Bayesian Methods for Addressing Two Missing Data Problems

Fletcher G.W. Christensen

Department of Statistics
University of California, Irvine

November 24, 2017

Outline of Presentation

1. The Motivation
2. Joint Parameter Modeling
3. Detection-Dependent Binary Response Modeling
4. Conclusion

The Study

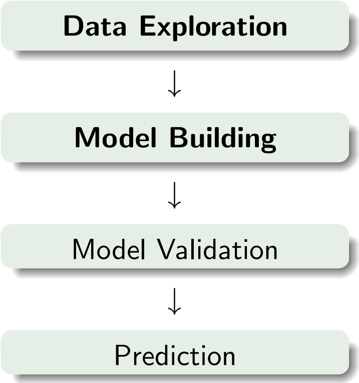
- The methods I'm going to discuss today arose while I was working on a problem with Dr. Ulrike Luderer from the UCI Center for Occupational and Environmental Health.
- We looked at the effect of poly-aromatic hydrocarbons (PAHs, a form of atmospheric pollution) on various menstrual cycle outcomes.
- Full details are available in "Associations between urinary biomarkers of polycyclic aromatic hydrocarbon exposure and reproductive function during menstrual cycles in women" by U. Luderer et al. (2017), in *Environment International*.

Luderer et al, 2017

The Data

- Longitudinal data ($\sim$6 menstrual cycles per woman) collected on 51 women.
- Response variables are all functionals of cycle-long hormone trajectory vectors on luteinizing hormone (LH) and estrone 3-glucuronide ($E_1$3G).
- Longitudinal pollutant exposure data on 9 hydroxylated PAH compounds—the end result of the body's metabolization process on PAHs, excreted via urine.
- A range of demographic information recorded once for each woman at the start of the study.

The Goal



- Previous work had linked PAHs to fertility in animals, and more granular pollution measures to human fertility.
- This was the first major study examining how the human reproductive system is affected by pollutants on the PAH level.
- We wanted to establish whether PAHs are important to this topic and begin looking at their role.
- Response data were collected with a new research tool, and we hoped to show it could be used effectively.

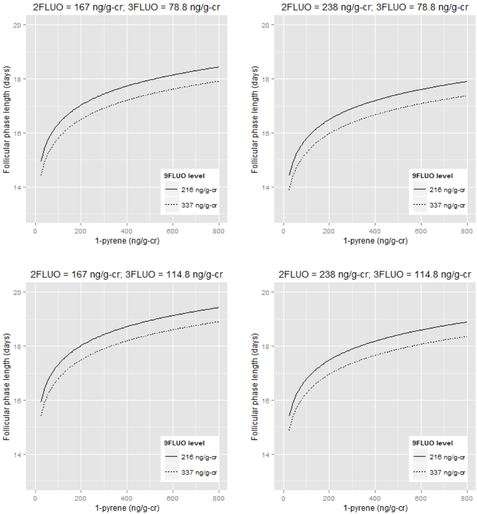
The Results

- Models including both demographic and PAH covariates generally outperformed models with demographic covariates alone.
- This included models involving smoking status, the most common avenue for environmental PAH exposure.
- One interesting finding is that the profile of related PAH metabolites may relate to endocrine outcomes.
- That is, differences in how women's bodies process environmental pollutants—rather than the quantity of the pollutants themselves—may have an important role in predicting the outcomes we considered.

Luderer et al, 2017

The Results

Figure 1. Follicular phase length as a function of OH-PAH levels.



Complication 1 – Missing Covariate Information

- PAH data are obtained through expensive laboratory analyses of urine or stool samples.
- Dr. Luderer and colleagues were able to collect samples for each menstrual cycle for each woman—but due to funding limitations, lab analyses are only available for about half of them.
- Philosophically, we don't believe in throwing away good data if it can be helped.
- Given that the relationships between response measures and PAH covariates are our primary scientific interest, we want a method that lets us take advantage of the observations where we don't have PAH data as well.

