Statistical Science
2020, Vol. 35, No. 3, 396–399
https://doi.org/10.1214/20-STS782
Main article: https://doi.org/10.1214/19-STS720
© Institute of Mathematical Statistics, 2020

Comment: Diagnostics and Kernel-based Extensions for Linear Mixed Effects Models with Endogenous Covariates

Hunyong Cho, Joshua P. Zitovsky, Xinyi Li, Minxin Lu, Kushal Shah, John Sperger, Matthew C. B. Tsilimigras and Michael R. Kosorok

Abstract. We discuss "Linear mixed models with endogenous covariates: modeling sequential treatment effects with application to a mobile health study" by Qian, Klasnja and Murphy. In this discussion, we study when the linear mixed effects models with endogenous covariates are feasible to use by providing examples and diagnostic tools as well as discussing potential extensions. This includes evaluating feasibility of partial likelihood-based inference, checking the conditional independence assumption, estimation of marginal effects, and kernel extensions of the model.

Key words and phrases: Linear mixed models, partial likelihood, conditional independence test, marginal effects, kernel mixed models.

The authors ("QKM") have made a significant breakthrough in data analysis problems by demonstrating the theoretical validity and utility of linear mixed effects models (LMMs) in the face of endogenous covariates. As the presence of endogenous covariates and the application of LMMs is common in microrandomized trials (MRTs), their method also makes an important contribution to precision medicine. We study when their method is feasible to use by providing examples and diagnostic tools as well as discussing potential extensions.

Feasibility of partial likelihood-based inference. We first discuss when QKM's partial likelihood-based inference is feasible. The authors claim that $p(X_{it}|H_{it-1}, A_{it-1}, Y_{it})$ in (11) of QKM does not involve $\xi \equiv (\alpha, \beta, \theta, \sigma_\epsilon)$ and can be ignored. However, when covariates are endogenous, this term might actually involve ξ or some parameters that are not orthogonal to ξ in general. In other words, when X_{it} contains some relevant information newer than Y_{it}, $p(X_{it}|H_{it-1}, A_{it-1}, Y_{it})$ may contain information about the prognostic effects α or delayed treatment effects β. This would make (13) of QKM not the scaled full-likelihood as the authors claim, but only a partial likelihood.

Wong (1986) illustrated the relative efficiency bound $k/(k+l)$ of partial likelihood estimators for autoregressive models of disconnected sequences, where k is the observed segment lengths and l is the missing segment lengths. Although the models proposed in QKM are distinct from the Wong et al. model since they involve treatments and are not fully autoregressive, the efficiency bound results provide intuition about the potential efficiency loss of the LMMs. In MRT settings, the information contained in the nuisance likelihood

Hunyong Cho is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: hunycho@live.unc.edu). Joshua P. Zitovsky is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: joshz@live.unc.edu). Xinyi Li is a postdoctoral fellow, Statistical and Applied Mathematical Sciences Institute (SAMSI) and University of North Carolina at Chapel Hill, Durham/Chapel Hill, North Carolina 27516, USA (e-mail: xli@samsi.info). Minxin Lu is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: mino12@live.unc.edu). Kushal Shah is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: kushshah@live.unc.edu). John Sperger is a graduate student, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: jsperger@live.unc.edu). Matthew C. B. Tsilimigras is a postdoctoral fellow, Department of Epidemiology, Department of Nutrition, Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: matthew_tsilimigras@unc.edu). Michael R. Kosorok is the W.R. Kenan, Jr. Distinguished Professor and Chair, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, USA (e-mail: kosorok@unc.edu).

$p(X_{it}|H_{it-1}, A_{it-1}, Y_{it})$ can be substantial when there is a structural relationship between the treatment effects on X_{it} and Y_{it}. For instance, when the treatment effects are not immediate but prolonged such as antidepressants (Artigas, Bortolozzi and Celada, 2018), are further assumed or known to be linear in time, and Y_{it} and X_{it} are the outcomes measured at different time points, then the information contained in the nuisance likelihood, $p(X_{it}|H_{it-1}, A_{it-1}, Y_{it})$, can be significant. A counterexample is the two-stage example in QKM where $X_{i2} = Y_{i2}$. In this case, there is no loss of information from ignoring the nuisance likelihood, and partial likelihood-based estimation is equivalent to full likelihood-based estimation regarding ξ.

