Model Selection in Longitudinal Distributed Lag Models with Variational AICs

M. J. Meyer
Georgetown University
Washington, DC

S. Carter
Carnegie Mellon Universiy
Pittsburgh, PA

E. J. Malloy
American University
Washington, DC

This guide should be used for preparation of the MATH 640 Final Paper. A .zip file with the files used to compile this document is on Canvas.

1 Introduction

Distributed lag models or DLMs are commonly used methods for analyzing lagged effects of exposure on various outcomes. Earlier work in DLMs consists of constrained polynomials and then penalized B-splines to model the effect of lagged pollution exposure on daily deaths (Schwartz, 2000; Welty et al., 2009). Further investigations include hierarchical DLMS to assess geographical effects (Baek et al., 2016) and DLMs with multiple, interacted lagged exposures (Wilson et al., 2017). Despite the volume of existing work on this class of models, data with more complex lag structures require additional methodological developments. Specifically, DLMs that handle lagged effects measured longitudinally, that is repeatedly over time, are limited to the random intercept case (Baek et al., 2016). However, longitudinally sampled lags may exhibit other induced covariance structures such that a random lag structure may be appropriate. Thus we propose a class of longitudinal DLMs (LDLMs) for both Gaussian and Poisson outcome data where the lag structure can arise from longitudinally sampled lags, including crossover trials. Our models use penalized splines to estimate lagged effects and to model random lags. For computational efficiency, we perform estimation using variational Bayesian inference, see Appendix A for a brief overview. We also derive variational AICs (VAICs) for our LDLMs and determine decision rules for model selection. We show in simulation that the VAICs performs well in identifying the correctly specified model and that our models have good properties in terms of mean integrated squared error (MISE). We also demonstrate our models on a crossover trial that examines the lagged effect of depressed blood oxygen saturation during simulated flight conditions on post-flight heart rate and post-flight heart rate variability.

2 Methods

Let Y_{ij} denote the jth response of the ith subject. We define the following linearized relationships:

$$g[E(Y_{ij})] = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{l}'_{ij}\boldsymbol{\gamma}_j + \mathbf{l}'_{ij}\mathbf{g}_i \text{ (random lag) and}$$
$$g[E(Y_{ij})] = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{l}'_{ij}\boldsymbol{\gamma}_j + b_i \text{ (random intercept)}$$

where $\mathbf{x}_{ij}$ and $\mathbf{l}_{ij}$ are $1 \times P$ and $1 \times \ell$ vectors, respectively, of fixed and lagged effects with corresponding vectors of coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_j$. We allow the lagged effects to vary by j for the crossover case where each subject gets both treatments giving each treatment its own estimated effect. For the longitudinal case, we drop the j index: $\boldsymbol{\gamma}_j = \boldsymbol{\gamma}$. The link function, $g(\cdot)$, depends on the distribution of Y_{ij} and can be either Gaussian or Poisson.

To reduce the number of parameters that need to be estimated and to model the functional form of the distributed lag, we perform a basis expansion of the lagged effects. Suppose Θ is the $\ell \times K$ matrix of known basis functions. The basis expanded coefficients are $\boldsymbol{\gamma}_j = \Theta\boldsymbol{\gamma}_j^S$ (for the crossover model), $\boldsymbol{\gamma} = \Theta\boldsymbol{\gamma}^S$ (longitudinal model), and $\mathbf{g} = (I_N \otimes \Theta)\mathbf{g}^S$ for some selection of knots, K—the basis expansions do not necessarily have to have the same number of knots. Under the crossover design, the spline-based models are

$$g[E(Y_{ij})] = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{l}'_{ij}\Theta\boldsymbol{\gamma}_j^S + \mathbf{l}'_{ij}\mathbf{g}_i \text{ (random lag) and}$$
$$g[E(Y_{ij})] = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{l}'_{ij}\Theta\boldsymbol{\gamma}_j^S + b_i \text{ (random intercept)}$$

and under the longitudinal design, $\boldsymbol{\gamma}_j^S = \boldsymbol{\gamma}^S$.