Complication 2 – Missing Response Information

- Response measures are all functionals of vectors of daily hormone measurements. These vectors are commonly called *trajectories*.
- Hormone trajectory data were collected by participants in their homes, using a urine-based fertility monitor each morning.
- There is considerable cycle-to-cycle variability in the proportion of days per cycle when data were collected.
- One of our response variables, ovulation, is binary. Because of the nature of the functional for calculating ovulation status, the probability on the outcome is dependent on the proportion of trajectory data observed. (A Yes is easier to see than a No.)

The Easy Way to Think About It – Two Models

- Consider fitting two linked models to these data: one to the observations with PAH information, and one to the observations without PAH information.
- These models will share parameters across the PAH and non-PAH groups when possible. These include coefficients on the demographic covariates, and subject-level random effects.
- Other parameters will be model-specific: coefficients for PAHs, and the additional error variance in the reduced model where no PAHs are observed.

The Easy Way to Think About It – Two Models

- Let $i = 1, \ldots, k$ index the women in our study and $j = 1, \ldots, n_{i_p}, \ldots, (n_{i_p} + n_{i_c})$ index the longitudinal observations on each woman. Let n_{i_p} be the number of observations per woman for which we lack PAH exposure data, and let n_{i_c} be the number of observations per woman for which we have such information.
- Then a simple two-model approach would give us

$$\begin{aligned} Y_{ij} &= X_{ij}\beta & &+\xi_i + \varepsilon_{ij} + \epsilon_{ij} & j &\in \{1, \ldots, n_{i_p}\} \\ Y_{ij} &= X_{ij}\beta + Z_{ij}\gamma & &+\xi_i + \varepsilon_{ij} & j &\in \{(n_{i_p}+1), \ldots, (n_{i_p} + n_{i_c})\} \end{aligned}$$

- Note that now β, ξ, and ε are elements of both models.

The Easy Way to Think About It – Two Models

A Note on Parameters:

$$\begin{aligned} Y_{ij} &= X_{ij}\beta & &+\xi_i + \varepsilon_{ij} + \epsilon_{ij} & j &\in \{1, ..., n_{i_p}\} \\ Y_{ij} &= X_{ij}\beta + Z_{ij}\gamma & &+\xi_i + \varepsilon_{ij} & j &\in \{(n_{i_p}+1), ..., (n_{i_p}+n_{i_c})\} \end{aligned}$$

- β is the collection of model parameters for the data we observe on every unit.
- γ is the collection of model parameters for the partial covariate data—the data we only observe on a subset of units.
- ξ is the subject-specific random effect.
- ε is the error in the complete model.
- ϵ is the additional error in the model when only partial covariate information is available.

The Matrix Representation

- Linear algebraic notation also helps us understand what's happening here. Assume woman i has six observations, with $n_{i_p} = n_{i_c} = 3$. Then we can write

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \\ Y_{i5} \\ Y_{i6} \end{bmatrix} = \begin{bmatrix} X_{i1} & 0 \\ X_{i2} & 0 \\ X_{i3} & 0 \\ X_{i4} & Z_{i4} \\ X_{i5} & Z_{i5} \\ X_{i6} & Z_{i6} \end{bmatrix} \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} \xi_i \\ \xi_i \\ \xi_i \\ \xi_i \\ \xi_i \\ \xi_i \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \\ \varepsilon_{i5} \\ \varepsilon_{i6} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The Matrix Representation

- Put another way, we can say that Y_i from the last slide has a multivariate normal distribution, with

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \\ Y_{i5} \\ Y_{i6} \end{bmatrix} \sim N\left(\begin{bmatrix} X_{i1}\beta \\ X_{i2}\beta \\ X_{i3}\beta \\ X_{i4}\beta + Z_{i4}\gamma \\ X_{i5}\beta + Z_{i5}\gamma \\ X_{i1}\beta + Z_{i6}\gamma \end{bmatrix}, \sigma^2\mathcal{I}_6 + \tau^2 \begin{bmatrix} \mathcal{I}_3 & 0_3^3 \\ 0_3^3 & 0_3^3 \end{bmatrix} + \rho\mathcal{J}_6^6 \right)$$

- Where $\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ is the error distribution for the full model, $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \tau^2)$ is the distribution of additional errors for the partial model, and $\xi_i \overset{\text{iid}}{\sim} N(0, \rho)$ is the random effect distribution.