Checking the conditional independence assumption. The authors argue that the conditional independence assumption must be verified from a domain science perspective. As the assumption may not be testable without having a further assumption, we posit a reasonable model to help with verification in practice and further discuss an ad hoc testing procedure.

Consider a regression model $X_{it+1} = (f_1(\mathbf{H}_{it}), A_{it}, Y_{it+1})^\top \gamma_1 + b_i^\top \gamma_2 + \eta_{it+1}$, where η_{it+1} is an independent mean zero error term. As the dimension of $f_1(\mathbf{H}_{it})$ increases, it is less likely that additionally having b_i provides a meaningful amount of information about X_{it+1}. In other words, as the dimension becomes larger, the amount of history and the proportion of variability explained by the history increase, and thus the explanatory power of additionally having the random effects decreases. Then conditional independence is more likely to hold under the posited model. This agrees with QKM's argument in the HeartSteps example that X_{it} is plausibly independent of the random effects given all earlier step counts and treatments because X_{it} can be largely explained through all the history. In contrast, if the history is relatively short, as in the early phases of an MRT, the random effects may provide additional significant information.

The length of history, however, is not a sufficient measure of the conditional independence. Even with a long history, if H_{it} and Y_{it} are highly noisy, and thus contain little information on X_{it}, b_i can still provide a significant amount of information on X_{it}. When evidence from domain science is not sufficient, a diagnostic procedure could be used to detect possible violations of the conditional independence assumption.

There are multiple challenges to directly testing the conditional independence hypothesis, $X_{it} \perp b_i | H_{it-1}, A_{it-1}, Y_{it}, \forall t$. First, as random effects are not observed in the data, testing conditional independence between random effects and the covariates may require additional modeling of $f(X_{it}|H_{it-1}, A_{it-1}, Y_{it}, b_i)$ and a more sophisticated estimation procedure. Second, conditioning on the full history can be problematic for conditional independence testing methods because they tend to lose power as the dimension of the conditioning random variable grows—the curse of dimensionality. Third, testing the assumption at every time point will lead to a multiple testing problem. A more practical diagnostic measure may be an ad hoc test of $X_{it} \perp \hat{b}_i | s_d(H_{it-1}), A_{it-1}, Y_{it}, \forall t \in \mathcal{T}$ based on the estimated random effects from the fitted model and where s_d controls the maximum time window by truncating the history to the last d time points. One nonparametric possibility for assessing this hypothesis is the Conditional Distance Independence Test (CDIT) (Wang et al., 2015).

Conducting this diagnostic procedure requires a choice of the history window d, a set of time points $\mathcal{T}$ to test on, and a strategy for summarizing multiple test statistics. A large history window may result in a loss of power, but too narrow a window may invoke false positives by bringing undue dependence that would not have existed if conditioning on the full history. Choice of the time periods at which to conduct the test, $\mathcal{T}$, is another important question. The simplest approach is picking one time point, for example, $\mathcal{T} = \{T\}$, which does not require summarization of the test. However, it does not make use of the full information available at the other time points. Ideally, one would test at every possible time point to be faithful to the original hypothesis. In this case, however, the individual test statistics could be considerably correlated among neighboring time points, and hence tests based on those statistics may not be powerful. To balance between loss of information and duplicity, one could pick every rth time point for testing: $\mathcal{T} = \{T - rk : k = 0, 1, 2, \dots, \lfloor T/r \rfloor\}$. The choice of r could be $d + 1$, for example. If the level of conditional dependence does not change much between neighboring time points, this approach does not lose much information, while still covering most of the time domain. Finally, summarization of the test can be done in multiple ways. Let $\boldsymbol{M} = (M_1, \dots, M_{|\mathcal{T}|})$ be the vector of the individual test statistics. A vector norm (e.g., $\|\mathbf{M}\|_2$ or $\|\mathbf{M}\|_\infty$) can be used to summarize the test, and the p-value can be obtained either analytically if the dependence structure of the individual tests can be reasonably posited, or through resampling: $p = B^{-1} \sum_{b=1}^{B} 1(\|\mathbf{M}\|_2 \leq \|\mathbf{M}^{(b)}\|_2)$, where $\mathbf{M}^{(b)}$ denotes the vector of statistics based on the bth resample. Alternatively, the p-values of the individual statistics can be summarized. The p-values could first be adjusted using a multiple test correction strategy such as Bonferroni or Benjamini–Hochberg (Benjamini and Hochberg, 1995), with the minimum value then chosen as the global p-value.

