Bayesian Latent Variable Modelling of Longitudinal Family Data for Genetic Pleiotropy Studies

Radu Craiu

Department of Statistical Sciences
University of Toronto

Joint with Andrew Paterson, Lei Sun and Lizhen Xu (Toronto)

INRA, Toulouse
February, 2015

Outline

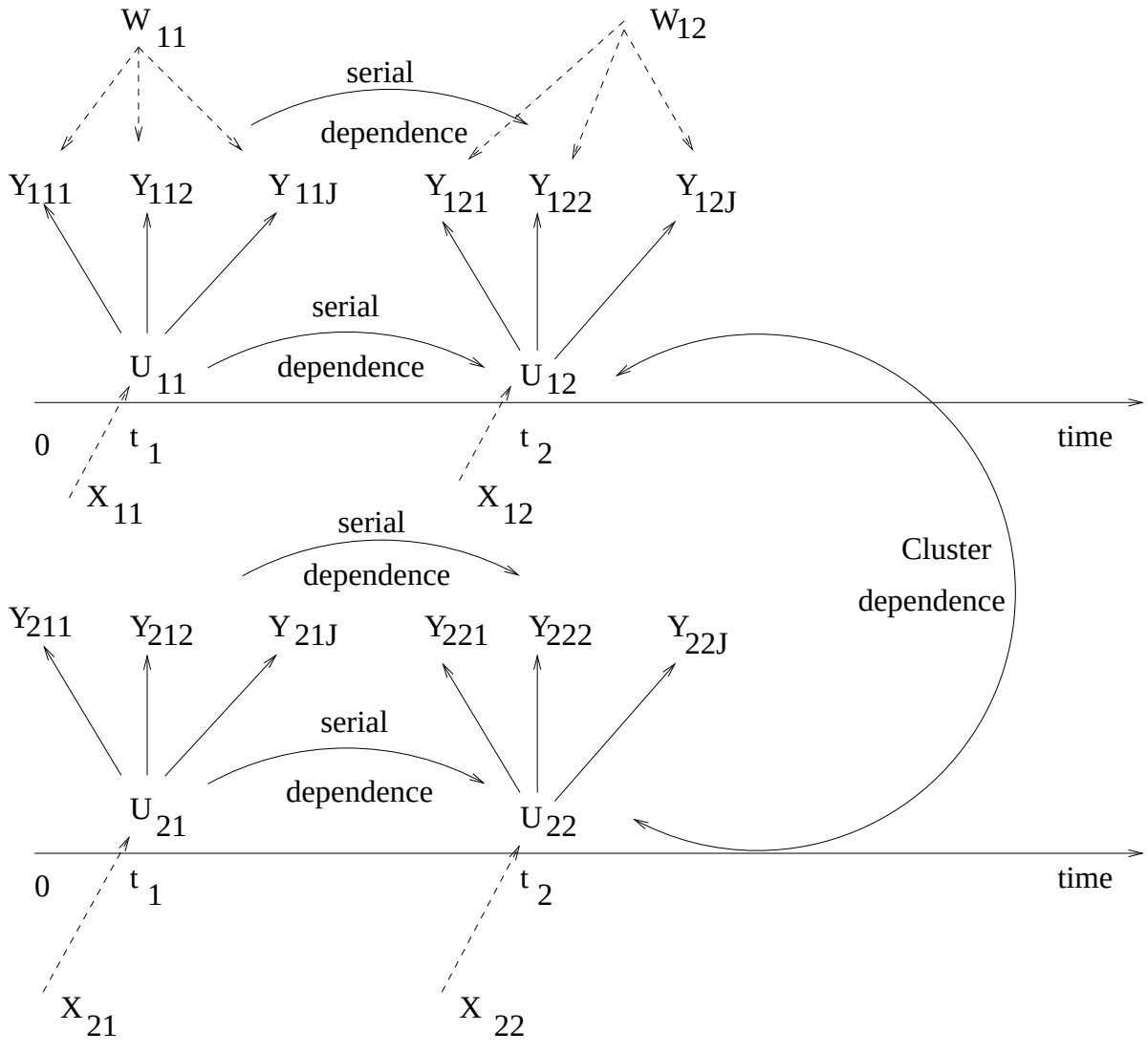
Pleiotropy

- For many complex human diseases, the trait of interest ("state of disease") is not directly observable (e.g. diabetes, hypertension, cardiovascular disease).
- Instead we observe a set of surrogate phenotypes (physical manifestations of the disease) which may be continuous or discrete.
- These response variables (phenotypes or outcomes) measure the underlying trait from different perspectives.
- In order to increase statistical efficiency, it is desirable to model these outcomes jointly.
- Many studies also involve repeated measures over time in samples that include clusters (e.g., families) $\Rightarrow$ complex dependence structures in the data.
- We are considering here continuous and binary phenotypes.

The Data and Model

- Let $\mathbf{Y}_{cit} = (\mathbf{Y}_{cit}^{c\,T}, \mathbf{Y}_{cit}^{b\,T})^T$ be the $J \times 1$ vector of responses (e.g. phenotypes) measured at the t^{th} time on the i^{th} individual from the c^{th} family (or cluster) for $c = 1, 2, ..., C$, $i = 1, 2, ..., N_c$, $t \in \{t_{ci1}, t_{ci2}, ..., t_{ciM_{ci}}\}$, and $j = 1, 2, ..., J$, where C denotes the total number of families, N_c is the number of individuals within the c^{th} family, M_{ci} is the total number of repeated measurements for individual i in cluster c and J is the total number of responses.
- The cluster (i.e., family pedigree) structure is known.
- Covariate measurements are available on all items at all times.
- The dependence patterns are modelled via random effects.
- The trait of interest is introduced as a latent variable U_{cit}.