Assumptions of Shared-Parameter Modeling

$$\begin{array}{llr} Y_{ij} = X_{ij}\beta & +\xi_i + \varepsilon_{ij} + \epsilon_{ij} & j \in \{1, \ldots, n_{i_p}\} \\ Y_{ij} = X_{ij}\beta + Z_{ij}\gamma & +\xi_i + \varepsilon_{ij} & j \in \{(n_{i_p}+1), \ldots, (n_{i_p}+n_{i_c})\} \end{array}$$

- In addition to the standard linear modeling assumptions, two additional assumptions are implied by the shared-parameter model. These assumptions ensure that β plays the same role in both equations.
  1. Z is centered, or includes an intercept – Without this, β_0 in the complete-covariate and partial-covariate models will differ. BLUEs for each subset of the data would necessarily estimate different values for β, rather than a shared value.
  2. $X \perp\!\!\!\perp Z$ – More generally, this assumption ensures that the βs estimated are equivalent between models. This is a strong assumption, however, and we will look at how robust the method is to violation.

Effects of Shared-Parameter Modeling

- The largest benefits of shared-parameter modeling are in estimating subject-specific random effects, which leads to substantial improvement in out-of-sample prediction accuracy on previously observed subjects.
- There are smaller benefits to the precision of estimates of the partial data parameters.
- When $X \perp\!\!\!\perp Z$ holds, there are also substantial improvements in the precision and accuracy of estimates of complete data parameters.

Monte Carlo Sample Construction

- By examining the linear algebraic mechanics of the shared parameter model, we developed suspicions about the scenarios under which it would show the greatest improvements over alternative modeling strategies.
- We generate longitudinal data for k units, with n_p partial observations and n_c complete observations per unit.
- Two covariates are generated for each of the $k(n_p + n_c)$ observations, and information on one of the covariates is removed for the kn_p partial observations after response values are calculated.
- Response data are generated based on the two covariates, the observational unit, and an additional error term.

Monte Carlo Sample Construction

- In our Monte Carlo analysis, we considered three candidate models:
  1. The shared-parameter model that I've discussed.
  2. A model where observations without the partial data are dropped from the sample.
  3. A model where the partial covariates themselves are removed, leaving the full sample size.
- Here we compare fit only under the constraint that the full model is true.

What We're Looking For

- We are primarily interested in the out-of-sample prediction error for these techniques—does shared-parameter modeling give appreciably better predictions than we would get if we simply ignored the partial observations?
- We are also interested in how parameter estimates change (in both value and precision) between our two candidate models.
- We want to know how simulation parameters (within-subject sample sizes n_p and n_c, within-subject correlation ψ, and correlation between covariates ϕ) affect estimator precision and predictive accuracy.

Predictive Accuracy Depends on Correlation

ψ	ϕ	Shared-Parameter Model	Observations Removed	Covariates Removed
0.0	0.0	1.015	1.022	2.031
	0.3	1.046	1.040	1.925
	0.7	1.112	1.022	1.543
0.3	0.0	1.162	1.218	2.167
	0.3	1.178	1.216	2.102
	0.7	1.235	1.218	1.657
0.7	0.0	1.217	1.331	2.262
	0.3	1.222	1.309	2.180
	0.7	1.280	1.312	1.733

Average prediction error when $k = 50$ and $n_c = n_i = 3$. True parameter values are $\beta = \gamma = 1$.

Predictive Accuracy Depends on Amount of Data

n_c	n_p	Shared-Parameter Model	Observations Removed	Covariates Removed
3	3	1.217	1.331	2.262
	10	1.145	1.339	2.171
	30	1.072	1.337	2.051
5	3	1.167	1.213	2.291
	10	1.111	1.199	2.146
	30	1.052	1.194	2.049
10	3	1.096	1.110	2.179
	10	1.067	1.094	2.079
	30	1.005	1.059	2.035

Average prediction error when $k = 50$, $\psi = 0.7$, and $\phi = 0.0$. True parameter values are $\beta = \gamma = 1$.

