Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Rejoinder

Blakeley B. McShane [a], Shane T. Jensen [b], Allan I. Pack [c] & Abraham J. Wyner [b]
[a] Kellogg School of Management, Northwestern University, Evanston, IL, 60611
[b] The Wharton School, University of Pennsylvania, Philadelphia, PA, 19104
[c] Center for Sleep and Circadian Neurobiology, University of Pennsylvania, Philadelphia, PA, 19104
Published online: 19 Dec 2013.

To cite this article: Blakeley B. McShane, Shane T. Jensen, Allan I. Pack & Abraham J. Wyner (2013) Rejoinder, Journal of the American Statistical Association, 108:504, 1165-1172, DOI: 10.1080/01621459.2013.844021

To link to this article: http://dx.doi.org/10.1080/01621459.2013.844021

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

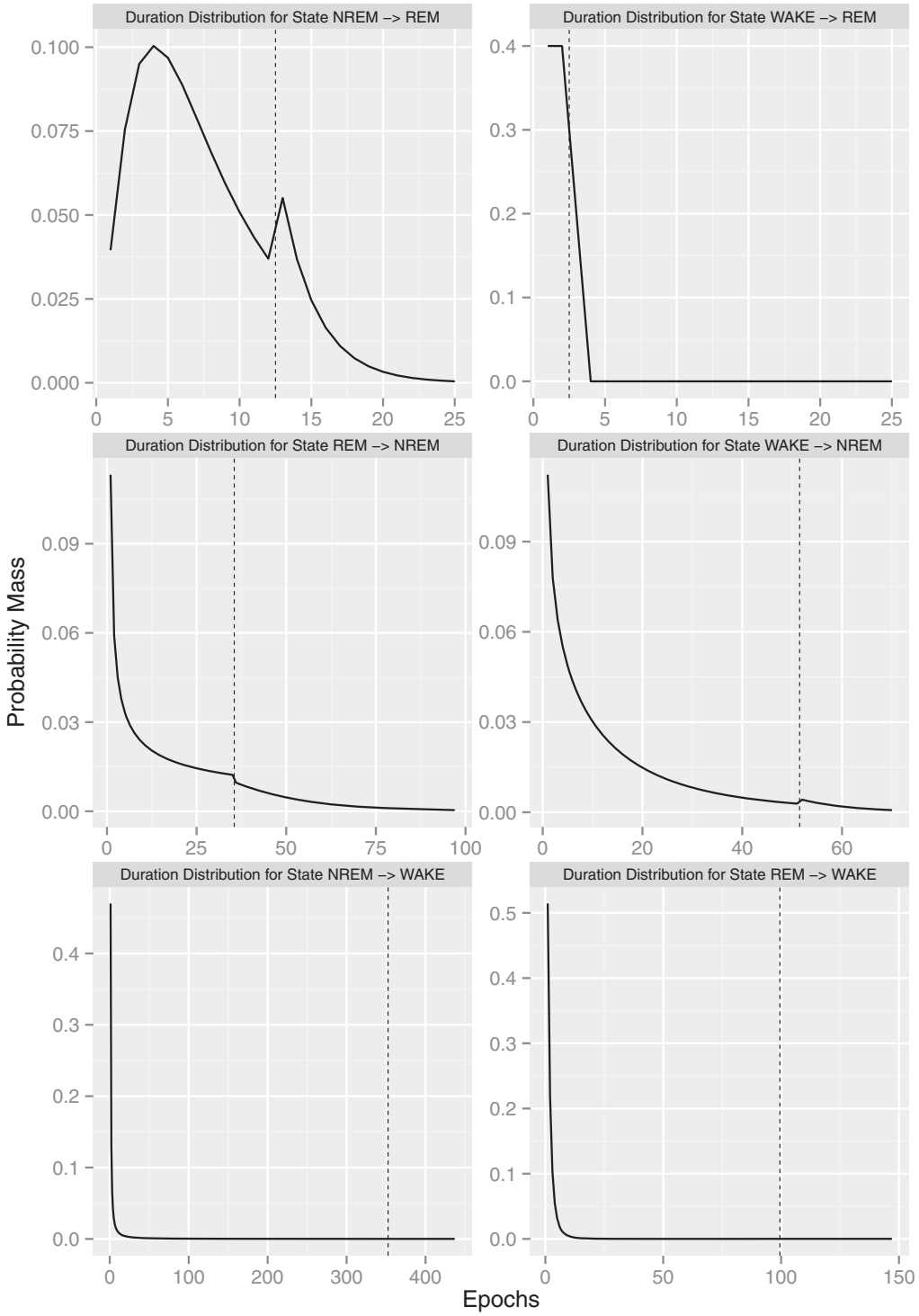

Figure 1. Transition-dependent duration distributions for the mouse simulation. We estimate a beta-negative binomial distribution with geometric tail for each conditional distribution using the procedure outlined in the online supplementary materials of our article. We plot distributions so that over 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

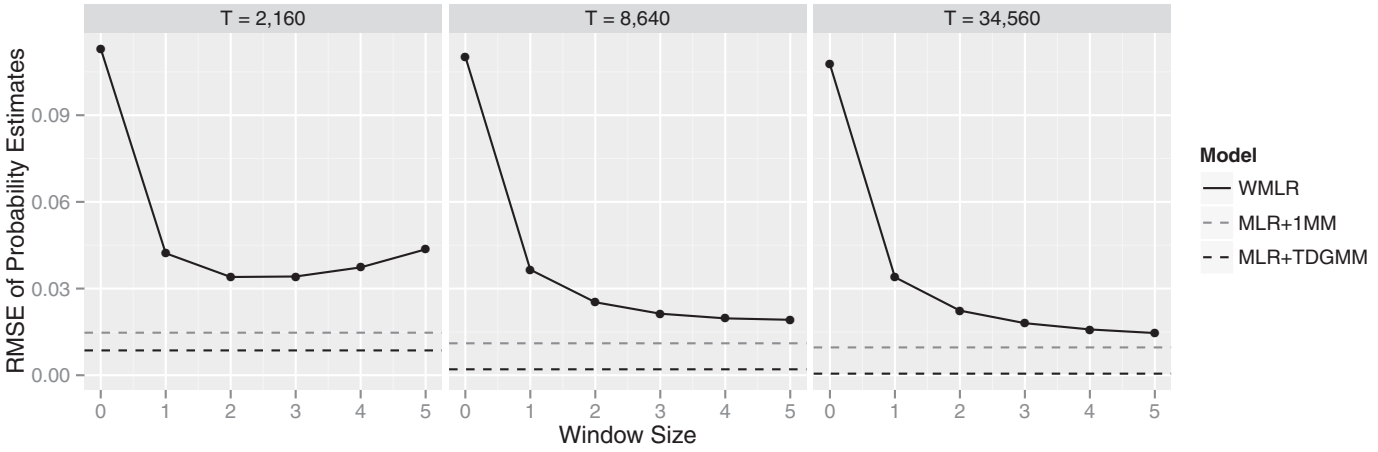

Figure 2. Root mean square error of probability estimates for the mouse simulation. MLR+TDGMM performs best while the WMLR approach asymptotes in w relatively quickly despite the rather longer-term patterns of time dependence in the simulated data. The results for classification error and classification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

