Statistical Science
2020, Vol. 35, No. 3, 400–403
https://doi.org/10.1214/20-STS794
Main article: https://doi.org/10.1214/19-STS720

Rejoinder: Linear Mixed Models with Endogenous Covariates: Modeling Sequential Treatment Effects with Application to a Mobile Health Study

Tianchen Qian, Predrag Klasnja and Susan A. Murphy

We thank the Editors for this opportunity and the discussants for their insightful comments. The discussions added much depth and breadth to this work. In the following, we address the comments by each discussant.

1. KRISTIN LINN

1.1 Potential Violation to the Conditional Independence Assumption

The conditional independence assumption in display (10) of the main manuscript says that the time-varying covariate is conditionally independent of the random effects given all the observed history prior to the covariate (i.e., all the previous outcomes, previous treatment assignments and previous covariates). Linn commented that the conditional independence assumption may require further justification for our data example, HeartSteps. In particular, she noted that latent factors, such as levels of depression and anxiety, can be associated with endogenous time-varying covariates such as location, and these latent factors can be part of the random effects. We would like to clarify that this possible association does not necessarily violate the conditional independence assumption; a violation would be such association *after* conditioning on the observed history. For our data example, it is possible that even after conditioning on the observed history, the current location of the user may still be correlated with the latent factors. Because this assumption is not testable without additional modeling assumptions, it is critical to carefully scrutinize the assumption both through scientific knowledge and through sensitivity analyses. Under additional modeling assumptions, it is possible to test for this assumption, and another discussant, Cho et al., provided a potential solution to test empirically the conditional independence assumption (see Section 3 of this rejoinder).

Linn also pointed out that the latent factors, such as levels of depression and anxiety, can change over time and be impacted by prior treatments. For such settings, latent variable models such as partially observed Markov decision processes may prove useful (e.g., Ross et al., 2011).

1.2 Choice Between Marginal Effect and Conditional Effect

Linn commented that a marginal effect estimate may be preferred over a conditional effect, due to reasons including the interest on population-level intervention and ethical and privacy concerns. We agree that there are settings where the marginal effect is of primary interest, such as in the primary analysis of an MRT. For these settings, we recommend use methods for estimating marginal causal effect for MRT (Boruvka et al., 2018, Qian et al., 2019). There are, however, settings where the person-specific effect is of interest. For example, in exploratory analyses behavioral scientists may want to understand how individuals respond to the treatment differently in addition to that explained by the difference in their observed covariates. Such analyses has the potential to aid the development of individualized mobile health interventions (Nahum-Shani et al., 2018).

1.3 Two Practical Suggestions

We wish to echo two other comments by Linn, as they have very important practical implications.

One of our main contributions is to show that under the conditional independence assumption, standard linear mixed model (LMM) software can be used to estimate the conditional effect even if there are endogenous covariates. As a cautionary note, Linn commented that "[e]ase of implementation also opens the door for quick application of the proposed estimation approach by nonstatisticians or time-deprived statisticians who may not take time to fully consider whether the critical assumptions seem reasonable." We reiterate the two caveats when applying standard LMM software with endogenous covariates: (i) When the conditional independence assumption

Tianchen Qian is Postdoctoral Fellow, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: qiantianchen@fas.harvard.edu). Predrag Klasnja is Assistant Professor, School of Information, University of Michigan, Ann Arbor, Michigan 48109, USA (e-mail: klasnja@umich.edu). Susan A. Murphy is Professor, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: samurphy@fas.harvard.edu).

is believed to hold, the resulting estimated treatment effect can only be interpreted conditional on the random effects, not marginally, unless there is no interaction between the random effects and the treatment indicator. (ii) The conditional independence assumption needs to be checked theoretically or through sensitivity analyses; otherwise, the treatment effect estimated via standard LMM can be biased.

Linn also brought up that the tech industry is leveraging mobile health data to attract users and improve engagement. We believe there will be great benefits if industry and academia work together on developing mobile health interventions to promote healthy behavior.

