Biometrics 68, 23–30
March 2012
DOI: 10.1111/j.1541-0420.2011.01645.x

Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification

Tyler H. McCormick,[1,*] Adrian E. Raftery,[2] David Madigan,[1] and Randall S. Burd[3]

[1]Department of Statistics, Columbia University, 1255 Amsterdam Avenue, New York, New York 10025, U.S.A.
[2]Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195-4322, U.S.A.
[3]Children's National Medical Center, 111 Michigan Avenue NW, Washington, District of Columbia 20010, U.S.A.
[*]*email:* tylermc@u.washington.edu

Summary. We propose an online binary classification procedure for cases when there is uncertainty about the model to use and parameters within a model change over time. We account for model uncertainty through dynamic model averaging, a dynamic extension of Bayesian model averaging in which posterior model probabilities may also change with time. We apply a state-space model to the parameters of each model and we allow the data-generating model to change over time according to a Markov chain. Calibrating a "forgetting" factor accommodates different levels of change in the data-generating mechanism. We propose an algorithm that adjusts the level of forgetting in an online fashion using the posterior predictive distribution, and so accommodates various levels of change at different times. We apply our method to data from children with appendicitis who receive either a traditional (open) appendectomy or a laparoscopic procedure. Factors associated with which children receive a particular type of procedure changed substantially over the 7 years of data collection, a feature that is not captured using standard regression modeling. Because our procedure can be implemented completely online, future data collection for similar studies would require storing sensitive patient information only temporarily, reducing the risk of a breach of confidentiality.

Key words: Bayesian model averaging; Binary classification; Confidentiality; Hidden Markov model; Laparoscopic surgery; Markov chain.

1. Introduction

We describe a method suited for high-dimensional predictive modeling applications with streaming, massive data in which the data-generating process is itself changing over time. Specifically, we propose an online implementation of the dynamic binary classifier, which dynamically accounts for model uncertainty and allows within-model parameters to change over time.

Our model contains three key statistical features that make it well suited for such applications. First, we propose an entirely *online implementation* that allows rapid updating of model parameters as new data arrive. Second, we adopt an *ensemble approach* in response to a potentially large space of features that addresses overfitting. Specifically we combine models using *dynamic model averaging* (DMA), an extension of Bayesian model averaging (BMA) that allows model weights to change over time. Third, our *autotuning algorithm* and Bayesian inference address the dynamic nature of the data-generating mechanism. Through the Bayesian paradigm, our adaptive algorithm incorporates more information from past time periods when the process is stable, and less during periods of volatility. This feature allows us to model local fluctuations without losing sight of overall trends.

In what follows we consider a finite set of candidate logistic regression models and assume that the data-generating model follows a (hidden) Markov chain. Within each candidate model, the parameters follow a state-space model. We present algorithms for recursively updating both the Markov chain and the state-space model in an online fashion. Each candidate model is updated independently because the definition of the state vector is different for each candidate model. This alleviates much of the computational burden associated with hidden Markov models. We also update the posterior model probabilities dynamically, allowing the "correct" model to change over time.

"Forgetting" eliminates the need for between-state transition matrices and makes online prediction computationally feasible. The key idea within each candidate model is to center the prior for the unobserved state of the process at time t on the center of the posterior at the $(t-1)$th observation, and to set the prior variance of the state at time t equal to the posterior variance at time $(t-1)$ inflated by a forgetting factor. Forgetting is similar to applying weights to the sample, where temporally distant observations receive smaller weight than more recent observations.

Forgetting calibrates or tunes the influence of past observations. Adaptively calibrating the procedure allows the amount of change in the model parameters to change over time. Our procedure is online and requires no additional data storage, preserving our method's applicability for large-scale problems and for cases where sensitive information should be discarded as soon as possible.

Our method combines components of several well-known dynamic modeling schemes (see Smith, 1979, or Smith, 1992,

for a review) with a procedure for calibrating the influence of past observations through forgetting. For the single-model setting, West and Harrison (1997), for example, describe a forgetting strategy similar to ours. The engineering literature also considers various forgetting schemes (sometimes referred to as "discounting")—see, for example, Fagin (1964), Jazwinsky (1970), and Kulhavý and Zarrop (1993)—although the degree of forgetting is typically selected a priori rather than in a data-adaptive way. Discretized versions of forgetting have also been used previously in the literature (see Kárný and Halousková, 1994).

In the multiple model setting, a related literature exists in signal processing with applications to tracking and filtering—see, for example, Kreucher, Hero, and Kastella (2004) or Bar-Shalom, Li, and Kirubarajan (2001). Our method is also related to recent advances in sparse dynamic graphical modeling (see Carvalho and West, 2007; Carvalho, Polson, and Scott, 2010; or Wang, Reeson, and Carvalho, 2010). As far as we know, ours is the first work to propose an autotuning method to calibrate the degree of forgetting and to couple this with a model averaging approach to account for model uncertainty.

We have applied our methodology to a pediatric surgery problem—determining features that distinguish the type of appendectomy a child might undergo. We will assume that data arrive as patients receive a procedure, and are then discarded. Although the entire dataset is available to use, we have modeled it as if data were only available at the time of the procedure. As electronic medical records become more prevalent, instantaneous reporting of data on procedures will likely become more prevalent, making this approach feasible and practical. Because medical records contain sensitive information, discarding records as soon as they are used for modeling reduces the risk of breaching patient confidentiality.

