Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Statistical Learning With Time Series Dependence: An Application to Scoring Sleep in Mice

Blakeley B. McShane, Shane T. Jensen, Allan I. Pack, and Abraham J. Wyner

We develop methodology that combines statistical learning methods with generalized Markov models, thereby enhancing the former to account for time series dependence. Our methodology can accommodate very general and very long-term time dependence structures in an easily estimable and computationally tractable fashion. We apply our methodology to the scoring of sleep behavior in mice. As methods currently used to score sleep in mice are expensive, invasive, and labor intensive, there is considerable interest in developing high-throughput automated systems which would allow many mice to be scored cheaply and quickly. Previous efforts at automation have been able to differentiate sleep from wakefulness, but they are unable to differentiate the rare and important state of rapid eye movement (REM) sleep from non-REM sleep. Key difficulties in detecting REM are that (i) REM is much rarer than non-REM and wakefulness, (ii) REM looks similar to non-REM in terms of the observed covariates, (iii) the data are noisy, and (iv) the data contain strong time dependence structures crucial for differentiating REM from non-REM. Our new approach (i) shows improved differentiation of REM from non-REM sleep and (ii) accurately estimates aggregate quantities of sleep in our application to video-based sleep scoring of mice. Supplementary materials for this article are available online.

KEY WORDS: Categorical; Classification; Machine learning; Markov; REM; Sequence.

1. INTRODUCTION: MOTIVATION AND CHALLENGES

1.1 The Science and Scoring of Sleep

Roughly 70 million Americans suffer from chronic sleep loss or sleep disorders, costing the nation approximately $16 billion per year in medical expenses and a further $50 billion per year in lost productivity (Patlak 2005). In total, over 70 such disorders afflict about 40 million Americans (Patlak 2005). As knowledge of these problems grows among the populace, sleep is becoming an increasingly important field of medical inquiry.

A major area of focus is in determining the genetic basis of sleep behaviors: nonrapid eye movement (NREM) sleep, rapid eye movement (REM) sleep, and wakefulness (WAKE). Of particular interest is REM sleep which occurs much less frequently than either NREM sleep or WAKE and has an important role in learning and memory (Stickgold et al. 2001). Sleep scientists have shown that various aspects of sleep are heritable including sleep duration (Partinen et al. 1983; Heath et al. 1990), the timing of sleep [i.e., circadian phase (Heath et al. 1990; Vink et al. 2001)], and the response to sleep deprivation (Franken, Chollet, and Tafti 2001). While some gene variants affecting these traits have been identified in human studies (He et al. 2009), present knowledge is limited because it is too costly and labor intensive to phenotype a large number of human subjects.

The principal alternative to human studies is to study gene variants in mice. Mice are specially bred to allow the mapping of quantitative trait loci to small regions of the genome (Churchill et al. 2004; Chesler et al. 2008) and their behavior is observed. Unfortunately, studies in mice are also quite limited in size (and consequently in scope) because the "gold standard" methodology for studying sleep even in mice is expensive, invasive, and time consuming. First, electrodes are surgically implanted into the head of a mouse. Then, after waiting 10–14 days for the mouse to recover from surgery, electroencephalographic (EEG) and electromyographic (EMG) waves are recorded at 256 Hz for 24 hr. The EEG/EMG data are then broken into distinct 10-sec blocks (termed "epochs") and each (i.e., 8640 epochs per mouse per day) is *manually* scored as NREM, REM, or WAKE by specially trained technologists.

Not only expensive and laborious, this process can also be quite inconsistent. If the same mouse is scored independently by two different technologists, there is disagreement on approximately 6% of epochs and up to 15% within the important REM stage (Guan et al. 2010; McShane et al. 2012). Moreover, when a given technologist revisits the data at a later time period, his original scores and new scores fail to match at rates that are only mildly better than those for two different technologists.

Consequently, sleep scientists have sought high-throughput automated systems that would avoid both the surgery and the labor associated with EEG/EMG-based manual scoring. Indeed, sleep scientists have already made initial efforts toward this end. A leading approach, termed the "40-second Rule" (Pack et al. 2007), uses video recordings (or, alternatively, electronic beam splits) to determine whether a mouse is moving or not. Any duration of inactivity lasting 40 sec or more is considered sleep. The 40-second Rule has been validated by comparison to manual scores based on EEG/EMG recordings in both young (Pack et al. 2007) and old (Naidoo et al. 2008) mice of the strain C57BL/6J. An alternative approach (Flores et al. 2007) also relies on mouse movements but instead uses piezoelectric sensors implanted into the floor of the mouse cage to detect movement. The data recorded by such sensors contain patterns that are characteristic of sleep versus wakefulness.

Blakeley B. McShane is Assistant Professor, Kellogg School of Management, Northwestern University, Evanston, IL 60611 (E-mail: *b-mcshane@kellogg.northwestern.edu*). Shane T. Jensen is Associate Professor, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *stjensen@wharton.upenn.edu*). Allan I. Pack is John Miclot Professor, Center for Sleep and Circadian Neurobiology, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *pack@mail.med.upenn.edu*). Abraham J. Wyner is Professor, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *ajw@wharton.upenn.edu*). This research was supported in part by the National Institutes of Health grants T32 HL07713, P01 AG17628, and MH081491.

Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Applications and Case Studies
DOI: 10.1080/01621459.2013.779838

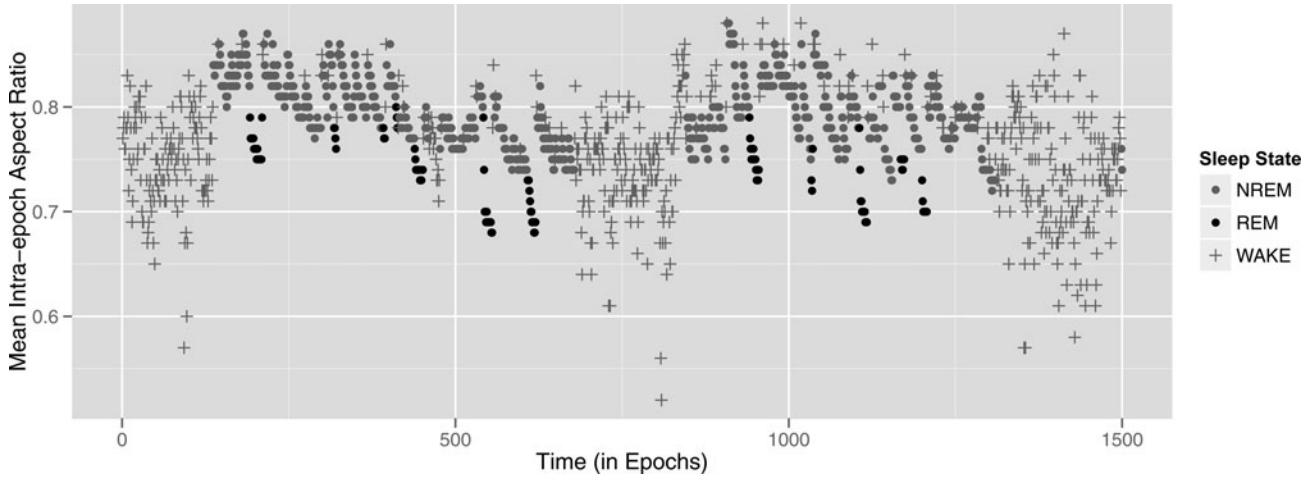

Figure 1. Mean intra-epoch aspect ratio. A time-series plot of the mean intra-epoch aspect ratio for one mouse.

Although these methods are able to differentiate sleep from wakefulness, they have no ability to detect the substages of sleep (i.e., NREM sleep and REM sleep). However, there are known physical manifestations differentiating NREM sleep and REM sleep that should induce subtle signals in digital video recordings [for a full review, see Steriade (2005)]. For example, while sleep in general is associated with a reduction in muscle tone (atonia), this reduction is far more pronounced in REM sleep as compared to NREM sleep with a near complete absence of muscle tone in REM sleep. Consequently, as the mouse transitions from NREM sleep to REM sleep, there is an onset of near complete atonia leading to a change in the shape of the mouse. In particular, the mouse "flattens out" such that there is a decrease in its aspect ratio (i.e., the ratio of the mouse's length to its width) and an increase in its area (i.e., the size of the mouse in a two-dimensional video recording). Both of these features are detectable by video. For example, the decrease in aspect ratio as the mouse transition from NREM sleep to REM sleep is visible in Figure 1, which plots the aspect ratio (derived from video recordings) and sleep state (determined by EEG/EMG-based manual scoring) of a mouse over several hours. While this change and similar ones are subtle, they are detectable and it is these features in combination with the striking temporal dependencies evident in Figure 1 that allow us to achieve our goal, namely, the development of a model that can identify NREM sleep versus REM sleep in mice based on digital video recordings.

1.2 Challenges for Statistical Learning Methods

To date, sleep scientists have focused on movement-based measures (e.g., activity versus inactivity) obtained from video cameras, electronic beams, or piezoelectric sensors. However, Figure 1 and similar plots suggest that additional covariates such as aspect ratio are relevant for classifying sleep stages. A natural approach, therefore, would be to (i) cull additional covariates from video data and (ii) employ statistical learning procedures such as logistic regression, AdaBoost (Freund and Schapire 1996), or random forests (RF) (Breiman 2001).

While statistical learning procedures have proven extremely successful at minimizing classification error on a wide variety of problems, they are known to have reduced performance in certain more difficult classification situations. For example, AdaBoost can produce "overfit" conditional class probability estimates that tend toward zero or one (Mease, Wyner, and Buja 2007; Mease and Wyner 2008), a problem that is particularly acute in noisy environments where the Bayes' error is high. The high rate of disagreement among manual scorers and the similarity of the NREM and REM states in video recordings both suggest that the Bayes' error is high in our setting.

Another difficulty is that many learning methods were developed for the binary classification setting and do not naturally accommodate a multiclass setting like ours. Even approaches that do (e.g., RF which does so by using a voting rule over the generated trees) can fail in situations where multiple states have similar observed covariates or when one or more states are rare. In our application, we face both of these concerns: not only are REM and NREM relatively undifferentiated in terms of the observed covariates but also REM occurs in a mere 5% of epochs (compared to about 45% for NREM and 50% for WAKE).

A final difficulty posed by our application—and the one that will be the focus of our efforts—concerns the fact that most statistical learning methods assume independently and identically distributed data. Specifically, the conditional class probability function $\mathbb{P}(Y_t|\boldsymbol{X}_t)$ for outcome Y_t given covariates $\boldsymbol{X}_t$ is assumed to be independent of the rest of the data $(\boldsymbol{Y}_{-t}, \mathbf{X}_{-t})$. As illustrated in Figure 1, our data form a non-iid time series with strong dependencies in both the response (i.e., the sleep states) and the covariates. It should be possible to gain additional discriminatory power by modeling these dependencies.

In summary, our application presents three challenges to standard statistical learning procedures: we require estimates of the conditional state probabilities in a (i) noisy, (ii) multiclass setting with rare and undifferentiated states, and (iii) time dependencies. We address these challenges by combining conventional classification methods with a very general form of the Markov model. In particular, we introduce methodology that enhances the conditional class probability estimates produced by standard statistical learning procedures so that they are able to take account of a very general class of time dependence structures. Nevertheless, by embedding these general dependence structures in a first-order Markov structure, our approach remains computationally feasible, parsimonious, and easily estimable.

The more general approach considered here is critical for our application, as the time dependence structures present

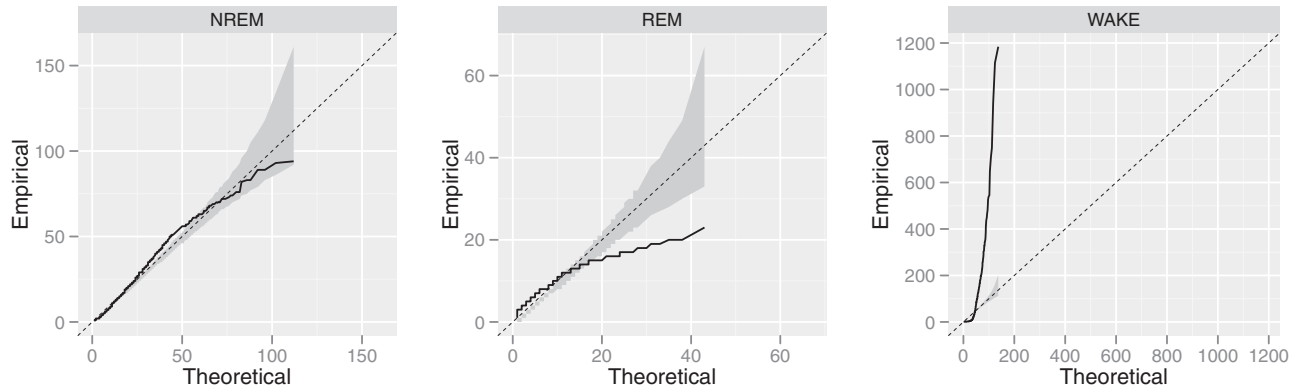

Figure 2. Q-Q plots for the geometric distribution fit to NREM, REM, and WAKE. The Q-Q line is given by the solid line and the $y = x$ line by the dashed line. The gray region is a null interval created using parametric bootstrap samples based on the maximum likelihood estimate.

in sleep data are not suitably addressed by prior techniques. For example, modeling strategies that combine classification methods with a first-order Markov model (Rabiner 1989; Smyth 1994) are ill-suited to sleep data as the first-order Markov model implies that bout durations (i.e., the individual state holding times) are (i) geometrically distributed and (ii) do not depend on the previous state. Prior literature has found both of these assumptions untenable (McShane et al. 2010) and, indeed, the geometric distribution provides a poor fit to our data as evidenced by the fact that the black Q-Q lines in Figure 2 exceed the gray null bands for each of the three states. Further, the bout duration distributions of each state depend strongly on the previous state (i.e., they are "transition-dependent") as the black Q-Q lines in Figure 3 exceed the gray null bands in each plot. In contrast to the first-order Markov model, our more general methodology can easily accommodate both of these features.

The empirical duration distributions plotted in Figures 2 and 3 point to a further difficulty posed by sleep data: in addition to their nongeometric shape and their transition-dependence, these distributions have extremely long (if not unbounded) support thus suggesting that a high-order Markov model may be necessary to adequately reflect the time dependence structures present in the data. A challenge, then, is to allow for the extension of the basic first-order Markov model to higher orders without the concomitant explosion in the number of parameters associated with standard higher-order Markov models. Building on related techniques such as non-stationary Markov models (Sin and Kim 1995; Vaseghi 1995; Djuric and Chun 2002), semi-Markov models (Janssen and Limnios 1999), variable duration Markov models (Ferguson 1980; Levinson 1986), and variable length Markov models (Buhlmann and Wyner 1999), our methodology overcomes this challenge by using a principled and pragmatic parsimonious parametric approach that naturally allows for the long state durations present in sleep data.

The remainder of our article is organized as follows. We introduce our methodology in Section 2. In Section 3, we evaluate its performance relative to simpler alternatives in an extensive simulation study. Section 4 demonstrates that exploiting time dependencies substantially adds to our ability to detect the signal extant in our video recordings of mice, particularly given the high level of noise inherent in the problem; our procedure is able to identify sleep and its substages with reasonable accuracy when validated by comparison to EEG/EMG assessments in C57BL/6J male mice. Section 5 discusses the principal benefits of our new sleep scoring methodology, including its ability (i) to detect the rare and subtle REM state that is of special interest to sleep researchers and (ii) to analyze the large numbers of mice required for genetic analysis.

2. METHODS

2.1 Data Structure

In the standard classification setting, we have a response variable Y, which takes on a finite set of values $Y \in S = \{1, \ldots, k\}$

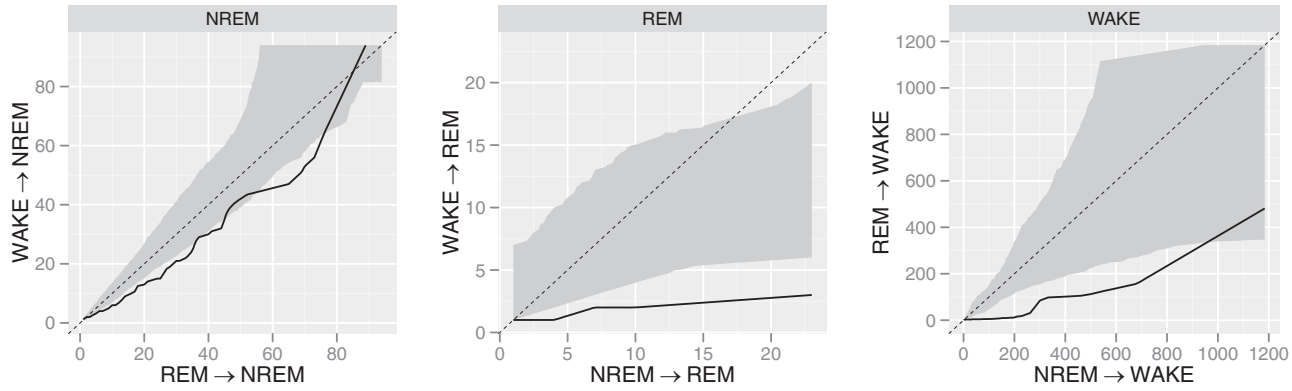

Figure 3. Bout duration Q-Q plots for each state conditional on the previous state. The Q-Q line is given by the solid line and the $y = x$ line by the dashed line. The gray region is a null interval created using nonparametric bootstrap samples which permute the true labels.

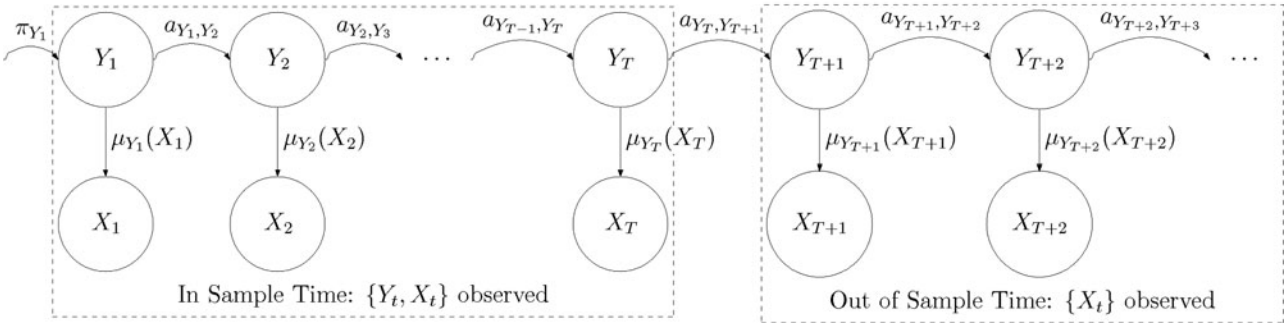

Figure 4. A first-order Markov model. In the first part, we observe both the Y_t and the $\boldsymbol{X}_t$. In the second part, we observe the $\boldsymbol{X}_t$ and must predict the Y_t.

depending on a set of covariates $\boldsymbol{X}$. Typically, the goal is to estimate the conditional class probability function $\mathbb{P}(Y = y|\boldsymbol{X} = \boldsymbol{x})$ based on iid training data $(y_i, \boldsymbol{x}_i)_{i=1}^N$ drawn from the joint distribution $\mathbb{P}(Y, \boldsymbol{X})$. Our sleep data contain strong sequential dependencies, forcing us to move beyond this iid assumption.

In particular, our dataset consists of (i) a vector of length T of response variables $\boldsymbol{Y}_{1:T} = (Y_1, \ldots, Y_T)$ with $Y_t \in S$ and (ii) a $T \times p$ matrix of covariates $\mathbf{X}_{1:T}$ whose rows are the video-based covariates $\boldsymbol{X}_t$ for time t. The goal is to either (i) estimate the joint distribution $\mathbb{P}(\boldsymbol{Y}_{1:T}, \mathbf{X}_{1:T})$ or (ii) predict a new response sequence $\boldsymbol{Y}^\star$ given a new covariate matrix $\mathbf{X}^\star$ based on the conditional distribution $\mathbb{P}(\boldsymbol{Y}_{1:T}|\mathbf{X}_{1:T})$.

2.2 Markov Models for Local Time Dependence

A standard approach for sequential data is the first-order Markov model (1MM; Rabiner 1989) shown in Figure 4. This model decomposes the joint distribution $\mathbb{P}(\boldsymbol{Y}_{1:T}, \mathbf{X}_{1:T})$ into three components:

1. An initialization distribution, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ where $\pi_i = \mathbb{P}(Y_1 = i)$.
2. A time-homogeneous transition matrix $\mathbf{A} = (a_{i,j})$ where $a_{i,j} = \mathbb{P}(Y_{t+1} = j|Y_t = i)$.
3. A set of multivariate time-homogeneous covariate emission distributions $\boldsymbol{\mu} = (\mu_1(\boldsymbol{x}), \ldots, \mu_k(\boldsymbol{x}))$ where $\mu_i(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X}_t = \boldsymbol{x}|Y_t = i)$.

