirical study), c_i is the censoring variable caused by the firm's exit from the study because of merger, acquisition or other failure, and $\delta_{i,t}$ is the default indicator (taking the value 0 or 1) so that $\delta_{i,t} = 1$ if $T_i \wedge (t - e_i) = \tau_i$. Assuming τ_i

and c_i to be independent, the likelihood function can be written as

$$g_{\boldsymbol{\theta}}(\mathbf{Y}_t|F_t) = \prod_{i=1}^{n}(\lambda_i(T_i,\boldsymbol{\theta}))^{\delta_{i,t}}e^{-\Lambda_i(T_i,\boldsymbol{\theta})}, \tag{3.6}$$

in which $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \kappa, \mu, \sigma)$ denotes the parameter vector and $\Lambda_i(t;\boldsymbol{\theta}) = \int_0^t \lambda_i(s;\boldsymbol{\theta})ds$ is the cumulative hazard function. Duffie et al. (2009) use a stochastic EM algorithm to estimate $\boldsymbol{\theta}$ and MCMC methods involving both Gibbs sampling and Metropolis-Hastings steps to estimate the latent frailty process. Lai and Bukkapatanam (2013) propose to use a faster adaptive particle filter instead, which enables us to carry out simulation studies in Examples 1 and 2.

The assumption of independent intensity processes for the default and exit times τ_i and c_i is called "doubly stochastic". Duffie, Saita and Wang (2007, p.637) acknowledged that "the doubly-stochastic assumption is overly restrictive" and that previous work has shown this assumption "does not fit the data well". A better way is to use the *competing risks* approach that classifies failures into types (e.g., failure from the disease process and from non-disease related causes). This approach considers the cause-specific hazard rate $\lambda_i^j(t) = \lim_{h\to 0} h^{-1}P(t \le T_i \le t+h, J_i = j|T_i \ge t)$, in which J_i is the cause of failure of subject i (Andersen et al., 1993, pp 298-304). It can be easily combined with dynamic EB modeling via the life-table method, leading to multinomial logistic (or multilogit) mixed models that we describe below.

Partitioning time into disjoint intervals $I_0 = [0, t_1), \cdots, I_k = [t_k, t_{k+1}), \cdots$, as in Section 3.1, let $\pi_{i,k;1}$ denote the conditional probability of default of firm i in the time interval I_k given that it has neither defaulted nor exited up to time t_k. Similarly, let $\pi_{i,k;2}$ denote the conditional probability of firm i exiting in the time interval I_k, and note that default, exit and surviving are mutually exclusive events. Let $Y_{i,k}$ be the trinomial variable taking the value 0, 1, or 2 for the event of surviving, default, or exit in the time interval I_k. Let

$$\eta_{i,k;j} = \log\left(P\{Y_{i,k} = j|Y_{i,k-1} = 0\}/P\{Y_{i,k} = 0|Y_{i,k-1} = 0\}\right),\ j = 1, 2. \tag{3.7}$$

The multilogit mixed model, which is a generalization of the logistic mixed model (3.3), can be applied to the trinomial outcomes $Y_{i,k}$:

$$\eta_{i,k;j} = \beta_{0j} + b_{0j} + \rho_j \log\left(\bar{Y}_k^{(j)}/\bar{Y}_k^{(0)}\right) + \boldsymbol{\beta}_{1j}'\mathbf{X}_{i,t_k} + \boldsymbol{\beta}_{2j}'\mathbf{U}_{t_k},\ j = 1, 2, \tag{3.8}$$

where $\bar{Y}_k^{(j)} = (\sum_{i\in\mathcal{H}_k} I_{\{Y_{i,k}=j\}})/|\mathcal{H}_k|$ and $\mathcal{H}_k$ is the set of firms that have neither defaulted nor exited up to time k.

Although using the event indicator $Y_{i,k}$ in the time interval I_k loses some information contained in observed event times T_i, the loss is relatively minor, as shown in Example 2. Moreover, the quantity of interest in credit risk management is the probability of default in the next month (or year), rather than forecasting the actual time to default of firm i. Besides being considerably simpler, an advantage of (3.8) is that it dispenses with the assumption of independence between the default and exit times. In fact, similar multilogit models and multilogit mixed models have been widely used in studying large portfolios of mortgage loans, with default and prepayment as competing risks for mortgage terminations; see Calhoun and Deng (2002), Clapp, Deng and An (2006), and Chapter 7 of Lai and Xing (2014) where the issue of evaluation of the performance of these probability forecasts is also addressed. The dynamic EB approach that includes the term $\rho_j \log\left(\bar{Y}_k^{(j)}/\bar{Y}_k^{(0)}\right)$ in (3.8) can be included to enhance these models.

