Clustering of Variants based on GWAS Summary Statistics Using a Gibbs Sampling Approach

Jerome Freudenberg and David Tang

March 18, 2021

Abstract

We apply Gibbs sampling to the summary statistics of a multi-trait GWAS of 35 blood and urine biomarkers to cluster variants on their effects. We implement two models: a simple binary model that only takes into account genome-wide significance and a continuous model that incorporates effect size magnitude and direction. We then use these clusters to compare variant cluster proportions across traits to analytically determine trait groupings. We find distinct variant clusters and biologically interpretable trait groupings that are not completely aligned with the pre-defined trait categories.

1 Introduction

The rapidly increasing amount of individual health and genetic data has resulted in a substantial rise in variant-phenotype associations; however, individual variant function and contributions to phenotype are not always clear. Our project leverages information from the phenotypic associations of each variant to examine which variants may have connected functionality. We apply Gibbs sampling to the summary statistics of a recent multi-trait GWAS, which identified 1,857 loci associated with at least one of 35 blood and urine biomarkers in UK Biobank data [3], to cluster variants based on their effects.

Additionally, the traits in the UKB paper were assigned to six categories: Bone and Joint, Cardiovascular, Diabetes, Hormone, Liver, and Renal, but it is not clear how these labels were decided. We use the cluster assignments to examine which traits have similar proportions of variants in each cluster to analytically determine trait categories.

2 Methods and Data

2.1 Data

We used summary statistics from the recently published GWAS for 35 different blood and urine biomarkers in the UKBiobank [3]. For the clustering, we focused on 5754 protein altering and non-coding variants that achieved genome wide significance for at least one of the 35 biomarkers.

2.2 Graphical Model

To cluster SNPs into different groups, we considered the following graphical model. Here, x is the observed data, z is the latent states, $\pi = (\pi_1, \dots, \pi_k)$ is the mixture proportions, and $\theta = (\theta_1, \dots, \theta_k)$ is the k parameters for the k mixture component likelihoods. Our goal was to find the joint posterior of (z, π, θ) given the data. To do this, we used the hierarchical model depicted in figure 1.

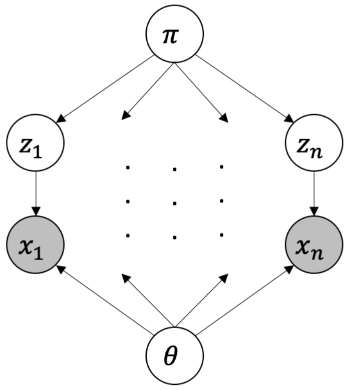

Figure 1: The graphical model representing the conditional relationships between variants and clusters

In order to fully specify this model, we needed to specify the priors on π and θ as well as the conditional distributions that relate the variables to each other. To this extent, we considered two different models: a binary model based on genome-wide significance and a continuous model that leverages the specific effect sizes.

2.3 Binary Model

In the binary model, the data $x_i = (x_{i1}, \dots, x_{ir})$ is a vector of 1's and 0's where x_{ij} is an indicator for whether or not SNP i reached genome wide significance for biomarker j. Intuitively, we would expect similar variants to act on related traits. This means that we would expect similar variants to reach genome wide significance on a set of related traits.

In terms of modeling, this set up is equivalent to the allele frequency models we considered for haploid data, making the simplifying assumption that biomarkers are independent. Specifically, we can fit the following hierarchical model:

$$\begin{aligned}
\pi &\sim \mathrm{Dirichlet}(1,\ldots,1)\\
f_{lj} &\sim \mathrm{Beta}(1,1)\\
x_{ij}|z_i = l, f &\sim \mathrm{Bernoulli}(f_{lj})
\end{aligned}$$

In relation to the graphical model above, we have that $\theta_l = f_l = (f_{l1},\ldots,f_{lr})$ is a vector of probabilities giving the probabilities of a variant from cluster l passing the genome wide significance level for the different biomarkers. We placed Dirichlet and Beta priors on π and f respectively to leverage the conjugacy properties of Bernoulli likelihoods.

