Logistic mixed model regression

Problem statement

The aim of this project is to develop a logistic mixed model formulation for a scalable implementation in Hail (hail.is) framework. General from of a logistic mixed model is $y = X\beta + Z\gamma$, where $\gamma \sim \mathcal{N}(0, \delta_g)$..

Generative model for this logistic model would be defined as

$$y_i \sim \text{Bern}(\sigma(p_i))$$

$$\sigma(p_i) = \frac{\exp(p_i)}{1 + \exp(p_i)}$$

$$p_i \sim \mathcal{N}(x_i\beta, \delta_g K)$$

where K is Kinship matrix $GG^\intercal$, δ_g is genetic additive variance, β is coefficient estimates for fixed effects, X and Z are design matrices of fixed and random effects respectively.

Our approach relies on Polson *et al.* [1] in deriving a data augmentation scheme for logistic regression similar to one for probit regression proposed by Albert and Chib [2]. The key to this approach is the derivation of the Polya-Gamma (PG) distribution $PG(b, c)$. The expectation of random variables (ω) from this distribution have a closed from of $\mathbf{E}(\omega) = \frac{b}{c}\tanh(c/2)$.

A posterior distribution for coefficients β with prior $\beta \sim \mathcal{N}(b, B)$ is obtained from iterative Gibbs sampler of the following form

$$(\omega_i \mid \beta) \sim \mathcal{PG}(n_i, x_i^\intercal)$$

$$(\beta \mid \omega, y) \sim \mathcal{N}(m_\omega, V_\omega)$$

where

$$V_\omega = (X^\intercal \Omega X + B^{-1})^{-1}$$

$$m_\omega = V_\omega(X^\intercal \kappa + B^{-1}b$$

Note that extension of this framework from fixed to mixed effects is done by incorporating random effects along with fixed effects into design and coefficient matrices so that ω latent variables are sampled in the same step. Priors of such distinct elements can combine various covariance and initial mean estimates in block structure. R package BayesLogit contains implementations for both random and fixed effect solutions with the Gibbs sampler [1]. A drawback for scalability is the unknown and possibly large number of iterations required for stable posterior estimates. Our aim is to increase the speed of computations so that the implementation in Hail would be feasible for large scale genetic association analyses.

Logistic Polya-Gamma mixed model with expectation-maximisation

This approach adapts the Logistic PG mixed model from using a Gibbs sampler in iterative parameter adjustment to an expectation-maximisation (EM) algorithm. Let's define $\alpha = (\beta, \gamma)$ as the joint vector of fixed random parameters and $A = [XZ]$ is the design matrix of all included effects. Additionally, V refers to block-covariance matrix of all effects and v to initial mean estimates. An EM algorithm is considerably faster but converges to a single point estimate instead of forming a posterior distribution. Changes to the original algorithm include replacing the sampling steps with an E step for expected values of ω based on closed form solutions in Polya-Gamma distribution and M step for α estimates using the maximum likelihood estimates (fixed+random parameter) coefficients as in Bayesian linear regression. Specifically the EM algorithm iterates until convergence between

$$\mathbf{E}(\omega) = \frac{1}{A\alpha} \tanh((A\alpha)/2)$$

$$\hat{\alpha} = ((A^{\intercal}\Omega A + V^{-1})^{-1})A^{\intercal}\kappa + V^{-1}v$$

where Ω is the diagonal matrix of ω_i, $i = 1 \ldots n$, n being number of variants

Alternative logistic Polya-Gamma mixed model with expectation-maximisation

This variation of the algorithm additionally divides the maximisation step into two separate phases for fixed effects β and γ. Let's define a working response that allows the switch into Gaussian framework through Polya-Gamma distributed variables as

$$\Omega^{-1}\kappa \sim \mathcal{N}(X\beta, \Omega^{-1}) + Z\gamma = \mathcal{N}(X\beta, \Omega^{-1}) + Z\mathcal{N}(0, \sigma_g) = \mathcal{N}(X\beta, \Omega^{-1} + \sigma_g K)$$

This leads to a 3 step iteration process of β, γ and ω. In each iteration, solutions for the first two parameters come from Bayesian maximum likelihood estimates of normally distributed variables.

