Week 7 Notes: Models for Dependent Data

Antonio R. Linero
University of Texas at Austin

Motivation

Question: How do we deal with dependent data?

Examples:

1. Y_i's are measurements taken over space or time.
2. Y_i's are *clustered*, e.g., measurements taken on the same individual (repeated measures) or on different individuals in the same hospital/school/family.

Motivation

Question: How do we deal with dependent data?

Examples:

1. Y_i's are measurements taken over space or time.
2. Y_i's are *clustered*, e.g., measurements taken on the same individual (repeated measures) or on different individuals in the same hospital/school/family.

Options:

1. Random effects models (good for *conditional inference* or if dependence structure is of interest)
2. Estimating equations (good for *marginal inference* or if dependence is a nuisance)

Generalized Linear Mixed Effects Models

A generalized linear mixed effects model (GLMM) modifies the systematic component of a GLM by setting

$$g(\mu_i) = X_i^\top \beta + Z_i^\top U$$

where U is now modeled as *random*, with density $U \sim f(u \mid \gamma)$.

Example: Random Intercepts

Exercise: Simple Example

Consider the hierarchical model

$$Y_{ij} = \mu + \alpha_j + X_{ij}^\top \beta + \epsilon_{ij}$$

where $\alpha_j \sim \text{Normal}(0, \sigma_\alpha^2)$. Show that this model can be written in the GLMM form for some choice of U_i and Z_i.

Example: Spatial Models

Suppose $Y_i = Y(s_i)$ is a spatially indexed process. A common model for such processes sets

$$g\{\mu(s_i)\} = X_i^\top \beta + Z(s_i)^\top U$$

where $Z(s_i) = (Z_1(s_i), \ldots, Z_B(s_i))^\top$ is a *basis function expansion* of a spatial process.

Random Effects Distribution

Usually we take

$$U \sim \text{Normal}(0, \Sigma_\gamma),$$

where Σ_γ usually has some known form.

Alternatively: use NPMLE or DPs to estimate $f(u \mid \gamma)$ nonparametrically.

Fitting GLMMs for Bayesians

```
if(is_small(my_data)) {
  run_mcmc(my_data, my_model)
} else {
  run_vb(my_data, my_model)
}
```

Won't go into MCMC (`Stan` works well), and don't have time for a digression on variational Bayes (VB). VB is much faster than MCMC but can produce poorly-calibrated uncertainty estimates.

Fitting GLMMs for Frequentists

Goal: maximize

$$L(\beta, \gamma) = \int \prod_i f(Y_i \mid \theta_i, \phi/\omega_i)\ f(u \mid \gamma)\ du.$$

This is hard for GLMMs! Unlike LMMs, cannot usually compute this analytically.

Why Integrate?

Exercise: Neyman-Scott Problem

One might wonder why we feel the need to integrate out the random effects instead of (say) maximizing over them. Suppose that we have paired responses (Y_{i1}, Y_{i2}) such that

$$Y_{ij} = \mu_i + \epsilon_{ij} \qquad \text{where} \qquad \epsilon_{ij} \sim \text{Normal}(0, \sigma).$$

Think of Y_{ij}'s as representing two measurements of individuals on some test; our interest is in σ, which describes variation in individuals, but not the μ_i's (we don't care about learning about the particular individuals we sampled, but about the population).

Suppose we use a flat prior for the random effects, $f(\mu_i) \propto 1$ (this is done to make the computations easier, and doesn't affect the qualitative conclusions).

(a) Compute the MLE of σ obtained from optimizing the joint likelihood

$$\ell(\mu_1, \ldots, \mu_N, \sigma) = \sum_{i=1}^{N} \sum_{j=1}^{2} -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(Y_{ij} - \mu_i)}{2\sigma^2}.$$

Does this seem like a good estimate?

(b) Compute the MLE of σ obtained from optimizing the integrated log-likelihood

$$\ell(\sigma) = \log \prod_{i=1}^{N} \int \prod_{j=1}^{2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(Y_{ij} - \mu_i)^2}{2\sigma^2} \right\} d\mu_i.$$

How does this compare to our other answer?