Illustration of the Data Structure



The Statistical Model

- The latent variable model

$$\mathbf{U}_{ci} = \mathbf{X}_{ci}\boldsymbol{\alpha} + g_c \mathbf{1}_{M_{ci}} + \mathbf{Z}_{ci}^T \otimes \mathbf{1}_{M_{ci}} \mathbf{a}_c + \boldsymbol{\epsilon}_{ci}, \qquad (1)$$

where:

- $\mathbf{U}_{ci} = (U_{ci1}, \ldots, U_{ciM_{ci}})^T$ is the vector of the longitudinal LV at times $\mathbf{t}_{ci} = (t_{ci1}, \ldots, t_{ciM_{ci}})^T$
- $\boldsymbol{\epsilon}_{ci} = (\epsilon_{ci1}, \ldots, \epsilon_{ciM_{ci}})^T$ is the vector of error terms and $\mathbf{X}_{ci} = (X_{ci1}^T, \ldots, X_{ciM_{ci}}^T)^T$ is a $M_{ci} \times p_2$ design matrix for the fixed effects α
- $\mathbf{Z}_c = (Z_{c1}^T, \ldots, Z_{cN_c}^T)^T$ is the Cholesky decomposition of the kinship coefficient matrix of the c^{th} family, $\mathbf{K}_c$, i.e., $\mathbf{Z}_c \mathbf{Z}_c^T = \mathbf{K}_c$.

The Statistical Model

- $\mathbf{a}_c = (a_{c1}, \ldots, a_{cN_c})^T$ account for common genetic factors.
- g_c account for environmental factors.
- $\epsilon_{ci} \sim N_{M_{ci}}(0, \sigma^2_\epsilon \mathbf{H}_{ci})$, where $\mathbf{H}_{ci}$ is a $M_{ci} \times M_{ci}$ matrix with the $(r, k)^{th}$ entry equal to $\rho^{|t_r - t_k|}$ (ρ is the correlation between the within-subject error terms that are one time unit apart).
- This allows for unequal number of observations between clusters and varying interval between measurements.
- We are particularly interested in the regression coefficient for the SNP's genotype (α) and factor loadings (λ's).
- Pleiotropy is detected if the SNP's genotype effect on U and at least two factor loadings are statistically significant.

The Statistical Model

- The continuous response model

$$y^c_{citj} = \beta_{0j} + b_{cij} + \mathbf{W}^T_{cit}\beta_j + \lambda_j U_{cit} + e_{citj}, \qquad (2)$$

where $e_{citj} \overset{\text{iid}}{\sim} N(0, \sigma^2_j)$, $\mathbf{W}_{cit}$ is a p_1-dimensional vector of direct effect covariates.

- The λ's are the factor loadings that quantify the effect of the latent variable on each phenotype.
- The random component b_{cij} captures the family-specific within-subject serial correlations.
- We assume $b_{cij} \overset{\text{iid}}{\sim} N(0, \tau^2_j)$, and e_{citj} and b_{cij} are mutually independent for $c = 1, ..., C$, $i = 1, ..., N_c$, $t = 1, ..., M_{ci}$ and $j = 1, ..., J$.

The Statistical Model

- If a response is binary, a generalized linear mixed model is assumed,

$$\mu_{citj} = \beta_{0j} + \mathbf{W}_{cit}^T \boldsymbol{\beta}_j + \lambda_j U_{cit} + b_{cij},$$

with a probit link,

$$E\left[y_{citj}^b | \mu_{citj}\right] = p(y_{citj}^b = 1 | \mu_{citj}) = \Phi(\mu_{citj}).$$

Statistical Complications - Direct or Indirect Covariate?

- Important: Splitting the available covariates into two disjoint sets that correspond to direct and indirect effects.
- Dependent variables of primary interest $\longrightarrow$ Indirect effects.
- A larger set of indirect effects leads to a more parsimonious model.
- Matter is complicated by lack of symmetry...

Statistical Complications - Direct or Indirect Covariate?

- Define the LV $U^*_{cit} = U_{cit} - \mathbf{X}^T_{cit}\boldsymbol{\alpha}$
- Switching $\mathbf{X}$ from the indirect to the direct set leads to an equivalent model.
- Switching covariates from direct to indirect effect does lead to a very different model and may produce different conclusions along with ...
- ... a significant increase in the deviance information criterion (DIC).

Statistical Complications - Identifiability

- For any $Q \in \mathbf{R}\backslash\{0\}$ we get an equivalent model

$$y^c_{citj} = \beta_{0j} + W^T_{cit}\beta_j + \lambda_j Q^{-1} Q U_{cit} + b_{cij} + e_{citj}, \qquad (3)$$

- Without any restriction on λ and the variance of U_{cit}, an infinite number of equivalent models can be created.
- We assume that:
  - The variance of U_{cit} is equal to 1 and that λ_j is non-negative.
  - The direct-effect covariates (W_{cit}) and the indirect-effect covariates (X_{cit}) are distinct.

Statistical Complications - Effect of Ignoring Cluster Correlation

- Individuals from the same family are genetically related resulting in correlation between their latent disease status.
- If familial dependence is ignored inference is biased.
- Consider the case of continuous only phenotypes and no repeated measurements.

Statistical Complications - Effect of Ignoring Cluster Correlation

- Model 1 (correct):

$$y_{cij} = \beta_{0j} + W_{ci}^T \beta_j + \lambda_j U_{ci} + e_{cij}, \text{ and } U_{ci} = \mathbf{X}_{ci}^T \boldsymbol{\alpha} + g_c + \mathbf{Z}_{ci}^T \mathbf{a}_c + \epsilon_{ci},$$

where $e_{cij} \sim N(0, \sigma_j^2)$ and $\epsilon_{ci} \sim N(0,1)$, $\lambda_j > 0$,
$g_c \sim N(0, \sigma_g^2)$ and $\mathbf{a}_c \sim N(0, \sigma_a^2 \mathbf{I}_{N_c})$.

- Model 2 (misspecified):

$$y_{hj} = \beta_{0j} + W_h^T \beta_j + \widetilde{\lambda}_j \widetilde{U}_h + e_{hj}, \quad \text{and} \quad \widetilde{U}_h = X_h^T \widetilde{\alpha} + \epsilon_h.$$

- It can be shown that

$$\widetilde{\lambda}_j > \lambda_j$$

and

$$|\widetilde{\alpha}| = \frac{\lambda_j}{\widetilde{\lambda}_j} |\alpha| < |\alpha|$$

Bayesian Model

- We consider a Bayesian framework for inference.
- If conditional conjugate priors are defined for the model parameters $\boldsymbol{\Theta}$, then a *standard Gibbs (SG) sampler* can be used to analyze the posterior distribution.
- The implementation requires introducing the random effects as latent variables/missing data. The set of all latent variables is denoted Ω.