Although adult studies suggest that laparoscopic surgery can reduce recovery time and lower cost compared to traditional open appendectomy (Hagendorf et al., 2007), the evidence is less compelling in children. Despite this difference, the overall proportion of appendectomies done laparoscopically in children has increased (Nguyen et al., 2004; Hagendorf et al., 2007). This suggests that the overall rate of laparoscopy may be driven by features other than direct patient benefit, such as diffusion of technology and training in hospitals, or nonmedical patient factors such as insurance type or race. We analyze data from the Nationwide Inpatient Sample from 1996 to 2002 to explore how the medical and nonmedical factors associated with the type of procedure a child receives have changed over time.

In Section 2, we describe DMA for binary classification. We demonstrate the performance of our proposed method using simulation studies in Section 3. In Section 4, we apply the method to data obtained from the National Inpatient Sample of children undergoing surgery for appendicitis between 1996 and 2002. Section 5 offers a discussion and potential extensions of this method.

2. Dynamic Model Averaging for Binary Classification

2.1 *Bayesian Dynamic Logistic Regression*

We propose a dynamic logistic regression method. We first describe the recursive estimation procedure for a single model in this section. We then account for model uncertainty by extending the DMA method of Raftery, Kárný, and Ettler (2010) to binary classification in Section 2.2.

2.1.1 *State-space model for within-model estimation.* Recursive estimation allows for sequential, online processing and is done in two steps: updating and prediction. Consider a binary response, y_t and a set of predictors $\mathbf{x}_t = (x_{1,t}, x_{2,t}, \ldots, x_{d,t})$ such that at time t:

$$y_t \sim \text{Bernoulli}\,(p_t),$$

where

$$\text{logit}\,(p_t) \;=\; x_t^T\,\theta_t,$$

where θ_t is a d-dimensional vector of regression coefficients.

At a given time, t, the procedure takes the posterior mode of θ from time $(t-1)$ and uses it to construct the prior for time t. We do this by first using the information up to time $(t-1)$ to construct an estimate of the parameters for time t, yielding the *prediction equation*. This equation predicts the value of the observation at time t based on the estimated parameter using data up to time $(t-1)$. The prediction equation is then combined with the observed data at time t, and the new information factors into updated parameter estimates via the *updating equation*. This process happens first within each model and then predictions between models are combined using the procedure described in the next section to produce predictions that account for model uncertainty.

We first develop the prediction equation and, as in Raftery et al. (2010), assume the state equation $\theta_t = \theta_{t-1} + \delta_t$, where the δ_t's are independent $N(0,\ W_t)$ random vectors. For a set of past outcomes, $Y^{t-1} = y_1, \ldots, y_{t-1}$, and reasonable starting values, recursive estimation begins by supposing

$$\theta_{t-1}\,|Y^{t-1} \sim N\,(\hat{\theta}_{t-1},\, \hat{\Sigma}_{t-1}).$$

Then the prediction equation is

$$\theta_t\,|Y^{t-1} \sim N\,(\hat{\theta}_{t-1},\, R_t), \tag{1}$$

where

$$R_t \;=\; \hat{\Sigma}_{t-1}\,/\lambda_t. \tag{2}$$

Forgetting is specified by equation (2) with λ_t typically taking a value slightly less than one. The model could also be specified without forgetting by using a covariance matrix W_t so that we would instead have $R_t = \hat{\Sigma}_{t-1} + W_t$. This approach, however, would require specifying the entire covariance matrix W_t, which would often be quite large.

We now combine the prediction equation (1) with the additional observation at time t to construct updated estimates. The posterior distribution of the updated estimate, θ_t, having observed y_t, can be written as follows:

$$p\,(\theta_t\,|Y^t) \;\propto p\,(y_t\,|\theta_t) p\,(\theta_t\,|Y^{t-1}). \tag{3}$$

Equation (3) is the product of the prediction equation (1) and the likelihood at time t. The term $p(\theta_t|\,Y^{t-1})$ now acts as the prior. The likelihood from the logistic regression model is not conducive to computing a closed-form expression for equation (3).

We approximate the right-hand side of equation (3) with a normal distribution where the mean of the approximating

normal distribution is the mode of equation (3). We use $\hat{\theta}_{t-1}$ as the starting value, yielding

$$\hat{\theta}_t \;=\; \hat{\theta}_{t-1} \;-\; D^2\ell\,(\hat{\theta}_{t-1})^{-1}\, D\ell\,(\hat{\theta}_{t-1}), \tag{4}$$

where $\ell(\theta) = \log p\;(y_t\,|\,\theta)\;p\;(\theta\,|\,Y^{t-1})$. Now

$$D\ell\,(\hat{\theta}_{t-1}) \;=\; (y_t - \hat{y}_t)\;x_t, \tag{5}$$

where $\text{logit}(\hat{y}_t) = \mathbf{x}_t^T\,\hat{\theta}_{t-1}$. Also

$$D^2\ell\,(\hat{\theta}_{t-1}) \;=\; R_t^{-1} \;+\; \hat{y}_t(1-\hat{y}_t)\;x_t\;x_t^T\,. \tag{6}$$

This relationship means that we can substitute equations (5) and (6) into equation (4) to get the updated estimate, $\hat{\theta}_t$. We update the state variance (i.e., the variance in the approximating normal distribution) using $\hat{\Sigma}_t = \{-D^2\ell\,(\hat{\theta}_{t-1})\}^{-1}$.