4. Applications to Prediction of Baseball Batting Averages

Batting average is an important performance measure for baseball players. For non-pitchers, a seasonal batting average is considered to be excellent if it is above 0.3, and is regarded unsatisfactory if it is below 0.2. It is defined as the ratio of "hits" (number of successful attempts) to "at bats" (number of qualifying attempts). The problem of predicting the batting performance of baseball players was first studied by Efron and Morris (1975, 1977), who used the batting averages from the first $m = 45$ at-bats of a small sample of $n = 18$ batters in 1970 to predict their batting averages for the remainder of the season. Specifically, let X_i and p_i denote the observed batting average after 45 at bats and the actual seasonal batting average, respectively, of player i ($1 \le i \le 18$). Assuming X_i to be independently distributed with $mX_i \sim \text{Bin}(m, p_i)$, Efron and Morris (1975, 1977) applied the variance-stabilizing transformation

$$Y_i = m^{1/2} \arcsin(2X_i - 1) \tag{4.1}$$

so that Y_i is approximately $N(\mu_i, 1)$, where $\mu_i = n^{1/2} \arcsin(2p_i - 1)$. They used the James-Stein (1961) estimator of μ_i to demonstrate the benefit of shrinkage and linear EB. Applying a different variance-stabilizing transformation, Brown

(2008) used the batting records of Major League players from an earlier part of the 2005 regular season to estimate each player's hitting probability p_i by different methods that are "motivated from empirical Bayes and hierarchical Bayes interpretations" and thereby to compare how well they predict the batting performance of the players for the remainder of the season.

In this section we apply the dynamic EB approach in Section 2 to the prediction of batting performance. We consider data from the five regular Major League seasons 2006-2010. Each regular season runs from late March to early October, so six "monthly" (Mar/Apr, May, Jun, Jul, Aug, Sep/Oct) data sets are collected for each of the 5 years. The batting averages, as well as other useful baseball statistics, are available for download from the website `http://www.fangraphs.com/leaders.aspx`. In order to reduce variability and to compare with Brown's results, the monthly data are aggregated into semi-seasonal (3-month) data, resulting in 10 semi-seasonal periods, labeled by $t = 1, \cdots, 10$. To apply EB and dynamic EB methods, we want the n individual players to be structurally similar. Since baseball players are categorized into batters, pitchers and fielders and since batting average is one of the key performance measures for batters but not for pitchers and fielders, we only consider batters and record the number $H_{i,t}$ of "hits" and the number $N_{i,t}$ of "at bats" for batter i in period t.

Following Efron and Morris (1975, 1977) and Brown (2008), we assume that $H_{i,t} \sim \text{Bin}(N_{i,t}, p_{i,t})$ when $N_{i,t} > 0$, where $p_{i,t}$ is the hitting probability of the ith batter in the tth period. Unlike these references that consider a single season and assume $p_{i,t}$ to be constant over the season, we allow $p_{i,t}$ to vary over the semi-seasons. To be comparable to their results, we consider predicting the hitting probabilities at $t = 6, 8, 10$ (i.e., for the second half of the $2006, 2008, 2010$ season) based on $(N_{i,s}, H_{i,s})$ for $s \le t-1$ and i belonging to the group of batters included in the study. Brown (2008) used the transformation

$$Y_{i,t} = \arcsin\left(\sqrt{\frac{H_{i,t}+1/4}{N_{i,t}+1/2}}\right),\ \mu_{i,t} = \arcsin\left(\sqrt{p_{i,t}}\right), \tag{4.2}$$

so that Y_{it} is approximately $N(\mu_{i,t}, 1/(4N_{i,t}))$. This is a refinement of (4.1) so that the normal approximation, with variance not depending on $p_{i,t}$, can still hold for smaller values of $N_{i,t}$ than those required by (4.1). Although the accuracy

of the normal approximation actually depends on $N_{i,t}p_{i,t}$, most of the batters have batting averages between 0.2 and 0.3 and therefore it suffices to focus on $N_{i,t}$ instead. Brown (2008) includes in his study players "having more than 10 at-bats," i.e., $N_{i,t} \geq 11$. Since the study includes a training period corresponding to the first half of the season and a test set corresponding to the second half of the season, it actually requires

$$N_{i,t} \geq 11 \text{ and } N_{i,t-1} \geq 11 \tag{4.3}$$

for $t = 6, 8, 10$ in our setting. Batters satisfying (4.3) will be called "eligible" in period t. In Section 4.1, we use this criterion for including batters into our study that applies linear EB and dynamic EB methods to $Y_{i,t}$. In Section 4.2, we relax the inclusion criterion and apply the dynamic EB approach via GLMM directly to $(N_{i,t}, H_{i,t})$ and show how the methodology recently developed by Lai, Gross and Shen (2011) can be applied to evaluate the prediction of $p_{i,t}$ based on data up to $t-1$.