Takeaways on Prediction

- Predictive accuracy is up to 20% better for shared-parameter modeling when between-subject variability is comparable to within-subject variability.
- Predictive accuracy also improves more when n_c is small and n_p is large.
- Although collinearity creates bias in parameter estimates, predictive accuracy still improves under shared-parameter modeling for mild collinearity.

Parameter Distributions Depend on Correlation

		Shared-Parameter Model				Observations Removed			
ψ	ϕ	$E[\beta\|\mathbb{X}]$	$sd[\beta\|\mathbb{X}]$	$E[\gamma\|\mathbb{X}]$	$sd[\gamma\|\mathbb{X}]$	$E[\beta\|\mathbb{X}]$	$sd[\beta\|\mathbb{X}]$	$E[\gamma\|\mathbb{X}]$	$sd[\gamma\|\mathbb{X}]$
0.0	0.0	1.003	0.067	0.991	0.083	1.000	0.083	0.990	0.083
	0.3	1.111	0.070	0.960	0.087	0.998	0.087	0.994	0.088
	0.7	1.409	0.084	0.701	0.104	1.006	0.116	0.987	0.116
0.3	0.0	1.000	0.072	1.000	0.089	1.007	0.093	0.995	0.091
	0.3	1.105	0.074	0.957	0.094	0.989	0.096	0.988	0.098
	0.7	1.406	0.088	0.720	0.110	1.012	0.130	0.990	0.129
0.7	0.0	1.001	0.073	0.993	0.093	1.002	0.098	0.992	0.098
	0.3	1.115	0.076	0.962	0.096	1.002	0.103	1.001	0.104
	0.7	1.413	0.089	0.707	0.114	1.007	0.137	0.988	0.137

Posterior characteristics of β and γ when $k = 50$ and $n_c = n_i = 3$.
True parameter values are $\beta = \gamma = 1$.

Parameter Distributions Depend on Amount of Data

		Shared-Parameter Model				Observations Removed			
n_c	n_p	$\mathrm{E}[\beta\|\mathbb{X}]$	$\mathrm{sd}[\beta\|\mathbb{X}]$	$\mathrm{E}[\gamma\|\mathbb{X}]$	$\mathrm{sd}[\gamma\|\mathbb{X}]$	$\mathrm{E}[\beta\|\mathbb{X}]$	$\mathrm{sd}[\beta\|\mathbb{X}]$	$\mathrm{E}[\gamma\|\mathbb{X}]$	$\mathrm{sd}[\gamma\|\mathbb{X}]$
3	3	1.001	0.073	0.993	0.093	1.002	0.098	0.992	0.098
	10	1.008	0.052	1.003	0.087	1.015	0.099	1.000	0.098
	30	1.003	0.034	1.000	0.085	1.013	0.975	0.996	0.098
5	3	1.006	0.060	1.001	0.069	1.005	0.071	1.000	0.071
	10	1.002	0.046	1.003	0.067	1.002	0.070	1.003	0.070
	30	0.995	0.032	1.002	0.065	0.992	0.070	1.000	0.071
10	3	1.005	0.043	1.000	0.047	1.003	0.047	1.001	0.047
	10	0.995	0.038	0.996	0.047	0.997	0.047	0.996	0.047
	30	1.001	0.029	0.998	0.046	0.996	0.047	0.998	0.047

Posterior characteristics of β and γ when $k = 50$, $\psi = 0.7$, and $\phi = 0.0$. True parameter values are $\beta = \gamma = 1$.

Takeaways on Parameter Estimation

- Even small amounts of collinearity create bias in the parameter estimates.
- Correlation between observations on the same unit increases the precision of the posteriors on the parameter distributions.
- Precision improves more when n_c is small and n_p is large.
- Considerable improvement can be obtained in precision on the βs. Improvement is made on the precision of the γs, but this is considerably smaller.

Where Might This Be Useful?

