APPENDIX FOR GCPBayes: An R package for studying Cross-Phenotype Genetic Associations with Group-level Bayesian Meta-Analysis

by T. Baghfalaki, P.E. Sugier, Y. Asgari, T. Truong, and B. Liquet

Appendix A: Material and methods

Let $\hat{\boldsymbol{\beta}}_k,\ k = 1, \cdots, K$ be a m-dimensional vector of the regression coefficients for the k^{th} study and $\hat{\boldsymbol{\Sigma}}_k$ be its estimated covariance matrix. Also, let $\hat{\boldsymbol{\beta}}_k|\boldsymbol{\beta}_k, \hat{\boldsymbol{\Sigma}}_k \sim N_m(\boldsymbol{\beta}_k, \hat{\boldsymbol{\Sigma}}_k),\ k = 1, \cdots, K$. Therefore, the summary statistics for each group is $\hat{\boldsymbol{\beta}}_k$ and $\hat{\boldsymbol{\Sigma}}_k,\ k = 1, \cdots, K$.
We consider three different priors for $\boldsymbol{\beta}_k$: continuous spike (CS), Dirac spike (DS) and hierarchical spike (HS). As mentioned before, the CS and DS are designed to detect pleiotropic effect at group level, but HS is designed to detect pleiotropy at both group and variable level. In the following, these methodologies are described briefly. For more details, see Baghfalaki et al. (2021).

Continuous spike

The hierarchical set-up of CS prior, by considering summary statistics as the input of the method, is given by:

$$\begin{aligned}
&\boldsymbol{\beta}_k|\xi_k, \sigma^2, \tau^2 \overset{ind}{\sim} (1-\xi_k)N_m(\mathbf{0}, \sigma^2\mathbf{I}_m) + \xi_k N_m(\mathbf{0}, \tau^2\mathbf{I}_m), \tau^2 > \sigma^2 > 0,\ k = 1, \cdots, K,\ m \geq 1,\\
&\xi_k|\kappa \overset{ind}{\sim} Ber(\kappa),\\
&\kappa|a_1, a_2 \sim Beta(a_1, a_2),\\
&\tau^2|c_1, c_2 \sim \text{I}\Gamma(c_1, c_2),
\end{aligned} \tag{A.1}$$

where σ^2 is a fixed value and the latent variable ξ_k is considered for taking into account the association of studies. If $\xi_k = 0$, then $\boldsymbol{\beta}_k \sim N_m(\mathbf{0}, \sigma^2\mathbf{I}_m)$. Otherwise, $\boldsymbol{\beta}_k \sim N_m(\mathbf{0}, \tau^2\mathbf{I}_m)$ and its components can be considered as non-zero values for large values of τ^2. The notation $\text{I}\Gamma(c_1, c_2)$ denotes an inverse gamma distribution with parameters c_1 and c_2. Note that the value of σ^2 should be small (e.g. 10^{-3} or 10^{-4}).

Dirac spike

The hierarchical set-up of DS prior is given by:

$$\begin{aligned}
&\boldsymbol{\beta}_k|\sigma^2, \kappa \overset{ind}{\sim} (1-\kappa)\delta_0(\boldsymbol{\beta}_k) + \kappa N_m(\mathbf{0}, \sigma^2\mathbf{I}_m),\ k = 1, \cdots, K,,\ m \geq 1,\\
&\kappa|a_1, a_2 \sim Beta(a_1, a_2),\\
&\sigma^2|d_1, d_2 \sim \text{I}\Gamma(d_1, d_2),
\end{aligned} \tag{A.2}$$

where $\delta_0(\boldsymbol{\beta}_k)$ denotes a point mass at $\mathbf{0} \in R^m$, such that $\delta_0(\boldsymbol{\beta}_k) = 1$ if $\boldsymbol{\beta}_k = \mathbf{0}$ and $\delta_0(\boldsymbol{\beta}_k) = 0$ if at least one of the m components of $\boldsymbol{\beta}_k$ is non-zero, that is $\boldsymbol{\beta}_k \neq \mathbf{0}$.