4.1. Linear and dynamic linear EB predictors of Y_{it}

Here we apply the dynamic linear EB approach to the prediction of $Y_{i,t}$ for eligible batters (i.e., those who satisfy (4.3)) in periods $t = 6, 8, 10$. The number of eligible players is 495 at $t = 10$, 497 at $t = 8$ and 500 at $t = 6$. The dynamic linear EB model we consider is of the form (2.4) with $p = 2$, $x_{i,t} = 1$ and without the term $\mathbf{b}_i'\mathbf{z}_{i,t}$. We choose $p = 2$ because $X_{i,t-2}$ is the batting average at the end of the past season and $X_{i,t-1}$ is that at the half-season for $t = 6, 8, 10$. Model selection using BIC further reduces the LMM to

$$Y_{i,t} = \begin{cases} \rho\bar{Y}_{t-1} + \beta + a_i & \text{for } t = 8, 10 \\ \rho\bar{Y}_{t-2} + \beta + a_i & \text{for } t = 6, \end{cases} \tag{4.4}$$

in which $a_i \sim N(0, \sigma^2)$. We can use those batters in the training sample who satisfy $\min(N_{i,s}, N_{i,s-1}, N_{i,s-2}) \geq 11$ for some $s \leq t-1$ to fit the "full model" in which $x_{i,s}$ above is augmented to $\mathbf{x}_{i,s} = (1, Y_{i,s-1}, Y_{i,s-2})'$, allowing autoregression of the batter's successive batting averages. Including this full model for model selection using BIC in the case $t = 10$ still chooses the model in the preceding paragraph that does not have the $Y_{i,s-1}$ and $Y_{i,s-2}$ terms.

Assuming $\mu_{i,t} = \mu_{i,t-1}$, Brown (2008) considered three linear EB estimators of $\mu_{i,t-1}$ (and therefore also of $\mu_{i,t}$) based on $Y_{j,t-1}$ for all batters with $N_{j,t-1} \geq 11$;

this includes batter i in view of (4.3). The linear EB estimators are EB(MM) which uses the method of moments (MM) to estimate the hyperparameters in the Bayes estimator, EB(ML) that estimates the hyperparameters by maximum likelihood instead, and the James-Stein estimator denoted by JS. Besides these linear EB estimators, he also considered for comparison the mean estimator $\bar{Y}_{t-1}$ and the "naive" estimator $Y_{i,t-1}$ of $\mu_{i,t-1}$. These estimates can be used to predict $\mu_{i,t}$ and are denoted by $\hat{Y}_{i,t}$. Because the actual $\mu_{i,t}$ is unknown and may not equal $\mu_{i,t-1}$ as assumed, an obvious way to evaluate prediction performance is to consider the discrepancy between $Y_{i,t}$ and its predictor $\hat{Y}_{i,t}$. Since $Y_{i,t}$ is approximately $N(\mu_{i,t}, 1/(4N_{i,t}))$, Brown (2008) proposed to use the estimated total squared error

$$\widehat{\mathrm{TSE}} = \sum\nolimits_{i\text{: batter } i \text{ is eligible at } t} \left\{ (Y_{i,t} - \hat{Y}_{i,t})^2 - \frac{1}{4N_{i,t}} \right\} \tag{4.5}$$

as a measure of prediction performance. This is an unbiased estimate of the squared-error loss $\sum_{i\text{: batter } i \text{ is eligible at } t}(\mu_{i,t} - \hat{Y}_{i,t})^2$, and is the same as the *adjusted Brier score* proposed by Lai, Gross and Shen (2011, Section 6.1) since the variance of the arcsin-transformed sum $Y_{i,t}$ is $1/(4N_{i,t})$. Brown (2008) also considered the normalized estimated squared error $\widehat{\mathrm{NSE}} = \widehat{\mathrm{TSE}}/\widehat{\mathrm{TSE}}_0$, where $\widehat{\mathrm{TSE}}_0$ is the estimated total squared error for the naive predictor $\hat{Y}_{i,t} = Y_{i,t-1}$. Table 3 gives the $\widehat{\mathrm{TSE}}$ and $\widehat{\mathrm{NSE}}$ of these predictors of $Y_{i,t}$ for $t = 6, 8, 10$ and those of the LMM (4.4). Also given each predictor are the 5-number summaries of the absolute errors $|Y_{i,t} - \hat{Y}_{i,t}|$ for $t = 6, 8, 10$. The results show the advantages of dynamic EB via LMM over the linear EB methods considered by Brown.