Strategies

Possibilities:

1. Numeric integration: use Gaussian quadrature, or similar. (`glmer`, sometimes)
2. Monte Carlo integration: use expectation-maximization (EM), with expectations computed using MCMC.
3. Laplace approximation: $\log f(y \mid u) \approx \log f(y \mid \widehat{u}) - \frac{1}{2}(u-\widehat{u})^\top \mathcal{J}(u-\widehat{u})$, and now my integral is easy. (`glmer`, other times)

Strategies

Possibilities:

1. Numeric integration: use Gaussian quadrature, or similar. (`glmer`, sometimes)
2. Monte Carlo integration: use expectation-maximization (EM), with expectations computed using MCMC.
3. Laplace approximation: $\log f(y \mid u) \approx \log f(y \mid \widehat{u}) - \frac{1}{2}(u - \widehat{u})^\top \mathcal{J} (u - \widehat{u})$, and now my integral is easy. (`glmer`, other times)

Problems:

1. Useless if integral is higher than one or two dimensions.
2. Just go full Bayes, you coward.
3. Only as good as the Laplace approximation. (INLA and VB are attempts at improving this.)

Conditional Versus Marginal Effects

What does β represent?

$$\begin{aligned} g\{\mathbb{E}(Y)\} &= g[\mathbb{E}\{\mathbb{E}(Y \mid U)\}] \neq \mathbb{E}[g\{\mathbb{E}(Y \mid U)\}] \\ &= X^\top \beta + Z^\top E(U) = X^\top \beta. \end{aligned}$$

It represents the effect of X *conditional on* U! When are these the same?

A Damping Effect

Exercise: Conditional vs. Marginal

Suppose that Y is binary and let Φ be the cdf of a Normal$(0, 1)$ random variable. Consider the mixed effects probit model

$$\Pr(Y = 1 \mid z, x, \gamma, \beta) = \Phi(x^\top \beta + z^\top \gamma)$$

where $\gamma \sim \text{Normal}(0, \Sigma)$ is a random effect.

(a) Show that, in this case, the marginal model is also a probit model

$$\Pr(Y = 1 \mid z, x, \beta, \Sigma) = \Phi\left(\frac{x^\top \beta}{\sqrt{1 + z^\top \Sigma z}}\right)$$

Hence, the marginal model for the probit is also a probit model in which the covariate effect β is *dampened* in the marginal model by a factor of $\sqrt{1 + z^\top \Sigma z}$.
Hint: we can write the left hand side as

$$\int \Phi(x^\top \beta + z^\top \gamma)\, f(\gamma)\; d\gamma = \int \int I(\epsilon \leq x^\top \beta + z^\top \gamma)\, \phi(\epsilon)\, f(\gamma)\; d\gamma\; d\epsilon$$

and the right-hand-side is the expectation of $I(\epsilon - z^\top \gamma \leq x^\top \beta)$ where $\epsilon \sim \text{Normal}(0, 1)$ and $\gamma \sim \text{Normal}(0, \Sigma)$; what is the distribution of $\epsilon - z^\top \gamma$?

(b) Consider the special case where the conditional success probability is given by $\Phi(\beta_0 + \beta_1 x + \gamma)$ where $\gamma \sim \text{Normal}(0, 4)$. Let $\beta_0 = 0$ and $\beta_1 = 2$. First, plot the conditional success probability as a function of x for 20 randomly sampled values of γ as dashed lines. Then, plot the marginal success probability as a solid line. Comment on what you see.

Practice

Exercise: Polls

In `polls.csv` you will find the results of several political polls from the 1988 U.S. presidential election. The outcome of interest is whether someone plans to vote for George Bush. There are several potentially relevant demographic predictors here, including the respondent's state of residence. The goal is to understand how these relate to the probability that someone will support Bush in the election. You can imagine that this information would help a great deal in poll re-weighting and aggregation.

Using `STAN` (or the `stan_glmer` function in `rstanarm`), fit a hierarchical logit model of the form

$$Y_{ij} \sim \text{Bernoulli}(p_{ij}),$$

$$\pi_{ij} = \frac{\exp(\mu_j + X_{ij}^\top \beta)}{1 + \exp(\mu_j + X_{ij}^\top \beta)}$$

to this dataset. Here, Y_{ij} is the response (Bush = 1, other = 0) for respondent j in state i, μ_i is a state-level intercept, X_{ij} is a vector of respondent-level demographic predictors, and β is a state-invariant regression coefficient vector.