In our examples, we initialize the process with coefficients from standard logistic regression models fit to the first one third of the data.

We propose an online, adaptive tuning procedure using the predictive likelihood:

$$f\,(y_t\,|Y^{t-1}) \;=\; \int_{\theta_t} p\,(y_t|\,\theta_t,\,Y^{t-1})p\,(\theta_t\,|Y^{t-1})\,d\theta_t. \tag{7}$$

This integral is not available in closed form so instead we use a Laplace approximation yielding

$$\begin{aligned} f\,(y_t\,|Y^{t-1}) &\approx (2\pi)^{d/2}\;\mid\{D^2(\hat{\theta}_t)\}^{-1}\;\mid^{1/2} p\,(y_t\,|Y^{t-1},\hat{\theta}_t) \\ &\times p\,(\hat{\theta}_t\,|Y^{t-1}). \end{aligned} \tag{8}$$

Lewis and Raftery (1997) suggest that this approximation should be quite accurate. The Laplace approximation makes the computation feasible and fast because $p(y_t|Y^{t-1},\hat{\theta}_t)$ is the logistic likelihood function evaluated at $\hat{\theta}_t$ and $(\mathbf{x}_t,y_t)$, and $p(\hat{\theta}_t|Y^{t-1})$ is a normal density with mean $\hat{\theta}_t$ and variance $\hat{\Sigma}_{t-1}/\lambda_t$ evaluated at its mean. We choose the value of λ_t that maximizes equation (8). In other words,

$$\lambda_t \;=\; \arg\max_{\lambda_t} \int_{\theta_t} p\,(y_t\,|\theta_t, Y^{t-1})\,p\,(\theta_t\,|Y^{t-1})\,d\theta_t.$$

Note that λ_t also enters equation (8) via $\hat{\theta}_t$. Also, this approach implies that equation (7) depends on the past trajectory of the forgetting factors. An alternative approach would be to treat λ_t in a fully Bayesian way (see Section 12.3 of West and Harrison, 1997) and maximize the full posterior, $f(\lambda_{1:t}|\,Y^t)$, where $\lambda_{1:t} = (\lambda_1, \ldots, \lambda_{t-1}, \lambda_t)$. This approach would allow potentially informative current observations, y_t, to be used for updating $\lambda_{1:t-1}$, but would be more difficult, computationally.

Because the θ_t's may be changing at different rates, we suggest that each parameter within each model at each time have its own forgetting parameter. Though it adds flexibility, this approach could be computationally burdensome. In our application we selected separate tuning factors for each continuous variable at each time. For categorical variables, we selected the same tuning factor for all levels of the variable.

One way to approximately maximize the above quantity would be to evaluate it for multiple values of λ_t. We found that the following, simpler, alternative performed comparably while requiring minimal computational effort. At each observation, consider two options: (i) no forgetting ($\lambda_t = 1$) or (ii) some forgetting ($\lambda_t = c(<1)$). The constant c must still be chosen by the user, but we now also consider not forgetting at each step. If the process is believed more variable (detecting a spike, e.g.), one should choose a smaller value whereas a value closer to one works better for a more stable process or a smooth trend. We found that our results were not sensitive to the chosen constant. In fact, simulation studies showed that the algorithm maintains similar performance across different tuning values by choosing to forget more or less often depending on the scale of the tuning factor. A similar method was proposed for the single-model case by Kárný and Halousková (1994).

The forgetting/no-forgetting approach is also more computationally feasible for selecting a different tuning factor for each parameter in a given model. For our application in Section 4.1, the data have around 15 candidate variables with some being indicators with multiple levels. If we assume that all indicators for a categorical variable share the same forgetting factor, then the combinatorics are manageable, even using a standard desktop computer.

2.2 *Dynamic Model Averaging*

In the multimodel case we have K candidate models (M_1, M_2, …, M_K). A key feature of this method is that model probabilities are also dynamic, allowing flexibility through time while avoiding overfitting at each observation.

We define L_t as a model indicator so that if $L_t = k$, the process is governed by model M_k at time t. For the multimodel case, we have

$$y_t\,|L_t \;=\; k \sim \text{Bernoulli}\,(p_t^{(k)}), \quad \text{where}$$

$$\text{logit}\,(p_t^{(k)}) \;=\; \mathbf{x}_t^{(k)T}\;\theta_t^{(k)}. \tag{9}$$

Notice from equation (9) that both the values of $\theta_t^{(k)}$ and the dimension of the vector are model specific. Following Raftery et al. (2010), we update $\theta_t^{(k)}$ conditionally on $L_t = k$.

As in the single-model case, estimation occurs in two steps—prediction and updating. In the multimodel case, however, the state space at each time now consists of the pair $(L_t, \boldsymbol{\theta}_t)$, where $\boldsymbol{\theta}_t = (\theta_t^{(1)}, \ldots, \theta_t^{(k)})$. Recursive estimation now acts on the pair $(L_t, \boldsymbol{\theta}_t)$:

$$\sum_{\ell=1}^{K} p\,(\theta_t^{(\ell)}\,|L_t \;=\; \ell,\, Y^{t-1})\,p\,(L_t \;=\ell\,|Y^{t-1}). \tag{10}$$

The key feature of equation (10) is that the $\theta_t^{(\ell)}$ term is present only *conditionally*, given $L_t = \ell$. We can now proceed with our prediction and updating steps separately for the model, L_t, and for the parameters within a given model.