4.2. Dynamic EB prediction of $p_{i,t}$ via GLMM

Brown (2008, p.32) has treated (4.2) as $N(\mu_{i,t}, 1/(4N_{i,t}))$ random variables "as long as $N_{i,t} \geq 12$." Although he relaxes this to $N_{i,t} > 10$ for the inclusion of players in his empirical study, Section 7 of his paper imposes the stronger constraint to develop tests of independence between the players' batting averages in the two halves of a season. Note that the GLMM approach developed in Section 2.2 can be applied directly to $H_{i,t} \sim \mathrm{Bin}(N_{i,t}, p_{i,t})$ without relying on the normal approximation via the transformation (4.2). Specifically, $\mathrm{Bin}(N_{i,t}, p_{i,t})$ belongs to the exponential family (2.5) with $\theta_{i,t} = \log(p_{i,t}/(1-p_{i,t}))$ and $g(\theta_{i,t}) = -N_{i,t}\log(1-p_{i,t})$. Therefore, instead of transforming $H_{i,t}$ to $Y_{i,t}$ via (4.2) and

Table 3. Estimated total squared error, normalized squared error, and five-number summaries of absolute prediction errors (multiplied by 10^3) for different predictors.

	Naive ($Y_{i,t-1}$)			Mean ($\bar{Y}_{t-1}$)			EB(MM)		
	$t=6$	$t=8$	$t=10$	$t=6$	$t=8$	$t=10$	$t=6$	$t=8$	$t=10$
$\widehat{\text{TSE}}$	1.93	2.36	1.57	1.78	1.45	1.96	1.10	1.07	1.68
$\widehat{\text{NSE}}$	1	1	1	0.918	0.613	1.25	0.567	0.453	1.07
Min	0	0.459	0	0.0144	0.0506	0.293	0.116	0.267	0.0502
Q_1	20.2	18.2	18.9	26.1	22.4	19.7	17.3	17.1	18.9
Med	43.2	46.3	41.6	48.8	44.7	41.0	38.0	37.9	36.5
Q_3	80.0	89.0	78.8	81.0	76.8	71.7	69.9	71.5	67.6
Max	449	445	391	413	368	416	396	362	401
	EB(ML)			JS			LMM		
	$t=6$	$t=8$	$t=10$	$t=6$	$t=8$	$t=10$	$t=6$	$t=8$	$t=10$
$\widehat{\text{TSE}}$	0.975	0.820	1.21	0.962	1.01	1.48	0.440	0.393	0.344
$\widehat{\text{NSE}}$	0.504	0.347	0.770	0.497	0.426	0.941	0.228	0.166	0.219
Min	0.256	0.116	0.124	0.167	0.240	0.185	0.381	0.180	0.0466
Q_1	18.2	18.5	18.5	16.7	16.0	17.8	18.4	15.6	16.5
Med	41.0	37.6	38.0	36.5	35.4	35.2	37.3	33.4	36.0
Q_3	71.2	68.2	65.6	67.4	67.0	62.7	66.8	61.1	58.7
Max	383	379	383	390	373	404	318	399	356

applying LMM (4.4) to $Y_{i,t}$, we can model $H_{i,t}$ directly by the GLMM

$$\text{logit}(p_{i,t}) = \alpha + \beta_1 \text{logit}(\bar{p}_{t-1}) + \beta_2 \text{logit}(\bar{p}_{t-2}) + b_i, \tag{4.6}$$

where $b_i \sim N(0, \sigma^2)$ is the subject-specific random effect and $\bar{p}_s$ is the average of $H_{i,s}/N_{i,s}$ over i in the training sample.

We next apply the GLMM (4.6) to predict $p_{i,t}$ for the subgroup of *relatively infrequent* batters, defined by those with

$$2 \le N_{i,t} \le 32 \text{ and } 0 < \bar{N}_{i,t-} \le 32 \tag{4.7}$$

in period t, where $\bar{N}_{i,t-}$ denotes the average number-at-bats of batter i over the periods $s \le t-1$ when $N_{i,s} \ge 2$, so $\bar{N}_{i,t-} > 0$ means that there is at least one such period. The choice of the threshold 32 will be explained later. If the batter does not play in period t, there is no information on his batting ability in that period. Batting only once also does not yield a meaningful average as it is either 0 or 1, and therefore we impose the lower bound 2 for $N_{i,t}$ in defining relatively

infrequent batters. Moreover, since a major difference between EB and a purely Bayesian approach is that it combines the individual's data with the data from other structurally similar subjects to come up with an estimate of the individual's latent parameter, a batter in the test sample must also belong to the training sample to obtain his EB estimate. This explains why we also require $\bar{N}_{i,t-} \le 32$ in (4.7) to reflect that the batter also bets infrequently, on average whenever he bats at least twice, from period 1 to $t-1$.