A sequence of data is generated from this first-order Markov model by (i) drawing an initial state $Y_1 = i$ from $\boldsymbol{\pi}$, (ii) drawing $\boldsymbol{X}_1 = \boldsymbol{x}$ from μ_i, (iii) transitioning to state $Y_2 = j$ according to the transition probabilities given by the ith row of $\mathbf{A}$, and (iv) repeating Steps (ii)–(iii) *mutatis mutandis* until time T is reached (for full details, see the online supplementary materials).

Collecting our model parameters into $\boldsymbol{\Delta} = (\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\mu})$, the likelihood is given by

$$\begin{aligned} L(\boldsymbol{Y}_{1:T}, \mathbf{X}_{1:T}|\boldsymbol{\Delta}) &= \mathbb{P}(Y_1)\mathbb{P}(\boldsymbol{X}_1|Y_1) \\ &\quad \times \mathbb{P}(Y_2|Y_1)\mathbb{P}(\boldsymbol{X}_2|Y_2)\cdots\mathbb{P}(Y_T|Y_{T-1})\mathbb{P}(\boldsymbol{X}_T|Y_T) \\ &= \pi_{Y_1}\left[\prod_{t=2}^{T} a_{Y_{t-1},Y_t}\right]\left[\prod_{t=1}^{T}\mu_{Y_t}(\boldsymbol{X}_t)\right]. \end{aligned} \tag{1}$$

Given observed $\boldsymbol{Y}_{1:T}$, one typically estimates the values of $\mathbf{A} = (a_{i,j})$ by the empirical rate of transitions from state i to state j. To estimate $\boldsymbol{\pi}$, one usually uses either (i) the empirical frequency of each of the k classes or (ii) the stationary distribution of the estimated transition matrix[1]. Finally, one can estimate each $\mu_i(\boldsymbol{x})$ independently using the $\boldsymbol{X}_t$ for which $Y_t = i$.

The overall purpose of this model is the scoring of "out-of-sample" sleep states. That is, given a covariate sequence $\mathbf{X}^\star = (\boldsymbol{X}_{T+1}, \ldots, \boldsymbol{X}_{T+T^\star})$ and estimated parameters $\widehat{\boldsymbol{\Delta}}$, we want to calculate the conditional class probabilities for $\boldsymbol{Y}^\star = (Y_{T+1}, \ldots, Y_{T+T^\star})$. The probabilities, denoted by $\widehat{\gamma}_t(i) \equiv \mathbb{P}(Y_t = i|\mathbf{X}^\star, \widehat{\boldsymbol{\Delta}})$, can be calculated using the forward–backward algorithm (Rabiner 1989). Our actual classification $\widehat{Y}_t$ for the unobserved time periods is the state with maximum probability, $\widehat{Y}_t = \arg\max_i \widehat{\gamma}_t(i)$; classification by the modal state $\widehat{Y}_t$ is optimal in terms of minimizing the out-of-sample classification error. While we also calculate the most likely path $\widehat{\boldsymbol{Y}}^\star = \arg\max_{\boldsymbol{Y}^\star} \mathbb{P}(\boldsymbol{Y}^\star|\mathbf{X}, \widehat{\boldsymbol{\Delta}})$ using the Viterbi algorithm, we do not use it for classification purposes as $\widehat{\boldsymbol{Y}}^\star$ is suboptimal under classification error loss (Rabiner 1989).

In the context of our sleep application, the first-order Markov model suffers from two serious limitations. First, the estimation of the k multivariate probabilities $\mu_i(\boldsymbol{x})$ is a difficult task, particularly since we have a large number of covariates and there are rare states such as REM. Second, long-term dependencies in the Y_t cannot be captured due to the first-order Markov property: any relationship between Y_t and Y_{t+k} must be "mediated" by $Y_{t+1}, \ldots, Y_{t+k-1}$. The first-order Markov assumption also implies that the holding times in each state are geometrically distributed, an assumption at odds with both data (see Figure 2) and common sense since sleep bouts cannot be memoryless.

Our proposed methodology addresses both of these limitations. In Section 2.3, we address the estimation of $\mu_i(\boldsymbol{x})$ by estimating the model in a *discriminative* fashion. In Sections 2.4 and 2.5, we introduce longer-term dependence structures and more general holding time distributions directly into the Y_t themselves.

2.3 Discriminative Markov Models

The estimation of the k multivariate covariate emission distributions $\mu_i(\boldsymbol{x})$ is difficult, particularly when the number of covariates p is large, the distributional form of μ_i is unknown,

[1]Estimating $\boldsymbol{\pi}$ may or may not be necessary for out-of-sample prediction. When the out-of-sample sequence "continues on" from the in-sample sequence as in Figure 4, the initialization distribution for Y_{T+1} is simply the row of the transition probability matrix $\mathbf{A}$ corresponding to the observed Y_T and thus no estimate of $\boldsymbol{\pi}$ is required. When the out-of-sample sequence is an entirely new sequence, an estimate of $\boldsymbol{\pi}$ is necessary.

or some of the states i are rare. However, it is potentially much easier to address the inverse problem of classifying categorical Y_t based on high-dimensional $\boldsymbol{X}_t$. This is the setting of various classification methods such as logistic regression and RF. In particular, Bayes' theorem links classification methods to the covariate emission distributions μ_i as

$$\mu_i(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X}_t = \boldsymbol{x} | Y_t = i) = \frac{\overbrace{\mathbb{P}(Y_t = i \mid \boldsymbol{X}_t = \boldsymbol{x})}^{\text{Classification Methods}} \cdot \overbrace{\mathbb{P}(\boldsymbol{X}_t = \boldsymbol{x})}^{\text{Constant wrt } Y_t}}{\underbrace{\mathbb{P}(Y_t = i)}_{\text{Marginal Probabilities}}} \propto \frac{f_i(\boldsymbol{X}_t)}{p_i},$$

where $f_i(\boldsymbol{X}_t) = \mathbb{P}(Y_t = i|\boldsymbol{X}_t)$ and $p_i = \mathbb{P}(Y_t = i)$. We can thus rewrite the likelihood given in Equation (1) as

$$L(\boldsymbol{Y}_{1:T}, \mathbf{X}_{1:T}|\boldsymbol{\Delta}) \propto \pi_{Y_1} \left[\prod_{t=2}^{T} a_{Y_{t-1},Y_t}\right] \left[\prod_{t=1}^{T} \frac{f_{Y_t}(\boldsymbol{X}_t)}{p_{Y_t}}\right] = L(\boldsymbol{Y}_{1:T}, \mathbf{X}_{1:T}|\boldsymbol{\Theta}),$$

where $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{f}, \boldsymbol{p})$ with $\boldsymbol{f} = (f_1(x), \ldots, f_k(x))$ and $\boldsymbol{p} = (p_1, \ldots, p_k)$.

We have thus transformed the difficult problem of estimating k multivariate probability distributions $\boldsymbol{\mu}$ into the more straightforward problem of estimating (i) a conditional class probability vector $\boldsymbol{f}$ and (ii) a marginal probability vector $\boldsymbol{p}$. For estimating $\boldsymbol{f}$, any classification method that gives conditional class probability estimates will suffice, though—depending on the data—some may be more appropriate than others; in this article, we focus on using either logistic regression or RF. Probabilities $\boldsymbol{p}$ are estimated by either (i) the empirical frequency of each of the k states or (ii) the stationary distribution corresponding to the estimate of $\mathbf{A}$. The forward–backward algorithms require only slight modifications to calculate $\widehat{\gamma}_t(i) = \mathbb{P}(Y_t = i|\mathbf{X}, \widehat{\boldsymbol{\Theta}})$ under our new parameterization (see the Appendix for details).

In summary, rather than estimating the Markov model *generatively*, we estimate the model *discriminatively* (Smyth 1994). The discriminative Markov model has many advantages including allowing us (i) to avoid the estimation of k multivariate probability emission distributions, (ii) to enhance *any* standard classification methodology to accommodate sequential data, and (iii) to save computational costs as compared with alternative sequential data methods [e.g., conditional random fields (Lafferty, McCallum, and Pereira 2001)].

Nonetheless, our discriminative Markov model faces the limitation that only local time dependence is modeled. Any relationship between Y_t and Y_{t+k} must be communicated via $Y_{t+1}, Y_{t+2}, \ldots, Y_{t+k-1}$ and holding times in each state must be geometrically distributed. We extend our discriminative Markov approach to accommodate more general and longer-range dependencies in the next two sections.

2.4 Generalized Markov Models

The first-order Markov model presented in the previous sections has geometrically distributed duration times: the probability of staying in state i for τ epochs conditional on having arrived in state i is $\mathbb{P}_i(\tau) = a_{i,i}^{\tau-1}(1 - a_{i,i})$. This distribution, and its memoryless property, provides a poor fit to our sleep data (see Figure 2). While it is conceptually straightforward to generalize the model to accommodate more general duration distributions, efficient computational algorithms (i.e., forward–backward and Viterbi) are only available for first-order Markov models. We overcome this limitation by embedding a *generalized* Markov model (denoted by GMM) with nongeometric durations into a 1MM structure, thereby retaining the use of the efficient forward–backward and Viterbi algorithms.

A GMM is parameterized by $\boldsymbol{\Delta} = (\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\mu}, \boldsymbol{\delta})$ in the generative case and $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{f}, \boldsymbol{p}, \boldsymbol{\delta})$ in the discriminative case. All parameters remain the same as in the 1MM case but (i) each $a_{i,i} = 0$ and (ii) holding times within state i are governed not by $a_{i,i}$ but by an arbitrary duration distribution $\delta_i(\tau) = \mathbb{P}_i(\tau)$, which gives the probability of spending τ periods in state i conditional on having arrived in state i. By $\boldsymbol{\delta}$, we denote the collection $(\delta_1(\tau), \ldots, \delta_k(\tau))$ of these distributions.

A sequence of data is generated from a GMM by (i) drawing an initial state $Y_1 = i$ from $\boldsymbol{\pi}$; (ii) drawing a duration τ_1 from δ_i and setting $Y_1 = \cdots = Y_{\tau_1} = i$; (iii) drawing $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{\tau_1} \overset{\text{iid}}{\sim} \mu_i$; (iv) transitioning to state $Y_{\tau_1+1} = j \neq i$ according to the transition probabilities given by the ith row of $\mathbf{A}$; and (v) repeating steps (ii)–(iv) *mutatis mutandis* until time T is reached (for full details, see the online supplementary materials). Hence, the likelihood is given by

$$\begin{aligned}
&L(\boldsymbol{Y}_{1:T}, \mathbf{X}_{1:T}|\boldsymbol{\Delta}) \\
&= \left[\mathbb{P}(Y_1)\mathbb{P}_{Y_1}(\tau_1)\right]\left[\prod_{t|Y_t \neq Y_{t-1}} \mathbb{P}(Y_t|Y_{t-1})\mathbb{P}_{Y_t}(\tau_t)\right]\left[\prod_{t=1}^{T} \mathbb{P}(\boldsymbol{X}_t|Y_t)\right] \\
&= \left[\pi_{Y_1}\delta_{Y_1}(\tau_1)\right]\left[\prod_{t|Y_t \neq Y_{t-1}} a_{Y_t|Y_{t-1}}\delta_{Y_t}(\tau_t)\right]\left[\prod_{t=1}^{T} \mu_{Y_t}(\boldsymbol{X}_t)\right] \\
&\propto \left[\pi_{Y_1}\delta_{Y_1}(\tau_1)\right]\left[\prod_{t|Y_t \neq Y_{t-1}} a_{Y_t|Y_{t-1}}\delta_{Y_t}(\tau_t)\right]\left[\prod_{t=1}^{T} \frac{f_{Y_t}(\boldsymbol{X}_t)}{p_{Y_t}}\right] \\
&= L(\boldsymbol{Y}_{1:T}, \mathbf{X}_{1:T}|\boldsymbol{\Theta}). \qquad (2)
\end{aligned}$$

To embed the GMM in a 1MM, we must model the $\delta_i(\tau)$ as

$$\delta_i(\tau) = d_i(\tau) \cdot \mathbf{I}(\tau \leq M_i) + (1 - q_i) \cdot s_i^{\tau - M_i - 1} \cdot (1 - s_i) \cdot \mathbf{I}(\tau > M_i), \qquad (3)$$

which is a mixture of (i) a distribution $d_i(\tau)$, which gives the probability mass for $\tau = 1, 2, \ldots, M_i$, and (ii) a shifted geometric distribution, which gives the probability mass for values of τ greater than M_i where M_i separates the "head" and "tail" of the distribution. More specifically, the head of the distribution has total mass given by the mixing proportion $q_i = \sum_{\tau=1}^{M_i} d_i(\tau)$ and this mass is spread over arbitrary shape $d_i(\tau)$ and arbitrary length M_i; on the other hand, the tail has total mass $1 - q_i$ and this mass is of geometric shape and unbounded length.

We write the full set of parameters for our GMM as $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{f}, \boldsymbol{p}, \boldsymbol{d}, \boldsymbol{s}, \boldsymbol{q})$ replacing $\boldsymbol{\delta}$ by the three equivalent parameters (i) $\boldsymbol{d} = (d_1(\tau), \ldots, d_k(\tau))$ where each d_i is a vector of length M_i giving the probability of remaining in state i conditional on having arrived there for $\tau = 1, \ldots, M_i$ periods, (ii) $\boldsymbol{s} = (s_1, \ldots, s_k)$ where each s_i is the geometric distribution parameter for state i, and (iii) $\boldsymbol{q} = (q_1, \ldots, q_k)$ where each $q_i = \sum_{\tau=1}^{M_i} d_i(\tau)$ is the mixing proportion between the head and tail of the duration distribution for state i.

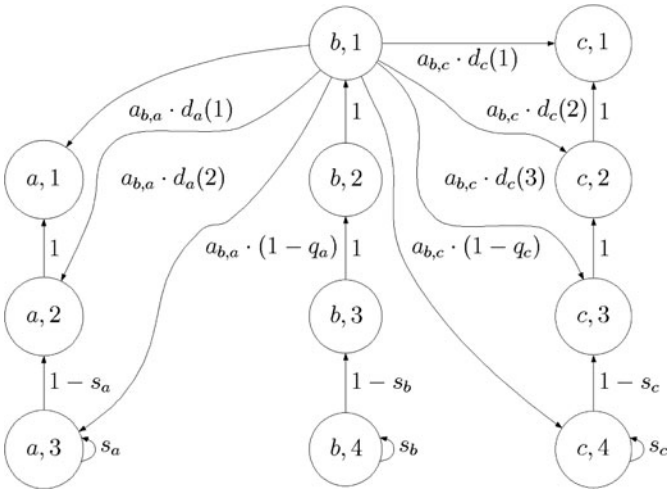

Figure 5. GMM transition diagram. The original state space is $S = \{a, b, c\}$ and the head sizes are $M_a = 2$, $M_b = M_c = 3$. All transitions within the "head" and "tail" durations are given. Of transitions away from states, only transitions away from $(b, 1)$ are given; similar transitions from $(a, 1)$ to each (b, n) and (c, n) and from $(c, 1)$ to each (a, n) and (b, n) are omitted for aesthetic reasons.

We can embed our GMM into a 1MM through a set of augmented variables. Given an observed sequence $\boldsymbol{Y}_{1:T} = (Y_1, \ldots, Y_T)$, we construct $R_t = \arg\max_\tau \{Y_t = Y_{t+1} = \cdots = Y_{t+\tau-1} \neq Y_{t+\tau}\}$ and $Z_t = \min(R_t, M_{Y_t} + 1)$. R_t gives how much longer the sequence remains in the current state and Z_t caps R_t at the maximal size. We denote our augmented variable as $Y'_t = (Y_t, Z_t)$ and our augmented state space as $S' = \{(i, n) | i \in S, n = 1, \ldots, M_i + 1\}$ (i.e., all possible values of the couplet Y'_t).

As a simple example, consider the observed sequence $\boldsymbol{Y}_{1:T} = \{a, a, a, a, b, b, b, b, b, c, c, a\}$ on $S = \{a, b, c\}$ with $M_a = 2$, $M_b = M_c = 3$. Our process converts $\boldsymbol{Y}_{1:T}$ to the augmented sequence $\boldsymbol{Y}'_{1:T} = \{(a, 3), (a, 3), (a, 2), (a, 1), (b, 4), (b, 4), (b, 3), (b, 2), (b, 1), (c, 2), (c, 1), (a, 1)\}$.

Our augmented state space S' requires an augmented transition probability matrix $\mathbf{A}'$, which we construct as follows (see Figure 5 for a simple example based on the above):

1. *Transitions away from states*: For $Z_t = 1$, there are three cases:
   (a) Transitions from $Y'_t = (i, 1)$ to $Y'_{t+1} = (i, n)$ have probability zero for $n = 1, \ldots, M_i$.
   (b) Transitions from $Y'_t = (i, 1)$ to $Y'_{t+1} = (j, n)$ have probability $a_{i,j} \cdot d_i(n)$ for $i \neq j$ and $n = 1, \ldots, M_j$.
   (c) Transitions from $Y'_t = (i, 1)$ to $Y'_{t+1} = (j, M_j + 1)$ have probability $a_{i,j} \cdot (1 - q_j)$ for $i \neq j$.
2. *Transitions within "head" duration of states*: For $Z_t = 2, \ldots, M_i$, the transition from $Y'_t = (i, Z_t)$ to $Y'_{t+1} = (i, Z_t - 1)$ has probability one; all other transitions have probability zero.
3. *Transitions within "tail" duration of states*: For $Z_t = M_i + 1$, the self-transition from $Y'_t = (i, M_i + 1)$ to $Y'_{t+1} = (i, M_i + 1)$ has probability s_i and the transition from $Y'_t = (i, M_i + 1)$ to $Y'_{t+1} = (i, M_i)$ has probability $(1 - s_i)$; all other transitions have probability zero.

We can also easily augment the remaining parameters for our new state space S'. Marginal probabilities $\boldsymbol{p}'$ are obtained from the stationary distribution of $\mathbf{A}'$. To obtain the conditional class probabilities $\boldsymbol{f}'$ and the initialization distribution $\boldsymbol{\pi}'$, let us first define $Y(i)$ and $Z(i)$ as the first and second elements, respectively, of the state couplet $i = (Y_t, Z_t) \in S'$. Then, the $\boldsymbol{f}'$ are obtained by setting $f'_i(\boldsymbol{x}) = \mathbb{P}(Y'_t = i | \boldsymbol{X}_t = \boldsymbol{x}) = f_{Y(i)}(\boldsymbol{x}) \cdot p'_i / \sum_{j|Y(j)=Y(i)} p'_j$. Similarly, each $\pi'_i = \mathbb{P}(Y'_1 = i)$ is given by $\pi'_i = \pi_{Y(i)} \cdot d_{Y(i)}(Z(i))$ if $Z(i) = 1, \ldots, M_i$ or $\pi'_i = \pi_{Y(i)} \cdot (1 - q_{Y(i)})$ if $Z(i) = M_i + 1$. Thus, the f'_i and π'_i are simply reweighted versions of the original $f_{Y(i)}$ and $\pi_{Y(i)}$.

In summary, we have embedded a GMM on state space S parameterized by $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{f}, \boldsymbol{p}, \boldsymbol{d}, \boldsymbol{s}, \boldsymbol{q})$ into a 1MM on state space S' parameterized by $\boldsymbol{\Theta}' = (\boldsymbol{\pi}', \mathbf{A}', \boldsymbol{f}', \boldsymbol{p}')$. Because this is a true embedding, the likelihood $L(\boldsymbol{Y}_{1:T}, \mathbf{X}_{1:T} | \boldsymbol{\Theta})$ given in Equation (2) is equivalent to the likelihood under the new parameterization, $L(\boldsymbol{Y}'_{1:T}, \mathbf{X}_{1:T} | \boldsymbol{\Theta}')$.

The estimation of the parameters $\boldsymbol{\Theta}'$ is straightforward and we begin with the transition probability matrix $\mathbf{A}'$. While the maximum likelihood estimate of $\mathbf{A}'$ is given by the empirical frequencies of the Y'_t transitions, this can be an inefficient estimate since it ignores the known structure in $\mathbf{A}'$ (i.e., entries are known to be zero, one, or products of the original $\mathbf{A}$ and the parameters $\boldsymbol{d}$, $\boldsymbol{s}$, and $\boldsymbol{q}$). Instead, we prefer (i) to estimate the transition probability $a_{i,j}$ from state i to state j using empirical frequencies and (ii) to parameterize the $d_i(\tau)$ and estimate the parameters d_i, s_i, and q_i using the observed duration times for state i. We can then combine estimates (i) and (ii) as detailed above to form our estimate of $\mathbf{A}'$.