For both the fixed and random lag components, we implement penalized B-splines or P-splines using the mixed model representation. Thus, Θ is a matrix of B-spline basis functions of size dependent upon the number of knots selected and the sizes of L and W. The corresponding penalty matrix, $\mathcal{P} = \xi D_0 + (1 - \xi)D_2$, is weighted between the zeroth derivative matrix (D_0) and the second derivative matrix (D_2). The parameter ξ controls the desired tradeoff between shrinkage and smoothness, with values near 0 favoring shrinkage. Depending on the model, the penalty priors on $\boldsymbol{\gamma}_j^S$ and $\boldsymbol{\gamma}^S$ are $\boldsymbol{\gamma}_j^S \sim N(\mathbf{0}, \lambda_j \mathcal{P}_j^{-1})$ and $\boldsymbol{\gamma}^S \sim N(\mathbf{0}, \lambda \mathcal{P}^{-1})$, where λ_j and λ are tuning-parameters. The random lagged effects performed best using P-splines with $K_g = n$ knots, where n is the total number of subjects. The penalty prior on $\mathbf{g}^S$ is then $\mathbf{g}^S \sim N\left[\mathbf{0}, \lambda_g(I_N \otimes \mathcal{P}_g)^{-1}\right]$ where λ_g is a tuning-parameter. The size of the penalty matrix depends on the number of basis functions and can vary by penalized component.

We place weakly informative priors on the components $\boldsymbol{\beta}$, $\beta_p \sim N(0, \sigma_b^2)$ with σ_b^2 fixed at something large. For the random intercept model, the elements of $\mathbf{b}$ are independent normals, $b_i \overset{iid}{\sim} N(0, \sigma_u^2)$. Variance component

priors depend on the model. When $Y_{ij} \sim N[E(Y_{ij}), \sigma^2]$, we place inverse-gamma priors (and hyper-priors) on all variance components: $\sigma^2 \sim IG(a_e, b_e)$, $\lambda_j \sim IG(a_j, b_j)$ or $\lambda \sim IG(a, b)$, and $\sigma_u^2 \sim IG(a_u, b_u)$ or $\lambda_g \sim IG(a_g, b_g)$. When $Y_{ij} \sim Pois[\exp\{E(Y_{ij})\}]$, we use half-Cauchy priors: $\sqrt{\lambda_j} \sim HC(A_j)$ or $\sqrt{\lambda} \sim HC(A)$, and $\sigma_u \sim HC(A_u)$ or $\sqrt{\lambda_g} \sim HC(A_g)$. To implement the half-Cauchy prior, we employ the mixture representation where, for example, $\lambda_j \sim IG(1/2, 1/\alpha_j)$ and $\alpha_j \sim IG(1/2, 1/A_j^2)$ induces $\sqrt{\lambda_j} \sim HC(A_j)$. The mixture applies to the other components, λ, σ_u^2, and λ_g, in an analogous fashion.

By leveraging the mixed model representation of penalized splines, we are able to generate estimates from our LDLMs using two existing general frameworks for variational Bayesian inference (Ormerod and Wand, 2010; Luts and Wand, 2015). Based on these, we develop algorithms to estimate parameters from both the random lag and random intercept models, see Appendix B. These variational algorithms efficiently provide a way to perform model selection between two potential large models. The VAIC has good asymptotic properties and, in the case of multiple linear regression, converges in probability to the standard AIC (You et al., 2014). The general form of the VAIC is

$$VAIC \equiv -2\log p(y|\theta^*) + 2P_D^*$$

where $\theta^* = E_q(\theta)$ and $P_D^* = 2\log p(y|\theta^*) - 2E_q[\log p(y|\theta)]$. The expectation is taken with respect to the variational approximation densities. We derive $\log p(y|\theta^*)$ and $E_q[\log p(y|\theta)]$ for both the Gaussian and Poisson LDLM cases along with their respective VAICs.

The smallest VAIC often suggests the best model fit—a "minimum" decision rule. As parsimonious models are also desirable, other rules might impose a threshold on the difference in VAICs between candidate models. When the difference is under the threshold, the more parsimonious model, i.e. the random intercept model, would be selected. We examine the minimum rule along with thresholds of 2, 5 and 10 in a simulation study.