Computational Complications: Torpid Mixing

- Due to high dependence between the components of the Markov chain corresponding to the parameter vector $\boldsymbol{\Theta}$ and the latent data vector Ω, we observe a very slow mixing of the chain.
- For instance, a small variance τ_j^2 leads to small random effects b_{cij} and vice versa. Similar patterns develop between the factor loadings λ_j and the latent variable U.
- These lead to computational inefficiency because the chain gets stuck in various regions of the sample space ("bottlenecks").

Computational Complications: A simple calculation

- $y_{ij} = \mu + b_j + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, \sigma^2)$ for all $1 \leq i \leq n$, $1 \leq j \leq C$.
- Conjugate priors:

$$p(\mu) = N(0, B_3^2),\ p(\sigma^2) = IG(A_1, B_1),$$

$$p(b_j) = N(0, \eta^2),\ \textit{and}\ p(\eta^2) = IG(A_2, B_2).$$

- Conjugate posteriors:

$$\pi(\eta^2|\ldots) = IG\left(\frac{c}{2} + A_2, B_2 + \sum_{j=1}^{C} b_j^2\right),$$

$$p(b_j|\ldots) = N\left(\frac{\frac{\bar{x}_j - \mu}{\sigma^2}}{1/\sigma^2 + 1/(n\eta^2)}, \frac{1}{n/\sigma^2 + 1/\eta^2}\right).$$

- $E[\eta^2|\ldots] < \sum_j b_j^2$ and $V(\eta^2|\ldots) < \sum_j b_j^2$, when $c > 5$.

Parameter Expansion for Increased Computational Efficiency

- Parameter Expansion/Auxiliary Variable methods have a long tradition in MCMC (Besag and Green, JRSSB '93; Higdon, JASA '98; Liu and Wu, JASA '99; van Dyk and Meng, JCGS '01)
- These methods aim at eliminating "bottlenecks" in simulation experiments by expanding the parameter space or by introducing "missing" data/latent variables in the model.
- However, the parameter expansion guidelines need to be modified/adapted for each model.

The simple calculation revisited

- $y_{ij} = \mu + \xi \frac{b_j}{\xi} + \epsilon_{ij} = \mu + \xi b_j^* + \epsilon_{ij}$
- $p(\xi) = N(0, \psi^2)$, $p(b_j^*) = N(0, \eta^{*2})$.
- $\pi(b_j^* | \ldots) = N\left(\frac{\xi(\bar{x}_j - \mu)/\sigma^2}{\xi^2/\sigma^2 + 1/(n\eta^{*2})}, \frac{1}{1/\eta^{*2} + n\xi^2/\sigma^2}\right)$.
- $p(\xi | \ldots) = N\left(\frac{\sum_j b_j^*(\bar{x}_j - \mu)/\sigma^2}{1/(n\psi^2) + \sum_j b_j^{*2}/\sigma^2}, \frac{1}{1/\psi^2 + n\sum_j b_j^{*2}/\sigma^2}\right)$
- The model is over-parametrized and the chain $(\mu, \xi, \sigma, \eta^*, \{b_j^*\})$ may not perform better than the original one.
- But once we transform back to the original scale

$$b_j = b_j^* \cdot \xi, \ \eta = \eta^* \cdot \xi,$$

  we can notice a significant increase in efficiency.
- Notice that the induced prior for η is not the same as the one used in the original model.

A Parameter Expanded Model - Continuous Outcomes

- Original model is

$$y^c_{citj} = \beta_{j0} + \mathbf{W}^T_{cit}\boldsymbol{\beta}_j + \lambda_j U_{cit} + b_{cij} + e_{citj},$$

$$U_{cit} = \mathbf{X}^T_{cit}\boldsymbol{\alpha} + g_c + \mathbf{Z}^T_{ci}\mathbf{a}_c + \epsilon_{cit},$$

where $c = 1, \ldots, C; i = 1, \ldots, N_c, t = 1 \ldots M_{ci}, j = 1, \ldots, J.$

A Parameter Expanded Model - Continuous Outcomes

- Introduce auxiliary parameters μ^*, $\{\xi_j : 1 \leq j \leq J\}$ and ψ and reparametrise the model.
- Transformed model:

$$y^c_{citj} = \xi_j \left(\frac{\beta_{j0}}{\xi_j} - \mu^* \frac{\lambda_j}{\xi_j \psi} \right) + \mathbf{W}^T_{cit} \boldsymbol{\beta}_j + \frac{\lambda_j}{\psi}(\psi U_{cit} + \mu^*) + \xi_j \frac{b_{cij}}{\xi_j} + e_{citj},$$

$$\psi U_{cit} + \mu^* = \mu^* + \mathbf{X}^T_{cit} \boldsymbol{\alpha} \psi + g_c \psi + \mathbf{Z}^T_{ci} \mathbf{a}_c \psi + \epsilon_{cit} \psi,$$

A Parameter Expanded Model - Continuous Outcomes

- Transformed model:

$$y^c_{citj} = \beta^*_{j0} + \mathbf{W}^T_{cit}\boldsymbol{\beta}_j + \lambda^*_j U^*_{cit} + \xi_j b^*_{cij} + e_{citj},$$

$$U^*_{cit} = \mu^* + \mathbf{X}^T_{cit}\boldsymbol{\alpha}^* + g^*_c + \mathbf{Z}^T_{ci}\mathbf{a}^*_c + +\epsilon^*_{cit}.$$

- The parameters are linked via

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}^*/\psi, \quad U_{cit} = (U^*_{cit} - \mu^*)/\psi, \quad \sigma^2_a = \sigma^{*2}_a/\psi^2, \quad \sigma^2_g = \sigma^{*2}_g/\psi^2,$$

$$\lambda_j = \lambda^*_j\psi, \quad \beta_{j0} = \xi_j\mu^*_{bj} + \lambda^*_j\mu^*, \quad \tau^2_j = \xi^2_j\tau^{*2}_j, \quad \text{for all } 1 \leq j \leq J.$$