Batters who satisfy (4.7) may be relatively new (including rookies) or old (including those near retirement) or used as substitutes for regular batters when they need some rest. They form a structurally similar subgroup that differs from the subgroup of regular batters. Since $N_{i,t}$ can be as small as 2, $H_{i,t}$ may not have much information about $p_{i,t}$ and therefore it appears difficult to evaluate predictors of $p_{i,t}$ in this case. Lai, Gross and Shen (2011) have recently resolved this difficulty and have developed a comprehensive methodology for such evaluation. In particular, letting S_t denote the subgroup of infrequent batters in period t (i.e., those satisfying (4.7)), we can estimate consistently the squared-error loss

$$L_t = \sum\nolimits_{i\in S_t} N_{i,t}(p_{i,t} - \hat{p}_{i,t})^2/N_t \tag{4.8}$$

by the adjusted Brier score

$$\hat{L}_t = \Big[\sum\nolimits_{i\in S_t} N_{i,t}\{h_{i,t}(1-\hat{p}_{i,t})^2 + (1-h_{i,t})\hat{p}_{i,t}^2\} - \sum\nolimits_{i\in S_t} N_{i,t}v_{i,t}\Big]_+/N_t, \tag{4.9}$$

where $N_t = \sum_{i\in S_t} N_{i,t}$, $h_{i,t} = H_{i,t}/N_{i,t}$, $v_{i,t} = N_{i,t}h_{i,t}(1-h_{i,t})/(N_{i,t}-1)$ and $\hat{p}_{i,t}$ is a predictor of $p_{i,t}$ that depends on the observations up to $t-1$. Note that $v_{i,t}$ is well defined since $N_{i,t} \ge 2$ by (4.7). Moreover, Lai, Gross and Shen (2011) have shown that the Brier loss difference $L_t - \tilde{L}_t$ between two predictors $\hat{p}_{i,t}$ and $\tilde{p}_{i,t}$ of $p_{i,t}$ can be consistently estimated by

$$\Delta_t = \sum\nolimits_{i\in S_t} N_{i,t}[h_{i,t}\{(1-\hat{p}_{i,t})^2 - (1-\tilde{p}_{i,t})^2\} + (1-h_{i,t})(\hat{p}_{i,t}^2 - \tilde{p}_{i,t}^2)]/N_t. \tag{4.10}$$

A widely used alternative to the Brier loss is the *Kullback-Leibler loss* L_t^{KL} that replaces $(p_{i,t} - \hat{p}_{i,t})^2$ in (4.9) by the Kullback-Leibler divergence

$$p_{i,t}\log(p_{i,t}/\hat{p}_{i,t}) + (1-p_{i,t})\log[(1-p_{i,t})/(1-\hat{p}_{i,t})]. \tag{4.11}$$

The difference $L_t^{\mathrm{KL}} - \tilde{L}_t^{\mathrm{KL}}$ can also be consistently estimated by

$$\Delta_t^{\mathrm{KL}} = \sum\nolimits_{i\in S_t} N_{i,t}[h_{i,t}\log(\tilde{p}_{i,t}/\hat{p}_{i,t}) + (1-h_{i,t})\log\{(1-\tilde{p}_{i,t})/(1-\hat{p}_{i,t})\}]/N_t.$$

Let S_t^1 be the subset of S_t satisfying the additional condition

$$N_{i,t-1} \geq 11. \tag{4.12}$$

The linear EB methods in Section 4.1 can be applied to predict $\mu_{i,t}$ using the transformed variables $Y_{j,t-1}$ for $j \in S_t^1$. The predictor $\hat{\mu}_{i,t}$ can be transformed back to yield the predictor $\hat{p}_{i,t} = (\sin \hat{\mu}_{i,t})^2$ of $p_{i,t}$. Table 4 gives the adjusted Brier scores $\hat{L}_t$ of these predictors and of the predictor "Bin" which applies the GLMM (4.6) directly to $H_{j,s}$ without transforming it to $Y_{j,s}$ for $j \in S_t^1$ and $s \leq t-1$. It also gives the differences Δ_t and Δ_t^{KL} between each of these linear EB predictors and Bin, which corresponds to $\tilde{p}_{i,t}$ in (4.10).

The LMM (4.4) only requires

$$N_{i,s} \geq 11 \text{ for some } s \leq t-1, \tag{4.13}$$

which is weaker than (4.12). Let S_t^2 denote the subset of S_t satisfying (4.13). Table 4 also gives the adjusted Brier scores and Δ_t, Δ_t^{KL} values of the LMM and Bin predictors when they are based on S_t^2 instead of S_t^1. Since Bin can be applied to the larger set S_t, Table 4 also gives the adjusted Brier score of Bin when it is based on S_t. The cardinalities $\#(\cdot)$ of S_t^1, S_t^2 and S_t are also shown in the table, and so are the numbers of batters in the associated training sample $\mathcal{T}_{t-1}^1$, $\mathcal{T}_{t-1}^2$, and $\mathcal{T}_{t-1}$. We have chosen the threshold 32 in (4.7) because it corresponds to the 20th percentile, at $t = 10$, of $N_{i,t-1}$ for batters with $N_{i,t-1} \geq 11$. Table 4 shows that Δ_t and Δ_t^{KL} of the linear EB and LMM predictors based on S_t^1 are all positive, demonstrating the advantage of the Bin predictor. Note that S_t is a substantially larger set than S_t^2, and only Bin is applicable to $S_t - S_t^2$. Moreover, only LMM and Bin are applicable to $S_t^2 - S_t^1$, and Table 4 shows that there is negligible difference between their predictive performances based on S_t^2.