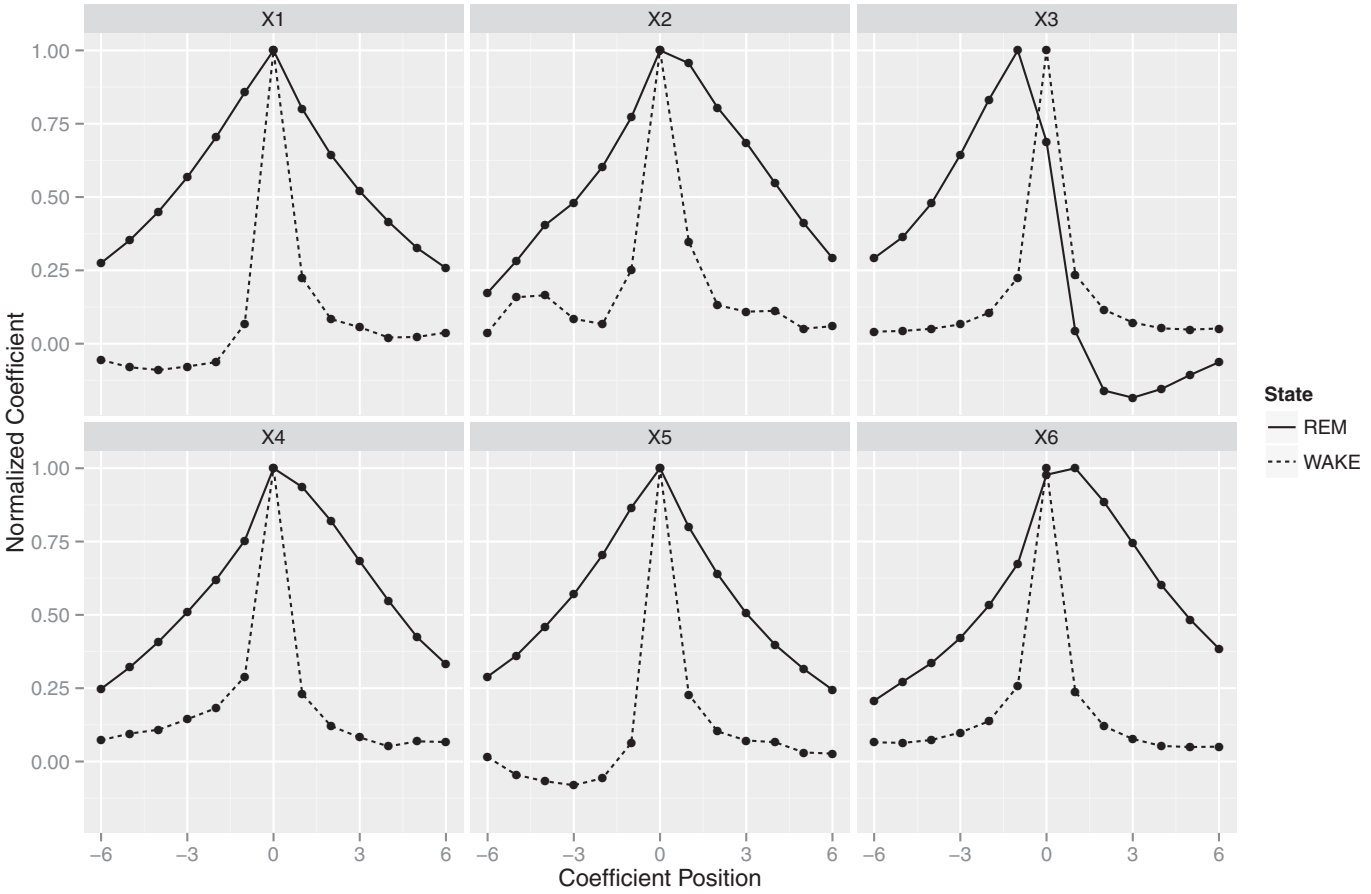

Figure 3. Normalized coefficients for the mouse simulation. The coefficients are normalized using the procedure outlined in KS and we set $\beta_{\text{NREM}} \equiv 0$ for identification. Many of the coefficients are not well-approximated by an exponentially weighted moving average and the coefficients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