A Parameter Expanded Model - Continuous Outcomes

- $b^*_{cij} \sim N(\mu^*_{bj}, \tau^{*2}_j)$, $g^*_c \sim N(0, \sigma^{*2}_g)$, $\mathbf{a}^*_c \sim N_{N_c}(0, \sigma^{*2}_a \mathbf{I}_{N_c})$ and $\boldsymbol{\epsilon}^*_{ci} \sim N_{K_{ci}}(0, \psi^2 \mathbf{H}_{ci})$.
- The conditional conjugate priors assigned to $\theta^* = (\alpha^*, \lambda^* \ldots, \psi)$ impose particular priors on $\theta = (\alpha, \lambda, \ldots)$.
- The parametrization is redundant and the algorithm is not efficient on the expanded state space, but it gains efficiency for the original set of parameters!

A Parameter Expanded Model - Mixed Outcomes

- When the traits are mixed denote $\{y^c_{citj} : 1 \leq j \leq J_1\}$ the continuous outcomes and $\{y^b_{citj} : J_1 + 1 \leq j \leq J\}$ the binary ones.
- The probit model is expanded using the latent variables y^{b*}_{citj} so that $y^b_{citj} = \mathbf{1}_{(0,\infty)}(y^{b*}_{citj})$.

A Parameter Expanded Model - Mixed Outcomes

- The continuous response models are expanded as before.

$$y^c_{citj} = W^T_{cit}\beta_j + \lambda^*_j U^*_{cit} + \xi_j b^*_{cij} + e_{citj},\ \ 1 \le j \le J_1,$$

$$p(y^b_{citj} = 1) = \Phi(W^T_{cit}\beta_j + \lambda^*_j U^*_{cit} + \xi_j b^*_{cij}),\ \ J_1 + 1 \le j \le J,$$

$$\mathbf{U}^*_{ci} = \mu^* \mathbf{1}_{K_{ci}} + \mathbf{X}_{ci}\boldsymbol{\alpha}^* + g^*_c \mathbf{1}_{K_{ci}} + \mathbf{1}_{K_{ci}}\mathbf{Z}_{ci}\mathbf{a}^*_c + \boldsymbol{\epsilon}^*_{ci},$$

where $b^*_{cij} \sim N(\mu^*_{bj}, \tau^{*2}_j)$, $g^*_c \sim N(0, \sigma^{*2}_g)$,
$\mathbf{a}^*_c \sim N_{N_c}(0, \sigma^{*2}_a \mathbf{I}_{N_c})$, $\boldsymbol{\epsilon}^*_{ci} \sim N_{K_{ci}}(0, \psi^2 \mathbf{H}_{ci})$.

- An additional level of parameter expansion is added via $\boldsymbol{\gamma} = (\gamma_{J_1+1}, \ldots, \gamma_J)^T \in \mathbf{R}^{J-J_1}$, a one-to-one mapping $\widetilde{y}^{b*}_{cikj} = \gamma_j y^{b*}_{cikj}$ and set $\tilde{\beta}_j = \gamma_j \beta_j$, $\tilde{\lambda}^*_j = \gamma_j \lambda^*_j$ and $\tilde{\xi}_j = \gamma_j \xi_j$. A priori, $\gamma_{J_1+1}, \ldots, \gamma_J$ are iid with prior distribution IG(0.1,0.1).

Variable Selection

- Of primary interest is the effect of a genetic marker on the latent variable.

$$U_{cit} = X_{cit}^T \alpha + Z_{cit}^T a_c + g_{ci} + \epsilon_{cit}.$$

- Of secondary interest is to determine whether the jth phenotype is indeed related to the latent disease status (i.e. $\lambda_j = 0$ or not).

$$y_{citj}^c = \beta_{0j} + b_{cij} + W_{cit}^T \beta_j + \lambda_j U_{cit} + e_{citj}.$$

Variable Selection

- We can use a spike-and-slab prior for λ_j^* (or α^*),

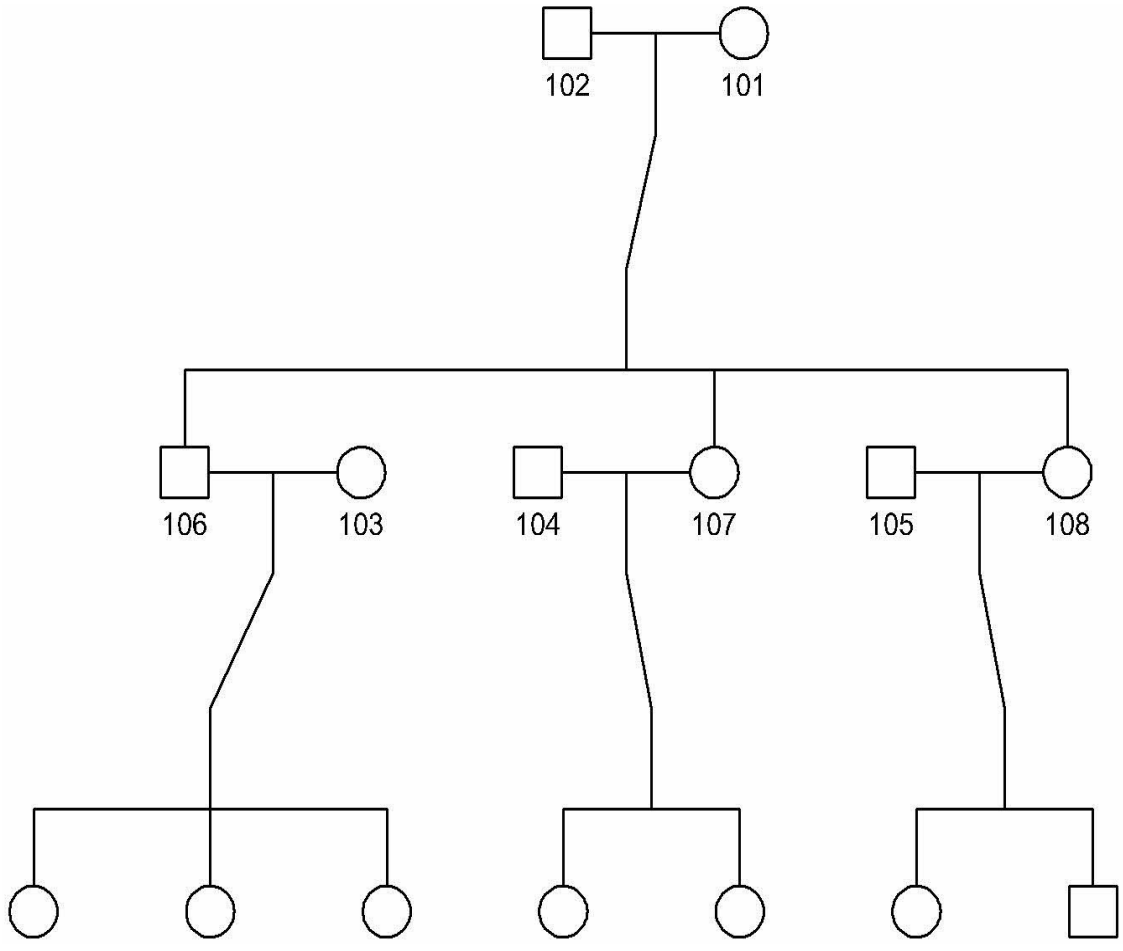
$$p(\lambda_j^*|\omega_j) = \omega_j \mathbf{1}_{\{0\}}(\lambda_j^*) + (1-\omega_j)\text{TN}_+(\lambda_j^*|0,1)$$