2. ERICA MOODIE AND DAVID STEPHENS

2.1 Use of Directed Acyclic Graphs

Moodie and Stephens (MS hereafter) provided a rather nice illustration of endogeneity and how one can assess confounding using directed acyclic graphs (DAGs). In fact, using their approach it is easy to see that the conditional independence assumption in display (10) in the main manuscript is not required for identifiability of the conditional on the random effects parameter, β. The main manuscript focuses on whether the estimator of β obtained by standard LMM software is consistent. The assumption in display (10) provides a sufficient condition for this.

2.2 A Time-Varying Confounder Need Not Be a Mediator

We agree with MS's point that a time-varying confounder need not be a mediator. Indeed, the literature we reviewed in Section 2.3 was not just for situations with feedback loops (where a time-varying confounder is also a mediator); rather, it was about correctly accounting for time-varying confounders (regardless of whether they are also mediators).

3. HUNYONG CHO ET AL.

3.1 Efficiency of the Partial Likelihood Approach

In establishing the feasibility of standard linear mixed models software to estimate the conditional effect, we used a likelihood factorization argument (see equations (11) through (13) of the main manuscript), and we claimed that in our setting any inference based on the full likelihood can be equivalently based on the partial likelihood. Cho et al. pointed out that when covariates are endogenous, the term dropped from the full likelihood when forming the partial likelihood, $p(X_{it} \mid H_{it-1}, A_{it-1}, Y_{it})$, may in fact contain additional information about the parameter of interest. Thus, efficiency loss may occur due to not using the full likelihood.

We agree that if there is additional structural information that implies that $p(X_{it} \mid H_{it-1}, A_{it-1}, Y_{it})$ contains additional information about the parameter of interest, then one should take advantage of this information. In fact, as discussed in Section 6 of the main manuscript, other authors have taken a different approach than ours by jointly modeling the time-varying covariates and time-varying outcomes when dealing with LMMs with endogenous covariates (albeit in settings different than MRTs); see, for example, Miglioretti and Heagerty (2004), Roy et al. (2006), Sitlani et al. (2012), Shardell and Ferrucci (2018). Another way to view this problem is that using the partial likelihood without modeling $p(X_{it} \mid H_{it-1}, A_{it-1}, Y_{it})$ versus using the full likelihood with additional models on $p(X_{it} \mid H_{it-1}, A_{it-1}, Y_{it})$ can be viewed as a bias-variance trade-off. It would be very interesting to operationalize this bias-variance trade-off so as to help data analysts in these settings decide if it is worthwhile to attempt to utilize prior knowledge about the variables in X_{it}.

3.2 An Empirical Test for the Conditional Independence Assumption

Cho et al. proposed sensitivity analyses to the conditional independence assumption via hypothesis tests of the independence between the covariates, X_{it}, and the predicted random effects, $\hat{b}_i$, conditional on the history information observed prior to X_{it}. To address the loss of power due to the curse of dimensionality from both the high dimensional H_{it} for fixed t and the multiple testing at different t's, they proposed to (i) construct a summary vector of the history, $s_d(H_{it})$, that ideally captures the role of H_{it} in the conditional independence assumption, and (ii) reduce the multiplicity in testing by only focusing on a subset of all the time points. This is an interesting area for further research given the desire by applied data analysts to use existing software.

3.3 A Coherent Marginal Effect Estimator

The treatment effect estimator proposed in the main manuscript is conditional on the random effects. As also pointed out by Linn in her discussion, Cho et al. mentioned the interest of estimating marginal effects. Although there are existing methods to estimate marginal causal effects for MRTs (Boruvka et al., 2018, Qian et al., 2019), these directly model the marginal effect, and as a result it is difficult to make their marginal models coherent with the conditional model considered in this manuscript. To construct a coherent marginal effect estimator, Cho et al. proposed a least squares estimator for $E(b_i \mid H_{it})$, a key term in the marginal model implied by the conditional model.

Two settings in which a marginal effect estimate might be of interest are in primary analysis, and in hypothesis generating analyses that are exploratory in nature. For the former, we prefer methods that rely on few assumptions, such as those in Boruvka et al. (2018), Qian et al. (2019). For the latter, methods such as the one proposed by Cho et al. are quite attractive, particularly as additional assumptions can be used to increase efficiency of the estimators.

