UPDATES TO BSFG

1. ABSTRACT

The linear mixed effect model is a workhorse of modern statistical genetics: including GWAS, QTL analysis, Genomic Prediction, Evolutionary Genetics, transcriptomics, growth curve analysis, etc. Recent advances in computational capacity and algorithms has made mixed model accessible to a wide range of researchers. Widely available software has made analyses with thousands (or millions) of individuals feasible. However, two key limitations of available methods are that most are limited to one (or a few) responses per individual, and most only allow a single random effect (besides the residual). Also, while methods appropriate for identically distributed Gaussian variables are common, methods applicable to other distributions, or non-linear response functions are less common. Here, we propose a general model for high-dimensional linear mixed effect models, which we call SLAMglmm and provide as an extendible R package. SLAMglmm builds on our earlier work (Runcie and Mukherjee 2013) that proposed using sparse factor models to efficiently estimate genetic covariance matrices for high dimensional traits from data on related individuals. We build on the earlier model in four key ways:

1. Fully generalize the mixed effect model, allowing multiple random effects for both the individual traits and the latent factor traits, and ”fixed” effects per-trait and per-factor.
2. Adapted the discrete prior structure for random effects used for the latent factor traits to the individual traits. This allows more intuitive prior elicitation, especially in models with multiple random effects.
3. Developed a new, more efficient Gibbs sampler for mixed effect models that greatly improves mixing and posterior convergence.
4. Added a new level between observed data and the linear mixed effect model based on a flexible link distribution. We develop link distributions for several disparate data types below including: partially missing observations, multiple-probe-per gene data, RNAseq data, time-series data.

The SLAMglmm package is written in R, and draws heavily from the following packages: *Matrix*, *Rcpp*, *RcppArmadillo*, *lme4*, *MCMCglmm*. The model syntax closely follows that of the widely used *lme4* package. Prior specification is similar to *MCMCglmm*.

2. Methods

2.1. **SLAMglmm model.** The SLAMglmm model is specified as:

$$
\begin{aligned}
\mathbf{y}_i &\sim g(\boldsymbol{\eta}_i, \Sigma_i, \theta_y) \\
[\eta_1, \eta_2, \ldots, \eta_n]^T = \boldsymbol{H} &= \boldsymbol{F}\boldsymbol{\Lambda}^T + \boldsymbol{X}\boldsymbol{B} + \sum \boldsymbol{Z}_i \boldsymbol{U}_{R_i} + \boldsymbol{E} \\
\mathbf{F} &= \boldsymbol{X}_F \boldsymbol{B}_F + \sum \boldsymbol{Z}_i \boldsymbol{U}_{F_i} + \boldsymbol{E}_F
\end{aligned} \tag{1}
$$

Here, $\boldsymbol{y}_i$ is a vector of observations for the ith individual. $\boldsymbol{\eta}_i$ is a vector of p potentially unobserved characteristics (traits) for the ith individual. These may correspond directly to the elements of $\boldsymbol{y}_i$, or may be parameters of a more complicated function / distribution g. $\boldsymbol{H}$ is an $n \times p$ matrix of these characteristics for the n individuals. In the original BSFG model, g was the identity function, so $\boldsymbol{Y} = \boldsymbol{H}$. The utility of this hierarchical specification will be demonstrated below.

In 1, $\boldsymbol{H}$ and $\boldsymbol{F}$ are $n \times p$ and $n \times k$ matrices of latent traits for the n individuals. These $p + k$ latent traits are related through a structural equation model given by the $p \times k$ factor loadings matrix $\boldsymbol{\Lambda}$. This structural equation model is the key feature of SLAMglmm (and BSFG), as the paths described by $\boldsymbol{\Lambda}$ explain all of the covariance among the $p + k$ traits; the k traits $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_k]$ are assumed to be independent, and the p traits $\boldsymbol{H} = [\boldsymbol{\eta}_{\bullet 1}, \boldsymbol{\eta}_{\bullet 2}, \ldots, \boldsymbol{\eta}_{\bullet p}]$ are conditionally independent, conditional on $\boldsymbol{F}$. Through priors on $\boldsymbol{\Lambda}$ (and k), we impose sparsity on the among-trait covariances for all random effects, ensuring that the model can scale efficiently with increasing numbers of traits.