  and $p(\omega_j) = \text{Beta}(a,b)$. The relevance of the jth phenotype is based on $P(\lambda_j > 0|\mathbf{Y})$. Easy
- We can consider comparing two models (almost identical, but one has $\lambda_j = 0$) via Bayes factor. Hard since it requires computing normalizing constants via Bridge/Path Sampling.
- Compare the two models via Deviance Information Criterion (DIC). Easy
- Inspect Hpdl's. Easy

Simulation Design

- We consider 100 families.
- The number of children in the third generation varies from one to five with probability $\{20\%, 40\%, 30\%, 7\%, 3\%\}$.
- For each individual, we assume that the probability of being observed longitudinally $\{1, 2, 3, 4\}$ times is $\{10\%, 30\%, 30\%, 30\%\}$
- The time of first measure is set as $\{0, 1, 1.5, 2\}$ with probability $\{50\%, 20\%, 20\%, 10\%\}$.
- The length of time between two consecutive measures is $\{1, 2, 3, 3.5\}$ with probability $\{50\%, 20\%, 20\%, 10\%\}$, respectively, resulting in an unbalanced design.

Sample Pedigree used in the Simulation Scenarios



Simulation Scenarios

M1 We consider $J = 3$ continuous response variables and set $\boldsymbol{\beta}_0 = (5, 5, 5)$, $\beta_{11} = \beta_{12} = \beta_{13} = 1$, $\alpha_1 = -1$, $\alpha_2 = 1$, $\boldsymbol{\lambda} = (5, 5, 5)$, $\boldsymbol{\tau}^2 = (0.3, 0.3, 0.3)$, $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$, $\sigma_a^2 = 0.3$, $\sigma_g^2 = 0.3$, and $\rho = 0.3$.

M2 We consider $J = 4$ and we simulate y_1, y_2 as continuous and y_3, y_4 as binary responses. We set $\boldsymbol{\beta}_0 = (1, 1, 1, 1)$, $\beta_{1j} = 1$ for all $j = 1, \ldots, 4$, $\alpha_1 = -1$, $\alpha_2 = 1$, $\boldsymbol{\lambda} = (2, 3, 1, 1)$, $\boldsymbol{\tau}^2 = (0.6, 0.6, 0.6, 0.6)$, $\sigma_1^2 = \sigma_2^2 = 1$, $\sigma_a^2 = 1$, $\sigma_g^2 = 1$, $\rho = 0.3$.

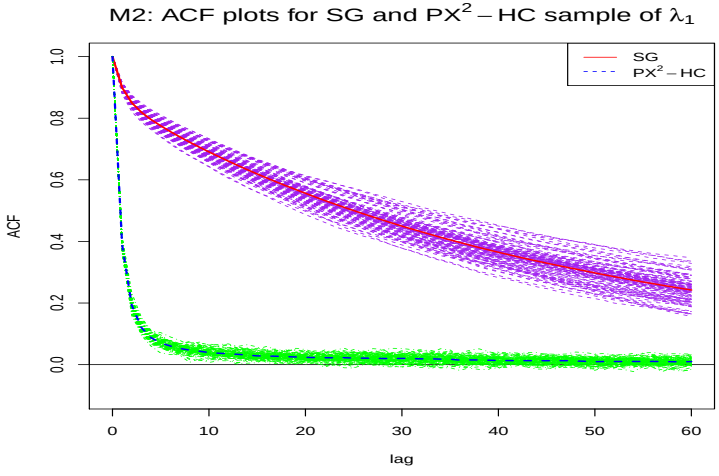
Measures of Efficiency

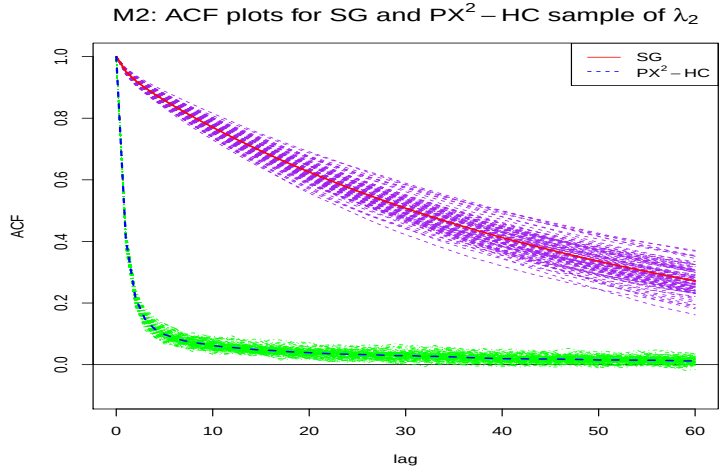
- When comparing algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ we compare the *effective sample size (ESS)* for each parameter via