The $\boldsymbol{p}'$ are estimated by using the stationary distribution of the estimate of $\mathbf{A}'$. The augmented conditional class probabilities $\boldsymbol{f}'$ are estimated by (i) estimating $\boldsymbol{f}$ as in the 1MM case (i.e., by using standard classification procedures such as logistic regression and RF) and then (ii) reweighting these estimates as above (i.e., $\hat{f}'_i(\boldsymbol{x}) = \hat{f}_{Y(i)}(\boldsymbol{x}) \cdot \hat{p}'_i / \sum_{j|Y(j)=Y(i)} \hat{p}'_j$). Similarly, the initialization parameters $\boldsymbol{\pi}$ are estimated by (i) estimating $\boldsymbol{\pi}$ as in the 1MM case and then (ii) reweighting these estimates as above (i.e., $\hat{\pi}'_i = \hat{\pi}_{Y(i)} \cdot \hat{d}_{Y(i)}(Z(i))$ if $Z(i) = 1, \ldots, M_i$ or $\hat{\pi}'_i = \hat{\pi}_{Y(i)} \cdot (1 - \hat{q}_{Y(i)})$ if $Z(i) = M_i + 1$). Given these estimates of a GMM embedded into a 1MM, we can apply the the forward–backward and Viterbi algorithms for the discriminative 1MM just as in Section 2.3

2.5 Transition-Dependent Generalized Markov Model

Although the GMM presented in Section 2.4 is much more general than a 1MM, certain properties of the GMM are not realistic in our application. Specifically, the duration distributions in a GMM are "unconditional": the duration distribution for state i does not depend on the prior state j. However, we have observed in our sleep data that the duration of, for example, WAKE bouts, which follow from NREM, are longer than WAKE bouts, which follow from REM (see Figure 3).

We thus introduce a transition-dependent generalized Markov model (TDGMM) where the duration distributions of the current state depend on the previous state. In this new model, $\boldsymbol{\delta}$ contains $k \cdot (k - 1)$ duration distributions $\delta_{i,j}(\tau)$ instead of the k duration distributions $\delta_i(\tau)$ of the GMM; each $\delta_{i,j}(\tau) = \mathbb{P}_{i,j}(\tau)$ is the probability of spending τ periods in state j conditional on having

arrived in state j from state i. The data-generation process and likelihood for the TDGMM are identical to those of the GMM with the appropriate $\delta_{i,j}$ used in place of δ_j.

We model the $\delta_{i,j}$ using the specification of Equation (3), and hence, our duration distribution parameters are again given by $(\boldsymbol{d}, \boldsymbol{s}, \boldsymbol{q})$. In contrast to the GMM where each of these parameters was of length k (i.e., one for each δ_i), in the TDGMM, each of these parameters is of length $k \cdot (k-1)$ (i.e., one for each $\delta_{i,j}$).

As with the GMM, we can embed the new TDGMM within the first-order Markov structure by augmenting the state space. Assuming we have observed $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$, we define $P_t = Y_{\sigma_t}$ where $\sigma_t = \arg\min_s(Y_s \neq Y_{s+1} = Y_{s+2} = \cdots = Y_t)$ (i.e., the state the sequence was in prior to the current state). We then construct R_t as we did in for the GMM (i.e., $R_t = \arg\max_\tau\{Y_t = Y_{t+1} = \cdots = Y_{t+\tau-1} \neq Y_{t+\tau}\}$) and again cap it with $Z_t = \min(R_t, M_{P_t,Y_t} + 1)$. Finally, we set the triplet $Y'_t = (P_t, Y_t, Z_t)$ thus collecting the previous state, the current state, and the capped length of time remaining in the current state. The augmented state space is $S' = \{(i, j, n) | i \in S, j \in S, i \neq j, n = 1, \ldots, M_{i,j} + 1\}$ (i.e., all possible values of the triplet Y'_t).

Consider again the example $\boldsymbol{Y}_{1:T} = \{a, a, a, a, b, b, b, b, b, c, c, a\}$ on $S = \{a, b, c\}$ with $M_{b,a} = M_{a,b} = M_{c,b} = 3$, $M_{c,a} = M_{a,c} = M_{b,c} = 2$. Supposing $Y_0 = c$, our process converts $\boldsymbol{Y}_{1:T}$ into the augmented sequence $\boldsymbol{Y}' = \{(c, a, 3), (c, a, 3), (c, a, 2), (c, a, 1), (a, b, 4), (a, b, 4), (a, b, 3), (a, b, 2), (a, b, 1), (b, c, 2), (b, c, 1), (c, a, 1)\}$.

As before, augmented state space S' requires an augmented transition probability matrix $\mathbf{A}'$, which can be constructed as follows (see Figure 6 for a simple example based on the above):

1. *Transitions away from states*: For $Z_t = 1$, there are three cases:
   (a) Transitions from $Y'_t = (i, j, 1)$ to $Y'_{t+1} = (i', j', n)$ have probability zero for $n = 1, \ldots, M_{i',j'}$ and $i' \neq j$.
   (b) Transitions from $Y'_t = (i, j, 1)$ to $Y'_{t+1} = (i', j', n)$ have probability $a_{i',j'} \cdot d_{i',j'}(n)$ for $n = 1, \ldots, M_{i',j'}$, $i' = j$, and $j' \neq j$.
   (c) Transitions from $Y'_t = (i, j, 1)$ to $Y'_{t+1} = (i', j', M_{i',j'} + 1)$ have probability $a_{i',j'} \cdot (1 - q_{i',j'})$ for $i' = j$ and $j' \neq j$.
2. *Transitions within "head" duration of states*: For $Z_t = 2, \ldots, M_{i,j}$, the transition from $Y'_t = (i, j, Z_t)$ to $Y'_{t+1} = (i, j, Z_t - 1)$ has probability one; all other transitions have probability zero.
3. *Transitions within "tail" duration of states*: For $Z_t = M_{i,j} + 1$, the self-transition from $Y'_t = (i, j, M_{i,j} + 1)$ to $Y'_{t+1} = (i, j, M_{i,j} + 1)$ has probability $s_{i,j}$ and the transition from $Y'_t = (i, j, M_{i,j} + 1)$ to $Y'_{t+1} = (i, j, M_{i,j})$ has probability $(1 - s_{i,j})$; all other transitions have probability zero.

We have thus embedded a TDGMM on state space S parameterized by $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{f}, \boldsymbol{p}, \boldsymbol{d}, \boldsymbol{s}, \boldsymbol{q})$ into a 1MM on state space S' parameterized by $\boldsymbol{\Theta}' = (\boldsymbol{\pi}', \mathbf{A}', \boldsymbol{f}', \boldsymbol{p}')$. Due to the similarity between the GMM and the TDGMM, we can again use the estimation strategy outlined in Section 2.4 (the only difference is that the duration distribution parameters $(\boldsymbol{d}, \boldsymbol{s}, \boldsymbol{q})$ are estimated based on the observed duration times for state j conditional on entering state j from state i rather than unconditionally and we use the above procedure to construct estimates of $\mathbf{A}'$; these differences propagate in the obvious manner into the estimates of $\boldsymbol{p}'$, $\boldsymbol{f}'$, and $\boldsymbol{\pi}'$).

3. SIMULATION EVALUATION OF MODELS

We evaluate the performance of several models presented in Section 2 via simulation where the true model is a TDGMM. Our simulation state space is $S = \{a, b, c\}$ and the initialization

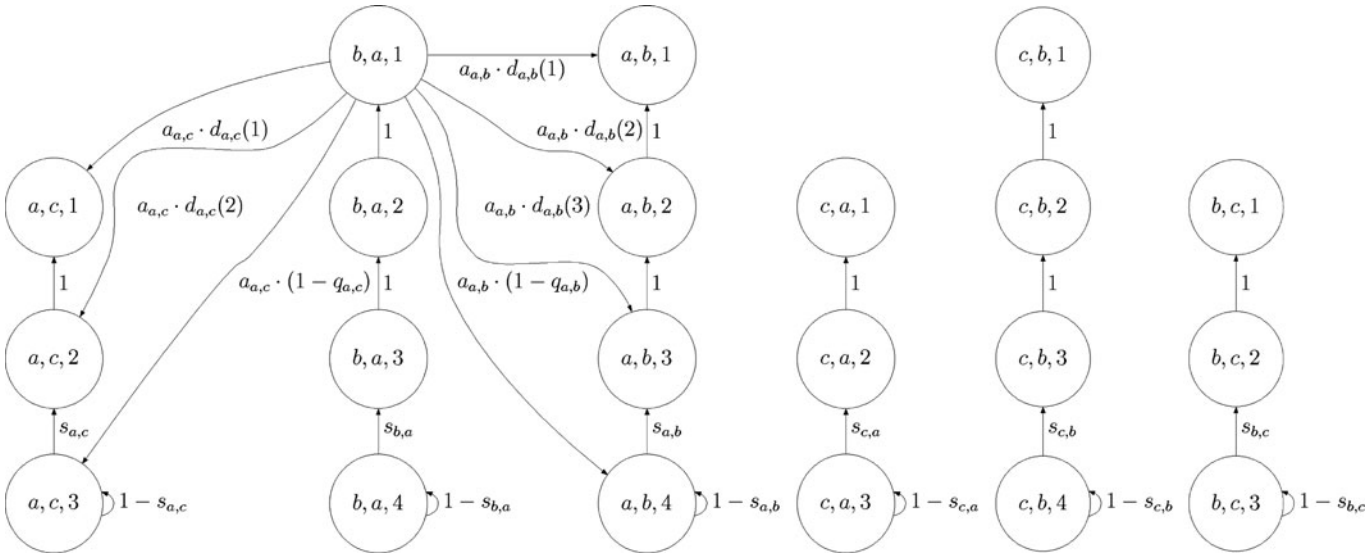

Figure 6. TDGMM transition diagram. The original state space is $S = \{a, b, c\}$ and the head sizes are $M_{b,a} = M_{a,b} = M_{c,b} = 3$, $M_{c,a} = M_{a,c} = M_{b,c} = 2$. All transitions within the "head" and "tail" durations are given. Of transitions away from states, only transitions away from $(b, a, 1)$ are given; similar transitions from $(a, c, 1)$ to each (c, a, n) and (c, b, n); from $(a, b, 1)$ to each (b, a, n) and (b, c, n); from $(c, a, 1)$ to each (a, b, n) and (a, c, n); from $(c, b, 1)$ to each (b, a, n) and (b, c, n); and from $(b, c, 1)$ to each (c, a, n) and (c, b, n) are omitted for aesthetic reasons.

distribution is the uniform distribution $\boldsymbol{\pi} = (1/3, 1/3, 1/3)$. We set the covariate emission distributions $\boldsymbol{\mu}$ equal to $X_t \sim N(\mu_{Y_t}, \sigma^2)$, where $\mu_a = 0$, $\mu_b = 1$, and $\mu_c = 2$ and σ is set to 13 different values ranging from 0.01 to 3. The transition probability distributions are given by

$$\mathbf{A} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 3/4 & 0 & 1/4 \\ 1/3 & 2/3 & 0 \end{pmatrix}.$$

The duration distributions $\delta_{i,j}$ are as in Equation (3) with $M_{i,j}$ set to $M = 10$ and $d_{i,j}$, the "head" components of the duration distributions, set to the Beta Negative Binomial distribution with finite support,

$$d(\tau|\alpha, \beta, r) = \frac{1}{c(\alpha, \beta, r, M)} \times \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+r)\Gamma(\tau+r-1)\Gamma(\tau+\beta-1)}{\Gamma(r)\Gamma(\alpha)\Gamma(\beta)\Gamma(\tau)\Gamma(\tau+r+\alpha+\beta-1)}, \quad \tau = 1, \ldots, M,$$

where $c(\alpha, \beta, r, M)$ is a normalizing constant ensuring that each $\delta_{i,j}$ sums to one (i.e., $d_{i,j}$ sums to $q_{i,j}$). The parameters (α, β, r) of the "head" components and (q, s) of the "tail" components of the duration distributions $\delta_{i,j}$ were set to

State	α	β	r	q	s
$a \to b$	0.00	1.00	442413	1.00	0.00
$a \to c$	0.00	1.65	0.45	1.00	0.00
$b \to a$	1808	148.41	33.12	1.00	0.00
$b \to c$	0.00	7.39	7.39	0.50	0.69
$c \to a$	0.00	22026	0.61	0.62	0.90
$c \to b$	0.00	1.00	442413	0.50	0.90

The parameters settings were chosen to provide a diversity of shapes as illustrated in Figure 7.

Our study uses three different training set sizes ($T = 100$, $T = 1000$, and $T = 10{,}000$). The test set size is always fixed at $T^\star = 200$ and "continues" from the training data as in Figure 4. For each value of σ and T, results are averaged over 1000 simulated datasets. The competing models we consider are (i) multinomial logistic regression (MLR), (ii) MLR enhanced by a 1MM (MLR+1MM), (iii) MLR enhanced by a GMM (MLR+GMM), and (iv) MLR enhanced by a TDGMM (MLR+TDGMM).

Performance is evaluated in three ways. First, we examine the classification error of the each of the four methods as well as that of the Bayes' Rule, which classifies based on the modal true marginal probability $\gamma_t(i) = \mathbb{P}(Y_t|\mathbf{X}^\star, \boldsymbol{\Delta})$. Second, since the Bayes' Rule gives the optimal classifications that minimize the error rate and since it is free of the noise inherent in the Y_t, we

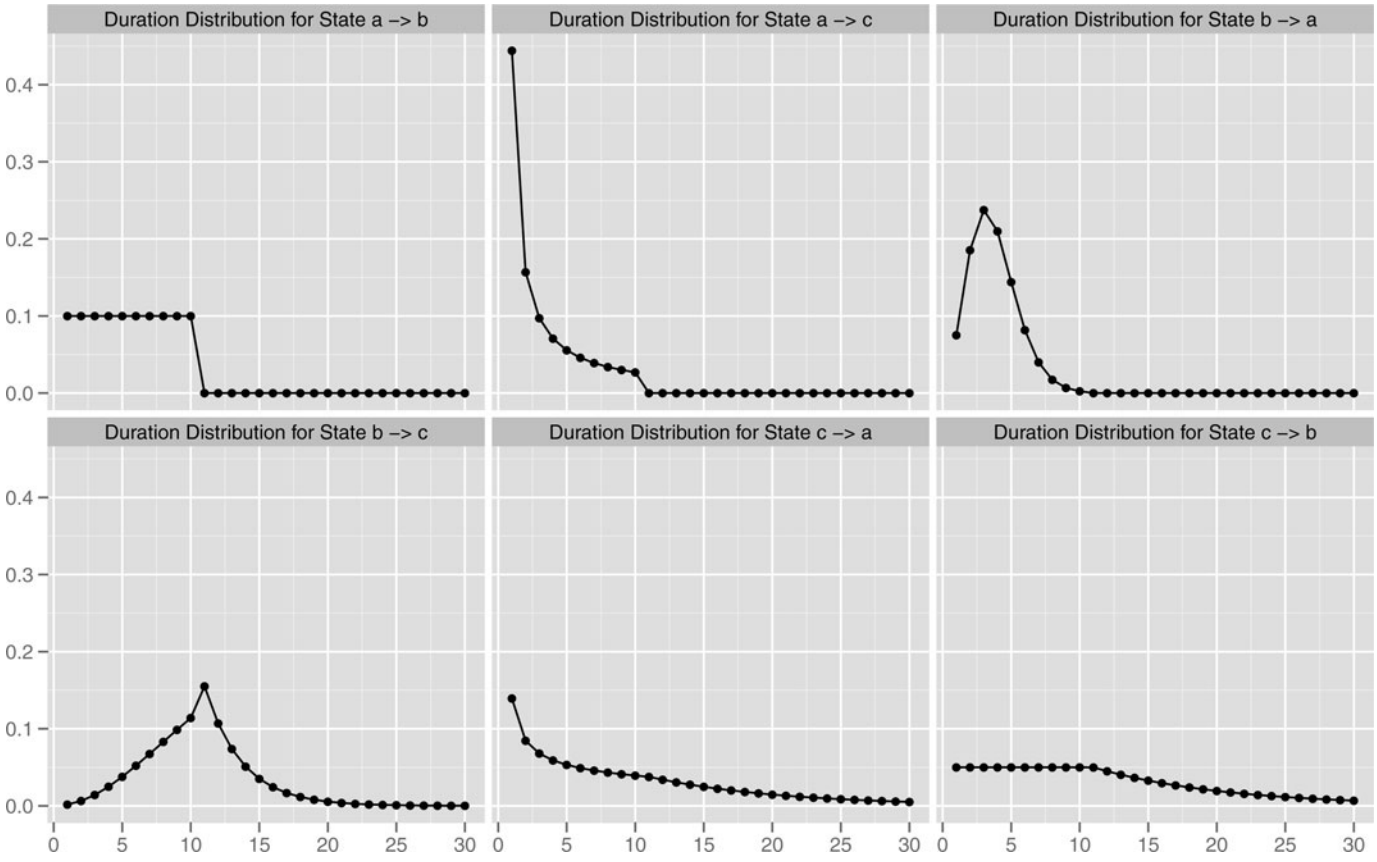

Figure 7. Transition-dependent duration distributions for the simulation.

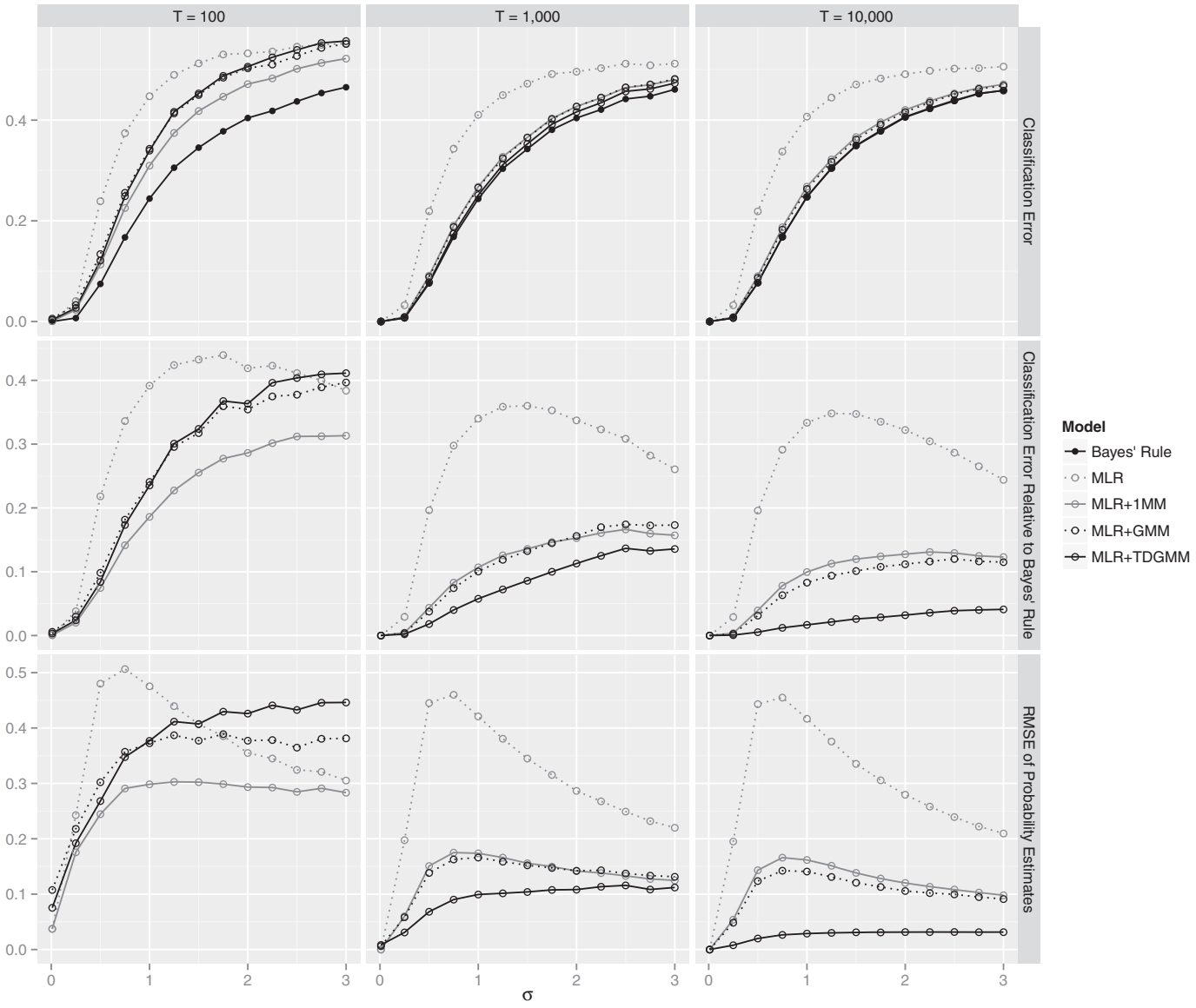

Figure 8. Simulation results. The x-axis is the noise level σ of the covariate emission distributions. The y-axis is the classification error (row 1), the classification error relative to Bayes' Rule (row 2), and the RMSE of the probability estimates (row 3). The columns give the three training set sizes.

also examine the error rate of each of the four methods relative to the classes predicted by the Bayes' Rule. Finally, we examine the root mean squared error (RMSE) of the marginal probability estimates $\widehat{\gamma}_t(i) \equiv \hat{\mathbb{P}}(Y = i|\mathbf{X}^\star, \widehat{\boldsymbol{\Theta}})$ from the four models[2].