2.1.1. *fixed effects.* The matrices $\boldsymbol{B}$ and $\boldsymbol{B}_F$ are $b \times p$ and $b_F \times k$ matrices of fixed effect coefficients for each of the $p + k$ latent traits, corresponding to the fixed effect design matrices $\boldsymbol{X}$ and $\boldsymbol{X}_F$. Fixed effects are used to model the effect of individual-level covariates and design features such as sex, gender, or environment. In BSFG, we allowed fixed effects only for the p observational-level traits (ie $\boldsymbol{B}$), and modeled them with a flat prior (independent Gaussian distributions with variance $= 10^6$). Here, we generalize this to allow fixed effects for the k. However, by simply introducing $\boldsymbol{X}_F$ with similarly flat priors, the $\boldsymbol{B}$ and $\boldsymbol{B}_F$ would not be identifiable, as all association between the covariates and $\boldsymbol{H}$ could be explained by $\boldsymbol{B}$. However, for each covariate, we propose that solutions in which the k latent factor traits $\boldsymbol{F}$ explain the covariate effects on several observation-level traits $\boldsymbol{H}$ should be favored, so we model the set of parameters $[\boldsymbol{b}_{j\bullet}, \boldsymbol{b}_{F_{j\bullet}}]^T$ as:

$$
\begin{aligned}
[\boldsymbol{b}_{j\bullet}, \boldsymbol{b}_{F_{j\bullet}}]^T &\sim \mathrm{N}_{p+k}(0, \sigma^2_{b_i} \boldsymbol{I}) \\
\sigma^2_{b_i} &\sim \mathrm{iG}(\alpha_b, \beta_b)
\end{aligned} \tag{2}
$$

with α_b and β_b chosen to given a small posterior mean.

In general, we allow the fixed effect design matrices $\boldsymbol{X}$ and $\boldsymbol{X}_F$ to be similar or different depending on context. However we impose two additional constraints. First, we assume that the first column of $\boldsymbol{X}$ corresponds to a global intercept, and set $\sigma^2_{b_1} = \mathrm{Inf}$ so as not to penalize this coefficient. Second, we force the intercept of each of the $\boldsymbol{f}_j$ traits to be zero by setting $\boldsymbol{B}_{F_{1\bullet}} = 0$ and the corresponding variance to be zero as well.

2.1.2. *random effects.* The matrices $\boldsymbol{U}_{R_i}$ and $\boldsymbol{U}_{F_i}$ are $r \times p$ and $r \times k$ coefficient matrices for the ith random effect. As in BSFG, these are modeled with Matrix normal distributions):

$$\begin{aligned} \mathbf{U}_{R_i} &\sim \mathrm{MN}_{r,p}(\mathbf{0}; \boldsymbol{K}_i, \boldsymbol{\Sigma}^R_{h^2_i}\boldsymbol{\Psi}^R) \\ \mathbf{U}_{F_i} &\sim \mathrm{MN}_{r,k}(\mathbf{0}; \boldsymbol{K}_i, \boldsymbol{\Sigma}^F_{h^2_i}) \end{aligned} \tag{3}$$

where the $\boldsymbol{K}_i, i \in 1 \dots R$ are $r_i \times r$ "kinship" matrices describing expected covariances of random effects. These replace the additive-genetic covariance matrix $\boldsymbol{A}$ used in BSFG, and could be any positive-definite matrices. The number of levels for each of the R random effects, r_i may not equal n. The design matrices $\boldsymbol{Z}_i$ are $n \times r_i$ matrices linking each random effect level to a corresponding individual. BSFG allowed only a single random effect for the factors, although a second random effect was used for the p observational-level traits in one example.

The among-trait covariance specification is also new relative to BSFG. $\boldsymbol{\Sigma}^F_{h^2_i}$ is equivalent to $\boldsymbol{\Sigma}_{h^2}$ from BSFG, ie a $k \times k$ diagonal matrix specifying the heritability (or percentage of variation explained by the additive-genetic covariance matrix) of each of the k latent factors. In SLAMglmm, $\boldsymbol{\Sigma}^F_{h^2_i}$ is also diagonal, specifying the percentage of variation in each of the k latent factors explained by each of the R random effects. $\boldsymbol{\Sigma}^R_{h^2_i}$ is a similar $p \times p$ matrix for the residuals of each of the p observational-level traits $\boldsymbol{H}$, accounting for the latent factor traits. $\boldsymbol{\Psi}^R$ is a $p \times p$ diagonal matrix of total residual variances for each for the p traits. This random effect specification is elaborated below in section **??**.

2.1.3. *link distribution.* Several link distributions are described below in section **??**. Beyond the identity $\boldsymbol{y}_i = \boldsymbol{\eta}_i$, the next simplest link function is of the form:

$$\mathbf{y}_i \sim \mathrm{N}(g(\boldsymbol{\eta}_i), \Sigma_y) \tag{4}$$

2.1.4. *prior on factors.* We use the same prior on $\boldsymbol{\Lambda}$ as in BSFG. This is the "infinite factor model" prior proposed by Bhattacharya and Dunson (2011) that imposes increasing column-wise shrinkage on higher order columns, which imposes sparsity by shrinking the number of important factors (ie paths), as well as element-wise shrinkage on each element of $\boldsymbol{\Lambda}$ through and ARD-type prior. Other priors such as the *TPD* prior of Engelhardt could be substituted.