We first describe prediction for the model indicator, L_t. The model prediction equation is

$$\begin{aligned} &P\,(L_t \;=\; k\,|Y^{t-1}) \\ &\quad=\; \textstyle\sum_{\ell=1}^{K} p(L_{t-1} \;=\; \ell\,|Y^{t-1})\,p\,(L_t \;=\; k\,|\,L_{t-1} \;=\; \ell). \end{aligned}$$

To avoid specifying the $K \times K$ transition matrix of $p(L_t = k|L_{t-1} = \ell)$ terms, we update using forgetting, specifically

$$P\,(L_t \;=\; k\,|\,Y^{t-1}) \;=\; \frac{P\,(L_{t-1} \;=\; k\,|\,Y^{t-1})^{\alpha_t}}{\sum_{\ell=1}^{K} P\,(L_{t-1} \;=\; \ell\,|\,Y^{t-1})^{\alpha_t}}, \tag{11}$$

where α_t is the forgetting factor, $\alpha_t \leq 1$. The α_t parameter in equation (11) increases uncertainty by flattening the distribution of L_t. In this way, only a single parameter must be set instead of an entire transition matrix.

The model updating equation is then:

$$P\left(L_t = k \,|\, Y^t\right) = \omega_t^{(k)} \Big/ \sum_{\ell=1}^{K} \omega_t^{(\ell)},$$

where

$$\omega_t^{(\ell)} = P(L_t = \ell \,|\, Y^{t-1})\, f^{(\ell)}\left(y_t \,|\, Y^{t-1}\right). \tag{12}$$

Notice that for each model, equation (8) provides the expression for $f^{(\ell)}(y_t|\, Y^{t-1})$. Furthermore, because the procedure for selecting the model-specific forgetting factor $\lambda_t^{(\ell)}$ already computes $f^{(\ell)}(y_t|\, Y^{t-1})$, no further computation is required for equation (12).

We also adjust α_t, the forgetting factor for the model indicator forgetting, using the predictive likelihood across our candidate models, $\boldsymbol{f(y_t|Y^{t-1})}$, so that

$$\begin{aligned}
&\boldsymbol{f\left(y_t \,|\, Y^{t-1}\right)} \\
&= \sum_{k=1}^{K} f^{(k)}\left(y_t|Y^{t-1}\right) P\left(L_t = k|Y^{t-1}\right) \\
&= \sum_{k=1}^{K} \left(\int_{\theta_t^{(k)}} P\left(y_t \,|\, \theta_t^{(k)}, Y^{t-1}\right) p\left(\theta_t^{(k)} \,|\, Y^{t-1}\right) d\theta_t^{(k)} \right) \\
&\quad \times \frac{P\left(L_{t-1} = k \,|\, Y^{t-1}\right)^{\alpha_t}}{\sum_{\ell=1}^{K} P\left(L_{t-1} = \ell \,|\, Y^{t-1}\right)^{\alpha_t}}.
\end{aligned}$$

We previously computed $f^{(k)}(y_t|\, Y^{t-1})$ so we can use it here with minimal additional computation. We select α_t as:

$$\arg\max_{\alpha_t} \sum_{k=1}^{K} f^{(k)}(y_t \,|\, Y^{t-1}) P(L_t = k|Y^{t-1}).$$

In practice, we choose between two candidate α_t values at each time (forgetting/no forgetting).

To predict y_t we then use $\hat{y}_t^{DMA} = \sum_{\ell=1}^{K} P(L_t = \ell \,|\, Y^{t-1}) \hat{y}_t^{(\ell)}$, where $\hat{y}_t^{(\ell)}$ is the predicted response for model ℓ at time t.

3. Simulation Results

Figure 1 shows parameter estimates for one parameter in a large logistic regression model. The mean Brier score for our model was 0.12, compared with 0.14 using a standard logistic regression model. While the data-generating mechanism changed gradually, the algorithm detected the trend and the adaptive tuning scheme accommodated the additional uncertainty. During this period, the algorithm chose tuning values corresponding to more diffuse prior distributions.

Near the midpoint of our simulations, we incorporated a spike in the data-generating mechanism. Our method quickly detected the drastic change in signal. The adaptive tuning algorithm again compensated (by reducing prior variance) in the stable period after the spike. The overall result was more precise estimation when behavior is more stable, and flexibility during more volatile periods.

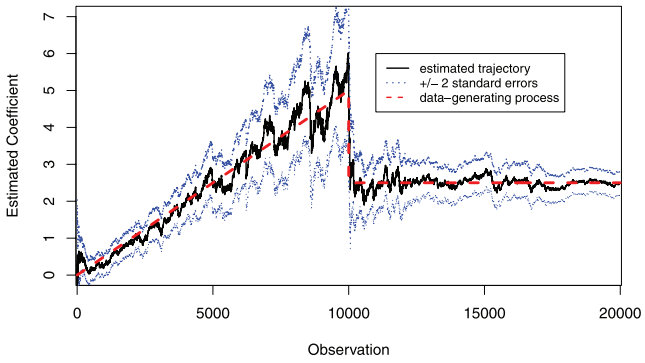

Figure 1. Parameter estimates ±2 standard errors for an example coefficient in logistic regression on simulated data. The dashed line represents the trajectory used to generate the data. We set the forgetting constant, c, as .99. We present additional simulation results for different values of c in Web Figures 1–4. We first generated a series of 20,000 coefficient values corresponding to the trajectory indicated by the dashed line, then generated predictor values and sampled our responses from the appropriate Bernoulli distribution. Our method detects more variable regions in the data-generating mechanism and compensates by increasing variance. When the trend is more consistent, the method responds by reducing variance.