Estimation of marginal effects. β in QKM's model only has a conditional-on-the-random-effects interpretation, and estimating marginal effects would be useful if one wishes to make inferences or predictions on future individuals not in the original study. For instance, estimation of marginal treatment effects is important when one aims to provide treatment recommendations to future patients based on LMMs. In such a precision medicine setting, we wish to estimate $E(Y_{it+1}|H_{it}, A_{it}=1) - E(Y_{it+1}|H_{it}, A_{it}=0)$, a quantity equal to $f_1(h_{it})^\top \beta + g_1(h_{it})^\top E(b_{1i}|H_{it})$. If we assume that the estimated random effects are asymptotically unbiased, then $E(b_{1i}| H_{it}) \approx E\{E(\hat{b}_{1i}|b_{1i}, H_{it})|H_{it}\} = E(\hat{b}_{1i}|H_{it})$. Thus, under this assumption, we need only to model $E(\hat{b}_{1i}|H_{it})$ to obtain asymptotically-valid marginal treatment effect estimates from our fitted LMM, as well as marginal interaction effect estimates between the treatment and other predictors.

Let $H^* = s(H_{it}) \subset H_{it}$, where s is a prespecified operator that summarizes H_{it} into a P-dimensional vector, while keeping $E(b_i|H_{it}) - E(b_i|s(H_{it}))$ negligible. Let $\hat{b}_{1i} = (\hat{b}_{1i1}, \dots, \hat{b}_{1iK})$, where $K = \dim(b_{1i})$. We posit the models $E(\hat{b}_{1ik}|H_{it}) = H_{it}^{*\top} \gamma_k$, $k = 1, 2, \dots, K$, and estimate γ_k by least squares. Let $\mathbf{H}_{nT\times P} = (H_{11}^*, H_{12}^*, \dots, H_{1T}^*, H_{21}^*, \dots, H_{nT}^*)^\top$, $\boldsymbol{\Gamma}_{P\times K} = (\gamma_1, \dots, \gamma_K)$, $\mathbf{b}_k = (\hat{b}_{11k}\mathbf{1}^\top, \dots, \hat{b}_{1nk}\mathbf{1}^\top)^\top$ and $\mathbf{B}_{nT\times K} = (\mathbf{b}_1, \dots, \mathbf{b}_K)$, where $\mathbf{1}$ is a $T \times 1$ vector of ones. We can rewrite our posited models as $E(\mathbf{B}|\mathbf{H}) = \mathbf{H}\boldsymbol{\Gamma}$, where the conditioning can be interpreted row-wise (i.e., $E(\mathbf{B}|\mathbf{H}) = [E(\hat{b}_{11}|H_{11}), E(\hat{b}_{11}|H_{12}), \dots, E(\hat{b}_{1n}|H_{nT})]^\top$). The least squares solution is $\hat{\boldsymbol{\Gamma}} = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\mathbf{B}$.