3. New concepts

As mentioned above, SLAMglmm proposes three new features that together greatly facility the analysis of high-dimensional linear mixed effect models. These are described more here:

3.1. **Structural equation models of among-trait covariances.** The central statistical challenge of large multivariate mixed effect models is that the number of parameters necessary to specify the among-trait covariance matrices grows as $p \times (p-1)/2 \times (R+1)$ with the number of traits and the number of random effects. Unconstrained estimates of all these covariance parameters requires an unrealistic number of observations for even moderately-sized trait vectors,

As proposed for BSFG, SLAMglmm uses a hierarchical sparse-factor structural equation model to explain the among-trait covariance structure. This model structure prioritizes the strongest, most important covariance signals in the data. A key development in BSFG was the idea that a single set of factors $\mathbf{\Lambda}$ could be used to explain both the additive-genetic and residual covariances, with the the factors re-weighted for each covariance matrix:

$$\begin{aligned}\mathbf{G} &= \mathbf{\Lambda}\mathbf{\Sigma}_{\boldsymbol{h^2}}\mathbf{\Lambda}^{\boldsymbol{T}} + \Psi_G \\ \mathbf{R} &= \mathbf{\Lambda}(\boldsymbol{I} - \mathbf{\Sigma}_{\boldsymbol{h^2}})\mathbf{\Lambda}^{\boldsymbol{T}} + \Psi_G\end{aligned} \tag{5}$$

This “sharing” of the factors between the covariance matrices does not force them to be similar: whenever one of the diagonal elements of $\mathbf{\Sigma}_{\boldsymbol{h^2}}$ ($\sigma_{h_j^2}$) equals 0(1), the column contributes only to the residual (additive-genetic) covariance matrix. But when the same factor contributes to both covariances, the same $\boldsymbol{\lambda}_j$ parameters can be re-used.

SLAMglmm uses this same strategy to efficiently estimate a set of $R+1$ covariance matrices for the R random effects and the residuals. Each diagonal element of the $\mathbf{\Sigma}^{\boldsymbol{F}}_{\boldsymbol{h_i^2}}$ matrices is allowed to equal 0 or 1, but can also take values in between, providing flexible sharing of covariance components among random effects:

$$\begin{aligned}\mathbf{\Sigma}_i &= \mathbf{\Lambda}\mathbf{\Sigma}^{\boldsymbol{F}}_{\boldsymbol{h_i^2}}\mathbf{\Lambda}^{\boldsymbol{T}} + \Sigma^R_{h_i^2}\Psi^R \\ \mathbf{\Sigma}_R &= \mathbf{\Lambda}(\boldsymbol{I} - \sum \mathbf{\Sigma}^{\boldsymbol{F}}_{\boldsymbol{h_i^2}})\mathbf{\Lambda}^{\boldsymbol{T}} + (\boldsymbol{I} - \sum \Sigma^R_{h_i^2})\Psi^R\end{aligned} \tag{6}$$

SLAMglmm also extends BSFG by leveraging the latent factors to estimate the fixed effects. If the the same sets of traits are associated with the fixed effect covariates as distinguish the random effect groupings (or residuals), the same factors can also be used to explain the fixed effect responses. These factor - covariate associations can be explored directly (ex sex or condition effects on the latent traits), or used to provide more robust fixed effect coefficients for the observation-level traits.

3.2. **prior elicitation.** Intuitive prior distributions are important for effective prior elicitation in Bayesian models. In most implementations of mixed effect models, inverse-Gamma priors are used for the variance components. This prior form is useful because of conjugacy to the likelihood. However, specifying hyperparameters for multiple variance components of this form can be difficult, especially when the prior information is highly diffuse. Gibbs samplers based on inverse-Gamma priors are known to have poor mixing properties, and so “parameter-expanded” forms are commonly used (ex. *MCMCglmm*, Gelman et al 2006?). These problems are exacerbated in the multivariate mixed model. The conjugate distribution of the covariance of multivariate random effects is the inverse-Wishart distribution, which is particularly difficult to use to specify vague knowledge and is generally not very flexible.

In BSFG, we used inverse-Gamma priors for the residual variances of each of the p observational-level traits, but used a recently proposed re-formulation of the random effect model in terms of heritability and total variance to specify priors for the factor variances

(Zhou?). Heritability (σ_a^2/σ_p^2) as a proportion of variance is more intuitive and easier to visualize than a variance, and is independent of the scale of the data, making prior elicitation easier. We used a discrete prior directly on heritability to permit highly flexible prior distributions.

In SLAMglmm, we extend this idea to the observational-level traits as well.