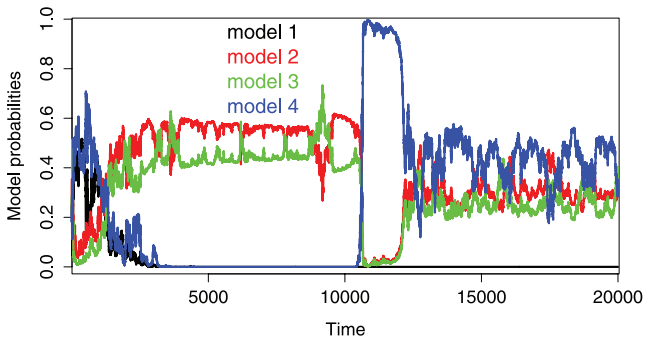

Figure 2. Posterior model probabilities. The data-generating model changes midway through the simulation, which is reflected in the estimated model probabilities. For presentation, we have included only four candidate models, though we found similar performance with much larger simulation experiments. Model 2, which generated the first 10,000 observations, consisted of four predictor variables: $\theta_{1,t} = 1$, $\theta_{2,t}$ increases gradually from 0 to 1.5, $\theta_{3,t}$ increases from 0 to 1 then is stable, $\theta_{4,t}$ increases from 0 to 1, then decreases to -1. Model 4, which generated the second 10,000 observations, contains only θ_1 and θ_2. Model 1 uses only of $\theta_{4,t}$ and model 3 contains all four predictors in model 2 along with two additional predictors. We set both forgetting constants to 0.999.

Figure 2 displays results for a simulation experiment for DMA. We used one model (model 2) to generate data for the first half of the simulation (10,000 observations). Then, we switched the data-generating mechanism to a different model (model 4) for another 10,000 observations. Our method quickly identified and adapted the estimated model probabilities. The approximations used here allow fast com-

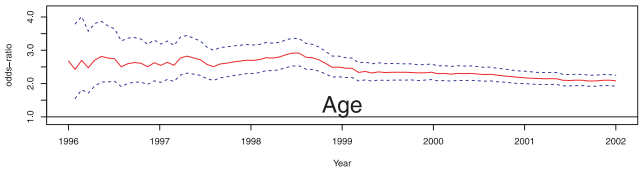

Figure 3. Coefficients for the age variable from the model using all predictors of laparoscopic or open appendectomy described in Section 4.1. The solid lines are coefficient estimates and the dashed lines are ±2 standard errors.

putation of posterior model probabilities, increasing the feasibility of large-scale, online implementations of DMA.

4. Pediatric Laparoscopic Appendectomies

4.1 *Background and Data*

We apply our method to a sample of children less than 15 years old from the Nationwide Inpatient Sample. The sample consists of 72,189 children who underwent appendectomies between 1996 and 2002. Previous research indicates that there are socioeconomic and racial differences in the presentation of appendicitis in children and that these differences may exist even when accounting for severity of appendicitis (Guagliardo et al., 2003; Smink et al., 2005). Although we have access to the entire dataset, we have modeled it as if it were only available at the time of the procedure.

The data also contain demographic and health information about the patient and information about the hospital. We observe the patient's race (white, African American, Hispanic, Asian, or other), age, sex, payment source (Medicaid/Medicare, private insurance, other), number of chronic diseases, and severity of appendicitis (no perforation, perforation without abscesses, abscess). Information about the hospital includes: teaching hospital or not, urban hospital or not, hospital volume of appendectomies, and children's hospital status (no children's hospital, children's hospital or children's unit in adult hospital). Of the 2449 hospitals in the study 659 did not report race and some states did not report hospital names, making it impossible to assign a children's hospital designation. All other variables were missing less than 3% of records. We agree with Hagendorf et al. (2007) that the data are missing at random and use complete-case analysis.

The overall rate of laparoscopic procedures in the sample increased from 6% in 1996 to 24% in 2002. We hypothesize that this change could also have been accompanied by a change in what predicts which children get a laparoscopic procedure.

4.2 *Results*

We first present results from one of the candidate models to compare our dynamic modeling strategy with static modeling results, using the same data as Hagendorf et al. (2007). We next address uncertainty in model choice through DMA. In both cases we present results from the online implementation of our dynamic logistic regression and model averaging procedure.

4.2.1 *Batch updating.* The data were reported monthly between 1996 and 2002. Because the temporal structure of the data is basic to our analysis, arbitrarily ordering the observations within each month could impact our estimates. Instead, we updated our model in batches where the observations within each batch were assumed to be independent and identically distributed. This assumption allows batch updating with minor modifications for vectorizing computations in the previously described model setting.