We present our simulation results in Figure 8. In terms of classification error, the correctly specified TDGMM model does best with a moderate to large amount of in-sample data. The incorrectly specified models improve only mildly as T is increased from 1000 to 10,000, whereas the correctly specified TDGMM improves markedly. However, even for $T = 10{,}000$, the TDGMM does not match the Bayes' Rule at high noise levels; this suggests that massive amounts of training data may be required in very high noise settings. The GMM model performs poorly relative to the TDGMM because it estimates a single duration distribution for each state when the reality is a mixture of transition-specific duration distributions.

In terms of the RMSE of the estimated probabilities, the TDGMM gives the best results in large training samples and appears to be converging on the true probabilities as the training set size grows. However, the simpler 1MM and GMM models seem to provide better estimates in small sample sizes and comparable estimates at high noise levels. Again, the incorrectly specified models improve only mildly as T is increased from 1000 to 10,000, whereas the correctly specified TDGMM improves markedly—an encouraging result for our approach.

4. APPLICATION TO SLEEP SCORING

4.1 Data Description and Summary Statistics

Our sleep data come from eight mice of the strain C57BL/6J. Each mouse was manually scored for sleep by the invasive and

[2]We also examined other proper scoring rules such as log loss and exponential loss (Savage 1971; Buja, Stuetzle, and Shen 2005). All yielded results that were qualitatively similar to those presented for squared error loss.

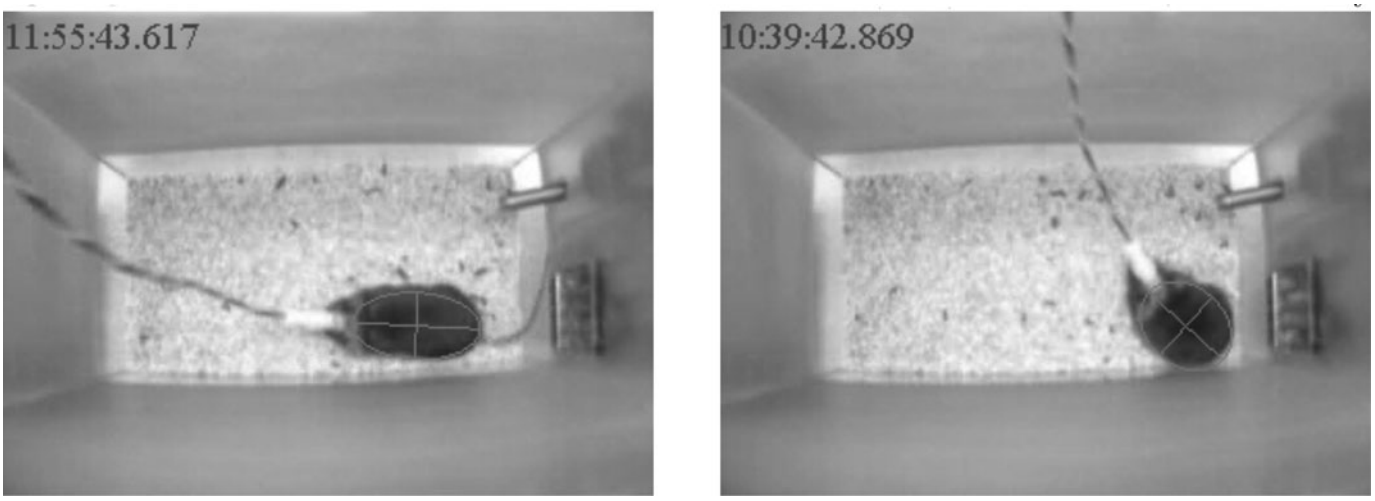

Figure 9. Two frames of video data. The left frame illustrates a typical awake posture, whereas the right frame illustrates a typical sleep posture. Our tracking software imposes an ellipse upon the mouse which is used to calculate its size (area), aspect ratio, and velocity.

laborious method described in Section 1.1. That is, the mice were surgically implanted with EEG/EMG electrodes and allowed a 10- to 14-day postsurgery recovery and a habituation period. Next, EEG/EMG signals were recorded for 24 hr, digitized at 256 Hz, and broken up into 10-sec epochs (8640 in total per mouse) for manual scoring. Each of the 8×8640 epochs was manually scored as NREM, REM, or WAKE by two independent scorers.

The independent scorers disagreed on about $\approx 6\%$ of the 10-sec epochs and disagreement rates were substantially higher among those epochs where the sleep stage was REM or transitional. For each epoch where the two scorers disagreed, we used an independent third scorer to break the tie and to thereby determine the "true" state.

The goal of our methodology is to predict sleep scores using video data that closely match the "gold standard" EEG/EMG-based manual scores. Our digital video consists of 100 frames per epoch and two example frames, one typical of wakefulness and one typical of sleep, are shown in Figure 9. Tracking software was used to fit an elliptical approximation to the mouse in each frame.

Using this ellipse, the tracking software calculates six continuous covariates for each epoch: the intra-epoch mean and standard deviation of the velocity, aspect ratio, and size (area) of the mouse. We also have a single binary covariate that indicates whether or not a light in the cage was turned on in each epoch (lights were on from 7 a.m. to 7 p.m.). Thus, we have seven covariates $\boldsymbol{X}_t$ with which to predict the sleep stage Y_t for each of 8640 epochs.

Table 1. Summary statistics by sleep state

Sleep state	Fraction of epochs (%)	Number of bouts	Average duration
NREM	43.99	2002	15.19
REM	4.82	452	7.38
WAKE	51.19	1907	18.55

NOTE: Average duration is given in number of 10-sec epochs.

Table 2. Summary statistics by conditional sleep state

Sleep state	Fraction of epochs (%)	Number of bouts	Average duration
NREM → REM	4.86	446	7.45
WAKE → REM	0.01	5	1.80
REM → NREM	3.12	102	20.94
WAKE → NREM	41.25	1,898	14.88
NREM → WAKE	48.19	1,552	21.26
REM → WAKE	2.58	350	5.04

NOTE: Average duration is given in number of 10-sec epochs.

To give an overall sense of the data, we provide summary statistics for the three states in Table 1. The mice spend about half of the day awake and the other half in the two stages of sleep. As mentioned previously, REM sleep is relatively rare, accounting for fewer than 5% of epochs. Furthermore, REM has many fewer bouts and a lower average bout duration.

As bout durations in a given state depend on the prior state (see Figure 3), we provide in Table 2 the same summary statistics for the three sleep states conditional on the previous state. We see that almost all bouts of REM come from NREM. This is not surprising since transitions from WAKE to REM are an extremely rare physiological occurrence[3]. For NREM, the bouts seem to be longer when entered from REM compared with when entered from WAKE. For WAKE, the bouts tend to be longer when entered from NREM rather than entered from REM. These characteristics of the data motivated our consideration of the TDGMM.

Another important feature of the duration distributions is heavy skewness, with some bouts lasting many times the length of the mean durations presented in Tables 1 and 2. As mentioned previously, sleep bout durations are distributed according to a

[3]Such transitions are indicative of sleep disorders, incorrectly scored epochs, or DREM. DREM is a direct transition from WAKE to REM that occasionally occurs in some mice. Such episodes occur almost exclusively during the lights on period and are the result of brief awakenings interrupting a sustained period of REM sleep (Fujiki et al. 2009).

"head and tail" distribution (McShane et al. 2010), which features (i) most of the probability mass in the first few epochs and (ii) the remaining mass in a long right tail. We found that the beta negative binomial distribution with geometric tail used in Section 3 could accommodate these features and provide a good fit to the conditional bout duration distributions of our TDGMM. Details of this estimation are provided in the supplementary materials.

4.2 Internal Model Evaluation

Our principal goal is to build a model based on video data, which best matches EEG/EMG-based manual scores on an epoch-by-epoch basis. We thus use a cross-validation approach to evaluating our various models: we take the full set of epochs for one mouse as our training data and test each model on the full set of epochs from the other seven mice. We then repeat this procedure using each mouse as the "in-sample" mouse and then average our results over all 483,840 out-of-sample epochs.

Before trying to discriminate REM from NREM, we first consider the simpler two-state problem of scoring SLEEP versus WAKE. We assign the "true" two-state score for a given epoch to be SLEEP if the manual scorers scored the epoch as NREM or REM and assign it to be WAKE otherwise (breaking ties, as above, by using an independent, third scorer when necessary). Our performance metric is the out-of-sample classification error rate of our predictions as compared to the manual scores.

For this two-state task, we compare the performance of six models: (i) MLR, (ii) MLR+1MM, (iii) MLR+GMM, (iv) RF, (v) random forests enhanced by a 1MM (RF+1MM), and (vi) RF enhanced by a GMM (RF+GMM). For data consisting only of two states, the TDGMM reduces to a GMM and so is excluded from this evaluation.

We also compare our model choices to the "40-second Rule" common in sleep science (Pack et al. 2007). The 40-second Rule judges a mouse "inactive" in a given epoch if the mean intra-epoch velocity is less than 3 pixels/s and scores an epoch as SLEEP when there are four or more consecutive inactive epochs (i.e., 40 or more seconds of inactivity); otherwise, it scores the epoch as WAKE.

The second column of Table 3 gives the classification error rate for each of these methods as well as the "gold-standard" rate of disagreement between manual scorers (which can be viewed as the minimal achievable error rate). We see that our 1MM and GMM enhancements of MLR and RF achieve error rates lower than the "40-second Rule" that is currently used by sleep scientists. The best overall error rates are achieved by the GMMs that have the most flexible model for the duration distributions. The MLR+GMM model provides a 15% raw improvement over MLR (which does not model time-series dependence); this equates to a 31% improvement relative to the minimum achievable error rate. It also provides a 19% raw improvement over the 40-second Rule or a 36% improvement relative to the minimum achievable. We also note that receiver operating characteristic (ROC) curves (not shown) demonstrate a uniform improvement of the GMMs relative to the 1MMs and base MLR and RF models.

We now consider the more difficult (and important) task of classifying each epoch into NREM versus REM versus WAKE.

Table 3. Out-of-sample classification error rates

Method	Two-state Error Rate (%)	Three-state Error Rate (%)	Three-state REM Rate (%)	Three-state REM FP (%)	Three-state REM FN (%)
MLR	9.7	14.9	1.4	1.2	95.3
MLR+1MM	8.4	25.1	18.3	16.8	52.4
MLR+(TD)GMM	8.2	21.9	14.5	13.0	55.8
RF	10.4	16.2	2.3	1.9	90.9
RF+1MM	8.9	24.7	17.4	15.9	53.0
RF+(TD)GMM	8.8	23.3	15.6	14.0	54.2
40-second Rule	10.1	NA	NA	NA	NA
Manual scores	4.8	5.8	NA	NA	NA

NOTE: The first column gives the methodology and the second column gives the overall classification error rate for the two-state SLEEP versus WAKE task. The third through sixth columns give, respectively, the overall classification error rate, the rate of REM prediction, the REM false positive rate, and the REM false negative rate for the three-state NREM versus REM versus WAKE task. We use a GMM for the two-state task and a TDGMM for the three-state task. FP = False Positive Rate; FN = False Negative Rate.

As discriminating between SLEEP and WAKE is comparably easy, the primary consideration is in classifying NREM versus REM. As mentioned, REM is a relatively rare state that is biologically important. Furthermore, it is the focus of many inquiries in sleep science. Consequently, in typical applications, the cost of misclassifying a NREM epoch as REM is much lower than the cost of misclassifying a REM epoch as NREM. Indeed, the rarity and importance of REM suggests the desirability of a relatively high false positive rate and a comparably low false negative rate.

For this three-state task, we compare the performance of (i) MLR, (ii) MLR+1MM, (iii) MLR+TDGMM, (iv) RF, (v) RF+1MM, and (vi) RF enhanced by a TDGMM (RF+TDGMM). We consider a TDGMM but not a GMM for this three-state task because the unconditional duration distributions of the GMM provided a poor fit to the data (see the online supplementary materials). We also exclude the "40-second Rule" from this comparison since this rule is incapable of differentiating NREM from REM sleep. In fact, there is no existing procedure in the sleep literature that classifies NREM versus REM sleep.

Classification error rates for the three-state task are presented in the third column of Table 3. The three-state task is complicated by the rarity of the REM state and the similarity between REM and NREM on the observed covariates, thus making overall error rates much higher than those in two-state task (i.e., column two versus column three). In addition to the overall error rates, we also include the rate of REM prediction (column four) as well as the false positive and false negative rate for REM (columns five and six, respectively) since this state is especially important to sleep researchers.

Table 3 gives some unsurprising results: the relatively rare REM state is difficult to classify correctly with REM false negative rates being especially large. It is a challenge to discover any method with power to detect REM sleep. That is, there is an inherent trade-off between (i) obtaining a low REM false negative rate accompanied by a higher overall error rate and a higher REM false positive rate or (ii) obtaining a lower overall error rate and a lower REM false positive rate while having a high REM false negative rate.

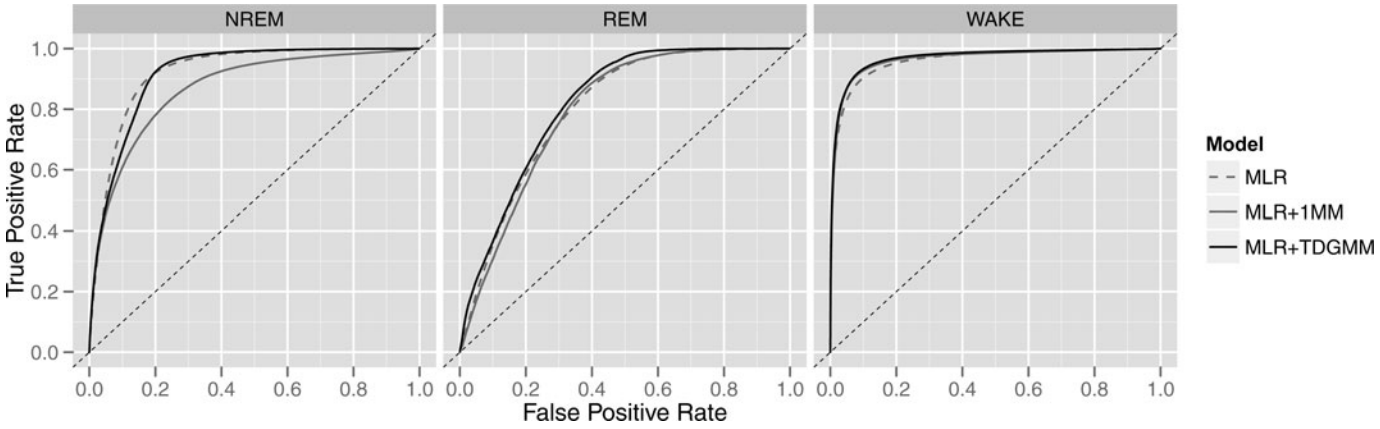

Figure 10. Out-of-sample ROC curves for the three-state REM versus NREM versus WAKE task. The curves are for the variants of multinomial logistic regression. Those for the variants of random forests are nearly identical.

The simple MLR and RF have the lowest overall error rates and REM false positive rates simply because they choose alternative (ii) and more or less ignore the REM state. Epochs are very rarely classified as REM by these simple methods, resulting in extremely high REM false negative rates. Due to the asymmetry in the cost of errors discussed above, sleep researchers would have a strong preference against such classifications.

In contrast, the TDGMM enhancements of MLR and RF are able to achieve a better balance for REM prediction: they have a much lower REM false negative rate while remaining competitive on the overall error rate and REM false positive rates. By accounting for the time dependence in the data, our TDGMM procedure is able to capture a greater proportion of the subtle REM signal.

We also examine ROC curves for the three states in Figure 10. MLR, MLR+1MM, and MLR+TDGMM all perform strongly at discriminating WAKE from the other states, though MLR+TDGMM uniformly dominates. On the other hand, there are substantial differences for REM and NREM. MLR+TDGMM uniformly dominates the other two models on the rare and important REM state. For NREM, MLR and MLR+TDGMM perform similarly and both uniformly dominate MLR+1MM.

4.3 External Model Evaluation

In this section, we compare our top methodology from the previous section (MLR+TDGMM) to several other classification techniques. Specifically, we consider AdaBoost (Freund and Schapire 1996), LogitBoost (Friedman, Hastie, and Tibshirani 2000), and Bagged Trees (Breiman 1996). We also examined two implementations of conditional random fields—the linear MALLET system (McCallum 2003) and the TreeCRF approach (Dietterich, Ashenfelter, and Bulatov 2004)—both of which are popular in the computer science literature[4].

[4]We used default parameter settings for MALLET. TreeCRF requires discrete data and so our covariates were binned using different numbers of quantiles (5, 10, 25, and 100). TreeCRF also requires that the number of leaves be specified (we tried 8, 16, 32, 64, 128, and 256). The TreeCRF results given are the best case over all these parameter settings (b and l are the corresponding number of bins and leaves).

Classification error rates for these methods applied to the three-state NREM versus REM versus WAKE task are presented in Table 4. In general, the alternative classification methods perform similarly to the simple MLR and RF models in Table 3 in that the important REM state is generally ignored. The result is lower overall error rates and REM false positive rates but extremely high REM false negative rates. Our MLR+TDGMM methodology achieves a much lower REM false negative rate with only modest increases in overall error rate and REM false positive rates.

4.4 Aggregate Measures of Sleep

Sections 4.2 and 4.3 focused on our ability to match the EEG/EMG-based manual scores on an epoch-by-epoch basis. However, sleep scientists are also interested in *aggregate* measures of sleep behavior such as the total number of minutes spent in each sleep state at various times of the day. While matching manual scores on an epoch-by-epoch basis is a sufficient condition for estimating these aggregate measures, it is not a necessary one and accurate estimates of aggregate quantities can be obtained from models that are less precise on an epoch-by-epoch basis.

A principal quantity of interest to sleep scientists is N_i^j, the average number of minutes mice spend in state i during each 2-hr block j of the day. This value can be calculated from EEG/

Table 4. Out-of-sample classification error rates for various external methods on the three-state NREM versus REM versus WAKE task

Method	Error rate (%)	REM rate (%)	REM FP (%)	REM FN (%)
AdaBoost	17.4	2.7	2.3	89.9
LogitBoost	20.7	4.0	3.7	88.7
Bagged Trees	15.2	1.5	1.2	93.4
MALLET CRF	15.1	3.9	3.3	85.2
TreeCRF 10*b*; 64*l*	16.9	3.7	3.4	89.3
MLR+TDGMM	21.9	14.5	13.0	55.8

NOTE: The first column gives the methodology, the second column the overall classification error rate, the third column the rate of REM prediction, the fourth column the REM false positive rate, and the fifth column the REM false negative rate.

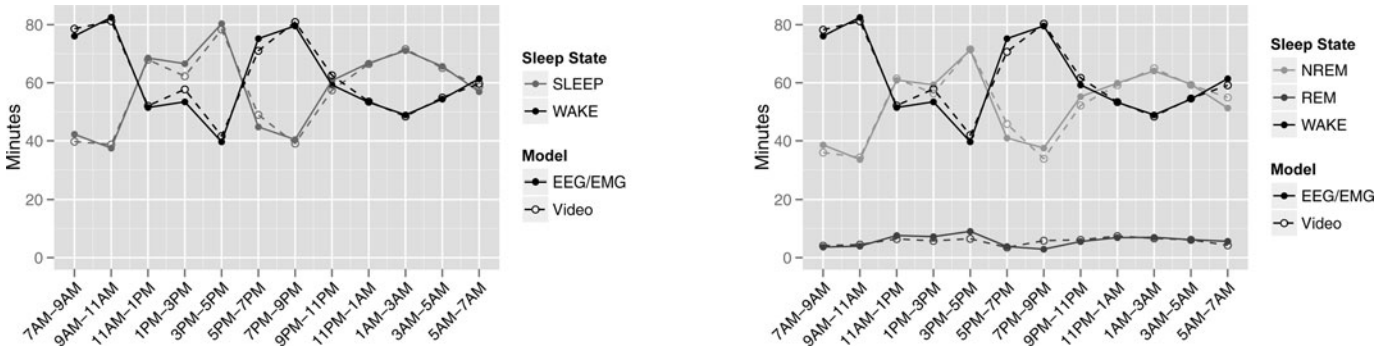

Figure 11. Minutes spent in each state. For the two-state SLEEP versus WAKE task (left panel), we compare EEG/EMG-based manual scores to our video-based MLR+GMM. For the three-state NREM versus REM versus WAKE task (right panel), we compare EEG/EMG-based manual scores to our video-based MLR+TDGMM.