3.4 An Extension to LMMs with Kernels

Cho et al. discussed an extension to incorporate radial basis kernels in order to accommodate possible violation to the linearity assumption of LMM, where they formulated the extension as a Bayesian LMM. We believe this is a useful extension to consider as there is much practical need of modeling the mean function with flexible functional forms. Pearce and Wand (2009) discussed the explicit connection between the generalized linear mixed models (GLMMs) and the kernel machines based on reproducing kernel Hilbert space (RKHS) theory, and they provided a principled way for constructing kernel-based extension to general mean curves. Other authors has also discussed the connection between random effects models and kernel methods (such as Gaussian process regression) or spline-based methods; see, for example, Speed (1991), Verbyla et al. (1999), Wand (2003), Pearce and Wand (2006). Below we discuss the connection between LMM and Gaussian process in a simplified setting and point out directions for further generalization of the proposal by Cho et al.

Consider the Bayesian version of the standard LMM with exogenous covariates presented in equation (1) of the manuscript, with prior $\beta \sim N(0, \Sigma_\beta)$. Following notation in Section 2.1 of the manuscript, this implies that marginally we have

$$\begin{aligned} &Y_i \mid X_i, \sigma_\epsilon, G, \Sigma_\beta \\ &\quad \sim N\big(0, X_i \Sigma_\beta X_i^T + Z_i G Z_i^T + \sigma_\epsilon^2 I_{T_i}\big), \end{aligned} \tag{1}$$

where I_{T_i} is a $T_i \times T_i$ identity matrix with T_i denoting the total number of decision points for individual i. Now consider a typical Gaussian process regression setup (e.g., Rasmussen and Williams, 2006). Suppose

$$\begin{aligned} Y_{it+1} \mid X_{it}, h, \sigma_\epsilon &\sim N\big(h(X_{it}), \sigma_\epsilon^2\big), \\ h &\sim \mathrm{GP}(0, k), \end{aligned}$$

where $\mathrm{GP}(0, k)$ denotes a Gaussian process with mean function 0 and covariance function k. Let K_i be a $T_i \times T_i$ matrix with (j_1, j_2)th entry $k(X_{ij_1}, X_{ij_2})$; K_i is known as the Gram matrix. Then by equation (2.30) of Rasmussen and Williams (2006), we have $Y_i \mid X_i, \sigma_\epsilon, k \sim N(0, K_i + \sigma_\epsilon^2 I_{T_i})$. Therefore, if we choose the kernel to be $k(X_{ij_1}, X_{ij_2}) = X_{ij_1} \Sigma_\beta X_{ij_2}^T + Z_{ij_1} G Z_{ij_2}^T$, then

$$Y_i \mid X_i, \sigma_\epsilon, k \sim N\big(0, X_i \Sigma_\beta X_i^T + Z_i G Z_i^T + \sigma_\epsilon^2 I_{T_i}\big),$$

coinciding with the likelihood of the Bayesian LMM in (1). Therefore, to generalize the Bayesian LMM, one can replace $X_{ij_1} \Sigma_\beta X_{ij_2}^T$ in the definition of $k(X_{ij_1}, X_{ij_2})$ by other kernel functions.

The above is a very simple example for the purpose of demonstrating the connection between Bayesian LMM and Gaussian process regression. The proposed model of Cho et al. is an extension to the LMM with endogenous covariates in equation (7) of the manuscript. Other such extensions can be formulated by substituting the Gaussian kernel used by Cho et al., $K_{it}(\boldsymbol{w}; \boldsymbol{Z}^*, \gamma)$, with other kernel functions such as the Matérn kernel. There has also been work in the literature that combines LMM with kernel methods with application to genetics (Liu, Lin and Ghosh, 2007) and to online reinforcement learning in mobile health studies (Tomkins et al., 2020).

Cho et al. proposed to tune hyperparameters via cross-validation. Another common approach to determine the value of hyperparameters in kernel methods is marginal maximum likelihood estimation (also known as empirical Bayes) (e.g., Rasmussen and Williams, 2006, Section 5). It has been suggested by some authors that cross-validation-based approach can be more robust against model misspecification (Wahba, 1990).