5. Discussion

While linear EB estimators such as (2.6) have provided basic credibility formulas in insurance rate-making, in practice an insurance policy is held over time and we propose herein a new dynamic EB approach to the prediction of future claims of an individual (or risk class) by pooling cross-sectional information over individual time series in a LMM or GLMM. There are many possibilities to model these time series data, allowing subject-specific random effects and using dynamics (through lagged variables) for the individual and cross-sectional time series.

Table 4. Adjusted Brier scores $\hat{L}_t$ and differential Brier and Kullback-Leibler scores, Δ_t and Δ_t^{KL}, for various predictors of p_{it} (all multiplied by 10^3).

	$t=10$			$t=8$			$t=6$		
	$\hat{L}_t$	Δ_t	Δ_t^{KL}	$\hat{L}_t$	Δ_t	Δ_t^{KL}	$\hat{L}_t$	Δ_t	Δ_t^{KL}
	(a) S_t^1-based predictors								
	$\lvert S_{10}^1\rvert=60,\ \lvert\mathcal{T}_9^1\rvert=104$			$\lvert S_8^1\rvert=52,\ \lvert\mathcal{T}_7^1\rvert=104$			$\lvert S_6^1\rvert=57,\ \lvert\mathcal{T}_5^1\rvert=103$		
EB(MM)	2.52	0.783	3.25	3.02	0.776	3.02	34.8	35.9	148
EB(ML)	2.71	0.974	3.96	3.09	0.843	3.29	0	0.429	1.74
JS	2.57	0.831	3.41	3.15	0.906	3.50	0	0.575	2.39
LMM	1.88	0.146	0.147	2.67	0.423	1.74	0	0.288	1.38
Bin	1.73	0	0	2.25	0	0	0	0	0
	(b) S_t^2-based predictors for LMM and Bin								
	$\lvert S_{10}^2\rvert=80,\ \lvert\mathcal{T}_9^2\rvert=286$			$\lvert S_8^2\rvert=77,\ \lvert\mathcal{T}_7^2\rvert=242$			$\lvert S_6^2\rvert=70,\ \lvert\mathcal{T}_5^2\rvert=190$		
LMM	1.21	-0.159	-0.845	3.18	-0.136	-0.643	0	0.100	0.506
Bin	1.37	0	0	3.32	0	0	0	0	0
	(c) S_t-based predictor using Bin								
	$\lvert S_{10}\rvert=112,\ \lvert\mathcal{T}_9\rvert=649$			$\lvert S_8\rvert=124,\ \lvert\mathcal{T}_7\rvert=563$			$\lvert S_6\rvert=110,\ \lvert\mathcal{T}_5\rvert=437$		
Bin	1.76	0	0	2.29	0	0	0	0	0

Model selection is important to avoid deterioration of prediction performance because of over-fitting. A subtle point noted in Section 2.3 is that for longitudinal data, subjects may be observed at different time-points, and an individual's predictor has to be developed by pooling information from subjects that have observations at these time-points. An important innovation of our dynamic EB approach is to replace μ_s by $\bar{Y}_s$ $(s<t)$ in the state-space model of Bühlmann and Gisler (2005), thereby providing flexible and computationally efficient models for evolutionary credibility. This is akin to using GARCH models instead of stochastic volatility models in financial econometrics; see Lai and Xing (2008). The dynamic EB approach pools cross-sectional information over individual time series to come up with flexible and computationally efficient methods for modeling longitudinal data and predicting future outcomes of the individuals. We have shown in Section 3 how this approach can be used to approximate an inherently complicated hidden Markov model of joint default intensities of multiple firms subject to the impact of observed and latent dynamic macroeconomic variables by a much simpler GLMM, the advantages of which are illustrated by the

simulation studies in Example 1 and 2.

By using the dynamic EB methods developed in Section 2, we are able to predict the batting performance of relatively infrequent batters to whom previous methods cannot be applied. One may ask why such prediction is of interest since their performance presumably has little effect on that of their team. In our analysis of these data, we also examined which batters produced the largest absolute prediction errors in Table 3. For $t = 8$, the batter is David Freese of the St. Louis Cardinals. Freese batted infrequently in the periods $t = 7$ and $t = 8$; his $N_{i,7} = 19$ and $N_{i,8} = 12$ are both ≤ 32. His batting average in period 7 was 0.158, while his batting average in period 8 was an astonishing 0.583 (he got 7 hits out of 12 at-bats), producing the large prediction error. Freese later became a starting third baseman of his team, with $N_{i,9} = 240$, but did not play in the period $t = 10$ because of season-ending injuries in June, 2010. He resumed playing in the 2011 season that is not included in the present study, and helped the Cardinals win the National League Championship Series, for which he was named the Most Valuable Player (MVP), and then the World Series, in which he was also named the MVP. He became the sixth player to win both MVP awards and also won the Babe Ruth Award as the 2011 postseason MVP.