EMG-based manual scores as

$$N_i^j = \frac{c}{8} \sum_{m=1}^{8} \sum_{t \in \text{Block}_j} \mathbf{I}(Y_{m,t} = i),$$

where $Y_{m,t}$ is the manual score for mouse m at epoch t and $c = 10/60$ converts time measured in 10-sec epochs to time measured in minutes.

Our methodology based on video data can be used to estimate N_i^j by summing our estimated probabilities $\widehat{\gamma}_t(i) = \mathbb{P}(Y_t = i|\mathbf{X}, \widehat{\boldsymbol{\Theta}})$ over all epochs t in block j. We average these probabilities over all combinations of one training (in-sample) mouse $\times$ seven testing (out-of-sample) mice

$$\widehat{N}_i^j = \frac{c}{8} \frac{1}{7} \sum_{m_{\text{in}}=1}^{8} \sum_{m_{\text{out}} \neq m_{\text{in}}} \sum_{t \in \text{Block}_j} \widehat{\gamma}_{m_{\text{in}},m_{\text{out}},t}(i),$$

where $\widehat{\gamma}_{m_{\text{in}},m_{\text{out}},t}(i) = \mathbb{P}(Y_{m_{\text{out}},t} = i|\mathbf{X}_{m_{\text{out}}}, \widehat{\boldsymbol{\Theta}}_{m_{\text{in}}})$ is the probability that mouse m_{out} is in state i at epoch t given covariates from mouse m_{out} and parameters estimated on mouse m_{in}.

This simple estimation is used for our two-state task of classifying SLEEP versus WAKE. For the more difficult three-state task of classifying NREM versus REM versus WAKE, we introduce additional parameters w_m, which partition the time spent REM and NREM for each mouse m. Specifically, we have

$$\widehat{N}_{\text{REM}}^j = \frac{c}{8} \frac{1}{7} \sum_{m_{\text{in}}=1}^{8} \sum_{m_{\text{out}} \neq m_{\text{in}}} \sum_{t \in \text{Block}_j} \hat{w}_{m_{\text{in}}} \cdot \widehat{\gamma}_{m_{\text{in}},m_{\text{out}},t}(\text{REM})$$

$$\widehat{N}_{\text{NREM}}^j = \frac{c}{8} \frac{1}{7} \sum_{m_{\text{in}}=1}^{8} \sum_{m_{\text{out}} \neq m_{\text{in}}} \sum_{t \in \text{Block}_j} \big[\widehat{\gamma}_{m_{\text{in}},m_{\text{out}},t}(\text{NREM}) + (1 - \hat{w}_{m_{\text{in}}}) \cdot \widehat{\gamma}_{m_{\text{in}},m_{\text{out}},t}(\text{REM})\big].$$

We let $\widehat{N}_{\text{WAKE}}^j$ remain as before. Each of the w_m parameters lies in $[0, 1]$ and allows the model to "give" a portion of the time spent in REM to NREM for mouse m. We estimate each w_m on only the in-sample mouse so that our evaluation remains entirely out of sample.

In Figure 11, we compare the N_i^j from EEG-/EMG-based manual scoring to the estimated $\widehat{N}_i^j$ from our video-based methodology. We examine the match for both the two-state task of classifying SLEEP versus WAKE (left panel) and the three-state task of classifying NREM versus REM versus WAKE (right panel). For both tasks, our estimated number of minutes spent in each state provides an excellent match to the number based on manual scores.

5. DISCUSSION

Our proposed methodology enhances standard statistical learning methods such as logistic regression and RF to account for time dependence. This approach is very general in that it can be used to enhance any statistical learning method that produces conditional class probability estimates. Our embedding methodology for Markov models makes a very rich and general class of dependence structures computationally feasible and easily estimable; although we have focused on first-order Markov models, generalized Markov models, and TDGMMs, our strategy can also accommodate higher-order Markov models, variable length Markov models, and other related models.

When applied to the classification of sleep states in mice based on video data, our procedure shows increased ability to discriminate NREM from REM sleep relative to alternative classification methods. This is a considerable advance over current approaches used by sleep scientists which can only discriminate SLEEP from WAKE (Flores et al. 2007; Pack et al. 2007). Furthermore, our procedure also allows for very accurate assessment of the total amount of NREM sleep, REM sleep, and wakefulness.

This is an improvement over the current approach to the assessment of this phenotype in mice which requires (i) surgery following anesthesia for the insertion of small screws in the scalp for the assessment of the electroencephalogram; (ii) insertion of a wire in the neck muscles for assessment of the electromyogram; (iii) recovery from surgery for several days; and (iv) assessment of stages of sleep or wakefulness by direct visual inspection and scoring of 8640 10-sec epochs per mouse per day. This is expensive, invasive, and time consuming thereby limiting the number of mice that can be studied. Thus, our approach has particular benefit when a large number of mice need to be studied for reasons of, for example, statistical power.

High-throughput automated scoring has several immediate applications. First, in studies in which mRNA changes or protein changes with sleep and wakefulness are being assessed, our approach is much more cost-effective for estimating sleep states. Second, there are several new mouse resources for studying the

genetic basis of behaviors such as sleep including (i) the panel of Collaborative Cross mice which consists of 300 lines of mice derived from eight founder strains (Churchill et al. 2004; Chesler et al. 2008; Collaborative Cross Consortium 2012); (ii) the panel of Diversity Outbred mice which is based on random mating of Collaborative Cross mice to provide genetic heterogeneity (Svenson et al. 2012); and (iii) other large panels of mice with individual genes knockout [e.g., Skarnes et al. (2011)]. Our methods could be applied to these large numbers of genetically heterogeneous mice to assess the genetic basis of various sleep behaviors. Third, high-throughput scoring has advantages in screening libraries of compounds to detect novel drugs that alter sleep and wake. Indeed, high-throughput scoring based on video data has already been employed in zebrafish to identify such compounds (Rihel et al. 2009) and the methodology developed here for mice facilitates this approach to mammalian species.

In many of aforementioned applications, the focus is on obtaining accurate estimates of aggregate measures of sleep as opposed to epoch-by-epoch scores. In addition to metrics such as the amount of time spent in each state at various points throughout the day, scientists are also interested in estimating quantities such as the number and typical length of bouts of each state to understand how they vary across the day (Saper, Scammell, and Lu 2005), particularly for mice that differ in terms of (i) genetic background (Mochizuki et al. 2011; Sehgal and Mignot 2011; Naidoo et al. 2012), (ii) age (Naidoo et al. 2008; Hasan et al. 2012), (iii) level of sleep deprivation (Franken, Malafosse, and Tafti 1999), and (iv) exposure to various compounds (Vienne et al. 2010). Our proposed methodology shows the ability to provide accurate assessments of these aggregate quantities in a high-throughput setting (McShane et al. 2012).

In addition to these immediate applications, sleep science suggests several potential future applications and enhancements of our proposed methodology. First, a three-dimensional recording of the mouse using high-definition video cameras as well as side-angle cameras in addition to overhead cameras would likely improve the assessment of mouse behavior; in particular, such a system could likely track breathing as well as the small twitches that occur in REM sleep, thus yielding improved differentiation of NREM sleep versus REM sleep on an epoch-by-epoch basis. Second, augmenting or substituting video data with piezoelectric data (i.e., sensors located in the floor of the mouse cage that measure pressure changes produced by movement of the mouse) could yield improved scores as there are (i) highly variable signals during wakefulness as the mouse moves around the cage and (ii) subtle signals during sleep and its substages that reflect breathing patterns (Friedman et al. 2004). Augmenting our model with covariates from both of these sources could potentially allow for the wholesale replacement of manual scoring as such covariates may be sufficiently precise to allow model-based error rates to approach the interhuman level of $\approx 5\%$.

In conclusion, video-based analyses show the ability to distinguish REM from NREM sleep in mice. While future elaborations of this technological approach could lead to further improvements in these estimates, high-throughput phenotyping of sleep in mice is feasible and will facilitate genetic studies and the investigation of chemical libraries to determine compounds that affect NREM, REM, and WAKE.

APPENDIX: DISCRIMINITIVE MARKOV ALGORITHMS

In Section 2.3, we introduced a discriminative approach to estimating a Markov model. Here we modify the forward–backward and Viterbi algorithms to accommodate our discriminative parameter set $\widehat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\pi}}, \hat{\mathbf{A}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{p}})$ rather than generative parameter set $\widehat{\boldsymbol{\Delta}} = (\hat{\boldsymbol{\pi}}, \hat{\mathbf{A}}, \hat{\boldsymbol{\mu}})$. We can use the results of the modified forward–backward algorithms (Algorithms 1 and 2) to estimate the conditional class probabilities $\widehat{\gamma}_t(i) = \mathbb{P}(Y_t = i|\mathbf{X}, \widehat{\boldsymbol{\Theta}})$ for each $i \in S$. In particular,

$$\widehat{\gamma}_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\mathbb{P}(\mathbf{X}_{1:T} = \mathbf{X}|\widehat{\boldsymbol{\Theta}})} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{k} \alpha_t(j)\beta_t(j)}.$$

As with conventional generative Markov models, our maximum likelihood estimate of Y_t is $\widehat{Y}_t = \arg\max_{i \in S} \widehat{\gamma}_t(i)$ and our maximum likelihood estimate of the sequence $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$ is given by the Viterbi Algorithm (Algorithm 3).

Algorithm 1 The Discriminative Forward Algorithm.

Begin with estimates $\widehat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\pi}}, \hat{\mathbf{A}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{p}})$ and define $\alpha_t(i) = \mathbb{P}(\mathbf{X}_{1:t}, Y_t = i|\widehat{\boldsymbol{\Theta}})$.

1. Initialization: $\alpha_1(i) = \hat{\pi}_i \frac{\hat{f}_i(\boldsymbol{x}_1)}{\hat{p}_i}$ for $i = 1, \ldots, k$.
2. Induction: $\alpha_{t+1}(j) = [\sum_{i=1}^{k} \alpha_t(i) \cdot \hat{a}_{i,j}] \frac{\hat{f}_j(\boldsymbol{x}_{t+1})}{\hat{p}_j}$ for $t = 1, \ldots, T-1$.
3. Termination: $\mathbb{P}(\mathbf{X}_{1:T} = \mathbf{X}|\widehat{\boldsymbol{\Theta}}) = \sum_{i=1}^{k} \alpha_T(i)$.

Algorithm 2 The Discriminative Backward Algorithm.

Begin with estimates $\widehat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\pi}}, \hat{\mathbf{A}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{p}})$ and define $\beta_t(i) = \mathbb{P}(\mathbf{X}_{t+1:T}|Y_t = i, \widehat{\boldsymbol{\Theta}})$.

1. Initialization: $\beta_T(i) = 1$ for $i = 1, \ldots, k$.
2. Induction: $\beta_t(i) = \sum_{j=1}^{k} \hat{a}_{i,j}\beta_{t+1}(j) \frac{\hat{f}_j(\boldsymbol{x}_{t+1})}{\hat{p}_j}$ for $t = T-1, \ldots, 1$

Algorithm 3 The Discriminative Viterbi Algorithm.

Begin with estimates $\widehat{\boldsymbol{\Theta}} = (\hat{\boldsymbol{\pi}}, \hat{\mathbf{A}}, \hat{\boldsymbol{f}}, \hat{\boldsymbol{p}})$ and define $\delta_t(i) = \max_{Y_{1:t-1}} \mathbb{P}(Y_{1:t-1}, Y_t = i, \mathbf{X}_{1:t}|\widehat{\boldsymbol{\Theta}})$.

1. Initialization: $\delta_1(i) = \hat{\pi}_i \frac{\hat{f}_i(\boldsymbol{x}_1)}{\hat{p}_i}$ and $\psi_1(i) = 0$ for $i = 1, \ldots, k$
2. Recursion:

$$\delta_t(j) = \max_{i=1,\ldots,k}[\delta_{t-1}(i)\hat{a}_{i,j}] \frac{\hat{f}_j(\boldsymbol{x}_t)}{\hat{p}_j}$$
$$t = 2, \ldots, T;\ j = 1, \ldots, k$$
$$\psi_t(j) = \operatorname{argmax}_{i=1,\ldots,k}[\delta_{t-1}(i)\hat{a}_{i,j}]$$
$$t = 2, \ldots, T;\ j = 1, \ldots, k.$$

3. Termination: $P^\star = \max_{i=1,\ldots,k} \delta_T(i)$ and $Y_T^\star = \operatorname{argmax}_{i=1,\ldots,k} \delta_T(i)$
4. Path (state sequence) backtracking: $Y_t^\star = \psi_{t+1}(Y_{t+1}^\star)$ for $t = T-1, \ldots, 1$.

SUPPLEMENTARY MATERIALS

Section 1: Generating Data from Markov Models
Section 2: Estimating Duration Distributions
Additional figures.

[Received August 2011. Revised May 2012.]

REFERENCES

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140. [1158]

——— (2001), "Random Forests," *Machine Learning*, 45, 5–32. [1148]

Buhlmann, P., and Wynser, A. J. (1999), "Variable Length Markov Chains," *The Annals of Statistics*, 27, 480–513. [1149]

Buja, A., Stuetzle, W., and Shen, Y. (2005), "Loss Functions for Binary Probability Estimation and Classification: Structure and Applications," available at *http://www-stat.wharton.upenn.edu/buja/PAPERS/paper-proper-scoring.pdf (Under review)*. [1155]

Chesler, E. J., Miller, D. R., Branstetter, L. R., Galloway, L. D., Jackson, B. L., Philip, V. M., Voy, B. H., Culiat, C. T., Threadgill, D. W., Williams, R. W., Churchill, G. A., Johnson, D. K., and Manly, K. F. (2008), "The Collaborative Cross at Oak Ridge National Laboratory: Developing a Powerful Resource for Systems Genetics," *Mammalian Genome*, 19, 382–389. [1147,1160]

Churchill, G. A., Airey, D. C., Allayee, H., Angel, J. M., Attie, A. D., Beatty, J., Beavis, W. D., Belknap, J. K., Bennett, B., Berrettini, W., Bleich, A., Bogue, M., Broman, K. W., Buck, K. J., Buckler, E., Burmeister, M., Chesler, E. J., Cheverud, J. M., Clapcote, S., Cook, M. N., Cox, R. D., Crabbe, J. C., Crusio, W. E., Darvasi, A., Deschepper, C. F., Doerge, R. W., Farber, C. R., Forejt, J., Gaile, D., Garlow, S. J., Geiger, H., Gershenfeld, H., Gordon, T., and Gu, J., Gu, W., de Haan, G., Hayes, N. L., Heller, C., Himmelbauer, H., Hitzemann, R., Hunter, K., Hsu, H. C., Iraqi, F. A., Ivandic, B., Jacob, H. J., Jansen, R. C., Jepsen, K. J., Johnson, D. K., Johnson, T. E., Kempermann, G., Kendziorski, C., Kotb, M., Kooy, R. F., Llamas, B., Lammert, F., Lassalle, J. M., Lowenstein, P. R., Lu, L., Lusis, A., Manly, K. F., Marcucio, R., Matthews, D., Medrano, J. F., Miller, D. R., Mittleman, G., Mock, B. A., Mogil, J. S., Montagutelli, X., Morahan, G., Morris, D. G., Mott, R., Nadeau, J. H., Nagase, H., Nowakowski, R. S., O'Hara, B. F., Osadchuk, A. V., Page, G. P., Paigen, B., Paigen, K., Palmer, A. A., Pan, H. J., Peltonen-Palotie, L., Peirce, J., Pomp, D., Pravenec, M., Prows, D. R., Qi, Z., Reeves, R. H., Roder, J., Rosen, G. D., Schadt, E. E., Schalkwyk, L. C., Seltzer, Z., Shimomura, K., Shou, S., Sillanpaa, M. J., Siracusa, L. D., Snoeck, H. W., Spearow, J. L., Svenson, K., Tarantino, L. M., Threadgill, D., Toth, L. A., Valdar, W., de Villena, F. P., Warden, C., Whatley, S., Williams, R. W., Wiltshire, T., Yi, N., Zhang, D., Zhang, M., and Zou, F. (2004), "The Collaborative Cross, a Community Resource for the Genetic Analysis of Complex Traits," *Nature Genetics*, 36, 1133–1137. [1147,1160]

Collaborative Cross Consortium. (2012), "The Genome Architecture of the Collaborative Cross Mouse Genetic Reference Population," *Genetics*, 190, 389–491. [1160]

Dietterich, T. G., Ashenfelter, A., and Bulatov, Y. (2004), "Training Conditional Random Fields via Gradient Tree Boosting," *Proceedings of the 21st International Conference on Machine Learning (ICML 2004)*, pp. 217–224. [1158]

Djuric, P. M., and Chun, J.-H. (2002), "An MCMC Sampling Approach to Estimation of Nonstational Hidden Markov Models," *IEEE Transactions on Signal Processing*, 50, 1113–1123. [1149]

Ferguson, J. D. (1980), "Variable Duration Models for Speech," in *Proceedings of Symposium on the Application of Hidden Markov Models to Text and Speech*, pp. 143–179. [1149]

Flores, A., Flores, J., Deshpande, H., Picazo, J., Xie, X., Franken, P., Heller, H., Grahn, D., and O'Hare, B. (2007), "Pattern Recognition of Sleep in Rodents Using Piezoelectric Signals Generated by Gross Body Movements," *IEEE Transactions on Biomedical Engineering*, 54, 225–233. [1147,1159]

Franken, P., Chollet, D., and Tafti, M. (2001), "The Homeostatic Regulation of Sleep Need is Under Genetic Control," *Journal of Neuroscience*, 21, 2610–2621. [1147]

Franken, P., Malafosse, A., and Tafti, M. (1999), "Genetic Determinants of Sleep Regulation in Inbred Mice," *Sleep*, 22, 155–169. [1160]

Freund, Y., and Schapire, R. (1996), "Experiments With a New Boosting Algorithm," in *Machine Learning: Proceedings of the Thirteenth International Conference*, pp. 148–156. [1148,1158]

Friedman, J., Hastie, T., and Tibshirani, R. (2000), "Additive Logistic Regression: A Statistical View of Boosting," *The Annals of Statistics*, 28, 337–374. [1158]

Friedman, L., Haines, A., Klann, K., Gallaugher, L., Salibra, L., Han, F., and Strohl, K. P. (2004), "Ventilatory Behavior During Sleep Among A/J and C57BL/6J Mouse Strains," *Journal of Applied Physiology*, 97, 1787–1795. [1160]

Fujiki, N., Cheng, T., Yoshino, F., and Nishino, S. (2009), "Specificity of Direct Transition From Wake to REM Sleep in Orexin/Ataxin-3 Transgenic Narcoleptic Mice," *Experimental Neurology*, 217, 46–54. [1156]

Guan, C., Ye, C., Yand, X., and Gao, J. (2010), "A Review of Current Large-Scale Mouse Knockout Efforts," *Genesis*, 48, 73–85. [1147]

Hasan, S., Dauvilliers, Y., Mongrain, V., Franken, P., and Tafti, M. (2012), "Age-Related Changes in Sleep in Inbred Mice Are Genotype Dependent," *Neurobiology of Aging*, 33, 195.e13–195.e26. [1160]

He, Y., Jones, C. R., Fujiki, N., Xu, Y., Guo, B., Holder, J. L. Jr, Rossner, M. J., Nishino, S., and Fu, Y. H. (2009), "The Transcriptional Repressor DEC2 Regulates Sleep Length in Mammals," *Science*, 325, 866–870. [1147]

Heath, A. C., Kendler, K. S., Eaves, L. J., and Martin, N. G. (1990), "Evidence for Genetic Influences on Sleep Disturbance and Sleep Pattern in Twins," *Sleep*, 13, 318–335. [1147]

Janssen, J., and Limnios, N. (eds.) (1999), *Semi-Markov Models and Applications*, Dordrecht, The Netherlands: Kluwer Academic. [1149]

Lafferty, J., McCallum, A., and Pereira, F. (2001), "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*, pp. 282–289. [1151]

Levinson, S. E. (1986), "Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition," *Computer Speech and Language*, 1, 29–45. [1149]

McCallum, A. (2003), "Efficiently Inducing Features of Conditional Random Fields," in *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, pp. 403–410. [1158]

McShane, B. B., Galante, R. J., Biber, M. P., Jensen, S. T., Wyner, A. J., and Pack, A. I. (2012), "Assessing REM Sleep in Mice Using Video Data," *Sleep*, 35, 433–442. [1147,1160]

McShane, B. B., Galante, R. J., Jensen, S. T., Naidoo, N., Pack, A. I., and Wyner, A. (2010), "Characterization of the Bout Durations of Sleep and Wakefulness," *Journal of Neuroscience Methods*, 193, 321–333. [1149,1157]