(a) Plot the mean and 95% credible interval for each state-level effect, ordered by their posterior mean.

(b) Which predictors appear to have the largest impact on the probability of an individual voting for Bush?

(c) (**Optional**) Consider making β a random effect, i.e., replace β with β_j. Is there any interesting variability in how the effect of the demographic predictors varies across states?

Practice

Exercise: Math Scores

The dataset in `mathtest.csv` shows the scores on a standardized math test from a sample of 10th grade students at 100 different U.S. urban schools, all having enrollment of at least 400 10th grade students. Let θ_i be the underlying mean test score for school i and let Y_{ij} be the score for the jth student in school i. You'll notice that the extreme school-level averages $\bar{Y}_i$ (both high and low) tend to be at schools where fewer students were sampled.

(a) Explain briefly why this would be.

(b) Consider a normal hierarchical model of the form

$$Y_{ij} \overset{\text{ind}}{\sim} \text{Normal}(\theta_i, \sigma^2)$$
$$\theta_i \sim \text{Normal}(\mu, \tau^2\sigma^2).$$

Write a function that fits this model by (approximately) sampling from the posterior distribution of $(\theta_1, \ldots, \theta_{100}, \mu, \sigma^2, \tau^2)$. Your function should be of the form (assuming the use of `R`)

```
fit_oneway_anova <- function(y, treatment,
                             num_warmup, num_save, num_thin) {
  ## Input:
  ##   Y: a vector of length n of observations
  ##   treatment: a vector indicating what treatment was
  ##                received (in this case, which school)
  ##   num_warmup: the number of iterations to discard to burn-in
  ##   num_save: the number of samples to collect
  ##   num_thin: the thinning interval of the chain
  ##
  ## Your code here...
  return(list(theta = theta_samples, mu = mu_samples,
              sigma = sigma_samples, tau = tau_samples))
}
```

Practice

Exercise: Baseball

In 1977, Efron and Morris analyzed data from the 1970 Major League Baseball (MLB) season. They took the batting average of 18 players over the first 45 at-bats. Let Y_i be the number of hits player i obtained over their first 45 attempts; then a sensible model for the number of hits might be

$$Y_i \sim \text{Binomial}(45, p_i), \qquad \text{where} \qquad p_i \sim \text{Beta}\{\rho\mu, \rho(1-\mu)\}.$$

(a) Write a function to fit a hierarchical model with $(\mu, \rho) \sim \pi(\mu, \rho)$. Specify whatever priors for μ and ρ you believe to be reasonable. Your code should be of the form:

```
fit_beta <- function(y, n, num_warmup, num_save, num_thin) {
  ## Your code here ...
  return(your_fitted_model) # nolint
}
```

(b) Compare the mean squared error of the UMVUE estimate $\widehat{p}_{i,\text{UMVUE}} = Y_i/45$ to the Bayes estimator of $\widehat{p}_{i,\text{Bayes}}$ that you get from the posterior. Which performs better?

(c) Interpret the hyperparameters μ and ρ; practically speaking, what information do these hyperparameters encode?

Digression: The Normal Means Problem

The *normal means problem* sets

$$[Y_i \mid \boldsymbol{\mu}] \overset{\text{ind}}{\sim} \text{Normal}(\mu_i, 1),$$

for $i = 1, \ldots, N$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$.

Goal: estimate $\boldsymbol{\mu}$ so as to make $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2$ as small as possible.

Digression: The Normal Means Problem

The *normal means problem* sets

$$[Y_i \mid \boldsymbol{\mu}] \overset{\text{ind}}{\sim} \text{Normal}(\mu_i, 1),$$

for $i = 1, \ldots, N$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)$.

Goal: estimate $\boldsymbol{\mu}$ so as to make $\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2$ as small as possible.

Applications:

1. ANOVA
2. Nonparametric estimation in the wavelet domain
3. High-dimensional multiple testing ($\boldsymbol{\mu}$ assumed sparse)

Surprising Fact

Definition (Admissibility)

An estimator $\widehat{\boldsymbol{\mu}}$ is called *admissible* if there does not exist a different estimator $\widetilde{\boldsymbol{\mu}}$ such that

$$\mathbb{E}_\mu\left(\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2\right) \leq \mathbb{E}_\mu\left(\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2\right)$$

for **all** values of μ.