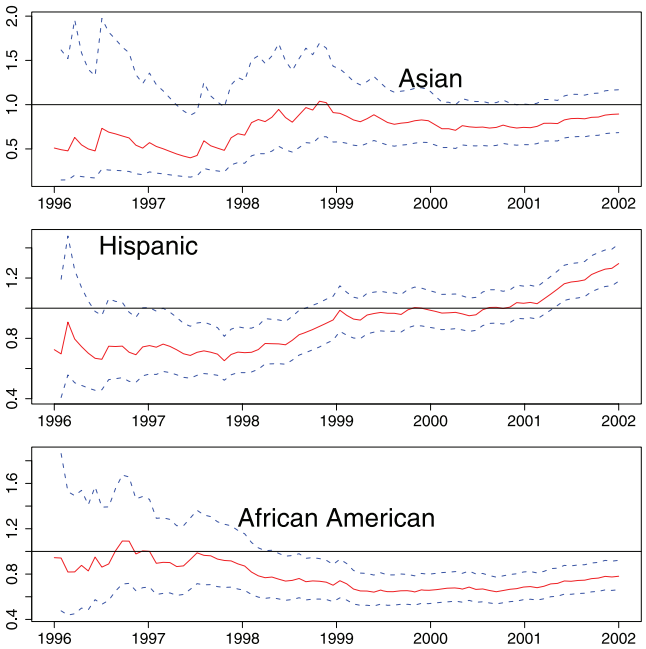

Figure 4. Coefficients for the race variables from the model using all predictors of laparoscopic or open appendectomy described in Section 4.1. The solid lines are coefficient estimates and the dashed lines are ±2 standard errors.

4.2.2 *Dynamic logistic regression.* Following Hagendorf et al. (2007) we considered first a model adjusting for all of the characteristics considered in Section 4.1 and found a surprisingly more complicated situation than Hagendorf et al. (2007) could have seen using static modeling. We set the forgetting constant to 0.99.

As patient's age increases, for example, Hagendorf et al. (2007) found an increased propensity for the procedure to be laparoscopic, which seems sensible because laparoscopic procedures became the normal treatment for acute appendicitis in adult over the study period. Considering Figure 3, however, we see that the coefficient for age increased steadily from 1996 until 1999 and then stabilized. This behavior is consistent with the notion that in the mid-1990s laparoscopic procedures were relatively new even for adults but had become common by the end of the decade.

Dynamic modeling also yields additional insights beyond static modeling when considering the role of a patient's race. Figure 4 shows the profiles of the race indicators, with white as the reference group. Taking the Hispanic indicator, for example, Hagendorf et al. (2007) did not find the coefficient to be statistically significant. Using dynamic modeling, however, the odds ratio was near 1 from around 1996 until 2001 but then increased to nearly 1.5 by the end of 2002. Considering the indicator for African Americans next,

Hagendorf et al. (2007) found the odds ratio for African American to be significantly less than 1. Dynamic modeling confirms this observation but adds the caveat that the trend was increasing and that by 2001 the odds ratio had nearly reached 1.

4.2.3 *Dynamic model averaging.* Hagendorf et al. (2007) considered both univariate and multivariate models and drew conclusions from the aggregate findings of all models fit. Our proposed model averaging, in contrast, acknowledges model uncertainty by considering results from multiple models.

The full model space in this problem consisted of 512 candidate models. Figure 5 displays the posterior model probabilities for the three models with highest posterior probability as well as for five additional candidate models selected using the conclusions based on static modeling from Hagendorf et al. (2007).

The models selected from Hagendorf et al. (2007) group predictors with a common theme (predictors related to hospital characteristics, e.g.). The models with highest posterior probability, however, contained combinations of these themes. Initially, the model with only the intercept had the highest posterior probability, indicating an overall rise in the rate of laparoscopic procedures. Near the end of 1996, a model that included medical factors and the prevalence of pediatric procedures at the hospital overtook the intercept-only model. As 1998 began, medical factors were no longer in the model with highest probability, being replaced by demographic and insurance type predictors.

In terms of specific coefficients, Hagendorf et al. (2007) suggested that nonmedical characteristics of the patient and the hospital significantly contribute to the propensity of a patient to get a laparoscopic procedure. Specifically, their work focused on race and payment type, both of which they found to be influential. Figure 6 displays the marginal probability for each predictor. Our dynamic model also found evidence for the importance of race and payment type, though it varied through time. The patient's race, for example, had low marginal probability during the initial 2 years of observation, but increased and remained high through 1998. We observe a similar pattern with payment type, though the marginal probabilities for payment type increased sooner, around the end of 1997.

We also compared the performance of DMA to each of the 512 potential static models. Each static model also included indicator variables for year, as in Hagendorf et al. (2007). The Brier score was approximately 0.14 using DMA and the best static regression model, even though our method has no prior knowledge about model performance. The overall misclassification rates at a threshold of 0.5 were also nearly equal. The static model overpredicted zero responses, yielding a lower sensitivity (proportion of predicted successes which were actually a success) score. The sensitivity score for DMA was 0.12, versus 0.02 for the best static regression. We found similar trends for additional threshold values.

5. Discussion and Conclusion

We have constructed an online, adaptive binary classification method that accounts for model uncertainty using DMA, thereby extending DMA to binary outcomes. We apply our model to data collected from children receiving either laparoscopic or open appendectomies. Our method provides insights into the treatment of appendicitis that were not clear using static logistic regression, even when year of presentation was considered as a covariate (Hagendorf et al., 2007). These data also included survey weights that we did not use in this analysis. Future models could include these weights, though we found little substantive difference in coefficients for static models with and without weights.