Mease, D., and Wyner, A. (2008), "Evidence Contrary to the Statistical View of Boosting," *Journal of Machine Learning Research*, 9, 131–156. [1148]

Mease, D., Wyner, A., and Buja, A. (2007), "Boosted Classification Trees and Class Probability/Quantile Estimation," *Journal of Machine Learning Research*, 8, 409–439. [1148]

Mochizuki, T., Arrigoni, E., Marcus, J. N., Clark, E. L., Yamamoto, M., Honer, M., Borroni, E., Lowell, B. B., Elmquist, J. K., and Scammell, T. E. (2011), "Orexin Receptor 2 Expression in the Posterior Hypothalamus Rescues Sleepiness in Narcoleptic Mice," *Proceedings of the National Academy of Sciences*, 108, 4471–4476. [1160]

Naidoo, N., Ferber, M., Galante, R. J., McShane, B. B., Hu, J. H., Zimmerman, J., Maislin, G., Cater, J., Wyner, A. J., Worley, P., and Pack, A. I. (2012), "Role of Homer Proteins in the Maintenance of Sleep-Wake States," *PLoS One*, 7, e35174. doi: 10.1371/journal.pone.0035174. [1160]

Naidoo, N., Ferber, M., Master, M., Zhu, Y., and Pack, A. I. (2008), "Aging Impairs the Unfolded Protein Response to Sleep Deprivation and Leads to Proapoptotic Signaling," *Journal of Neuroscience*, 28, 6539–6548. [1147,1160]

Pack, A. I., Galante, R. J., Maislin, G., Cater, J., Metaxas, D., Lu, S., Zhang, L., Smith, R. V., Kay, T., Lian, J., Svenson, K., and Peters, L. L. (2007), "Novel Method for High-Throughput Phenotyping of Sleep in Mice," *Physiological Genomics*, 28, 232–238. [1147,1157,1159]

Partinen, M., Kaprio, J., Koskenvuo, M., Putkonen, P., and Langinvainio, H. (1983), "Genetic and Environmental Determination of Human Sleep," *Sleep*, 6, 179–185. [1147]

Patlak, M. (2005), *Your Guide to Healthy Sleep (NIH Publication No. 06-5271)*, Bethesda, MD: U.S. Department of Health and Human Services. [1147]

Rabiner, L. R. (1989), "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, 77, 257–286. [1149,1150]

Rihel, J., Prober, D. A., Arvanites, A., Lam, K., Zimmerman, S., Jang, S., Haggarty, S. J., Kokel, D., Rubin, L. L., Peterson, R. T., and Schier, A. F. (2009), "Zebrafish Behavioral Profiling Links Drugs to Biological Targets and Rest/Wake Regulation," *Science*, 327, 348–351. [1160]

Saper, C. B., Scammell, T. E., and Lu, J. (2005), "Hypothalamic Regulation of Sleep and Circadian Rhythms," *Nature*, 437, 1257–1263. [1160]

Savage, L. J. (1971), "Elicitation of Personal Probabilities and Expecations," *Journal of the American Statistical Association*, 66, 783–801. [1155]

Sehgal, A., and Mignot, E. (2011), "Genetics of Sleep and Sleep Disorders," *Cell*, 146, 194–207. [1160]

Sin, B., and Kim, J. H. (1995), "Nonstationary Hidden Markov Model," *Signal Processing*, 46, 31–46. [1149]

Skarnes, W. C., Rosen, B., West, A. P., Koutsourakis, M., Bushell, W., Iyer, V., Mujica, A. O., Thomas, M., Harrow, J., Cox, T., Jackson, D., Severin, J., Biggs, P., Fu, J., Nefedov, M., de Jong, P. J., Stewart, A. F., and Bradley, A. (2011), "A Conditional Knockout Resource for the Genome-Wide Study of Mouse Gene Function," *Nature*, 474, 337–342. [1160]

Smyth, P. (1994), "Markov Monitoring With Unknown States," *IEEE Journal of Selected Areas in Communications, Special Issue on Intelligent Signal Processing for Communications*, 12, 1600–1612. [1149,1151]

Steriade, B. M. (2005), "Brain Electrical Activity and Sensory Processing During Waking and Sleep States," in *Principles and Practice of Sleep Medicine* (4th ed.), eds. M. H. Kryger, T. Roth, and W. C. Dement, Philadelphia, PA: Elsevier Saunders, pp. 101–119. [1148]

Stickgold, R., Hobson, J., Fosse, R., and Fosse, M. (2001), "Sleep, Learning, and Dreams: Off-Line Memory Reprocessing," *Science*, 294, 1052–1057. [1147]

Svenson, K. L., Gatti, D. M., Valdar, W., Welsh, C. E., Cheng, R., Chesler, E. J., Palmer, A. A., McMillan, L., and Churchill, G. A. (2012), "High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population," *Genetics*, 190, 437–448. [1160]

Vaseghi, S. V. (1995), "State Duration Modeling in Hidden Markov Models," *Signal Processing*, 41, 31–41. [1149]

Vienne, J., Bettler, B., Franken, P., and Tafti, M. (2010), "Differential Effects of GABAB Receptor Subtypes, Gamma-Hydroxybutyric Acid, and Baclofen on EEG Activity and Sleep Regulation," *Journal of Neuroscience*, 30, 14194–14204. [1160]

Vink, J. M., Groot, A. S., Kerkhof, G. A., and Boomsma, D. I. (2001), "Genetic Analysis of Morningness and Eveningness," *Chronobiology International*, 18, 809–822. [1147]

Comment

Kerby SHEDDEN

The article "Modeling Time Series Dependence for Sleep Scoring in Mice," henceforth MTD, provides us with the opportunity to contrast two approaches to prediction. On the one hand, we can model the joint distribution $P(\mathbf{Y}_{1:T}, \mathbf{X}_{1:T})$ of the observed data in detail, motivated by the fact that the Bayes' rule derived from P is optimal for making predictions. This approach is exemplified by the use in MTD of estimated conditional distributions $\hat{P}(Y_t = y|\mathbf{X}_{1:T})$ for prediction of Y_t from the observable $\mathbf{X}_{1:T}$, where $\hat{P}$ is an estimate of P. On the other hand, we can frame the problem directly in predictive terms, expressing Y_t as the response variable in a regression model. This is exemplified by the use in MTD of multinomial logistic regression (MLR) to capture the conditional relationship of Y_t given X_t.

There is a strong rationale for estimating the Bayes' rule via an estimate of the joint distribution P. However, many aspects of P may be difficult to estimate, and could have little influence on prediction. Regression approaches focus directly on the aspects of P that are most relevant for prediction, and hence can manage the impact of estimation variance on predictive performance. In MTD, the joint model for $\mathbf{Y}_{1:T}$ and $\mathbf{X}_{1:T}$ is used to allow for more flexible transition behavior between the sleep states. However, this flexibility is only relevant for prediction to the extent that it impacts the estimated prediction function $\hat{P}(Y_t = y|\mathbf{X}_{1:T})$. Alternatively, there are ways to directly increase the flexibility of a regression relationship, such as by expanding the predictor variables using basis functions. We will explore this avenue below.

In MTD, reconstruction of Y_t using the estimated Bayes' rule obtained from $\hat{P}$ is shown to substantially outperform logistic regression in simulation studies, but is only slightly advantageous, at best, in the real data example. This led us to seek to better understand the role of model specification in settings such as this where the primary goal is prediction.

1. WINDOWED MLR

The MLR approach used in MTD simply regresses Y_t on X_t. As noted in MTD, this ignores potentially useful information in the neighboring X_s values. One way to exploit this information while remaining within the MLR framework is to regress Y_t on a window of X_s values containing X_t. We assessed the performance of this "windowed MLR" using as predictors the symmetric window $X_{t-w}, \ldots, X_{t+w}$, containing $2w + 1$ consecutive X_s values. The training and validation data $\mathbf{Y}_{1:T}$, $\mathbf{X}_{1:T}$ were simulated according to the model described in Section 3 of MTD using $T = 100$ and 1000, with values of σ ranging from 0.5 to 1.5.

Our findings for $T = 1000$ are shown in Figure 1. The left panel shows the windowed MLR error rates, and the right panel shows the error rates relative to the Bayes' error rate. The error rates drop substantially when using $w > 0$ compared to $w = 0$, showing the importance of exploiting serial dependence. For example, in the high signal-to-noise setting $\sigma = 0.5$, the error rate drops from 0.22 when using $w = 0$ (consistent with MTD Figure 8) to 0.14 when using $w = 5$. Figure 8 of MTD shows that the model-based error rate for this setting is around 0.08. Thus, more than half of the distance to the model-based methods is recovered by using the windowed MLR instead of the simple MLR with $w = 0$. In the moderate signal-to-noise setting $\sigma = 1.5$, windowed MLR achieves an error rate of roughly 0.39, comparable to the error rate of roughly 0.37 achieved by the MLR+TDGMM method. It is not surprising that the advantage of using the correct data-generating model for prediction (i.e., the MLR+TDGMM method) is greater when the signal is stronger.

We carried out similar studies for $T = 100$. When $\sigma = 0.5$, MLR using $w = 0$ gives an error rate of 0.24 (consistent with MTD figure 8), but this can be reduced to 0.19 using $w = 1$. The model-based approaches of MTD produce error rates between 0.11 and 0.14, depending on the model (MTD Figure 8).

To further understand the windowed MLR approach, we examined the coefficient estimates for the case $w = 6$. The MLR approach fits the model $P(Y = k|X = x) \propto \exp(\beta_k' x)$ for $k = 1, 2, 3$, where $\beta_1 \equiv 0$ for identification. Here, X represents a window of 13 values centered around the value that is contemporaneous with Y, along with an intercept. This can be viewed as a two index model that captures the conditional probabilities of Y_t in terms of linear predictors $\beta_j' \tilde{X}_t$, $j = 2, 3$, where $\tilde{X}_t$ is the window of $2w + 1 = 13$ values of X_s centered on X_t. We first estimated $\beta_j^* \equiv E\hat{\beta}_j$, $j = 2, 3$ by averaging the $\hat{\beta}_j$ over 200 replicate samples obtained from the model described in Section 3 of MTD, with $T = 1000$. Since we wished to focus on the structure of the linear predictors, rather than the link to the

Kerby Shedden, Department of Statistics, University of Michigan, Ann Arbor, MI 48104, USA (E-mail: *kshedden@umich.edu*).

Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Applications and Case Studies
DOI: 10.1080/01621459.2013.836970

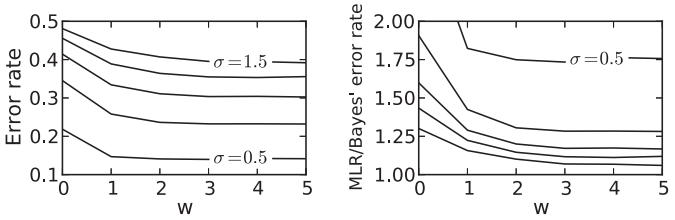

Figure 1. Left panel: Error rate of MLR using windows of width $2w + 1$ around X_t to predict Y_t. Right panel: Ratio of MLR error rate to Bayes' rule using $X_1, \ldots, X_T$. Curves corresponding to $\sigma = 0.5, 0.75, 1, 1.25$, and 1.5 are shown. The curves are strictly ordered according to σ.

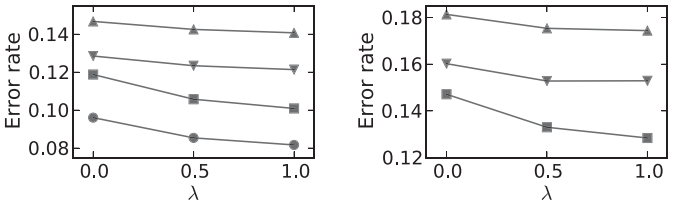

Figure 3. Error rates for varying degrees of heterogeneity in the dwell time distributions, determined by λ. The curves, from top to bottom, correspond to: windowed MLR with linear terms, windowed MLR with linear and quadratic terms, TDGMM, and the Bayes' rule. The left plot is based on data simulated according to the simulation model of MTD, the right panel is based on data with additional autoregressive dependence in the X_t.

probability scale, we then dropped the intercept and rescaled the remaining terms of β_2^* and β_3^* to have a maximum value of 1. The resulting weight vectors derived from β_2^* are shown in Figure 2. The weights derived from β_2^* and β_3^* (not shown) are nearly parallel, indicating that although given the option to use two distinct indices, MLR only finds use for a single index. The weights strongly resemble an exponential function centered at the origin (Figure 2, left panel). The exponential form of the weights is further supported by plots of $\log \beta_j^*(k+1) - \log \beta_j^*(k)$ (figure 2, right panel), which would be constant if the form were exactly exponential. The windowed MLR predictions are thus seen to very nearly result from an exponentially weighted moving average (EWMA) of the X_t. The exponential weight function has a wider bandwidth as σ increases, which reflects that ability of MLR to exploit the bias/variance tradeoff.

Since the windowed MLR reduces to using an EWMA analysis in the setting considered here, it seems that information in the transition structure of the Markov chain is not fully exploited. Using a windowed MLR or the equivalent EWMA approach imposes a homogeneous (in time) smoothing constraint on the predicted Y_t sequence. This smoothing substantially improves predictive performance over unwindowed MLR, and is adaptive, as indicated by the differing smoothing bandwidths seen in the left panel of Figure 2. However, it is unable to fully approximate the Bayes' rule, as reflected in its somewhat inferior performance relative to the model-based approaches developed in MTD. The cost of this bias will be greatest when the signal is strong relative to the noise.

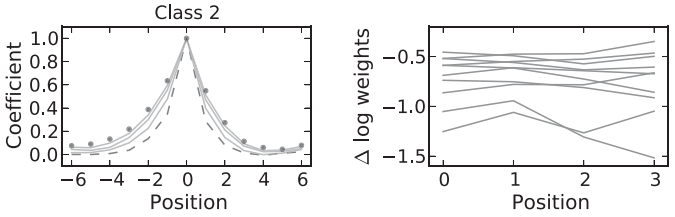

Figure 2. Rescaled versions of β_2^* (left) and the first differences of $\log \beta_2^*$ (right). The order of the differencing is reversed to the left of the mode. The dotted sequence corresponds to $\sigma = 1.5$, the broken line corresponds to $\sigma = 0.5$, and three intermediate values of σ are shown in solid gray lines.

2. IMPACT OF UNEQUAL DWELL TIME DISTRIBUTIONS

The simulation model presented in Section 3 of MTD has strong heterogeneity in the dwell time distributions, as depicted in their Figure 7. To explore the impact of the dwell time distributions on prediction performance, we carried out simulation studies using three related models. We first averaged the six dwell time distributions δ_{jk} $(j \neq k)$ in MTD to form a single distribution δ_A. We then formed dwell time distributions $\tilde{\delta}_{jk} = \lambda\delta_{jk} + (1-\lambda)\delta_A$ using $\lambda = 0, 1/2$, and 1. Note that when $\lambda = 1$, we have the dwell time distributions of MTD; and when $\lambda = 0$, the six dwell time distributions are identical. The TDGMM estimation procedure was implemented as described in MTD, except for the estimation of the dwell time distributions, which we estimated using the empirical frequencies of the observed dwell times.

Figure 3 shows the results of this analysis. In addition to the windowed MLR discussed above, we also considered MLR with quadratic terms $\tilde{X}_t(j)^2$ and $X_t \cdot \tilde{X}_t(j)$ in addition to the linear terms. We see when the dwell time distributions are heterogeneous ($\lambda = 1$), the error rate of the model-based approach is around 0.04 units lower than that of windowed MLR, and about 0.02 units lower than that of the windowed MLR with quadratic terms. When the dwell time distributions are identical ($\lambda = 0$), the performance differences narrow, and the error rate for windowed MLR with quadratic terms is less than 0.01 and greater than the error rate for the model-based approach.

A final consideration is that the various models developed in MTD are simplistic in that the observable values X_t are treated as being conditionally independent given the Y_t. To explore this, we generated the X_t as an autoregressive Gaussian process with autocorrelation 0.3, and with the same mean and variance structure as the simulation model from Section 3 of MTD. The prediction performance is summarized in the right panel of Figure 3 (model-based estimates are based on the misspecified TDGMM, and the Bayes' error rate is not shown since it is difficult to calculate). All prediction error rates increase relative to the case with independent errors, but the model-based analysis continues to slightly outperform the regression approach.

Comment

Donglin ZENG and Yuanjia WANG

This is an interesting article which proposes a novel time-series model to predict the rare and important state of REM sleep in mice. The article compares different alternatives in modeling time-dependence including the first-order Markov model, the discriminative Markov model, and the transition-independent or dependent Markov models. The last is shown to yield the best prediction performance in identifying the REM state.

The state-manifest covariate X_t contains six continuous covariates for each epoch: the within-epoch mean and standard deviation of the velocity, aspect ratio, and size of the mouse. The proposed methods impose two key assumptions: (1) X_t is independent of $Y_{t-1}, X_{t-1}, \ldots, Y_1, X_1$ given Y_t; (2) Y_t does not dependent on covariates before time t given states $Y_1, \ldots, Y_{t_1}$. These assumptions lead to a simple and efficient backward-and-forward algorithm to update the prediction probabilities. However, the validity of these assumptions may need to be justified. For example, the conditional independent in assumption (1) does not apply to time-independent covariates such as the size of the mouse and those covariates which may not be associated with Y_t but have serial correlations. The second assumption can be violated in the situation when the mouse baseline covariates have some long-term effects on sleeping states. One possible way to check assumption (1) empirically may be to perform simple regression of X_t on $(Y_t, Y_{t-1}, \ldots, Y_1, X_1)$ by testing whether all the regressors except Y_t have nonsignificant effects. Similar regression models can be used to check assumption (2) in the real data analysis.

In the estimation for the proposed methods, it seems to us that the article assumes stationary distribution of $P(Y_t = i|X_t)$ and $P(Y_t = i)$. A more efficient estimation procedure is to maximize the likelihood function by including constraint $p_i = \sum_j \pi_j a_{j,i}$. In the generalized Markov model and its transition-dependent version, durations spent on a particular state have been modeled. An alternative approach is to model a high transition probability from state i to itself if the number of the times in state i prior to this time point is less than M_i. In other words, the generalized Markov models try to incorporate the immediate state history information in modeling transition probabilities.

The true state labels are from two or three independent scores and the percent of disagreement is around 5%. One possible solution to account for the error in labeling gold states is to model the mis-specification of the true latent label. For example, let D_t be the true state then we model the scored label Y_t (multivariate if from multiple scorers) given D_t. Then the likelihood function with D_t known is

$$\pi_{D_1} P(Y_1|D_1) \prod_t \{P(Y_t|D_t)P(D_t|D_{t-1})P(D_t|X_t)/P_{D_t}\}.$$

Treating D_t as the missing data, one may call for the EM algorithm or MCMC algorithm for estimation.

Donglin Zeng is Professor, Department of Biostatistics, University of North Carolina, CB#7420, Chapel Hill, NC 27514-7420, USA. (E-mail: *dzeng@email.unc.edu*). Yuanjia Wang is Associate Professor, Department of Biostatistics, Mailman School of Public Health, Columbia University, 722 W. 168th St., New York, NY 10032, USA. (E-mail: *yw2016@columbia.edu*).

Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Applications and Case Studies
DOI: 10.1080/01621459.2013.836971

Rejoinder

Blakeley B. McShane, Shane T. Jensen, Allan I. Pack, and Abraham J. Wyner

We warmly thank editors Hal Stern and Joseph Ibrahim for selecting our article (McShane et al. 2013) for discussion. We are grateful for the opportunity to receive feedback on our work from discussants who possess a tremendous breadth of knowledge and expertise and we thank them for the great deal of time and effort they put into contemplating and responding to our article. Their careful and considered comments serve not only to further elucidate our findings but also to educe additional research questions. It is thus our hope that our humble article and the ensuing discussion will serve as a springboard for us and for other scholars.

In this rejoinder, we aim to do three things. First, we introduce a new simulation that is based on our mouse data. This new simulation, motivated by the discussion, helps us achieve our second aim, namely providing an in-depth response to each of the discussants. Finally, we introduce some additional findings that shed further light on model performance.

In the text that follows, we abbreviate the two discussions as KS (Shedden 2013) and ZW (Zeng and Wang 2013).

1. MOUSE SIMULATION

The simple simulation of Section 3 of our article was the focus of the discussion by KS. In order both to respond to several of his noteworthy points and to support some additional findings of our own, we propose a more complicated simulation that is based on our mouse data and thus better reflects the key features of our applied setting. In particular, we use our mouse data to estimate the parameters of a transition-dependent generalized Markov model (TDGMM) and then simulate data from a TDGMM conditional on these parameter values.