Acknowledgment

This research was presented by the first author in the Statistica Sinica Special Invited Session of the Joint Meeting of the 2011 Taipei International Statistical Symposium and the 7th Conference of the Asian Regional Section of the International Association for Statistical Computing. He thanks the Organizing Committee for the invitation. His research was supported by the National Science Foundation under DMS-1106535.

Appendix: A Hybrid Method for Likelihood Computation and Applications

Lai, Shih and Wong (2006a, b) have refined the computation of the MLEs in R and SAS software packages for GLMM by using a hybrid method that combines Laplace's approximation with Monte Carlo integration. Note that the likelihood function of the GLMM defined by (2.7) can be written as $\prod_{i=1}^{n} L_i(\phi, \boldsymbol{\alpha}, \boldsymbol{\beta})$, where

$$L_i(\phi, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \left\{ \prod_{t=1}^{T} f(y_{i,t}; \theta_{i,t}, \phi) \right\} \Phi_{\boldsymbol{\alpha}}(\mathbf{b}) d\mathbf{b}, \tag{A.1}$$

in which $\Phi_{\boldsymbol{\alpha}}$ denotes the normal density function with mean 0 and covariance matrix depending on an unknown parameter $\boldsymbol{\alpha}$. Denote $l_i(\mathbf{b}|\phi,\boldsymbol{\alpha},\boldsymbol{\beta})$ as the log-likelihood of $L_i(\phi,\boldsymbol{\alpha},\boldsymbol{\beta})$ and $\ddot{l}_i$ the Hessian matrix consisting of second partial derivatives of l_i with respect to the components of $\mathbf{b}$. Laplaces's asymptotic formula for integral yields the approximation

$$\int e^{l_i(\mathbf{b}|\phi,\boldsymbol{\alpha},\boldsymbol{\beta})}d\mathbf{b} \approx (2\pi)^{q/2}\left\{\det[-\ddot{l}_i(\hat{\mathbf{b}}_i|\phi,\boldsymbol{\alpha},\boldsymbol{\beta})]\right\}^{-1/2}\exp\left\{l_i(\hat{\mathbf{b}}_i|\phi,\boldsymbol{\alpha},\boldsymbol{\beta})\right\}, \tag{A.2}$$

where q is the dimension of $\mathbf{b}_i$, $\hat{\mathbf{b}}_i = \hat{\mathbf{b}}_i(\phi,\boldsymbol{\alpha},\boldsymbol{\beta})$ is the maximizer of $l_i(\mathbf{b}|\phi,\boldsymbol{\alpha},\boldsymbol{\beta})$. Let $\mathbf{V}_i = -\ddot{l}_i(\hat{\mathbf{b}}_i|\phi,\boldsymbol{\alpha},\boldsymbol{\beta})$. Since Laplace's asymptotic formula (A.2) may be a poor approximation to (A.1) when $\lambda_{\min}(\mathbf{V}_i)$ is not sufficiently large, Monte Carlo integration, whose error is independent of q, can be used as an alternative method to evaluate (A.1). Lai, Shih and Wong (2006a, b) use Monte Carlo integration instead of Laplace's asymptotic formula for those i with $\lambda_{\min}(\mathbf{V}_i) < c$, where c is a positive threshold. Specifically, instead of sampling $\mathbf{b}^{(h)}$ directly from $\Phi_{\boldsymbol{\alpha}}$ as in an earlier version of the hybrid method proposed by Lai and Shih (2003a), sample it from a mixture of the prior normal distribution with density $\Phi_{\boldsymbol{\alpha}}$ and the posterior normal distribution $N(\hat{\mathbf{b}}_i, [-\ddot{l}_i(\hat{\mathbf{b}}_i|\phi,\boldsymbol{\alpha},\boldsymbol{\beta}) + \varepsilon\mathbf{I}]^{-1})$, where ε is a small positive number to ensure that the covariance matrix is invertible. This has the advantage of further incorporating the essence of Laplace's method in the Monte Carlo step such that the method is less dependent (than direct Monte Carlo) on the choice of the threshold c. Lai et al. (2006a) suggest using $c = 10$ and a mixture distribution that assigns a weight of 0.2 to $\Phi_{\boldsymbol{\alpha}}$. We use this hybrid method for computing the information criterion that is also used in the following enhancement of (2.7).