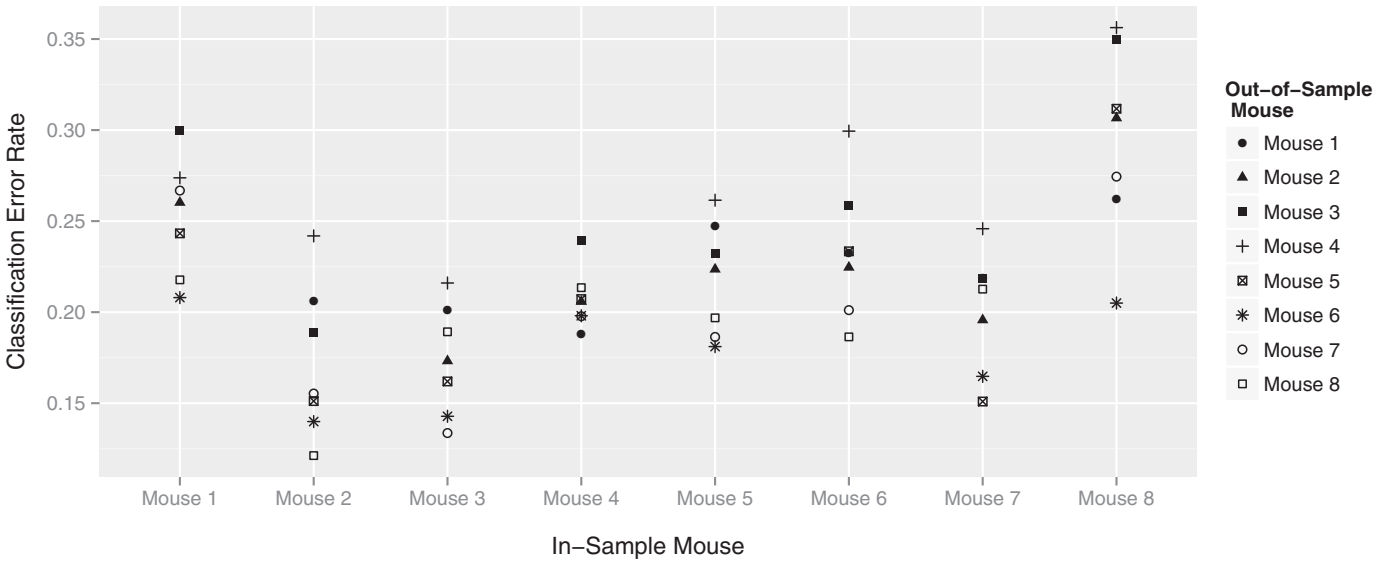

Figure 4. Classification errors by in-sample mouse and out-of-sample mouse for the mouse data. There is substantial heterogeneity in classification error rate both by in-sample mouse and by out-of-sample mouse and there are consistent patterns. The results for the rate of REM prediction, the REM false-positive 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, $\mathbb{P}(Y_t|\mathbf{X}^\star, \boldsymbol{\Delta})$) 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 $\mathbb{P}(Y_t|\boldsymbol{X}_t^\star, \boldsymbol{\Delta})$ 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, $\mathbb{P}(Y_t|\mathbf{X}^\star, \boldsymbol{\Delta})$) 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

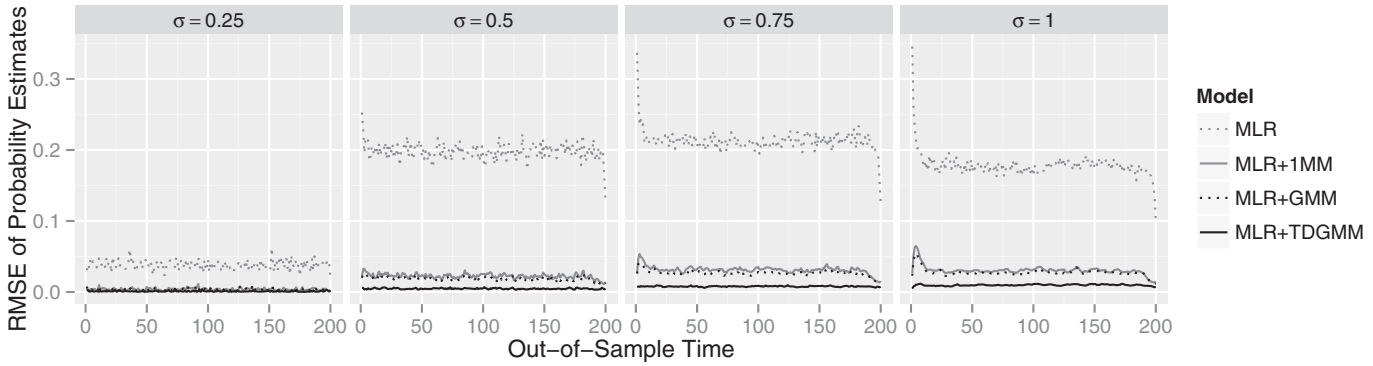

Figure 5. Root mean square error of probability estimates over out-of-sample time for the article simulation. The magnitude and pattern of error varies in time, covariate noise level σ, and model choice. The results for classification error and classification error relative to the Bayes' Rule are qualitatively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