3 Simulation Study and Application

The goal of the VAIC is to select a "best" model. Our simulation is designed to evaluate this by simulating under two different "true" models: one with a random lag and one with a random intercept only. We consider sample sizes of $n = 25, 50, 100$, sampling density on the lag of $\ell = 60, 120$, and three designs: crossover, longitudinal with two, and with three measurements. The true form of the lag varies between a peak effect, cyclical effect, and sigmoidal effect. We simulate 500 datasets under each combination and apply both random lag and random intercept LDLMs. Table 1 contains the percent of correctly specified Gaussian models, by decision rule, for the peak effect under a crossover design. These results are similar across effects and designs.

Our LDLM has good MISE when the model is correctly specified, see Table 2. Under the smallest setting,

Table 1: Percent correctly specified out of 500 simulated datasets, by decision rule. $Y_{ij} \sim N[E(Y_{ij}), \sigma^2]$.

| Truth | ℓ | n | Decision Rule | | | |
			min	< 2	< 5	< 10
Int.	60	25	85.2%	88.4%	91.8%	95.8%
		50	93.8%	94.4%	96.2%	98.6%
		100	96.6%	97.6%	98.8%	99.6%
	120	25	91.0%	93.8%	96.4%	98.6%
		50	96.2%	97.6%	98.4%	99.0%
		100	99.2%	99.4%	99.6%	99.8%
Lag	60	25	100%	100%	100%	99.2%
		50	100%	100%	100%	100%
		100	100%	100%	100%	100%
	120	25	99.8%	99.8%	99.6%	99.0%
		50	100%	100%	100%	100%
		100	100%	100%	100%	100%

Table 2: Average Model MISE, by specification. Values are averaged over 500 simulated datasets. $Y_{ij} \sim N[E(Y_{ij}), \sigma^2]$.

| Truth | ℓ | n | Specified Model | |
			Lag	Int.
Int.	60	25	0.0286	0.0228
		50	0.0121	0.0096
		100	0.0067	0.0058
	120	25	0.0149	0.0103
		50	0.0071	0.0059
		100	0.0042	0.0040
Lag	60	25	0.3583	0.7501
		50	0.0689	0.2994
		100	0.0330	0.1339
	120	25	0.2345	0.5559
		50	0.0734	0.2236
		100	0.0319	0.1110

when $n = 25$ and $\ell = 60$, average MISEs are similar, but the correctly specified model has has lower MISE as both n and ℓ increase. Model convergence is relatively a fast. Figure 1 contains the variational lower bound for one simulated dataset when $n = 25$, $\ell = 60$, the true structure is random intercept but the model is misspecified with a random lag.

The motivating data comes from a study of the effects of exposure to altitude on heart rate and heart rate variability (Meyer et al., 2019; Meyer et al., 2022). We use the LDLMs and VAICs to select the best covariance structure to describe the relationship between lagged effect of blood oxygen saturation and post-flight heart rate as well as the rate of post-flight supraventricular ectopic heartbeats. For the LDLM with heart rate as the outcome and lagged effect of blood oxygen saturation under both experimental and non-experimental conditions, we find the random lag model to have $VAIC = 404.8$ while the random intercept model has $VAIC = 432.2$. Regardless of decision rule, given that the difference is
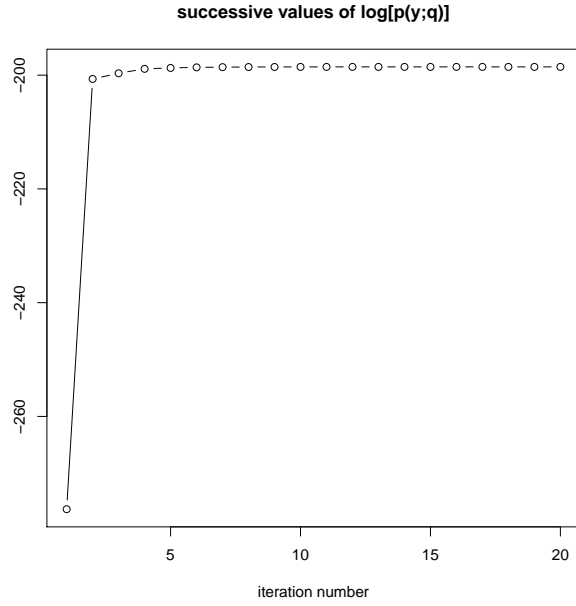
successive values of log[p(y;q)]