Analyses can be done online and estimates updated as additional data become available. This feature makes the method applicable to other situations where data must be collected and processed sequentially or where storing data is unappealing, as is often the case when sensitive data such as healthcare records are being analyzed. The absence of a closed-form posterior makes sequential updating challenging. West, Harrison, and Migon (1985) proposed an alternative approach using conjugate priors through a transformation to the natural parameter space in the exponential family. A key feature of both our approach and that of West et al. (1985) is specifying prior distributions that yield a tractable form for sequential updating. West et al. (1985) put conjugate priors on the natural parameter of the exponential family density. They then used the relationship between the natural parameter space (η_t in their notation) and the underlying state vector (these are equivalent in the Gaussian case because of the identity link) to derive updates for the first two moments of the state distribution (see (3.2) and (3.3) in West et al., 1985). Our approach updates the state vector directly but assumes the prediction equation, (1), is well approximated by a Gaussian distribution. As in the West et al. (1985) approach, our distributional assumptions ensure updates are based on calculating approximations of the first and second moments of the posterior.

We also offer a new possibility for adaptive tuning in this type of model. The tuning strategy we propose is adaptive and requires a fraction of the computation that would be required to compute the full transition matrices. Our method is similar in spirit to maximizing prequential probability (Dawid and Vovk, 1999), but differs in our choice to use the current data to update the parameters before evaluating the marginal likelihood.

Our method is designed for applications where no data are stored. If even a few observations were to be stored, however, we could use these data for tuning using more traditional crossvalidation type procedures. In that case, we propose updating the tuning factor to maximize the average one-step-ahead predictive score for the stored data. The choice of score would depend on the application. For logistic regression, for example, we suggest using Brier scores.

Specifically, we suggest updating the forgetting factor every n_0 observations to maximize the average score for the previous n_0 observations. Because tuning may be computationally intensive, we may not wish to update the forgetting factor with every new batch of stored observations. We suggest choosing n_0 based on available computational resources and the complexity of the problem. Similarly, we posit that n_0 will often be determined by necessity (limited storage space or computation time). If possible, this could also be determined empirically using the first portion of the data.

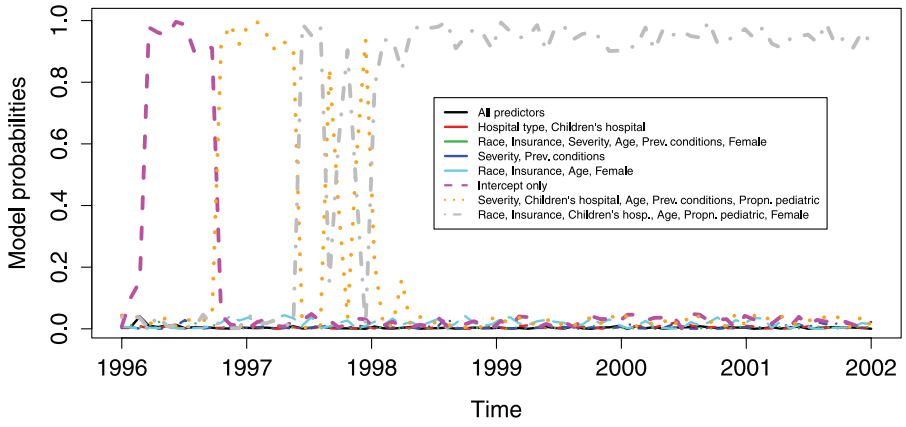

Figure 5. Posterior model probabilities for a subset of the 512 candidate models for predicting appendectomy type. The first five models in the legend were selected based on conclusions from Hagendorf et al. (2007), whereas the last three are models with dominant posterior probability. We used a parallel implementation where each model was updated independently on a separate CPU. We combined results at each step to compute model probabilities. We grouped similar indicator variables (Hispanic, African American, Asian, and other race were all included or all excluded, e.g.) to reduce dimensionality and preserve interpretability.

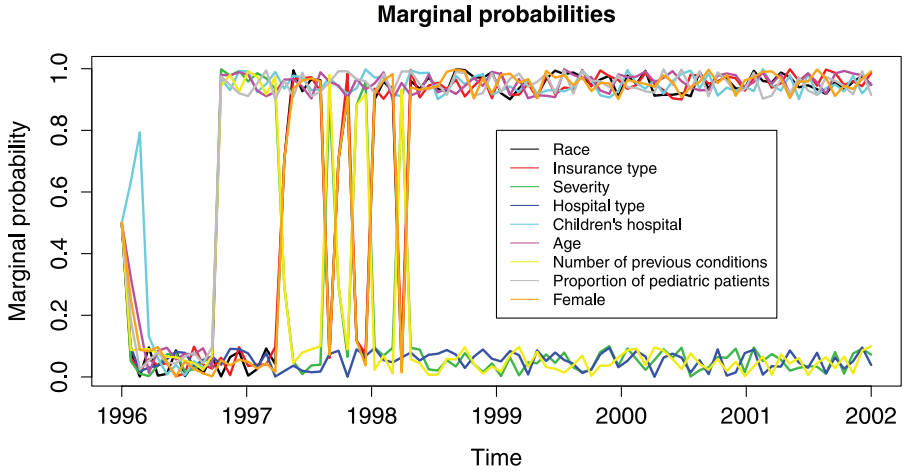

Figure 6. Marginal probabilities for the nine predictors.

The procedure described in this section could be applied to other regression problems by adapting the predictive score, possibly using the continuous ranked probability score (Gneiting and Raftery, 2007). We could then also test, using Bayes' factors, to see if our λ_t and α_t values are changing over time or if we should revert to a model with a constant forgetting factor.