The mouse simulation state space $S = \{\text{NREM}, \text{REM}, \text{WAKE}\}$ is the mouse data state space, the initialization distribution $\boldsymbol{\pi}$ is the observed marginal distribution of the mouse data, the transition probability distributions $\mathbf{A}$ are the observed transition probabilities of the mouse data, and the transition-dependent duration distributions $\boldsymbol{\delta}$ are beta-negative binomial with geometric tail fit to the observed mouse data and plotted in Figure 1 (Q–Q plots showing the fit of the estimated duration distributions to the mouse data appear in Figure 4 of the online supplementary materials of our article). The covariate emission distributions $\boldsymbol{\mu}$ are multivariate normal with state-specific means and a common covariance matrix; this choice of distribution results in a linear decision boundary and is fit to the observed mouse data for the six continuous covariates omitting, for obvious reasons, the powerful binary covariate that indicated whether or not the light in the mouse cage was on in epoch t. Full details of simulation parameters are provided in the Appendix.

Our study uses three different training set sizes ($T = 2160$, $T = 8640$, and $T = 34{,}560$ which are, respectively, one-fourth of the actual number of epochs observed for a given mouse, the actual number of epochs observed for a given mouse, and four times the actual number of epochs observed for a given mouse). The test set size is always fixed at $T^\star = 200$ (our results are not sensitive to this choice) and the test data "continue" from the training data as in Figure 4 of our article. All results are averaged over 1000 replicates of the simulation.

As in our article, we evaluate model performance in three ways: classification error, classification error relative to the Bayes' Rule, and the root mean square error of the probability estimates.

2. RESPONSE TO KS

KS notes that aspects of the joint distribution $\mathbb{P}(\boldsymbol{Y}_{1:T}, \mathbf{X}_{1:T})$ may be difficult to estimate while having little influence on prediction. This fact is of critical importance and we are remiss for not having emphasized it sufficiently in our article. We thank KS for having called attention to it. Nonetheless, since in our principal model we take a discriminative rather than a generative approach, we believe we focus on the exact aspects of the distribution most relevant for prediction.

We found the windowed multinomial logistic regression (WMLR) approach proposed by KS intriguing and we were gratified to see that our model-based TDGMM approach stood up in the additional simulations conducted by him. In fact, WMLR was the first model we employed on our mouse data. This approach did not yield satisfactory predictive power thus motivating our first-order Markov model (1MM), generalized Markov model (GMM), and TDGMM approaches.

To examine the performance of the WMLR approach relative to alternative model choices, we compare the performance of several models on the mouse simulation. The models we consider are (i) multinomial logistic regression (MLR), (ii) MLR enhanced by a 1MM (MLR+1MM), (iii) MLR enhanced by a TDGMM (MLR+TDGMM), and (iv) WMLR(w), that is, WMLR with window size w; we let w range from zero to five as in KS Figure 1 and note that WMLR(0) is simply MLR. The root

Blakeley B. McShane is Assistant Professor, Kellogg School of Management, Northwestern University, Evanston, IL 60611 (E-mail: *b-mcshane@kellogg.northwestern.edu*). Shane T. Jensen is Associate Professor, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *stjensen@wharton.upenn.edu*). Allan I. Pack is John Miclot Professor, Center for Sleep and Circadian Neurobiology, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *pack@mail.med.upenn.edu*). Abraham J. Wyner is Professor, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *ajw@wharton.upenn.edu*).

Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Applications and Case Studies
DOI: 10.1080/01621459.2013.844021

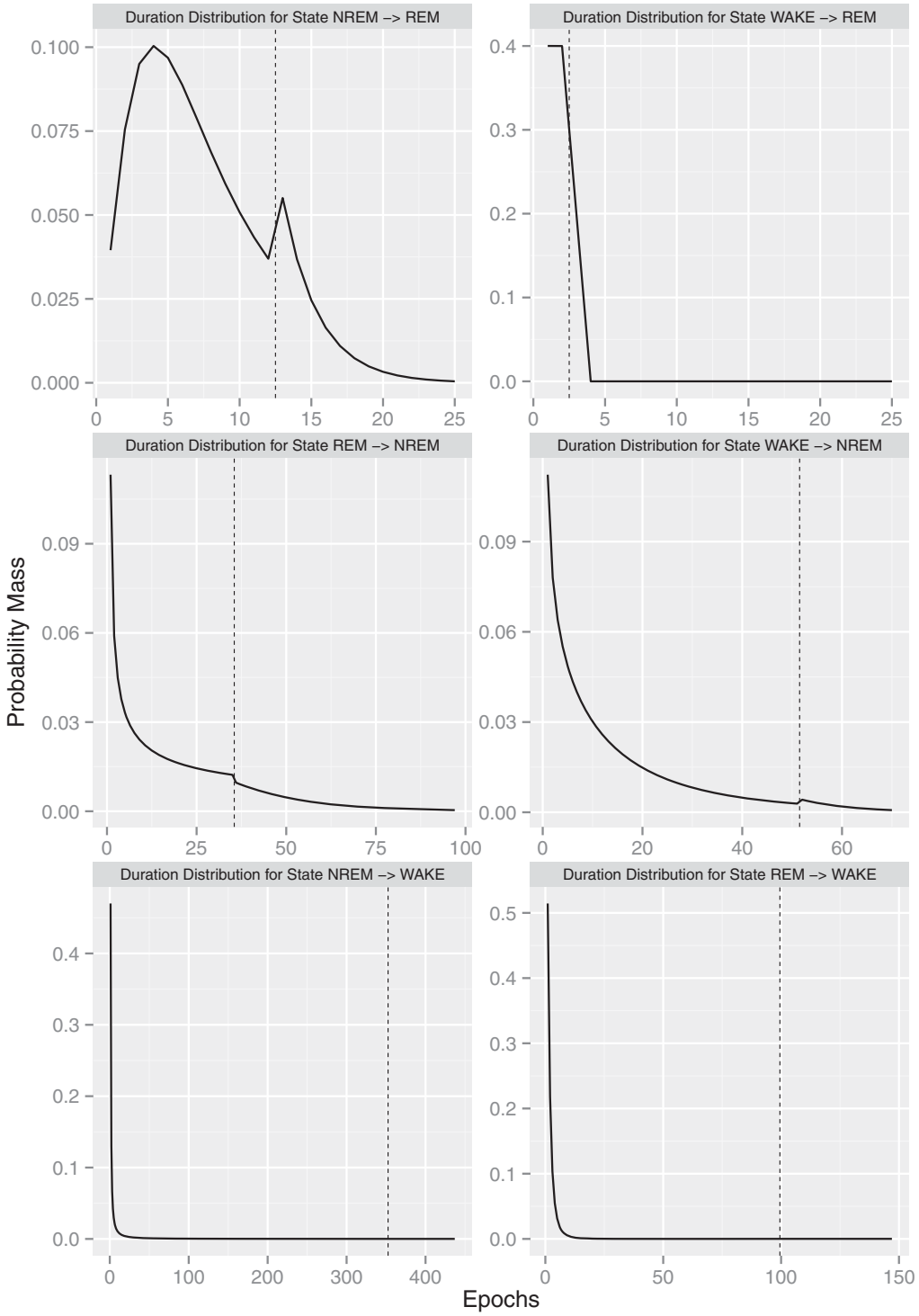

Figure 1. Transition-dependent duration distributions for the mouse simulation. We estimate a beta-negative binomial distribution with geometric tail for each conditional distribution using the procedure outlined in the online supplementary materials of our article. We plot distributions so that over 99% of the total mass appears in the plots and extend the plots so that 25 epochs at minimum appear on the x-axis. The dashed vertical lines separate the"head" and "tail" of the distributions.

mean square errors of the probability estimates (i.e., relative to the Bayes' Rule which uses the true probabilities, $\mathbb{P}(Y_t|\mathbf{X}^\star, \mathbf{\Delta})$) of each of the various models for each of the three training set sizes are plotted in Figure 2. As can be seen, the gains in performance for WMLR asymptote in w relatively quickly despite the longer-term patterns of time dependence indicated in Figure 1. Further, MLR+1MM dramatically outperforms WMLR—even for relatively high values of w; this is particularly notable because, while both are incorrectly specified, the latter (i) makes use of $(K-1)\cdot(1+p+2wp)$ coefficients, where $K=3$ is the size of the state space and $p=6$ is the number of covariates and (ii) can capture longer-term patterns of time dependence. Finally, it is clear that the MLR+TDGMM approach is dominant.

Abstracting from our data setting, we are concerned about the use of WMLR when either (i) there is long-term time-series

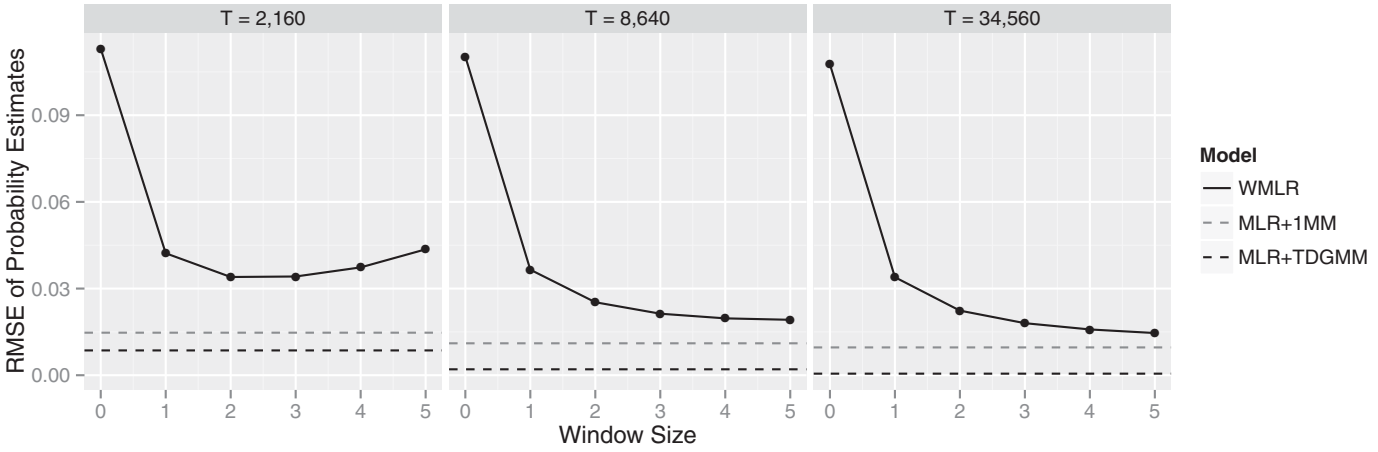

Figure 2. Root mean square error of probability estimates for the mouse simulation. MLR+TDGMM performs best while the WMLR approach asymptotes in w relatively quickly despite the rather longer-term patterns of time dependence in the simulated data. The results for classification error and classification error relative to the Bayes' Rule are qualitatively similar.

dependence in the response variable or (ii) p is large. In the case of the former, there is a risk that the prespecified window size w would not be large enough to capture the long-term dependence. Further, increasing w to sufficiently capture the long-term dependence increases the effective number of covariates rather dramatically to $(K-1)\cdot(1+p+2wp)$ and this increase is exacerbated when the size of the original covariate space p is already large. While the simulations of KS, adapted from the simple simulation of Section 3 of our article, are interesting, they do not address these particular concerns since, in these simulations, the time-series dependence in the response is relatively short-term and there is only $p = 1$ covariate.

Our concern about the effective size of the covariate space would be mitigated by KS's findings that (i) WMLR is nearly equivalent to an exponentially weighted moving average of the $\boldsymbol{X}_t$ (see KS Figure 2) and (ii) the β_k^* are parallel if both of these findings were empirically general across a wide variety of data settings. In such a case, WMLR would require the estimation of only $1+2p$ model parameters (i.e., an intercept and a coefficient and decay parameter for each covariate) rather than $(K-1)\cdot(1+p+2wp)$ model parameters. However, consider the coefficients from our mouse simulation normalized using the procedure outlined in KS and plotted in Figure 3 for $w = 6$ as in KS Figure 2. Clearly, many of the coefficients are not well-approximated by an exponentially weighted moving average. Further, they are not parallel either; indeed, the parallel coefficients found by KS are a direct consequence of the data generation process for the simple simulation (i.e., univariate normal covariate emission distributions with equally spaced state-specific means and common variance).

We appreciate the additional exploration of the duration (or dwell time) distributions provided by KS. It was gratifying to see that our model-based approach performed well in this setting. It would also have been interesting to see the performance of the GMM version of our model in the $\lambda = 0$ setting; while neither the TDGMM or GMM reflect the fact that the duration distributions are identical across all K states when $\lambda = 0$, the GMM would at least reflect the fact that the duration distributions are not transition-dependent.

When selecting (or estimating) δ_A, the common duration distribution in the KS simulation, we might recommend weighting the $\delta_{j,k}$ by the marginal frequencies of each conditional state rather than employing a straight average as in KS. We also caution that a large amount of data is necessary to obtain estimates of the duration distributions when using empirical frequencies as in KS; the parsimonious parametric approach employed in our article is more likely to perform better with little data.

Finally, we were intrigued by KS's final simulation which modified our simple simulation to include autoregressive errors in the observed covariate. We were pleased that the MLR+TDGMM outperformed both linear and quadratic WMLR even in this setting where it is misspecified. Further, we think it is important to note that, in our generative model, time-series dependence in Y_t can induce time-series dependence in the $\boldsymbol{X}_t$. In other words, although our model assumes that $\boldsymbol{X}_t$ is conditionally independent of the rest of the data $(\boldsymbol{Y}_{-t}, \mathbf{X}_{-t})$ given Y_t, the $\boldsymbol{X}_t$ considered unconditionally will undoubtedly be time-dependent. If, in a particular data setting, the time dependence in the $\boldsymbol{X}_t$ induced by the Y_t is not sufficient to capture the full extent of the time dependence in the $\boldsymbol{X}_t$, one could consider incorporating a WMLR within our TDGMM framework. While this might not be the most principled approach to capturing the "excess" autoregressive signal in the $\boldsymbol{X}_t$, the results presented in KS Figure 3 indicate that it could be promising.

3. RESPONSE TO ZW

We agree with ZW that our conditional independence assumption (i.e., that $\boldsymbol{X}_t$ is conditionally independent of the rest of the data $(\boldsymbol{Y}_{-t}, \mathbf{X}_{-t})$ given Y_t) would not be appropriate when either (i) X_t is not changing over time or (ii) X_t is unassociated with Y_t but has serial correlation. However, we should clarify that, in our application, each of the covariates contained in $\boldsymbol{X}_t$ does vary over time–including the size (area) of the mouse as suggested by the two video frames shown in Figure 9 of our article. With regards to a covariate that is not associated with Y_t, we wonder why such a covariate would be employed in a discriminative model designed to predict Y_t.

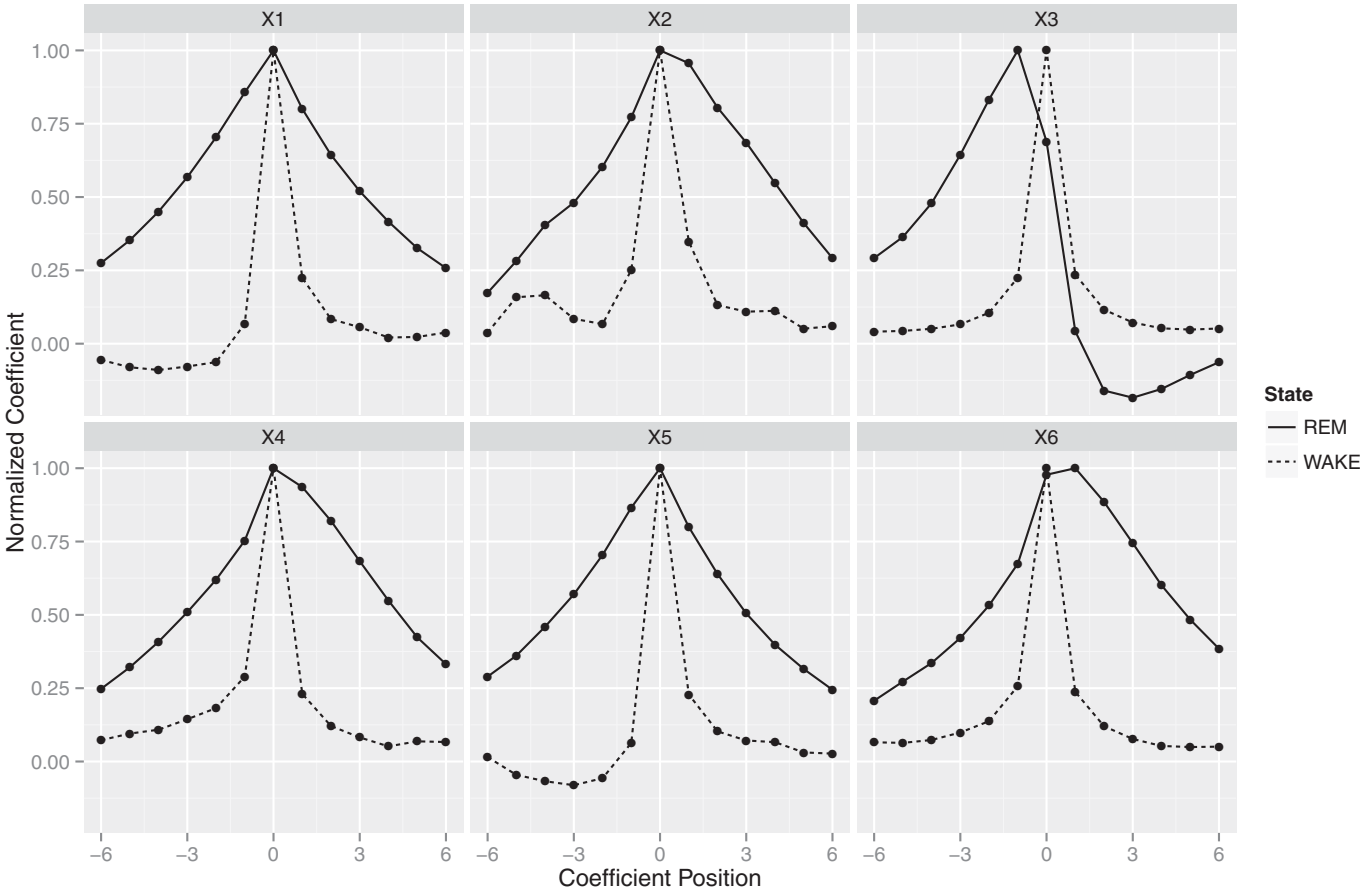

Figure 3. Normalized coefficients for the mouse simulation. The coefficients are normalized using the procedure outlined in KS and we set $\beta_{\text{NREM}} \equiv 0$ for identification. Many of the coefficients are not well-approximated by an exponentially weighted moving average and the coefficients for REM and WAKE are not parallel.

ZW propose a procedure to empirically evaluate the conditional independence assumption that $\boldsymbol{X}_t$ is conditionally independent of $(\boldsymbol{Y}_{-t}, \mathbf{X}_{-t})$ given Y_t by regressing $\boldsymbol{X}_t$ on $\boldsymbol{Y}_{1:(t-1)}$ and $\mathbf{X}_{1:(t-1)}$. However, we note that this procedure should also include $\boldsymbol{Y}_{(t+1):T}$ and $\mathbf{X}_{(t+1):T}$ as covariates to fully evaluate the conditional independence assumption. Further, in practice, this regression is likely difficult to implement given such a large covariate space and careful attention would need to be given to the issue of simultaneously testing so many covariates, especially given these covariates are likely to be highly collinear under our model. An alternative approach would be to consider a sliding window approach that regresses $\boldsymbol{X}_t$ on $\boldsymbol{Y}_{(t-w):(t+w)}$ and $\mathbf{X}_{(t-w):(t+w)}$; this approach nonetheless still suffers from having a large number of collinear covariates and further is useful only when the pattern of time dependence in the data is relatively short-term. ZW also suggest evaluating the conditional independence of Y_t and $\mathbf{X}_{-t}$ given $\boldsymbol{Y}_{-t}$ using a similar regression-based approach; we note this approach suffers from exactly the same issues as the approach suggested for the evaluation of the conditional independence assumption of $\boldsymbol{X}_t$.

ZW's alternative suggestion of using the local state history to model the transition probabilities (and, in particular, using "a high transition probability from state i to itself if the number of the times in state i prior to this time point is less than M_i") is intriguing. This approach, where the transition probabilities depend on not just the state but also on the duration of the state, is a particular form of a time-inhomogeneous Markov model known as a nonstationary Markov model (Djuric and Chun 2002) and it is equivalent to our GMM approach provided that the transition probabilities away from one state to a different state vary in the duration of the original state as a constant times one minus the duration-dependent self-transition probability of the original state.

ZW raise the issue of the disagreement between scorers. While this issue is entirely legitimate, we reiterate that, on epochs where the two scorers disagreed, a third scorer was brought in to break the tie; this strongly mitigates any concern about the accuracy of the classification for these epochs. While we agree with ZW's suggestion that a hidden Markov model (HMM) could be employed to infer the true underlying sleep state (modeling the two observed scores as a function of the true underlying sleep state), it is not clear the results from such a model would be particularly illuminating as the epochs on which the two scorers disagree are almost certainly going to have extremely high uncertainty under this HMM.