- This technique is best applied to longitudinal data where some of the covariates are only available on a subset of observations for each unit.
- These issues are most likely to arise in settings where some data are very expensive to obtain.
- One application area I believe this may be particularly well-suited for is in monitoring chronic conditions, as a way to develop retrospective subject-specific baselines immediately following diagnosis, or when some critical medical data will be difficult to collect frequently.

Revisiting Our Data

- Recall that in the PAH and endocrine function data, one of our response variables of interest was whether a given cycle was ovulatory or anovulatory.
- Recall also that in our data, ovulation status is a functional of the LH hormone trajectory, and that trajectory information is incomplete for many cycles.
- While our missing data under the first complication is (surprisingly) MCAR, missingness here is MNAR—and thus much more problematic.

How Did the Detection Dependence Arise?

- Our functional for ovulation status involves the detection of an "LH surge"—a distinct feature in the menstrual hormone trajectory that occurs when a woman's body releases the necessary hormones to provoke ovulation and the creation of a corpus luteum.
- The LH surge lasts approximately one day. With incomplete data collection, it is easy to miss the occurrence of this feature.
- We need to see nearly a full trajectory to decide ovulation didn't happen. We can make a decision that it did happen, however, with as little as two days of data.

Some Example Hormone Trajectories

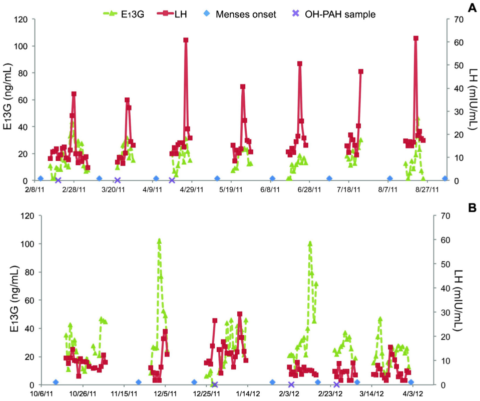


Figure 2. Representative urinary LH and $E_1$3G concentrations for two participants across multiple menstrual cycles.

What is Detection Dependence?

- Consider the setting of estimating prevalence, π, for some condition $\mathcal{C}$. (Let $\overline{\mathcal{C}}$ represent the absence of condition $\mathcal{C}$.)
- We have a test that, given sufficient data, can identify $\mathcal{C}$ with perfect accuracy.
- We know that in many cases our data are not sufficient to make a determination about $\mathcal{C}$—that is, we fail to detect either $\mathcal{C}$ or $\overline{\mathcal{C}}$.
- Let $\mathcal{D}$ represent detection—of either $\mathcal{C}$ or $\overline{\mathcal{C}}$. Let $\overline{\mathcal{D}}$ represent non-detection—the event that we are unable to determine presence or absence of the condition from our data.

What is Detection Dependence?

- Detection dependence arises when $\Pr[\mathcal{C}|\mathcal{D}]$ and $\Pr[\mathcal{C}]$ are not equal.
- In the case of perfect information, where $\Pr[\mathcal{D}] = 1$, then $\Pr[\mathcal{C}|\mathcal{D}] = \Pr[\mathcal{C}]$. But when detection is not assured, these probabilities differ.
- The present work shows how supplemental information on the adequacy of the data to make a detection can be used to enhance estimation of $\Pr[\mathcal{C}]$ in a Bayesian setting.

Detection Dependence in Ovulation Data

- Again, the key element here is that it is much easier to conclusively determine that an LH surge occurred—you only need to see two days, if you see the correct two days—than to determine that one did not occur.
- Technically, this is the relation $\Pr[\mathcal{D}|\mathcal{C}] \geq \Pr[\mathcal{D}|\overline{\mathcal{C}}]$, not the relation described in the definition of detection dependence. Their equivalence is straightforward to show, however.