We leave a mathematically rigorous proof of the consistency of our least squares estimates to future work, though we do give a heuristic argument in the Supplementary Material that could be helpful. It is worth noting that we can also include interactions between covariates and time in H_{it}^* to model changes in associations with time, or include higher-order terms to model nonlinear associations. In either case, the same estimation procedure would still apply. For future work, it may be possible to develop models that are more robust to model misspecification or more efficient. For example, as it is more realistic to assume that history is caused by random effects than the other way around, positing models for $E(Y_{it}|b_{1i})$ and $E(X_{it}|b_{1i})$ and inverting them to model $E(b_{1i}|H_{it})$ may lead to a more realistic model. However, there are a few challenges unique to positing and estimating models for the estimated random effects that will need to be taken into account. We leave a brief discussion of these challenges to the Supplementary Material for those interested.

Kernel extensions. Accurate estimates and valid inference from linear mixed effects models can only be obtained with assumptions of linearity, which are often violated in practice. To deal with this problem, we discuss an extension to LMMs with kernels and show how QKM's results apply in this case. The approach is loosely based on the mixed effects random forest (MERF) model by Hajjem, Bellavance and Larocque (2014). While the typical LMM assumes $Y_i = X_i\beta + Z_ib_i + \epsilon_i$, the MERF model assumes $Y_i = f(X_i) + Z_ib_i + \epsilon_i$, where Y_i, X_i and Z_i are a vector of responses, a matrix of fixed covariates, and a matrix of random covariates associated with the ith subject, respectively. f is a nonparametric fixed regression function, $b_i \overset{\text{iid}}{\sim} N(\mathbf{0}, G(\theta))$ is a linear random effect and $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2 I)$ is a vector of random errors. However, while the MERF model uses random forests to estimate $f(X_i)$, we use Gaussian kernels, and our optimization procedure also differs.

Let $\mathbf{Z}_{it} = (\mathbf{H}_{it}^\top, A_{it})^\top$ and $\mathbf{Z}_{it}^* = s(\mathbf{Z}_{it}) \in \mathbb{R}^{d+1}$, where s is a prespecified operator that removes all elements from $\mathbf{Z}_{it}$ except for A_{it} and some known d-dimension subset $\mathbf{H}_{it}^*$ of $\mathbf{H}_{it}^\top$. Let q be the number of the earliest time points for which s cannot be applied. For instance, if s extracts history from the last 3 time points, then s cannot be applied to $\mathbf{Z}_{it}$ for $t \in \{1, 2\}$, and $q = 2$. Define $h : \mathbb{R}^{d+1} \to \mathbb{R}^{n(T-q)}$ such that $h(\mathbf{w}; \mathbf{Z}^*, \gamma) = (K_{1\{q+1\}}(\mathbf{w}; \mathbf{Z}^*, \gamma), K_{1\{q+2\}}(\mathbf{w}; \mathbf{Z}^*, \gamma), \dots, K_{1T}(\mathbf{w}; \mathbf{Z}^*, \gamma), K_{2\{q+1\}} (\mathbf{w}; \mathbf{Z}^*, \gamma), \dots, K_{nT}(\mathbf{w}; \mathbf{Z}^*, \gamma))^\top$ and $K_{it}(\mathbf{w}; \mathbf{Z}^*, \gamma) = \exp(-\gamma\|\mathbf{w} - \mathbf{Z}_{it}^*\|^2)$ is a Gaussian kernel with a fixed bandwidth γ. Finally, let $f(\mathbf{Z}_{it}; \mathbf{Z}^*, \gamma) = h(s(\mathbf{Z}_{it}); \mathbf{Z}^*, \gamma)$. We assume the model

$$\begin{aligned} Y_{it+1} = {} & \alpha_0 + f(\mathbf{Z}_{it}; \mathbf{Z}^*, \gamma)^\top \alpha + g_0(\mathbf{H}_{it})^\top b_{0i} \\ & + A_{it} g_1(\mathbf{H}_{it})^\top b_{1i} + \epsilon_{it+1}, \end{aligned} \tag{1}$$