Hierarchical spike

For considering a prior with ability to have shrinkage effects at both the group-level and the variable-level, we cannot apply the prior distribution on $\boldsymbol{\beta}_k$ directly; instead, a reparameterization of $\boldsymbol{\beta}_k$ is considered as $\boldsymbol{\beta}_k = \boldsymbol{V}_k^{1/2}\boldsymbol{b}_k, \boldsymbol{V}_k^{1/2} = diag(\tau_{k1}, \cdots, \tau_{km})$. In order to define a prior with these properties, two spike and slab priors are considered in a hierarchical setup,

one of them for $\boldsymbol{b}_k$ and the other one for $\tau_{kj},\ j = 1, \cdots, m$, leading to a HS prior. The HS prior is as follows:

$$\begin{array}{l} \boldsymbol{\beta}_k = \boldsymbol{V}_k^{1/2}\boldsymbol{b}_k, \boldsymbol{V}_k^{1/2} = diag(\tau_{k1}, \cdots, \tau_{km}),\ k = 1, \cdots, K,\ m > 1, \\ \boldsymbol{b}_k | \kappa \overset{ind}{\sim} (1-\kappa)\delta_0(\boldsymbol{b}_k) + \kappa N_m(\boldsymbol{0}, \sigma^2 \mathbf{I}_m), \\ \tau_{kj} | \kappa^*, s^2 \overset{ind}{\sim} (1-\kappa^*)\delta_0(\tau_{kj}) + \kappa^* N^+(0, s^2), j = 1, 2, \cdots, m, \\ \kappa | a_1, a_2 \sim Beta(a_1, a_2), \\ \kappa^* | c_1, c_2 \sim Beta(c_1, c_2), \\ \sigma^2 | d_1, d_2 \sim \Pi\Gamma(d_1, d_2), \\ s^2 | e_1, e_2 \sim \Pi\Gamma(e_1, e_2), \end{array} \tag{A.3}$$

where $N^+(0, s^2)$ denotes a univariate truncated normal distribution at zero with mean 0 and variance s^2. The value of e_1 is set to be 1. The value of e_2 is estimated using the Monte Carlo EM algorithm (MCEM) which leads to an empirical Bayes Gibbs sampler, such that, for the t^{th} EM update, we have $e_2^{(t)} = \frac{1}{E_{e_2^{(t-1)}}(\frac{1}{s^2} | \hat{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\Sigma}}_k,\ k=1,\cdots,K)}$.

Appendix B: Simulated examples

Summary statistics level data for K=5

We consider the summary level data for one group with $m = 10$ variables for $K = 5$ studies. The regression coefficients are simulated using a distributional assumption as follow:

$$\hat{\boldsymbol{\beta}}_k \sim N_m(\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_k),\ k = 1, 2, ..., K, \tag{B.4}$$

where $\boldsymbol{\Sigma}_k = \boldsymbol{SRS}$ such that $\boldsymbol{S} = diag(\boldsymbol{s})$ is a scale diagonal matrix with the main diagonal entry a standard error vector $\boldsymbol{s}$. We assume that $\boldsymbol{s} = s\boldsymbol{1}_m$ and $s = 0.05$. Also, $\boldsymbol{R}$ is a compound-symmetry correlation matrix with non-diagonal components ρ with $\rho = 0.25$. We consider 30% intra-group sparsity using the form of the $\boldsymbol{\beta}_k$ as follows:

$$\boldsymbol{\beta}_k = (\underbrace{\beta, \cdots, \beta}_{7}, 0, 0, 0)',\ k = 1, 2, \tag{B.5}$$

$$\boldsymbol{\beta}_3 = \boldsymbol{\beta}_1, \tag{B.6}$$

$$\boldsymbol{\beta}_k = \boldsymbol{0},\ k = 4, 5, \tag{B.7}$$

where β is the magnitude of the effects value getting the value 0.4 and -0.4 with equal probability.

For this purpose, the following R commands can be considered:

```
> K <- 5
> m <- 10
> set.seed(12345)
> sign <- rbinom(m, 1, 0.5)
> sign[sign == 0] <- -1
> betat <- c(rep(0.4, m * 0.7), rep(0, m * 0.3))
> BETA <- matrix(0, K, m)
> betat <- betat * sign
> BETA[1, ] <- BETA[3, ] <- betat
> BETA[2, ] <- -betat
> corr <- 0.25
> S <- diag(0.05, m)
> R <- matrix(corr, m, m) + (1 - corr) * diag(m)
> Sigmat <- S %*% R %*% S
> simSIGMA <- list()
> simBeta <- list()
```

```
> for (tab in 1:K) {
+   simBeta[[tab]] <- mvtnorm::rmvnorm(1, BETA[tab, ], Sigmat)
+   simSIGMA[[tab]] <- Sigmat
+ }
> snpnames <- sprintf("SNP%s", seq(1:m))
> genename <- "simulated_example"
```

This data is embeded in the **GCPBayes** package and a user could extract it by the following commands and run any of the **GCPBayes** methods on it:

```
> library(GCPBayes)
> data(Simulated_summary)
> genename <- Simulated_summary$genename
> snpnames <- Simulated_summary$snpnames
> Betah <- Simulated_summary$simBeta
> Sigmah <- Simulated_summary$simSIGMA
```