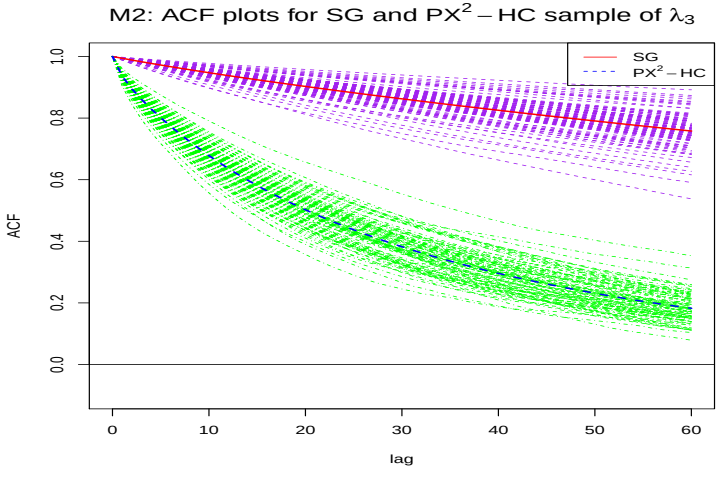
$$\Delta_{ESS}(\mathcal{A}_1, \mathcal{A}_2) = 100 \times \left(\frac{\text{ESS}_{A_2} - \text{ESS}_{A_1}}{\text{ESS}_{A_1}} \right)$$

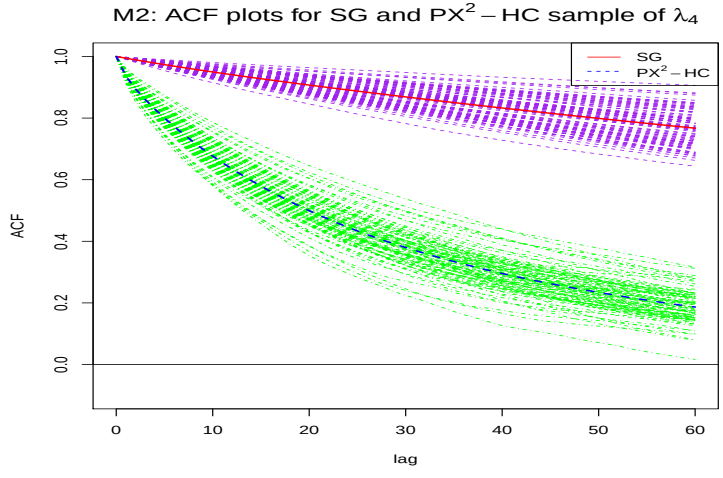
- ESS plays a central role in determining the number of iterations until a certain desired precision is attained.