Proposition 1

Let $Pr[\mathcal{D}|\mathcal{C}] \geq Pr[\mathcal{D}|\overline{\mathcal{C}}]$*, and define* $Pr[\mathcal{C},\mathcal{D}] + Pr[\mathcal{C},\overline{\mathcal{D}}] = \pi$*.*
Then $Pr[\mathcal{C}|\mathcal{D}] \geq \pi \geq Pr[\mathcal{C}|\overline{\mathcal{D}}]$*.*

Quantifying Detectability

- Let Y be a collection of test data, of various levels of adequacy for detection.
- We will index Y according to the data's adequacy to detect. Assume that there exist $\mathcal{I}$ distinct levels of adequacy, and let $\mathcal{G}_i, i \in \{1, ..., \mathcal{I}\}$ be used to index those levels of adequacy.
- Now, for each $\mathcal{G}_i$, define the following two-way table:

	$\mathcal{C}$	$\overline{\mathcal{C}}$	
$\mathcal{D}$	y_i	u_i	n_i
$\overline{\mathcal{D}}$	z_i	v_i	m_i
	o_i	p_i	N_i

Where N_i is the total number of tests in the $\mathcal{G}_i$ sufficiency level, and the other variables are as indicated by the table.

Probabilities of Interest

Now define the following probabilities of interest:

$$\begin{aligned}
\mathsf{Pr}[\mathcal{D}|\mathcal{G}_i] &= \delta_i \\
\mathsf{Pr}[\mathcal{C}|\mathcal{D},\mathcal{G}_i] &= \theta_i^{D} \\
\mathsf{Pr}[\mathcal{C}|\overline{\mathcal{D}},\mathcal{G}_i] &= \theta_i^{\overline{D}}
\end{aligned}$$

And recall that we have, from before:

$$\begin{aligned}
\mathsf{Pr}[\mathcal{C}] &= \mathsf{Pr}[\mathcal{C}|\mathcal{G}_i] \\
&= \mathsf{Pr}[\mathcal{C},\mathcal{D}|\mathcal{G}_i] + \mathsf{Pr}[\mathcal{C},\overline{\mathcal{D}}|\mathcal{G}_i] \\
&= \theta_i^{D}\delta_i + \theta_i^{\overline{D}}(1-\delta_i) \\
&= \pi
\end{aligned}$$

Identifiability Concerns

- Note that within the i-th table, we have three parameters but only two independent pieces of information. So within the i-th table, $\theta_i^{\overline{D}}$ is non-identifiable.
- Because of our formulation of the problem, however, knowing $\theta_i^{\overline{D}}$ is equivalent to knowing π. That is, if we know the probability of detection, and if we know the conditional probability of having the condition given detection, and if we know the population-level probability of having the condition, then we would also know $\theta_i^{\overline{D}}$ with perfect accuracy.
- This gives us a model with many constraints, but at most one non-identifiable parameter across all $\mathcal{I}$ tables.

Identifiability in Bayesian Modeling

- Bayesian models can be constructed, even for non-identifiable problems—but it is important to understand how identifiability impacts them.
- Here, we are modeling the parameters $\{\pi, \theta^D, \delta\}$ where θ^D and δ represent the appropriate vectors of θ_i^D and δ_i parameters.
- In our construction thus far, π is non-identifiable. That has an important meaning for how a Bayesian model will deal with π.

Posteriors for Non-Identifiable Parameters

Suppose the likelihood is free of π. Then:

$$\begin{aligned} p(\pi, \theta, \delta | Y) &= \frac{L(\pi, \theta, \delta | Y) p(\pi, \theta, \delta)}{f(Y)} \\ &= \frac{L(\theta, \delta | Y) p(\theta, \delta) p(\pi | \theta, \delta)}{f(Y)} \\ &= p(\theta, \delta | Y) p(\pi | \theta, \delta) \end{aligned}$$

- So the posterior distribution for π conditional on θ and δ does not involve the data in any way. It is equivalent to its prior.
- But the posterior for θ and δ does involve the data, so our beliefs about π can be updated indirectly.

Taking a Step Back

- Recall that our $\Pr[\mathcal{C}] = \Pr[\mathcal{C}|\mathcal{G}_i]$ constraint reduced us from $\mathcal{I} + 1$ non-idenifiable parameters to just one.
- Constraining models is a standard approach to reducing or eliminating identifiability issues, but there are other approaches.
- Hui and Walter (1980) resolved the problem of non-identifiability in diagnostic tests without a gold standard by adding information on a second test and a second population.
- Steinberg and Cardell (1992) addressed the non-identifiability created by observing only one level of a binary response by supplementing their data with an outside sample when finite-population sampling rates are available.