Individual level data for K=3

We consider the individual level data for continuous phynotypes of one group with $m = 30$ variables for $K = 3$ studies such that the sample sizes of the studies are $n_1 = 1200$, $n_2 = 1000$ and $n_3 = 2000$. For $n_k \times m$ (k=1,2,3) genotypes matrix $\boldsymbol{X}$, we follow the approach of Stanislas et al. (2017) and Broc et al. (2021). The approach is a simple strategy to control the minor allele frequency (MAF) of each SNP which are coded as minor allele counting $\{0, 1, 2\}$. Also, there is a group structure such that the SNPs of a groups are from the same linkage disequilibrium block. Each line of the genotype matrix $\boldsymbol{X}$ is a random sample from a multivariate random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{S}$. We consider a exchangeable structure $(s = 0.6)$. In order to have genotype data, each SNP is randomly assigned an MAF probability (we call it MAF_j, $\text{j} = 1, \cdots, \text{P}$). A value of $\text{MAF}_\text{j} = 0.25$ is assigned to all SNPs. The Hardy-Weinberg equation are then applied to discretize X_{ij} to 0, 1 and 2. In practice, X_{ij} is set to 0 if $X_{ij} < \text{Quartile}(\boldsymbol{X}_\text{i}, (1 - \text{MAF}_\text{j})^2)$, X_{ij} is set to 2 if $X_{ij} > \text{Quartile}(\boldsymbol{X}_\text{i}, \text{MAF}_\text{j}^2)$ and X_{ij} is set to 1 otherwise, where $\text{Quartile}(\boldsymbol{X}, \text{p})$ denote to p^{th} quantile of variable $\boldsymbol{X}$. We consider the following regression model for study k, $k = 1, 2, 3$,

$$\boldsymbol{Y}_k = \boldsymbol{X}_k \boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k,\ k = 1, 2, 3,$$

where $\boldsymbol{\varepsilon}_k \sim N(\mathbf{0}, \boldsymbol{I})$, $k = 1, 2, 3$, and

$$\begin{aligned}
\boldsymbol{\beta}_1 &= (\underbrace{0, \cdots, 0}_{m-2}, -0.5, 0.5)', \\
\boldsymbol{\beta}_2 &= (0.5, \underbrace{0, \cdots, 0}_{m-2}, 0.5)', \\
\boldsymbol{\beta}_3 &= (\underbrace{0, \cdots, 0}_{m-3}, 0.5, 0, 0)'.
\end{aligned}$$

For this purpose, the following R commands can be considered for the first study, the R commands for other studies are the same:

```
> library(mvtnorm)
> set.seed(12345)
> n <- 1200
> m <- 30
> rho <- 0.6
> Sigma <- diag(1 - rho, m) + matrix(rho, m, m)
> Xs <- rmvnorm(n, rep(0, m), Sigma)
> MAF <- 0.25
> q1 <- (1 - MAF)^2
> q2 <- (1 - MAF)^2 + 2 * (1 - MAF) * MAF
```

```
> Q1 <- quantile(Xs, q1)
> Q2 <- quantile(Xs, q2)
> X1 <- matrix(1, n, m)
> X1[Xs > Q2] <- 2
> X1[Xs < Q1] <- 0
> Beta <- c(rep(0, m - 2), -0.5, .5)
> Y1 <- c()
> for (i in 1:n) {
+   Y1[i] <- rnorm(1, X1[i, ] %*% Beta, 1)
+ }
> colnames(X1) <- sprintf("SNP%s", seq(1:m))
> Study1 <- cbind(Y1, X1)
```

This data is embeded in the **GCPBayes** package and a user could extract it by the following commands and run any of the **GCPBayes** methods on it:

```
> library(GCPBayes)
> data(Simulated_individual)
> Study1 <- Simulated_individual$Study1
> Study2 <- Simulated_individual$Study2
> Study3 <- Simulated_individual$Study3
```

Survival outcomes and gene expression data for K=2
We consider individual level data for survival outcomes of one group with $m = 10$ variables for $K = 2$ studies such that the sample sizes of the studies are $n_1 = 500$ and $n_2 = 600$. The gene expression data are generated the same as Bair et al. (2006) and Van Wieringen et al. (2009). Thus, the gene expression data for each study is distributed as:

$$\log(X_{ij}) = \begin{cases} 3 + \varepsilon_{ij} & i \leq \frac{n}{2}, j \leq 5, \\ 4 + \varepsilon_{ij} & i > \frac{n}{2}, j \leq 5, \\ 3.5 + \varepsilon_{ij} & j > 5, \end{cases}$$

where $i = 1, \cdots, n_k$, $j = 1, \cdots, m$, the ε_{ij} are drawn from a standard normal distribution. The survival and censoring times are exponentially distributed and 30% is considered censoring rate. Also, the signals are considered as follows:

$$\boldsymbol{\beta}_1 = (0.5, 0.25, \underbrace{0, \cdots, 0}_{m-2})',$$
$$\boldsymbol{\beta}_2 = (0, 0.25, 0.25, \underbrace{0, \cdots, 0}_{m-3})'.$$