We thank ZW for the additional citations beyond those contained in our article that pertain to adaptive multiclass weighted learning procedures. While these procedures are useful in settings with multiple classes and/or rare or unbalanced state (such as our REM state), they unfortunately do not address the most pertinent aspects of our application (i.e., long-term time dependence in a noisy setting with high Bayes error).

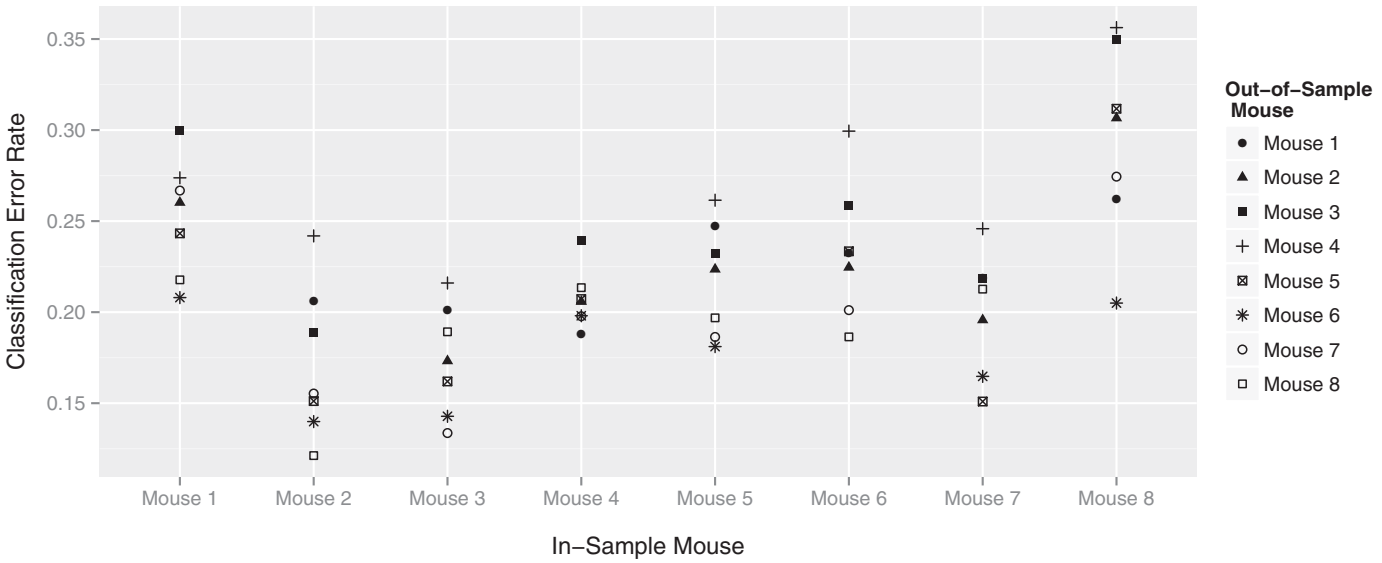

Figure 4. Classification errors by in-sample mouse and out-of-sample mouse for the mouse data. There is substantial heterogeneity in classification error rate both by in-sample mouse and by out-of-sample mouse and there are consistent patterns. The results for the rate of REM prediction, the REM false-positive rate, and the REM false-negative rate are qualitatively similar.

Finally, we completely agree with ZW that the issue of mouse-to-mouse variability merits additional attention. Returning to the data and fitting procedure of our article, recall that we evaluated our various models by taking the full set of data for one mouse (the "in-sample mouse") as our training data, testing on the full set of data from the other seven mice (the "out-of-sample mice"), and repeating this procedure over all combinations of in-sample mouse and out-of-sample mouse. Consequently, we can evaluate model performance for each pair of in-sample and out-of-sample mice and we do so for our classification error rate metric in Figure 4. As can be seen, there is substantial heterogeneity in classification error rate both by in-sample mouse and by out-of-sample mouse. Further, there are consistent patterns; for example, the sleep behavior of Mouse 6 appears comparably easy to classify regardless of the in-sample mouse while the sleep behavior of Mouse 4 appears comparably difficult. Qualitatively similar results hold for other metrics we examined including the rate of REM prediction, the REM false-positive rate, and the REM false-negative rate. Future work should clearly consider mouse-to-mouse variability and we have already made initial efforts in this direction by modeling the sleep behavior of each mouse using distinct parameters but pooling information across mice using a hierarchical structure.

4. ADDITIONAL FINDINGS

We have two additional findings that we demonstrate by returning to the simulation of Section 3 of our article. First, we return to our examination of out-of-sample prediction, but, rather than evaluating predictions averaged over all $T^\star = 200$ out-of-sample time points, we examine each time point individually. Second, we examine the performance of various "oracle-like" models.

The simulation of Section 3 of our article considered (i) 13 values of σ, the standard deviation of the covariate emission distributions, ranging from nearly zero to three and (ii) three values of T, the training set size, ranging from 100 to 10,000. Here, we focus on $\sigma \in \{0.25, 0.50, 0.75, 1.00\}$ and $T = 1000$. The marginal probability of each state, $\mathbb{P}(Y_t = i)$, is 0.288 for state a, 0.491 for state b, and 0.220 for state c; consequently, the classification error achieved by the model which always predicts the modal state (i.e., state b) is $1 - 0.491 = 0.509$. We can thus think of 0.509 as an upper bound on the classification error of a model. Further, we note that the Bayes' error (i.e., the classification error achieved by the model that uses the true probabilities, $\mathbb{P}(Y_t|\mathbf{X}^\star, \mathbf{\Delta})$) increases monotonically in σ taking on values 0.007, 0.077, 0.168, and 0.243, respectively. Similarly, the classification error of the model that uses the true conditional probabilities $\mathbb{P}(Y_t|\boldsymbol{X}_t^\star, \mathbf{\Delta})$ but ignores the time-series information in Y_t (i.e., by conditioning only on $\boldsymbol{X}_t$, the covariates at time t, rather than on $\mathbf{X}_t^\star$, the full set of covariates for all time periods) also increases monotonically in σ taking on values 0.032, 0.319, 0.338, and 0.403, respectively. Consequently, we can think of these four values of σ as varying the noise level from a relatively low level to a relatively high level.

In Figure 5, we plot the root mean square error of the probability estimates (i.e., relative to the Bayes' Rule which uses the true probabilities, $\mathbb{P}(Y_t|\mathbf{X}^\star, \mathbf{\Delta})$) at each out-of-sample time period for each of the four estimated models considered in Section 3 of our article; the root mean square error is taken over 1000 independent replicates of our simulation. Before discussing the results, we note that, while the signal in the data clearly attenuates as the out-of-sample time period increases (and thus, for example, the classification error of all models, except MLR which lacks a time-series component, increases), our evaluation is relative to the Bayes' Rule and thus model performance need not be monotone in time.

In the lowest noise setting where the covariates are strongly predictive of the response, we see the probability estimates are very close to the true ones—even for MLR which entirely ignores all time-series dependence in the data. However, as the noise level increases to the point where the covariates are not

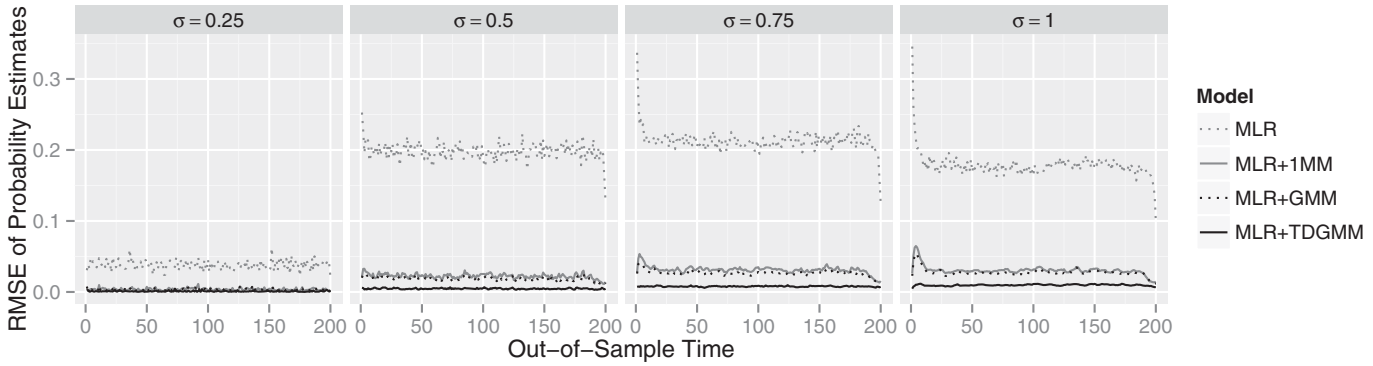

Figure 5. Root mean square error of probability estimates over out-of-sample time for the article simulation. The magnitude and pattern of error varies in time, covariate noise level σ, and model choice. The results for classification error and classification error relative to the Bayes' Rule are qualitatively similar.

particularly predictive and thus any signal in the data comes from harnessing time-series dependence, the various models begin to distinguish themselves (we note that, as both the noise level and the out-of-sample time period get arbitrarily large, the Bayes' Rule probabilities converge to the marginal probabilities given earlier).

Not surprisingly, MLR+TDGMM, the only correctly specified model, performs best with near-zero error at any given time point and little pattern in the errors. Also unsurprising is the fact that MLR performs worst as it ignores all time-series dependence in the data. Nonetheless, the pattern of error is striking: it performs relatively worst at time periods that are immediately out-of-sample and performs relatively best at time periods at the end of the out-of-sample data and this pattern is exacerbated as the noise level increases. This occurs both because (i) the evaluation is a relative one (i.e., of MLR which does not account for time-series dependence relative to the Bayes' Rule which does) and (ii) because the out-of-sample data "continues on" from the in-sample data; in this setting, models that account for time-series dependence can be very accurate immediately out-of-sample (i.e., locally) while, at the end of the out-of-sample period, such models have essentially no information other than that contained in the covariates. This pattern is exacerbated for high noise levels because, when the noise is high, the bulk of the signal in the data comes from harnessing time-series dependence and not from the covariates. Unsurprisingly, in the middle of the out-of-sample period, there appears to be a compromise between making use of information from both time-series dependence and the covariates.

The incorrectly specified MLR+1MM and MLR+GMM which take account of local time-series dependence but are not general enough to capture the full pattern of dependence in the data provide interesting contrasts to MLR and MLR+TDGMM. At relatively low values of noise, they perform almost as well as the MLR+TDGMM and there is no strong pattern in the errors. On the other hand, as the noise level increases to the point that the time-series dependence is providing the bulk of the information about the response, there is a strong pattern to the errors. These models (i) perform relatively well immediately out-of-sample when the more local patterns of time dependence captured by these models are reflective of the underlying data generation process; (ii) perform relatively more poorly in time periods that are moderately out-of-sample when there is strong time dependence in the data that these models cannot capture; (iii) perform relatively better as the ergodic patterns of time dependence in the Y_t wash out; and (iv) perform relatively very well at the end of the out-of-sample period where the covariates provide the bulk of the information about the Y_t.

Moving to our second finding, recall that the MLR+TDGMM is composed of two sets of estimates: (i) estimates of the conditional class probabilities $\mathbb{P}(Y_t|\boldsymbol{X}_t)$ and (ii) estimates of the time-series structure (i.e., the transition probability matrix and the duration distributions). We note that the noise level in the covariates impacts the quality of the former set of estimates while having no impact on the quality of the latter set. Given these two sets of estimates, one could imagine four versions of the MLR+TDGMM: (i) one that uses the true conditional class probabilities and the true time-series probabilities (i.e., the Bayes' Rule or "oracle" probabilities); (ii) one that uses estimated conditional class probabilities and estimated time-series probabilities (i.e., ordinary probability estimates); (iii) one that uses the true conditional class probabilities and estimated time-series probabilities (i.e., "conditional class semioracle" probability estimates); and (iv) one that uses estimated conditional class probabilities and the true time-series probabilities (i.e., "time-series semioracle" probability estimates). We consider the root mean square error of the ordinary probability estimates as well as those of the two semioracles (i.e., relative to the Bayes' Rule or the oracle probabilities) for the settings of our article simulation considered earlier averaged over all $T^\star = 200$ out-of-sample time points and over 1000 independent replicates of the simulation. Before proceeding to our results, we note that all four versions of the MLR+TDGMM have one minor "oracle-like" property, namely that they use the true value of the head size $M_{i,j}$ of each transition-dependent duration distribution.

We present our results in Table 1. Perhaps surprisingly, the ordinary probability estimates beat those of the two semioracles across a wide variety of simulation settings (neither semioracle is uniformly superior to the other). In other words, errors in estimating the conditional class probabilities seem to "cancel" with errors in estimating the time-series probabilities leading to superior combined estimates. When we examined the error by time as in Figure 5, there was relatively little pattern for small σ (i.e., when the Bayes' Rule probabilities are close to

Table 1. Root mean square error of probability estimates for the article simulation

Method	$\sigma = 0.25$	$\sigma = 0.50$	$\sigma = 0.75$	$\sigma = 1.00$
Ordinary	0.031	0.069	0.091	0.098
Conditional class semi-oracle	0.026	0.088	0.165	0.219
Time series semi-oracle	0.021	0.083	0.179	0.247

The ordinary probability estimates beat those of the two semioracles across a wide variety of simulation settings. The results for classification error and classification error relative to the Bayes' Rule are qualitatively similar.

the conditional class probabilities) but, as σ increases, the error of the two semioracle probability estimates was closest to the error of the ordinary probability estimates at the beginning and end of the out-of-sample time and worst in between.

While the fact that the ordinary probability estimates outperform those of the two semioracles may perhaps be vexing or even troubling, we analogize it to the case of simple linear regression where the ordinary estimators for the slope and intercept are $\hat{\beta} = r_{x,y}s_y/s_x$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$. For out-of-sample prediction using root mean square error as the loss function, the ordinary estimator $(\hat{\alpha}, \hat{\beta})$ is superior to the semioracle estimator $(\alpha, \hat{\beta})$ that uses the true intercept α, exactly analogous to that given earlier. In this case, it is of course easy to see that the semioracle can do better by estimating β in light of known α and that the resulting estimator $(\alpha, \tilde{\beta} = \sum_i (y_i - \alpha)x_i / \sum_i x_i^2)$ is superior to both $(\hat{\alpha}, \hat{\beta})$ and $(\alpha, \hat{\beta})$; a similar result holds, *mutatis mutandis*, for the semioracle which knows the true slope β and which should use $\tilde{\alpha} = \bar{y} - \beta\bar{x}$ in place of $\hat{\alpha}$. Coming back to our case, this suggests that aspects of the joint distribution $\mathbb{P}(\boldsymbol{Y}_{1:T}, \mathbf{X}_{1:T})$ known by, for example, the time-series oracle (for instance, the marginal distribution of the Y_t) should be used when estimating the conditional class probabilities, and we confirm this does indeed result in probability estimates which outperform the ordinary ones. Nonetheless, this behavior is interesting because it would appear, in contradistinction to the simple linear regression case, that errors in estimating the conditional class probabilities (due to, for example, errors in estimating the marginal distribution of the Y_t) would compound–rather than cancel–with errors in estimating the time series probabilities (and, of course, vice versa) even though Table 1 does indeed indicate canceling.

These results concerning model performance as a function of the out-of-sample time period and model performance of the oracle-like models do not appear to be dependent on particular aspects of the design of the article simulation. In fact, similar results hold for the mouse simulation. Consequently, it is our belief that these findings are relatively general across a wide variety of data settings, and we thus believe that understanding these results more deeply is a potentially fruitful topic for future research.

5. DISCUSSION

The last decade has seen tremendous advances in statistical learning for classification and conditional class probability estimation in the iid setting. Nonetheless, recently developed methods applied without modification are a poor choice in our setting due to the strong patterns of time-series dependence contained in our data. Our goal has been to leverage the power of recent advances while simultaneously harnessing the signal provided by this time-series dependence. However, there is no single obvious way to do this. One approach, employed in the discussion by KS, is rather straightforward in that it applies the standard statistical learning methods used in the iid setting to an augmented set of covariates. Our approach, on the other hand, is both more model-based and more application-driven. We model the time-series dependence contained in the data separately from the conditional class probabilities by using a powerful and general form of the Markov model for the former and the standard statistical learning methods used in the iid setting for the latter. While our approach is more computationally challenging compared to the more straightforward approach, it nonetheless remains computationally feasible while also being more parsimonious and more easily estimable. It is also more adept at capturing the rather general and long-term patterns of time dependence frequently encountered in applied settings. Further, by employing a coherent and unified probability model for the data, we obtain genuine probability estimates that allow us to easily calibrate and optimize our model's performance across the wide variety of objectives faced in our application.

APPENDIX: MOUSE SIMULATION DETAILS

The mouse simulation state space $S = \{\text{NREM}, \text{REM}, \text{WAKE}\}$ is the mouse data state space, the initialization distribution $\boldsymbol{\pi} = (0.4400, 0.0483, 0.5117)$, and the transition probability distributions $\mathbf{A}$ are

$$\mathbf{A} = \begin{pmatrix} 0.0000 & 0.2234 & 0.7766 \\ 0.2239 & 0.0000 & 0.7761 \\ 0.9974 & 0.0026 & 0.0000 \end{pmatrix}$$

The transition-dependent duration distributions $\boldsymbol{\delta}$ are beta-negative binomials with geometric tails; in particular, the parameters (α, β, r) of the "head" components and (q, s) of the "tail" components of each duration distribution $\delta_{i,j}$ were set to

State	α	β	r	q	s
NREM → REM	4.5076	5.4133	5.4094	0.8341	0.6682
WAKE → REM	0.3330	2.3036	2.0218	0.8000	0.0000
REM → NREM	0.0000	0.5214	36099000	0.8119	0.9488
WAKE → NREM	5.7763	0.7369	108.53	0.9551	0.9066
NREM → WAKE	0.4596	0.7288	0.7282	0.9897	0.9962
REM → WAKE	3.0829	1.6451	1.6434	0.9886	0.9911

The head sizes corresponding to $\boldsymbol{q}$ are $M_{\text{NREM}\to\text{REM}} = 12$, $M_{\text{WAKE}\to\text{REM}} = 2$, $M_{\text{REM}\to\text{NREM}} = 35$, $M_{\text{WAKE}\to\text{NREM}} = 51$, $M_{\text{NREM}\to\text{WAKE}} = 352$, and $M_{\text{REM}\to\text{WAKE}} = 99$, and the probability mass functions resulting from these parameter estimates appear in Figure 1.

The covariate emission distributions $\boldsymbol{\mu}$ are multivariate normal with state-specific means and a common covariance matrix; this choice of distribution results in a linear decision boundary and is fit to the observed mouse data for the six continuous covariates omitting, for obvious reasons, the powerful binary covariate that indicated whether or not the light in the mouse cage was on in epoch t. The state-specific

means are

State	X_1	X_2	X_3	X_4	X_5	X_6
NREM	0.3529	−0.6786	−0.8167	−0.7739	0.2445	−0.7887
REM	−0.0920	−0.6968	−0.8328	−0.7946	0.5787	−0.7508
WAKE	−0.2947	0.6493	0.7809	0.7405	−0.2649	0.7491

while the common covariance matrix is

$$\begin{pmatrix} 0.9003 & -0.1764 & -0.2179 & -0.2273 & -0.1968 & -0.1849 \\ -0.1764 & 0.5581 & 0.2793 & 0.2996 & -0.0670 & 0.3194 \\ -0.2179 & 0.2793 & 0.3609 & 0.3517 & -0.0499 & 0.2901 \\ -0.2273 & 0.2996 & 0.3517 & 0.4253 & -0.0163 & 0.3080 \\ -0.1968 & -0.067 & -0.0499 & -0.0163 & 0.9215 & -0.0574 \\ -0.1849 & 0.3194 & 0.2901 & 0.3080 & -0.0574 & 0.4118 \end{pmatrix}.$$

REFERENCES

Djuric, P. M., and Chun, J.-H. (2002), "An MCMC Sampling Approach to Estimation of Nonstational Hidden Markov Models," *IEEE Transactions on Signal Processing*, 50, 1113–1123. [1168]

McShane, B. B., Jensen, S. T., Pack, A. I., and Wyner, A. J. (2013), "Statistical Learning With Time Series Dependence: An Application to Scoring Sleep in Mice" (with discussion), *Journal of the American Statistical Association* 108, 1147–1162. [1165]

Shedden, K. (2013), Discussion of "Statistical Learning With Time Series Dependence: An Application to Scoring Sleep in Mice," *Journal of the American Statistical Association* 108, 1162–1163. [1165]

Zeng, D., and Wang, Y. (2013), Discussion of "Statistical Learning With Time Series Dependence: An Application to Scoring Sleep in Mice," *Journal of the American Statistical Association* 108, 1164. [1165]