Surprising Fact

Definition (Admissibility)

An estimator $\widehat{\boldsymbol{\mu}}$ is called *admissible* if there does not exist a different estimator $\widetilde{\boldsymbol{\mu}}$ such that

$$\mathbb{E}_\mu \left(\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \right) \leq \mathbb{E}_\mu \left(\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \right)$$

for **all** values of μ.

Fact: The obvious estimator $\widehat{\boldsymbol{\mu}} = \boldsymbol{Y}$ (which is the UMVUE, best equivariant estimator, and MLE) is NOT ADMISSIBLE when $N \geq 3$.

James-Stein Estimator

For $N \geq 3$, the estimator

$$\widehat{\boldsymbol{\mu}}_{\text{JS}} = \left(1 - \frac{(N-2)}{\|\boldsymbol{Y}\|_2^2}\right) \boldsymbol{Y}$$

dominates $\boldsymbol{Y}$.

(**Optional:**) Show that this estimator arises from a *random effects model* with $\mu_i \overset{\text{iid}}{\sim} \text{Normal}(0, \sigma_\mu^2)$ where the shrinkage factor is replaced with an unbaised estimator

$$\frac{1}{\sigma_\mu^2 + 1} \approx \frac{N-2}{\|\boldsymbol{Y}\|^2}.$$

Exercise

Exercise: Random or Fixed Effects?

Consider $n = 5$ and $\mu = (1, 1, 3, 3, 5)/5$ and let $Y \sim \text{Normal}(\mu, \text{I})$. Conduct a simulation experiment comparing the mean squared error in estimating μ, $\mathbb{E}\{\|\mu - \widehat{\mu}\|^2\} = \sum_j \mathbb{E}\{(\mu_j - \widehat{\mu}_j)^2\}$ of the following estimators:

1. The maximum likelihood estimator $\widehat{\mu} = Y$.
2. The *predicted value* of μ_j given by $\mathbb{E}(\mu_j \mid Y) = \frac{\nu^2 Y_j}{1+\nu^2}$ with the random effect distribution $\mu_j \sim \text{Normal}(0, \nu^2)$. Estimate ν with its MLE after integrating out μ. *Hint:* the MLE of ν^2 is $\max\{\frac{\|Y\|^2}{5} - 1, 0\}$, but you need to show this.

Repeat this over 1000 replications for each estimator. How do the methods compare? Note that the MLE has many "desirable" properties — it is minimax optimal, it is the UMVUE, and it is the best invariant estimator.

Exercise

Exercise: Normal Means in High Dimensions

Consider the model $Z_i \sim \text{Normal}(\mu_i, 1)$ (conditional on μ_i) where $i = 1, \ldots, P$ and P is very large (say, 10,000). A-priori, we expect many of the μ_i's to be zero; this might be reasonable, for example, in genomic problems where the Z_i's represent test statistics corresponding to P different genes, where we expect that most genes are unrelated to the response we are interested in. We consider a hierarchical model

$$\mu_i \sim p \cdot \text{Normal}(0, \tau^2) + (1 - p) \cdot \delta_0,$$

where δ_0 is a point mass distribution at 0. That is, with probability p, μ_i is non-zero (in which case it has a normal distribution) and, with probability $1 - p$, μ_i is identically zero.

(a) Suppose that p is known. Show that the marginal distribution of Z_i is a mixture of two normal distributions,

$$m(Z_i) = p \cdot \text{Normal}(0, 1 + \tau^2) + (1 - p) \cdot \text{Normal}(0, 1).$$

(b) Given $Z_i = z$, show that the posterior probability that $\mu_i = 0$ is

$$\Pi(\mu_i = 0 \mid Z_i = z) = \frac{(1 - p) \cdot \text{Normal}(z \mid 0, 1)}{p \cdot \text{Normal}(0 \mid \tau^2 + 1) + (1 - p) \cdot \text{Normal}(z \mid 0, 1)}.$$

(c) One might be tempted to use an "uninformative prior" in this setting, taking $\tau \to \infty$. What happens to the posterior probability in part (b) if you do this? Explain.

(d) Find the value of τ^2 which minimizes $\Pi(\mu_i = 0 \mid Z_i = z)$. Show that the posterior odds of $\mu_i \neq 0$ is given by

$$O = \frac{p}{|z|(1 - p)} \exp\left\{\frac{1}{2}(z^2 - 1)\right\}$$

Generalized Estimating Equations

What if I don't care about conditional inference?