Figure 1: Sample successive variational lower bound for random lag model, when random intercept is the truth. $Y_{ij} \sim N[E(Y_{ij}), \sigma^2]$.

27.3, the VAIC suggests that the random lag model induces a better fitting covariance structure for the data when post-flight heart rate is the outcome. For post-flight heart rate variability, we first perform a log-transformation first. Based on this outcome, the random lag model has $VAIC = 120.1$ while the random intercept model has $VAIC = 123.5$. Using the decision rule of selecting the more parsimonious model when the difference in VAICs is less than five, we'd select the random intercept model as the best representation of the covariance structure.

4 Discussion

Variational AICs have good discrimination for LDLMs under a number of different decision rules—particularly when the true model is a random lag. The models also exhibit good properties in terms of MISE. Applying the LDLMs and VAIC to the motivating data, we can efficiently select a covariance structure and build inferential models to assess the impact of blood oxygen desaturation, due to altitude exposure, on post-flight heart rate and heart rate variability, measured with rMSSD.

References

J. Baek, E. V. Sanchez-Vaznaugh, and B. N. Sánchez. 2016. Hierarchical distributed-lag models: Exploring varying geographic scale and magnitude in associations between the built environment and health. *American Journal of Epidemiology*, 183:583–592.

J. Luts and M. P. Wand. 2015. Variational inference for count response semiparametric regression. *Bayesian Analysis*, 10:991–1023.

M. J. Meyer, I. Mordukhovich, G. A. Wellenius, M. A. Mittleman, J. P. McCracken, B. A. Coull, and E. McNeely. 2019. Changes in heart rate and rhythm during a crossover study of simulated commercial flight in older and vulnerable participants. *Frontiers Physiology*, 10:1339.

M. J. Meyer, I. Mordukhovich, B. A. Coull, J. P. McCracken, R. Germany, G. A. Wellenius, M. A. Mittleman, and E. McNeely. 2022. Impact of simulated flight conditions on supraventricular and ventricular ectopy. *JAMA Network Open*, Submitted & Under Review.

J. T. Ormerod and M. P. Wand. 2010. Explaining variational approximations. *The American Statistician*, 64:140–153.

J. Schwartz. 2000. The distributed lag between air pollution and daily deaths. *Epidemiology*, 11:320–326.

L. J. Welty, R. D. Peng, S. L. Zeger, and F. Dominici. 2009. Bayesian distributed lag models: Estimating effects of particulate matter air pollution on daily mortality. *Biometrics*, 65:282–291.

A. Wilson, Y.-H. M. Chiu, H-.H. L. Hsu, R. O. Wright, R. J. Wright, and B. A. Coull. 2017. Bayesian distributed lag interaction models to identify perinatal windows of vulnerability in children?s health. *Biostatistics*, 18:537–552.

C. You, J. T. Ormerod, and S. Müller. 2014. On variational bayes estimation and variational information criteria for linear regression models. *Australian & New Zealand Journal of Statistcs*, 56:73–87.

A Variational Bayesian Inference

Let $\boldsymbol{\theta}$ be a vector of parameters and $\mathbf{y}$ denote a vector of observed data. The posterior distribution for $\boldsymbol{\theta}$ given $\mathbf{y}$ is $p(\boldsymbol{\theta}|\mathbf{y}) = p(\mathbf{y}, \boldsymbol{\theta})/p(\mathbf{y})$. Typically, $p(\mathbf{y})$ is intractable and posterior estimates need to be obtained algorithmically. For the arbitrary density q, $p(\mathbf{y})$ is bounded below by $\underline{p}(\mathbf{y}; \boldsymbol{\theta})$ where

$$\underline{p}(\mathbf{y}; q) = \exp\left[\int q(\boldsymbol{\theta}) \log\left\{\frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\} d\boldsymbol{\theta}\right].$$