A Combination of Methods

- We augment our binary response data with information from the non-detection set and a measure of data adequacy to address identifiability concerns.
- In addition to the constraint previously discussed, we can also constrain our model such that the detection rate, δ_i increases as a function of the data's adequacy for detection.
- We operationalize adequacy for detection using a finite measure, $\mu(\mathcal{G}_i) \equiv w_i$.
- We will use $\mu(\mathcal{G}_i)$ in our prior specification for δ. The key feature of the measure is that it should accurately reflect an ordering among the $\mathcal{G}_i$'s according to data adequacy for detection.

An Initial Model

Recall our two-way table for the information contained in $\mathcal{G}_i$:

	$\mathcal{C}$	$\overline{\mathcal{C}}$	
$\mathcal{D}$	y_i	u_i	n_i
$\overline{\mathcal{D}}$	z_i	v_i	m_i
	o_i	p_i	N_i

Then we propose the model:

$$y_i \sim Bin(n_i, \theta_i^{D})$$

$$m_i \sim Bin(N_i, 1 - \delta_i)$$

$$z_i \sim Bin(m_i, \theta_i^{\overline{D}})$$

An Initial Model

Our choice of priors reflects the logical constraints we place on the model:

$$\begin{aligned}
\operatorname{logit}(\pi) &\sim N(a_0, b_0) \\
\operatorname{logit}(\delta_1) &\sim N(a_1, b_1) \\
\operatorname{logit}(\delta_{i+1}) - \operatorname{logit}(\delta_i) &\sim Exp\left(\frac{c_1}{w_{i+1} - w_i}\right) \\
\operatorname{logit}(\theta_i^D) - \operatorname{logit}(\pi) &\sim Exp\left(\frac{c_2}{1 - w_i}\right) \\
\theta_i^{\overline{D}} &= \frac{\pi - \theta_i^D \delta_i}{1 - \delta_i}
\end{aligned}$$

Where We Are Now

- We are currently in the process of developing coherent models that approximate satisfying the set of constraints we impose.
- We are looking at relaxing some of our assumptions to allow the model to be used with standard Gibbs software, and seeing if we'll still get the results we want.
- The method should be tractable with a hand-coded Gibbs or Metropolis algorithm, though this could limit its usefulness for applied researchers.

Marginal DIC Estimation

- Model comparison in the original paper used the marginalized deviance information criterion (DIC).
- Accurate DIC calculation requires careful attention to focal and non-focal stochastic elements of a Bayesian model, something not provided by naive software implementations.
- We are working toward a computationally efficient method for taking existing MCMC iterates and improving DIC estimation based on focal element choice.
- We also want to better understand what, precisely, is being calculated as DIC by common software packages, and how it relates to the properly marginalized value.

Classification Based On Longitudinal Trajectories

- We're also interested in using Bayesian non-parametrics to explore classification methods that rely on full hormone trajectory information instead of just functionals applied to those trajectories.
- This would give us a flexible tool for pattern detection that doesn't rely on parametric modeling assumptions.
- Up to now, classification and discrimination have not involved the simultaneous use of multiple biological processes to distinguish individuals.
- The Bayesian approach that we will use will provide a coherent method to accomplish this goal.

References

- S.L. Hui & S.D. Walter (1980) "Estimating the error rates of diagnostic tests." *Biometrics* **36** 167-171.
- U. Luderer, F. Christensen, W.O. Johnson, J. She, H.S.S. Ip, J. Zhou, J. Alvaran, E.F. Krieg Jr, & J.S. Kesner (2017), "Associations between urinary biomarkers of polycyclic aromatic hydrocarbon exposure and reproductive function during menstrual cycles in women," *Environment International*, **100**, 110-120.
- D. Steinberg & N.S. Cardell (1992), "Estimating logistic regression models when the dependent variable has no variance," *Communications in Statistics – Theory and Methods*, **21**, 423-450.

That’s all, folks!

Thank you!