First

$$\beta \mid \omega \sim \mathcal{N}(m_\omega, V_\omega)$$

$$V_\omega = (X^\intercal(\Omega^{-1} + \delta_g K)^{-1} X + B^{-1})^{-1}$$

$$m_\omega = V_\omega(X^\intercal(\Omega^{-1} + \delta_g K)^{-1}\Omega^{-1}\kappa + B^{-1}b$$

Second

$$\omega_i \mid \beta, \gamma \sim \mathcal{PG}(1, \psi)$$

$$\psi = A\alpha, \psi \sim \mathcal{N}(X\beta, \delta_g K)$$

$$\mathbf{E}(\omega) = \frac{1}{A\alpha}\tanh((A\alpha)/2)$$

Third

$$\gamma \mid \omega \sim \mathcal{N}(m'_\omega, V'_\omega)$$

$$V'_\omega = (Z^\intercal(\Omega^{-1} + \Sigma)^{-1} Z + \Sigma^{-1})^{-1}$$

$$m'_\omega = V_\omega(Z^\intercal(\Omega^{-1} + \Sigma)^{-1}\Omega^{-1}\kappa + \Sigma^{-1}b'$$

where Σ and b' are a diagonal covariance matrix and a random effect prior mean vector.

R package: hailLogitMMSupport

Preliminary implementations of the above mentioned methods were implemented in R and packaged in a R library hailLogitMMSupport. This small library consists of 4 algorithms that solve for logistic regression fixed effects. In genetic analyses, our primary interest lies in estimates of just the fixed effect parameters. Although random effects may have a skewing effect. hese Fixed effect parameter estimates are used in predicting probabilities of the phenotype $p(y = 1)$ which are employed in variant-wise GWAS through a likelihood ratio test as described by Chen et al. [3]. In addition to two above described models we have implemented an extra two models that disregard random effects and only estimate fixed effects. These are meant to be used either for reference in assessing extent of random effect induced change on a prospective dataset but also for just estimating standard logistic regression models. The implemented models in hailLogitMMSupport are

- *logit.PG.EM.mm.gen*. The model described in subsection "Logistic Polya-Gamma mixed model with expectation-maximisation".
- *logit.PG.EM.fastlmm.ksisamp*. The model described in Alternative logistic Polya-Gamma mixed model with expectation-maximisation
- *logit.EM.R*. A simplification of *logit.PG.EM.mm.gen* that only deals with fixed effects.
- *logit.PG.EM.mm.gen.gibbs*. A Gibbs Sampling mixed effect algorithm similar to the analogous method in BayesLogit R package. It differs by assuming normal distribution of the random effect parameters instead of original Gamma distribution. It returns a posterior of fixed effect parameters.

Note that in a mixed effect setting the (co)variances of fixed (c) and random parameters (phi) are unknown. Adjustment to these parameters affects the final parameter estimates. The package contains a very simple method for data generation that can be used for trying out these methods (function: *Data()*) and a method (function: *generateBenchmarks()*) that applies a dataset to 3 of the above functions (in addition to standard glm in R) and outputs the fixed effect estimates given varying covariance parameters.

The package is available in github through https://github.com/ttasa/hailLogitMMSupport and can be installed using *devtools::install_github* function.

Kirjandus

[1] N. G. Polson, J. G. Scott, and J. Windle, “Bayesian inference for logistic models using Polya-Gamma latent variables,” *arXiv:1205.0310 [stat]*, May 2012. arXiv: 1205.0310.

[2] J. H. Albert and S. Chib, “Bayesian Analysis of Binary and Polychotomous Response Data,” *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[3] H. Chen, C. Wang, M. P. Conomos, A. M. Stilp, Z. Li, T.Šofer, A. A. Szpiro, W. Chen, J. M. Brehm, J. C. Celedón, S. Redline, G. J. Papanicolaou, T. A. Thornton, C. C. Laurie, K. Rice, and X. Lin, “Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models,” *American Journal of Human Genetics*, vol. 98, pp. 653–666, Apr. 2016.