ACF plots for **M2**: $\lambda_1 - \lambda_4$

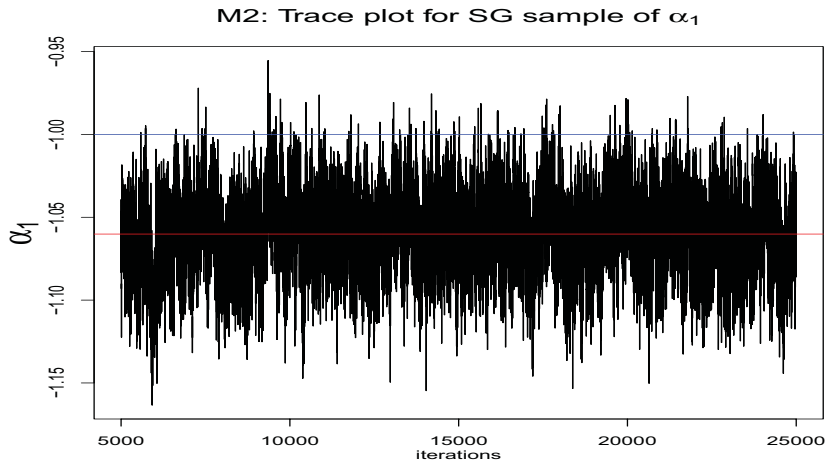


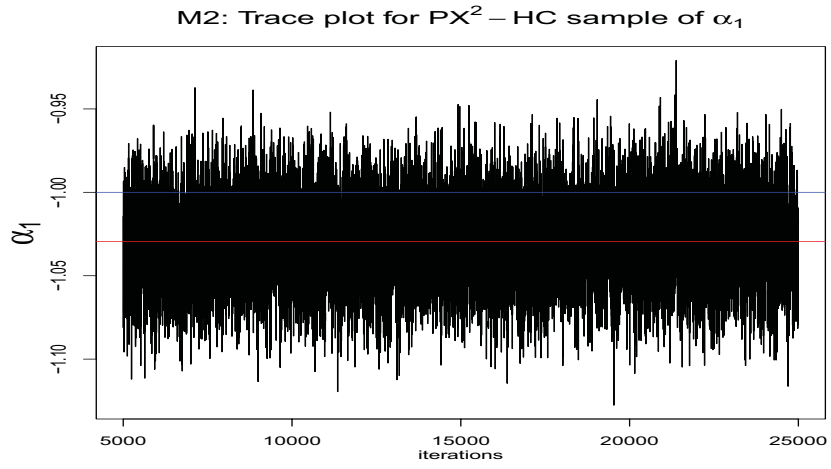


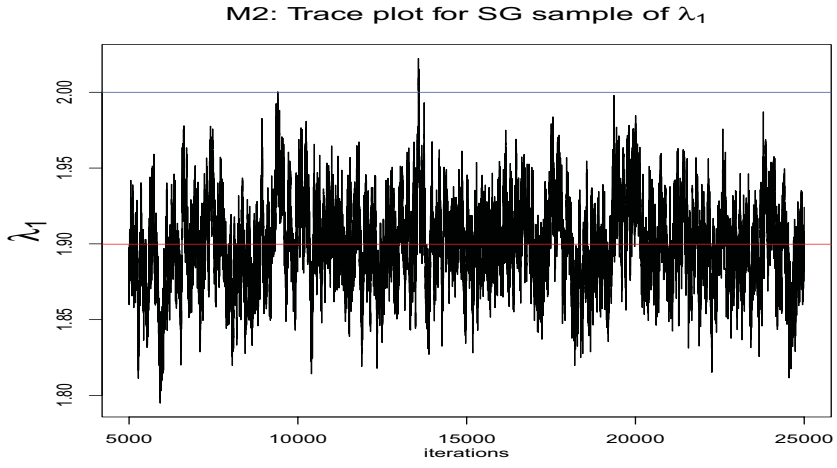


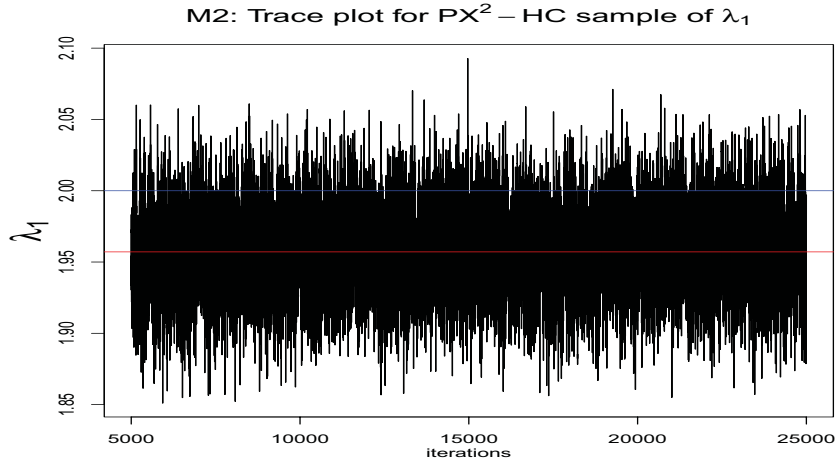


Trace plots for **M2**: α_1, λ_1









M2: Simulation Results

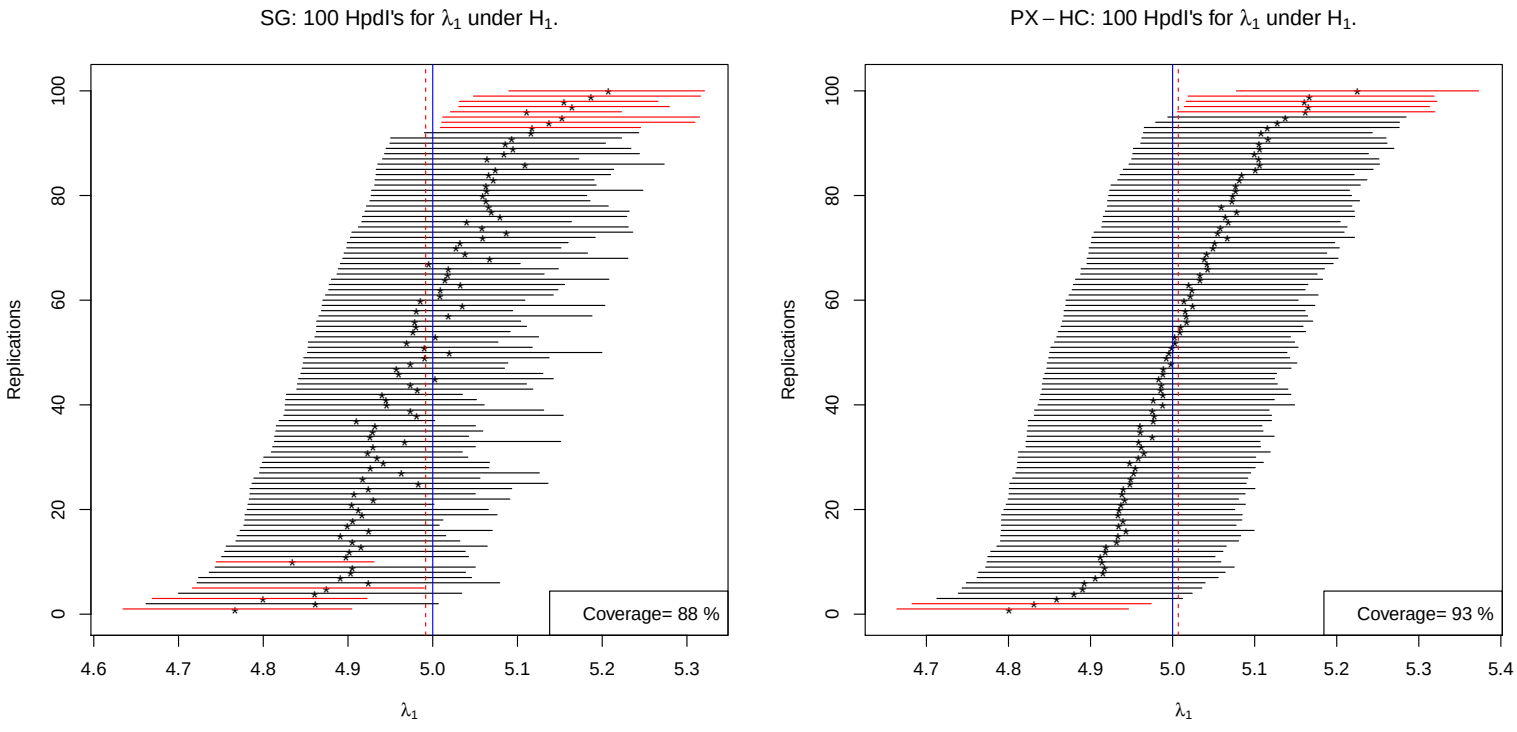
Parameters		Value	SG		PX^2-HC		Δ_{ESS}
			Est.	RMSE	Est	RMSE	
α	α_1	-1.0	-1.003	0.024	-1.002	0.024	923
	α_2	1.0	1.000	0.050	1.000	0.050	83
λ	λ_1	2.0	2.001	0.039	2.001	0.036	1124
	λ_2	3.0	3.001	0.060	3.002	0.057	1145
	λ_3	1.0	1.010	0.054	1.001	0.051	361
	λ_4	1.0	1.017	0.062	1.009	0.057	381
σ_a^2		1.0	1.021	0.140	1.019	0.136	166
σ_g^2		1.0	1.024	0.188	1.022	0.190	34