Variational algorithms work by maximizing $\underline{p}(\mathbf{y}; q)$ over a class of densities, q, that are tractable. This in turn minimizes the Kullback-Leibler Divergence between $q(\boldsymbol{\theta})$, and the posterior itself, $p(\boldsymbol{\theta}|\mathbf{y})$. Given a partition of $\boldsymbol{\theta}$ in to M subcomponents, $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$, we use the mean field approximation to construct q, specifically $q(\boldsymbol{\theta}) = \prod_{m=1}^{M} q_m(\boldsymbol{\theta}_m)$. Optimal densities are obtained iteratively with convergence, and therefore minimization, occurring when changes in the variation lower bound, $\underline{p}(\mathbf{y}; q)$, become negligible. For a thorough introduction to variational Bayesian techniques with examples, see (Ormerod and Wand, 2010).

B Model Derivations and Algorithms

The variational lower bounds for the Gaussian model with a random distributed lag is

$$\log[\underline{p}(\mathbf{y};q)] = \frac{1}{2}(P + K_0 + K_1 + K_g) - \frac{N}{2}\log(2\pi)$$
$$- \frac{P}{2}\log(\sigma_b^2) + \frac{1}{2}\log\left(|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{\gamma}^S,\mathbf{g}^S)}|\right)$$
$$- \frac{1}{2\sigma_b^2}\left[\boldsymbol{\mu}_{q(\boldsymbol{\beta})}'\boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \operatorname{tr}\left\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right\}\right]$$
$$- a_e \log(b_e) - \left(a_e + \frac{N}{2}\right)\log(B_{q(\sigma^2)})$$
$$+ \log\left(\Gamma\left(a_e + \frac{N}{2}\right)\right) - \log\left(\Gamma(a_e)\right)$$
$$+ a_0 \log(b_0) - \left(a_0 + \frac{K_0}{2}\right)\log(B_{q(\lambda_0)})$$
$$+ \log\left(\Gamma\left(a_0 + \frac{K_0}{2}\right)\right) - \log(\Gamma(a_0))$$
$$+ a_1 \log(b_1) - \left(a_1 + \frac{K_1}{2}\right)\log(B_{q(\lambda_1)})$$
$$+ \log\left(\Gamma\left(a_1 + \frac{K_1}{2}\right)\right) - \log(\Gamma(a_1))$$
$$+ a_g \log(b_g) - \left(a_g + \frac{K_g}{2}\right)\log(B_{q(\lambda_g)})$$
$$+ \log\left(\Gamma\left(a_g + \frac{K_g}{2}\right)\right) - \log(\Gamma(a_g)).$$

The lower bound for the random intercept Gaussian model is

$$\log[\underline{p}(\mathbf{y};q)] = \frac{1}{2}(P + K_0 + K_1 + K_u) - \frac{N}{2}\log(2\pi)$$
$$- \frac{P}{2}\log(\sigma_b^2) + \frac{1}{2}\log\left(|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{\gamma}^S,\mathbf{g}^S)}|\right)$$
$$- \frac{1}{2\sigma_b^2}\left[\boldsymbol{\mu}_{q(\boldsymbol{\beta})}'\boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \operatorname{tr}\left\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right\}\right]$$
$$- a_e \log(b_e) - \left(a_e + \frac{N}{2}\right)\log(B_{q(\sigma^2)})$$
$$+ \log\left(\Gamma\left(a_e + \frac{N}{2}\right)\right) - \log\left(\Gamma(a_e)\right)$$
$$+ a_0 \log(b_0) - \left(a_0 + \frac{K_0}{2}\right)\log(B_{q(\lambda_0)})$$
$$+ \log\left(\Gamma\left(a_0 + \frac{K_0}{2}\right)\right) - \log(\Gamma(a_0))$$
$$+ a_1 \log(b_1) - \left(a_1 + \frac{K_1}{2}\right)\log(B_{q(\lambda_1)})$$
$$+ \log\left(\Gamma\left(a_1 + \frac{K_1}{2}\right)\right) - \log(\Gamma(a_1))$$
$$+ a_u \log(b_u) - \left(a_u + \frac{K_u}{2}\right)\log(B_{q(\lambda_u)})$$
$$+ \log\left(\Gamma\left(a_u + \frac{K_u}{2}\right)\right) - \log(\Gamma(a_u)).$$