For clustered data $\{Y_{ij} : i = 1, \ldots, N_j, j = 1, \ldots J\}$ with marginal model $g\{E(Y_{ij} \mid X_{ij}, \beta)\} = X_{ij}^\top \beta$, consider a *generalized estimating equation* (GEE):

$$\sum_{j=1}^{J} \frac{\partial \boldsymbol{\mu}_j^\top}{\partial \beta} V_j^{-1} (\boldsymbol{Y}_j - \boldsymbol{\mu}_j) = \mathbf{0}_P.$$

where $V_j = V_j(\boldsymbol{\mu}_j, \alpha)$ is a *working covariance matrix* for the $\boldsymbol{Y}_j$'s.

GEE Properties

- If $\boldsymbol{\mu}_j$ is correctly specified, estimator usually is consistent/satisfies a CLT.
- If $V_j(\boldsymbol{\mu}_j, \alpha)$ is also correctly specified, estimator will be *efficient*.
- Robust standard errors (based on sandwich matrix) used for inference, works even if V_j is totally misspecified.
- Generalizes the quasi-likelihood estimating equation

$$\sum_{i=1}^{N} \frac{\omega_i \left(Y_i - \mu_i\right)}{\phi \, V(\mu_i) \, g'(\mu_i)} = \mathbf{0}$$

Form of Working Variance

Working variance is parameterized as

$$V_j = D_j^{1/2} \, R_j(\alpha) \, D_j^{1/2}$$

where

$$D_j = \text{diag}\left(\frac{\phi}{\omega_{ij}} V(\mu_{ij}) : i = 1, \ldots, N_j \right)$$

and $R_j(\alpha)$ is a *working correlation matrix*.

Choices of Working Correlation

Take $R_j(\alpha)$ to be

$$\underbrace{\begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix}}_{\text{exchangeable}} \quad \text{or} \quad \underbrace{\begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{N_j-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{N_j-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{N_j-1} & \alpha^{N_j-2} & \alpha^{N_j-3} & \cdots & 1 \end{pmatrix}}_{\text{AR1}}$$

depending on context.

Note: our motivation for getting these nearly correct is efficiency!

Exercise

Exercise: Ticks Revisited

Recall the `ticks` dataset from the previous set of notes and consider a model with a linear predictor containing the terms `YEAR`, `HEIGHT`, and `YEAR:HEIGHT`. In this exercise, we will compare a model that takes

$$\mathbb{E}_\theta(Y_{ij} \mid X_i = x) = \exp(\alpha + x^\top \beta) \tag{1}$$

to a hierarchical model that takes

$$[Y_{ij} \mid X_{ij} = x, \alpha_j, \beta] \overset{\text{ind}}{\sim} \text{Poisson}\{\exp(\alpha + b_j + x^\top \beta)\}, \qquad b_j \sim \text{Normal}(0, \sigma_b^2). \tag{2}$$

Note that the first model does not explicitly make a statement about the dependence structure within clusters, while the second specifies the full joint distribution of the Y_{ij}'s.

a. Show that the Poisson random effects model (2) is a special case of the model (1) in the sense that if (2) is true then (1) is also true. **NOTE:** this is very important, as otherwise we would not be able to apples-to-apples comparisons of the inferences between the two models.

b. Use the `stan_glmer` function in the `rstan` package to fit (2) to the `ticks` data. Then, plot the posterior distribution of the `HEIGHT` coefficient; is there evidence that this coefficient is non-zero?

c. Briefly, state which correlation structure seems best suited to this data (AR1 or exchangeable)? Justify your answer.

d. Using the exchangeable correlation structure, use the `geeglm` function in the `geepack` package to fit (1) using a GEE. Compare the standard error reported here with the standard error of the Poisson GLMM for the `HEIGHT` coefficient.

e. Repeat part (d), but use the independence correlation structure. How does the standard error compare across the two models? Which correlation structure would you recommend. **NOTE:** Of course, you shouldn't choose correlation structures according to which inferences you prefer after the fact...

Cons of GEEs

- Marginal, rather than conditional, inference (not necessarily a bad thing)
- No estimate of the data generating mechanism
- Can be difficult to check/critique
- Not easy to do a Bayesian version