To allow for more flexible modeling of the fixed effects $\boldsymbol{\beta}'\mathbf{x}_{i,t}$ in (2.7), we relax the linear assumption and use instead univariate regression splines of degree r and their tensor products as basis functions, thereby extending (2.7) to

$$h(\mu_{i,t}) = \sum_{j=1}^{p}\theta_j h(\bar{Y}_{t-j}) + \beta_0 + \sum_{m=1}^{M}\beta_m B_m(\mathbf{x}_{i,t}) + \mathbf{b}_i'\mathbf{z}_{i,t}, \tag{A.3}$$

in which $B_m(\mathbf{x}_{i,t})$ is a product of terms of the form $x_{i,t,k}^l$ or $(x_{i,t,k}-\xi_{m,k})_+^r$ for some $1 \le l \le r$, $1 \le k \le d$ and some suitably chosen knots $\xi_{m,k}$, where $t_+ = \max(0,t)$. Lai et al. (2006b) propose to place the knots $\xi_{m,k}$ at certain quantiles of the

kth components of the d-dimensional covariate variables $\mathbf{x}_{i,t}$, and to use the following stepwise procedure to choose the spline basis for (A.3). A forward addition step chooses the basis function, among those not already included in the model, with the largest absolute value of the Rao statistic. Forward stepwise addition continues until an information criterion such as

$$\text{BIC} = -2\sum_{i=1}^{n} \log L_i(\hat{\theta}_1, \cdots, \hat{\theta}_p, \hat{\phi}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) + (\log n)(\text{number of parameters}) \quad \text{(A.4)}$$

does not decrease further or when there is no more candidate basis function to be included. Then stepwise backward elimination proceed until the information criterion does not improve; each elimination step removes the basis function in the model with the smallest value of the Wald statistic.

References

Andersen, P.K., Borgan Ø., and Gill R.D. (1993). *Statistical models based on counting processes.* Springer Series in Statistics, Springer, New York.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9-25.

Brown, L.D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *Ann. Appl. Statist.* **2**, 113-152.

Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bull.* **4**, 199-207.

Bühlmann, H. and Gisler, A. (2005). *A Course in Credibility Theory and its Applications.* Springer, New York.

Calhoun, C.A. and Deng, Y. (2002). A dynamic analysis of fixed and adjustable rate mortgage terminations. *J. Real Estate Finance & Econ.* **24**, 9-33.

Clapp, J.M., Deng, Y. and An, X. (2006). Unobserved heterogeneity in models of competing mortgage termination risks. *Real Estate Econ.* **34**, 243-273.

Crosbie, P.J. and Bohn, J.R. (2002). Modeling default risk. *Technical Report*, KMV Corporation, `http://www.ma.hw.ac.uk/~mcneil/F79CR/Crosbie_Bohn.pdf`

Duffie, D., Eckner, A., Horel, G. and Saita, L. (2009). Frailty Correlated Default. *J. Finance.* **64**, 2089-2123

Duffie, D., Saita, L., and Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *J. Financ. Econ.* **83**, 635-665

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70**, 311-319.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American.* 119-127.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1**, 361-379. Univ. Calif. Press.

Lai, T.L. and Bukkapatanam, V. (2013). Adaptive filtering, non-linear state-space models, and applications to finance and econometrics. In *State-Space Models and Applications in Economics and Finance* (S. Wu and Y. Zeng, eds.). Springer, New York. To appear.

Lai, T.L., Gross, S.T. and Shen, D.B. (2011). Evaluating probability forecasts. *Ann. Statist.* **39**, 2356-2382.

Lai, T.L. and Shih, M.C. (2003a). Nonparametric estimation in nonlinear mixed effects models. *Biometrika* **90**, 1-13.

Lai, T.L. and Shih, M.C. (2003b). A hybrid estimator in nonlinear and generalised linear mixed effects models. *Biometrika* **90**, 859-879.

Lai, T.L., Shih, M.C. and Wong, S.P. (2006a). A new approach to modeling covariate effects and individualization in population pharmacokinetic. *J. Pharmacokinetics & Pharmacodynamics* **33**, 49-74.

Lai, T.L., Shih, M.C. and Wong, S.P. (2006b). Flexible modeling of fixed and random effects in generalized mixed models for longitudinal data. *Biometrics* **62**, 159-167.

Lai, T.L., Sun, K.H. and Wong, S.P. (2010). Information sets and excess zeros in random effects modeling of longitudinal data. *Statistics in Biosciences* **2**, 81-94

Lai, T.L. and Xing, H. (2008). *Statistical Models and Methods for Financial Markets.* Springer, New York.

Lai, T.L. and Xing, H. (2014). *Active Risk Management: Financial Models and Statistical Methods.* Chapman & Hall/CRC, Baton Rouge.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1**, 157-163. Univ. Calif. Press.

Robbins, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.* **11**, 713-723.

Ross, S. (2013). *Simulation*, 5th edition. Elsevier, Burlington, MA.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1**, 197-206. Univ. Calif. Press.

Zeger, S.L. and Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics* **44**, 1019-1031.