Our algorithm is presented in Table 3 where C is the full design matrix, $C = [X\ L(I_2 \otimes \Theta)\ W(I_n \otimes \Theta)]$ for the

---

Table 3: The algorithm for Gaussian outcome with a random distributed lag where $\mathcal{D} =$ blockdiag $\left\{(\sigma_b^2)^{-1}I_{p\times p}, \frac{a_0 + \frac{1}{2}K_0}{B_{q(\sigma_0^2)}}\mathcal{P}, \frac{a_1 + \frac{1}{2}K_1}{B_{q(\sigma_1^2)}}\mathcal{P}, \frac{a_g + \frac{1}{2}K_g}{B_{q(\sigma_g^2)}}\mathcal{P}\right\}$ and $\mathcal{SSY} = \left\{Y - C\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{\gamma}^S,\mathbf{g}^S)}\right\}'\left\{Y - C\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{\gamma}^S,\mathbf{g}^S)}\right\}$. The algorithm for the random intercept model replaces $B_{q(\lambda_g)}$ with $B_{q(\sigma_u^2)}$ and sets $\mathcal{P} = I_{N/2 \times N/2}$.

Initialize:
$B_{q(\sigma^2)}, B_{q(\sigma_0^2)}, B_{q(\sigma_1^2)}, B_{q(\sigma_g^2)} > 0$
Iterate:
$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{\gamma}^S,\mathbf{g}^S)} = \left[\frac{a_e + \frac{N}{2}}{B_{q(\sigma^2)}}C'C + \mathcal{D}\right]^{-1}$
$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{\gamma}^S,\mathbf{g}^S)} = \left(\frac{a_e + \frac{N}{2}}{B_{q(\sigma^2)}}\right)\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{\gamma}^S,\mathbf{g}^S)}C'Y$
$B_{q(\sigma^2)} = b_e + \frac{1}{2}\left[\mathcal{SSY} + \operatorname{tr}\left\{C'C\right\}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{\gamma}^S,\mathbf{g}^S)}\right]$
$B_{q(\lambda_0)} = b_0 + \frac{1}{2}\left[\boldsymbol{\mu}_{q(\boldsymbol{\gamma}_0^S)}'\boldsymbol{\mu}_{q(\boldsymbol{\gamma}_0^S)} + \operatorname{tr}\left\{\frac{a_0 + \frac{1}{2}K_0}{B_{q(\sigma_0^2)}}\mathcal{P}\right\}\right]$
$B_{q(\lambda_1)} = b_1 + \frac{1}{2}\left[\boldsymbol{\mu}_{q(\boldsymbol{\gamma}_1^S)}'\boldsymbol{\mu}_{q(\boldsymbol{\gamma}_1^S)} + \operatorname{tr}\left\{\frac{a_1 + \frac{1}{2}K_1}{B_{q(\sigma_1^2)}}\mathcal{P}\right\}\right]$
$B_{q(\lambda_g)} = b_g + \frac{1}{2}\left[\boldsymbol{\mu}_{q(\boldsymbol{\gamma}_g^S)}'\boldsymbol{\mu}_{q(\boldsymbol{\gamma}_g^S)} + \operatorname{tr}\left\{\frac{a_g + \frac{1}{2}K_g}{B_{q(\sigma_g^2)}}\mathcal{P}\right\}\right]$
until changes in $\log[\underline{p}(\mathbf{y};q)]$ are negligible.

random distributed lag and $C = [X\ L(I_2 \otimes \Theta)\ U]$ for the random intercept.

C Replication Code

Our replication code can be found in the file `replication_code.R` alongside this manuscript.