where $1 \le i \le n$, $q+1 \le t \le T$, g_0, g_1 are known functions similar to s, $(b_{0i}^\top, b_{1i}^\top)^\top \overset{\text{iid}}{\sim} N(\mathbf{0}, G(\theta))$, $\epsilon_{it+1} \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)$, $A_{it} \perp (b_{0i}, b_{1i})|H_{it}$ and $X_{it} \perp (b_{0i}, b_{1i})| H_{it-1}, A_{it-1}, Y_{it}$. To avoid overfitting and nonidentifiability of α, we additionally impose a prior on the fixed effects, $\alpha \sim N(0, (\lambda\mathbf{K})^{-1})$, where $\mathbf{K} = (f(\mathbf{Z}_{1\{q+1\}}; \mathbf{Z}^*, \gamma), f(\mathbf{Z}_{1\{q+2\}}; \mathbf{Z}^*, \gamma), \dots, f(\mathbf{Z}_{1T}; \mathbf{Z}^*, \gamma), f(\mathbf{Z}_{2\{q+1\}}; \mathbf{Z}^*, \gamma), \dots, f(\mathbf{Z}_{nT}; \mathbf{Z}^*, \gamma))^\top$ is our $n(T-q) \times n(T-q)$ kernel matrix and λ is a fixed scalar penalty parameter. This becomes a standard Bayesian LMM where we assume treatments are randomized and conditional independence holds, and thus this model can be fit with off-the-shelf Bayesian software packages. Based on the results of QKM, it is not difficult to show that standard Bayesian software will take this model and maximize $\log \mathcal{L}(\alpha_0, \alpha, \theta, \sigma_\epsilon|X, Y, Z, \gamma) - \lambda\alpha^\top\mathbf{K}\alpha/2$ provided the conditional independence assumption holds, where $\mathcal{L}(\alpha_0, \alpha, \theta, \sigma_\epsilon|X, A, Y, \gamma) = \prod_i p(X_i, A_i, Y_i|\alpha_0, \alpha, \theta, \sigma_\epsilon, \gamma)$ is the full-data likelihood of model (1) and $-\lambda\alpha^\top \times \mathbf{K}\alpha/2$ is a penalty term from the log-prior.

λ and γ can be tuned by partitioning our data into observations from n_1 subjects for training and from n_2 subjects for testing. For a grid of (λ, γ) values, we can fit our model on the n_1 training subjects, giving us estimates

$\hat{\xi}^{(\lambda,\gamma)} = (\hat{\alpha}_0^{(\lambda,\gamma)}, \hat{\alpha}^{(\lambda,\gamma)}, \hat{\theta}^{(\lambda,\gamma)}, \hat{\sigma}_\epsilon^{(\lambda,\gamma)})$, and evaluate

$$\begin{aligned} &\log \mathcal{L}_1\big(\hat{\xi}^{(\lambda,\gamma)} | X^{\text{Test}}, A^{\text{Test}}, Y^{\text{Test}}, \gamma\big) \\ &\quad = \sum_{i=1}^{n_2} \log \int \prod_{t=q+1}^{\top} p\big(Y_{it+1}^{\text{Test}} | \\ &\qquad \mathbf{H}_{it}^{\text{Test}}, A_{it}^{\text{Test}}, b_i; \hat{\xi}^{(\lambda,\gamma)}, \gamma\big)\, d\hat{F}^{(\lambda,\gamma)}(b_i), \end{aligned} \tag{2}$$