Importantly, Cho et al. discussed three implications of the conditional independence assumption for their kernel mixed models in the presence of endogenous time-varying covariates, based on the likelihood factorization argument in Section 3 of the manuscript. Namely, that the kernel mixed models can be fit using existing software, that the estimates will maximize full data likelihood of the training set given fixed hyperparameters, and that the hyperparameters can be chosen to maximize the full data likelihood of the test set. These observations can facilitate a potentially wide application of kernel-based mixed models to mobile health and other domains with endogenous covariates.

ACKNOWLEDGMENTS

Research presented in this paper was supported by the National Heart, Lung and Blood Institute under award number R01HL125440; the National Institute on Alcohol Abuse and Alcoholism under award number R01AA023187; the National Institute on Drug Abuse under award number P50DA039838; and the National Institute of Biomedical Imaging and Bioengineering under award number U54EB020404. The authors thank Marianne Menictas, Liao Peng and Shuang Li for helpful suggestions.

REFERENCES

BORUVKA, A., ALMIRALL, D., WITKIEWITZ, K. and MURPHY, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. *J. Amer. Statist. Assoc.* **113** 1112–1121. MR3862343 https://doi.org/10.1080/01621459.2017.1305274

LIU, D., LIN, X. and GHOSH, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **63** 1079–1088, 1311. MR2414585 https://doi.org/10.1111/j.1541-0420.2007.00799.x

MIGLIORETTI, D. L. and HEAGERTY, P. J. (2004). Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics* **5** 381–398.

NAHUM-SHANI, I., SMITH, S. N., SPRING, B. J., COLLINS, L. M., WITKIEWITZ, K., TEWARI, A. and MURPHY, S. A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals Behav. Med.* **52** 446–462. https://doi.org/10.1007/s12160-016-9830-8

PEARCE, N. D. and WAND, M. P. (2006). Penalized splines and reproducing kernel methods. *Amer. Statist.* **60** 233–240. MR2246756 https://doi.org/10.1198/000313006X124541

PEARCE, N. D. and WAND, M. P. (2009). Explicit connections between longitudinal data analysis and kernel machines. *Electron. J. Stat.* **3** 797–823. MR2534202 https://doi.org/10.1214/09-EJS428

QIAN, T., YOO, H., KLASNJA, P., ALMIRALL, D. and MURPHY, S. A. (2019). Estimating time-varying causal excursion effect in mobile health with binary outcomes. Preprint. Available at arXiv:1906.00528.

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

ROSS, S., PINEAU, J., CHAIB-DRAA, B. and KREITMANN, P. (2011). A Bayesian approach for learning and planning in partially observable Markov decision processes. *J. Mach. Learn. Res.* **12** 1729–1770.

ROY, J., ALDERSON, D., HOGAN, J. W. and TASHIMA, K. T. (2006). Conditional inference methods for incomplete Poisson data with endogenous time-varying covariates: Emergency department use among HIV-infected women. *J. Amer. Statist. Assoc.* **101** 424–434. MR2256164 https://doi.org/10.1198/016214505000001203

SHARDELL, M. and FERRUCCI, L. (2018). Joint mixed-effects models for causal inference with longitudinal data. *Stat. Med.* **37** 829–846. MR3760452 https://doi.org/10.1002/sim.7567

SITLANI, C. M., HEAGERTY, P. J., BLOOD, E. A. and TOSTESON, T. D. (2012). Longitudinal structural mixed models for the analysis of surgical trials with noncompliance. *Stat. Med.* **31** 1738–1760. MR2947521 https://doi.org/10.1002/sim.4510

SPEED, T. (1991). Comment for "That BLUP is a good thing: The estimation of random effects". *Statist. Sci.* **6** 42–44.

TOMKINS, S., LIAO, P., KLASNJA, P., YEUNG, S. and MURPHY, S. (2020). Rapidly personalizing mobile health treatment policies with limited data. Preprint. Available at arXiv:2002.09971.

VERBYLA, A. P., CULLIS, B. R., KENWARD, M. G. and WELHAM, S. J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **48** 269–311.

WAHBA, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442 https://doi.org/10.1137/1.9781611970128

WAND, M. P. (2003). Smoothing and mixed models. *Comput. Statist.* **18** 223–249.