M2: Ignoring clusters

Parameters		True	Considering cluster			Ignoring cluster		
			bias	sd	RMSE	bias	sd	RMSE
α	α_1	-1.0	0.006	0.023	0.024	0.369	0.026	0.370
	α_2	1.0	-0.004	0.046	0.046	-0.370	0.066	0.376
λ	λ_1	2.0	0.009	0.036	0.037	1.176	0.113	1.181
	λ_2	3.0	0.016	0.054	0.056	1.754	0.165	1.761
	λ_3	1.0	0.017	0.056	0.058	0.608	0.099	0.616
	λ_4	1.0	0.008	0.058	0.058	0.595	0.103	0.604

M2: HPDI's for λ_1

HPDI's constructed under Standard Gibbs (left) and PX-DA (right):



GAW18: Genetic study of Hypertension

- Data included genotypes from a real human whole genome sequencing study (N = 483 individuals) and systolic and diastolic blood pressure phenotypes plus age, sex, medication use and cigarette smoking.
- The data were longitudinal, with three measurements for most participants at roughly 5-year intervals.
- Among the 464 individuals, 396 individuals have at least one blood pressure measures (90 have only one, 78 have two, 131 have three and 97 have four measurements).
- The length of time between two consecutive measurements ranges from 3 to 9 years, and the number of family members varies from 11 to 36.

GAW18: Genetic study of Hypertension

- We focused on a set of six SNPs that had been reported to be significantly associated with either DBP or the binary hypertension trait
- We applied the Bayesian LVM method to analyze one SNP at a time assuming an additive genetic model.
- The phenotypes are SBP and DBP, and the covariates include the genotype of the SNP, age and sex.

GAW18: Results for SNP rs9816772

Model	Covariates		DIC
	Direct	Indirect	
1	Age+Sex	SNP	15729.3
2	Age	Sex+SNP	15744.0
3	Sex	Age+SNP	15226.5
4	-	Age +Sex+SNP	15948.3

GAW18: Results for SNP rs9816772

- rs9816772 had been identified to be associated with DBP.

	Parameter	Estimate	logBF	95% HpdI
SBP	λ_1	13.15	255.3	(12.19, 14.11)
DBP	λ_2	7.60	139.6	(7.01, 8.14)
Sex for SBP	β_{11}	-0.66	-0.074	(-2.12, 0.81)
Sex for DBP	β_{21}	-1.79	2.017	(-2.92, -0.65)
rs9816772	α_1	-0.045	-0.653	(-0.208, 0.124)
Age	α_2	0.043	126.53	(0.036, 0.049)

Genetic study of type 1 diabetes (T1D) complications.

- The study sample consists of $n = 1300$ individuals with T1D from the Diabetes Control and Complications Trial (DCCT)
- Various phenotypes thought be to related to T1D complication severity, including glycosylated hemoglobin (HbA1c) and diastolic (DBP) and systolic blood pressure (SBP). We define hyperglycaemia $HPG = \mathbf{1}(HbA1C > 8)$.
- Previous studies have identified rs7842868 on chromosome 8 as a SNP significantly associated with DBP.
- Our goal here is to formally perform a multi-phenotype analysis, jointly analyzing the measured manifest variables using the proposed Bayesian LVM methodology. This approach allows us not only to determine if rs7842868 is associated with the latent conceptual T1D complication variable, but also to test if DBP and SBP are truly related to the LV.

Genetic study of type 1 diabetes (T1D) complications

Analysis of SNP rs7842868				
	Parameter	Estimate	95% HpdI	$\widehat{logBF}$
SBP	λ_1	6.621	(6.153, 7.077)	114.85
DBP	λ_2	3.842	(3.566, 4.110)	112.98
HPG	λ_3	0.011	$(\frac{2.19}{10^7}, \frac{2.98}{10^2})$	-1.05
rs7842868	α_1	-0.269	(-0.372, -0.164)	10.06
sex	α_2	-0.721	(-0.866, -0.584)	62.27
cohort	α_3	0.443	(0.299, 0.585)	20.15
treatment	α_4	0.128	(-0.004, 0.263)	0.366

This is just the beginning...

- When is the conjectured existence of the LV defensible? What does it really represent?
- Indirect/Direct Covariates dilemma: does assignation depend on the SNP or SNP/Environment interactions? Can we get more "clear cut" criteria?
- Evaluate the contribution of each phenotype to the model (rather than 0/1 decision). May be useful to reduce the number of phenotypes.
- Too computational for looking at thousands of genes. It currently takes about 2mins per SNP. Maybe a combination of Bayes/Frequentist methods can speed things up.

References

- Liu, J. S., Wu, Y. N. (1999) "Parameter Expansion for Data Augmentation.," *Journal of the American Statistical Association*, 94, 1264–1274.
- Hobert, J. P., Marchev, D. (2008), "A theoretical comparison of the data augmentation, marginal augmentation and PX-DA algorithms," *Ann. Statist.*, 36, 532–554.
- Ghosh, J., Dunson, D. B. (2009), "Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis.," *Journal of Computational and Graphical Statistics*, 18(2), 306–320.
- **Lizhen Xu[‡], Radu V. Craiu, Andriy Derkach, Andrew Paterson and Lei Sun (2013). Using a Bayesian Latent Variable Approach to Detect Pleiotropy in the Genetic Analysis Workshop 18 Data.** *BMC Proceedings*.
- **Lizhen Xu[‡], Radu V. Craiu, Lei Sun and Andrew Paterson (2015). Parameter expanded Algorithms for Bayesian Latent Variable Modeling of Genetic Pleiotropy Data.** *Journal of Computational and Graphical Statistics*, **to appear.**