where $p(Y_{it+1}^{\text{Test}} | \mathbf{H}_{it}^{\text{Test}}, A_{it}^{\text{Test}}, b_i; \xi, \gamma) =_d N(\mu_{it+1,\gamma}, \sigma_\epsilon^2)$, $\mu_{it+1,\gamma} = \alpha_0 + f(\mathbf{Z}_{it}^{\text{Test}}; \mathbf{Z}^*, \gamma)^\top \alpha + g_0(\mathbf{H}_{it}^{\text{Test}})^\top b_{0i} + A_{it} g_1(\mathbf{H}_{it}^{\text{Test}})^\top b_{1i}$ and $\hat{F}^{(\lambda,\gamma)}$ is the distribution of b_i given $\theta = \hat{\theta}^{(\lambda,\gamma)}$. As integration is over the estimated distribution function of b_i, which only depends on $\hat{\theta}^{(\lambda,\gamma)}$, the actual random effects from the test subjects need not be known, and the likelihood can easily be computed with off-the-shelf numerical integration packages. A good final choice of (λ, γ) is that which maximizes $\log \mathcal{L}_1(\hat{\xi}^{(\lambda,\gamma)} | X^{\text{Test}}, A^{\text{Test}}, Y^{\text{Test}}, \gamma)$. It can be shown that under conditional independence, the values of (λ, γ) which maximize $\log \mathcal{L}_1(\hat{\xi}^{(\lambda,\gamma)} | X^{\text{Test}}, A^{\text{Test}}, Y^{\text{Test}}, \gamma)$ also maximize the full test likelihood $\mathcal{L}(\hat{\xi}^{(\lambda,\gamma)} | X^{\text{Test}}, A^{\text{Test}}, Y^{\text{Test}}, \gamma) = \sum_{i=1}^{n_2} \log p(X_i^{\text{Test}}, A_i^{\text{Test}}, Y_i^{\text{Test}} | \hat{\xi}^{(\lambda,\gamma)}, \gamma)$.

Therefore, there are three important results that hold under conditional independence. First, we can reduce the kernel mixed effects model to a standard Bayesian LMM and fit it using existing software. Second, the resulting estimates will maximize the full-data likelihood of training cases for a given bandwidth γ and penalty parameter λ. Finally, λ and γ can be chosen to maximize full-data likelihood of the test cases. These results imply that under conditional independence, we can use standard software to fit kernel models that accurately model the population or underlying data-generating distribution of interest.

We performed several simulations to demonstrate the performance of LMMs in estimating treatment effect when the true effect is nonlinear. The details and results of the simulations can be found in the Supplementary Material (Cho et al., 2020). Our results show that LMM estimates are less reliable when the assumption of linearity is violated, and for certain nonlinear associations, the mean squared error can be quite high. Using Gaussian radial kernels would allow for much more flexible models of the conditional response distribution and lead to more accurate conditional-on-the-random-effect estimates. Furthermore, our model can easily be combined with our marginal effect estimation idea to achieve a more flexible marginal mean model and more accurately predict the difference in treatment effect for future individuals with particular histories. Finally, unlike inference from standard LMMs, Gaussian kernel mixed effects models can allow one to determine important variables even if they have nonlinear associations with treatment efficacy and response. For instance, relative importance of variables can be established by randomly permuting values for each variable and observing the decrease in likelihood after training.

SUPPLEMENTARY MATERIAL

Supplement to "Comment: Diagnostics and Kernel-based Extensions for Linear Mixed Effects Models with Endogenous Covariates" (DOI: 10.1214/20-STS782SUPP; .pdf). Supplementary information.

REFERENCES

ARTIGAS, F., BORTOLOZZI, A. and CELADA, P. (2018). Can we increase speed and efficacy of antidepressant treatments? Part I: General aspects and monoamine-based strategies. *Eur. Neuropsychopharmacol.* **28** 445–456. https://doi.org/10.1016/j.euroneuro.2017.10.032

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

CHO, H., ZITOVSKY, J. P., LI, X., LU, M., SHAH, K., SPERGER, J., TSILIMIGRAS, M. C. B. and KOSOROK, M. R. (2020). Supplement to "Comment: Diagnostics and Kernel-based Extensions for Linear Mixed Effects Models with Endogenous Covariates." https://doi.org/10.1214/20-STS782SUPP

HAJJEM, A., BELLAVANCE, F. and LAROCQUE, D. (2014). Mixed-effects random forest for clustered data. *J. Stat. Comput. Simul.* **84** 1313–1328. MR3169395 https://doi.org/10.1080/00949655.2012.741599

WANG, X., PAN, W., HU, W., TIAN, Y. and ZHANG, H. (2015). Conditional distance correlation. *J. Amer. Statist. Assoc.* **110** 1726–1734. MR3449068 https://doi.org/10.1080/01621459.2014.993081

WONG, W. H. (1986). Theory of partial likelihood. *Ann. Statist.* **14** 88–123. MR0829557 https://doi.org/10.1214/aos/1176349844