For this purpose, the following R commands can be considered for the first study, the R commands for the other study is the same:

```
> library(survival)
> library(BhGLM)
> set.seed(12345)
> n <- 500
> m <- 10
> X <- matrix(0, n, m)
> for (i in 1:n) {
+   for (j in 1:m) {
+     if (i <= n / 2 & j <= m / 2) (X[i, j] <- exp(3 + rnorm(1)))
+     if (i > n / 2 & j <= m / 2) (X[i, j] <- exp(4 + rnorm(1)))
+     if (j > m / 2) (X[i, j] <- exp(3.5 + rnorm(1)))
+   }
+ }
> B <- c(-0.5, -0.25, rep(0, m - 2))
```

```
> lambda <- exp(-X %*% B)
> Y <- rexp(n, rate = lambda)
> C <- quantile(Y, .7)
> Y[Y > C] <- C
> delta <- rep(0, n)
> delta[Y == C] <- 1
> T <- Surv(Y, delta)
> colnames(X) <- sprintf("Gene%s", seq(1:m))
> Study1 <- list(T = T, X = X)
```

This data is embeded in the **GCPBayes** package and a user could extract it by the following commands and run any of the **GCPBayes** methods on it:

```
> data(Simulated_individual_survival)
> Study1 <- Simulated_individual_survival$Study1
> Study2 <- Simulated_individual_survival$Study2
```

Also, the summary statistics data could be computed by the following commands:

```
> Fit1=BhGLM::bcoxph(Study1$T ~ Study1$X)
> Betah1 <- Fit1$coefficients
> Sigmah1 <- Fit1$var
> Fit2=BhGLM::bcoxph(Study2$T ~ Study2$X)
> Betah2 <- Fit2$coefficients
> Sigmah2 <- Fit2$var
> Betah <- list(Betah1,Betah2)
> Sigmah <- list(Sigmah1,Sigmah2)
```

Bibliography

T. Baghfalaki, P.-E. Sugier, T. Truong, A. N. Pettitt, K. Mengersen, and B. Liquet. Bayesian meta-analysis models for cross cancer genomic investigation of pleiotropic effects using group structure. *Statistics in Medicine*, 40(6):1498–1518, 2021. [p1]

E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006. [p4]

C. Broc, T. Truong, and B. Liquet. Penalized partial least squares for pleiotropy. *BMC bioinformatics*, 22(1):1–31, 2021. [p3]

V. Stanislas, C. Dalmasso, and C. Ambroise. Eigen-epistasis for detecting gene-gene interactions. *BMC bioinformatics*, 18(1):1–14, 2017. [p3]

W. N. Van Wieringen, D. Kun, R. Hampel, and A.-L. Boulesteix. Survival prediction using gene expression data: a review and comparison. *Computational statistics & data analysis*, 53 (5):1590–1603, 2009. [p4]

Taban Baghfalaki
Tarbiat Modares University
Faculty of Mathematical scinces,
Department of Statitics,
Tehran, Iran.
Université Paris-Saclay; UVSQ; INSERM, U1018; Gustave Roussy; CESP, Team Exposome and Heredity
Villejuif, France
ORCiD: 0000-0002-2100-4532
t.baghfalaki@modares.ac.ir

Pierre-Emmanuel Sugier
Universite de Pau et des Pays de l'Adour
Laboratoire de Mathématiques et de leurs Applications de Pau
UMR CNRS 5142, E2S-UPPA, France
Université Paris-Saclay; UVSQ; INSERM, U1018; Gustave Roussy; CESP, Team Exposome and Heredity
Villejuif, France
ORCiD: 0000-0002-5846-1104
pe.sugier@univ-pau.fr

Yazdan Asgari
Université Paris-Saclay; UVSQ; INSERM, U1018; Gustave Roussy; CESP, Team Exposome and Heredity
Villejuif, France
ORCiD: 0000-0001-6993-6956
yazdan.asgari@inserm.fr

Thérèse Truong
Université Paris-Saclay; UVSQ; INSERM, U1018; Gustave Roussy; CESP, Team Exposome and Heredity
Villejuif, France
ORCiD: 0000-0002-2943-6786
therese.truong@inserm.fr

Benoit Liquet
Macquarie University
School of Mathematics and Physical Sciences
NSW, Australia
Universite de Pau et des Pays de l'Adour
Laboratoire de Mathématiques et de leurs Applications de Pau
UMR CNRS 5142, E2S-UPPA, France
ORCiD: 0000-0002-8136-2294
benoit.liquet-weiland@mq.edu.au, benoit.liquet@univ-pau.fr