Our goal in tuning is to optimize a criterion related to our method's performance. The best way to maximize the predictive score would be through numerical optimization, but this is typically computationally burdensome. We propose evaluating two values at each update (forgetting and no forgetting). Instead, following Kárný and Halousková (1994), we could evaluate across a grid of tuning values (equally spaced on the scale of the effective sample size, $1/(1\text{-}\lambda_t)$, e.g.) or, further reducing computation, evaluate only three candidate values at each update—the previous tuning value and the values on either side in the grid. Evaluating across a grid would avoid the need to specify a 'no-forgetting' condition, but would instead require tuning the precision of the grid. The magnitude of the forgetting factor depends on the volatility of the process, which is not known in advance. A highly volatile process, for example, may be adequately described using a grid with ten equally spaced values between 0.9 and 1. A less volatile process, however, maybe be well described using a grid with more precision between 0.99 and 1. Using our approach, we found the frequency of forgetting adjusted to the size of the constant and the volatility of the process.

As shown in Section 4.2, our method can accommodate a model space with hundreds of models using a parallel computing environment. Many applications will have much larger model spaces, making it infeasible to evaluate every model at each time point. Even if the full set of models could be evaluated, it may be desirable to preserve computing resources by updating only a subset of models at each time point. The proposed method could be adapted through an 'Occam's window' approach (Madigan and Raftery, 1994), where we evaluate only an 'active' subset of the models at each time. The

active set of models includes the model with highest posterior probability and other models whose posterior probability is within a threshold of the best candidate. Models outside the set are reevaluated periodically and added if their predictive probability is sufficiently high. This approach is also known as *model set adaptation* in other contexts (Li, 2005; Li, Zhao, and Li, 2005).

6. Supplementary Materials

Web Figures referenced in Section 3 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

TM is supported by a Google Ph.D. Fellowship in Statistics. Raftery's research was supported by NIH grants R01 HD54511 and R01 GM084163, and by NSF grant ATM0724721. We gratefully acknowledge the helpful comments of Miroslav Kárný, two anonymous reviewers, the associate editor, and the editor.

References

Bar-Shalom, Y., Li, X. R., and Kirubarajan, T. (2001). *Estimation with Applications to Tracking and Navigation*. New York: John Wiley & Sons.

Carvalho, C. M. and West, M. (2007). Dynamic matrix-variate graphical models. *Bayesian Analysis* **2,** 69–97.

Carvalho, C., Polson, N., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97,** 465–480.

Dawid, A. P. and Vovk, V. G. (1999). Prequential probability: Principles and properties. *Bernoulli* **5,** 125–162.

Fagin, S. L. (1964). Recursive linear regression theory, optimal filter theory, and error analysis of optimal systems. *IEEE International Convention Record* **12,** 216–240.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102,** 359–378.

Guagliardo, M., Teach, S., Huang, Z., and Joseph, J. (2003). Racial and ethnic disparities in pediatric appendicitis rupture rate. *Academy of Emergency Medicine* **10,** 1218–1227.

Hagendorf, B. A., Liao, J. G., Price, M. R., and Burd, R. S. (2007). Evaluation of race and insurance status as predictors of undergoing laparoscopic appendectomy in children. *Annals of Surgery* **245,** 118–125.

Jazwinsky, A. W. (1970). *Stochastic Processes and Filtering Theory*. New York: Academic Press.

Kárný, M. and Halousková, A. (1994). Pre-tuning of self-tuners. In *Advances in Model-Based Predictive Control*, D. Clark (ed), 333–343. Oxford: Oxford University Press.

Kreucher, C. M., Hero III, A. O., and Kastella, K. (2004). Multiple model particle filtering for multitarget tracking, the twelfth annual workshop on adaptive sensor array processing. Lexington, MA.

Kulhavý, R. and Zarrop, M. B. (1993). On a general concept of forgetting. *International Journal of Control* **58,** 905–924.

Lewis, S. M. and Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Assocation* **92,** 648–655.

Li, X. R. (2005). Multiple-model estimation with variable structure—part ii: Model-set adaptation. *IEEE Transactions on Automatic Control* **45,** 2047–2060.

Li, X. R., Zhao, Z. L., and Li, X. B. (2005). General model-set design methods for multiple-model approach. *IEEE Transactions on Automatic Control* **50,** 1260–1276.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89,** 1535–1546.

Nguyen, N., Zainabadi, K., Mavandadi, S., Paya, M., Stevens, C. M., Root, J., and Wilson, S. E. (2004). Trends in utilization and outcomes of laparoscopic versus open appendectomy. *American Journal of Surgery* **188,** 813–820.

Raftery, A. E., Kárný, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics* **52,** 52–66.

Smink, D. S., Fishman, S. J., Kleinman, K., and Finkelstein, J. A. (2005). Effects of race, insurance status, and hospital volume on perforated appendicitis in children. *Pediatrics* **115,** 1068–1070.

Smith, J. Q. (1979). A generalisation of the Bayesian steady forecasting model. *Journal of the Royal Statistical Society, Series B* **41,** 375–387.

Smith, J. Q. (1992). A comparison of the characteristics of some Bayesian forecasting models. *International Statistical Review* **60,** 75–85.

Wang, H., Reeson, C., and Carvalho, C. (2010). *Dynamic financial index models: Modeling conditional dependencies via graphs*. Technical Report, Department of Statistical Science, Duke University, Durham, North Carolina.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag.

West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association* **80,** 73–83.

Received July 2010. *Revised April* 2011.
Accepted april 2011.