2.4 Multivariate Normal Model

Using a SNPs genome wide significance across multiple traits to cluster the data is a reasonable first approach, but doing so fails to incorporate any information about the effect size. We would like to not only cluster variants by the traits they are associated with, but also by the strength and directionality of the association. To do this, we instead modeled x_i as a vector of effect sizes, where x_{ij} gives the effect size of variant i on biomarker j in the GWAS.

In terms of modeling assumptions, we assumed that each data point x_i came from a mixture of multivariate normal distributions. In this model, $\theta = ((\mu_1,\Sigma_1),\ldots,(\mu_r,\Sigma_r))$ is a vector of pairs of mean and covariances that parameterize the multivariate normal distribution of the r different mixture components such that the conditional density of x_i is

$$x_i|z_i = l, \mu_l, \Sigma_l \sim N_r(\mu_l, \Sigma_l)$$

We once again leveraged conjugation by placing a Normal-inverse-Wishart prior on the pair (μ_l, Σ_l). The Normal-inverse-Wishart distribution is parameterized by 4 parameters, μ_0, λ, Ψ, and ν. In terms of these four parameters, we have that the posterior distribution of (μ_l, Σ_l) conditional on cluster membership is

$$\mu_l, \Sigma_l | x, z = l \sim NIW(\mu_n, \lambda_n, \Psi_n, \nu_n)$$

where μ_n, λ_n, Ψ_n, and ν_n are parameters given by the following

$$\begin{aligned}
\mu_n &= \frac{\lambda\mu_n + n_l\bar{x}}{\lambda + n_l}\\
\lambda_n &= \lambda + n_l\\
\Psi_n &= \Psi + S + \frac{\lambda}{\lambda + n_l}(\bar{x} - \mu_0)^T(\bar{x} - \mu_0)\\
S &= (x^l - \bar{x})^T(x^l - \bar{x})\\
\nu_n &= \nu + n_l
\end{aligned}$$

In these equations, n_l denotes the number of observations from mixture component l and x^l denotes the subset of the data from mixture component l. The parameters μ, λ, Ψ, and ν are chosen uniformly for all mixture components with $\mu = 0$, $\lambda = 1$, $\Psi = I_{r\times r}$, and $\nu = r$.

2.5 Trait Categorization

As a secondary goal, we wanted to find a way to use summary statistics to cluster traits into various categories analytically. In theory, we should be able to do this with a similar MCMC scheme, viewing the data as an m dimensional vector where m is the number of variants instead of an r dimensional vector where r is the number of traits. This sampler, however, would require significantly more computational resources. In the continuous model, we would need to store $m \times m$ matrices to parameterize the covariance of an m-variate normal distribution. This is infeasible with the number of variants we are considering and the limited computational resources we have.

Instead of directly clustering traits by fitting a mixture model, we can indirectly group them based on the cluster memberships of their genome wide significant variants. Specifically, we can represent each trait in k dimensional space such that the l^{th} component corresponds to the proportion of its genome wide significant SNPs that are assigned to cluster l. By looking at a subset of components, we can project of each trait into a 2 dimensional space where we expect similar traits to cluster together.

2.6 Rank Inverse Normal Transform

To better fit a normal distribution, we first performed a rank-based inverse normal transform on the effect sizes. This transformation clearly breaks our probability model of the effect sizes as a mixture of multivariate normal random variables, but we found this transformation necessary in order to prevent underflow error when fitting the model to the data.

2.7 Sampling

For both of the models described above, we obtained samples from the posterior distribution using a Gibbs sampler MCMC scheme. For the sake of time, we ran the sampler for 500 iterations and discarded the first 50 samples as part of the burn in phase. We assigned each variant x_i to its cluster based on the posterior mode of z_i.

3 Results

3.1 Binary Model

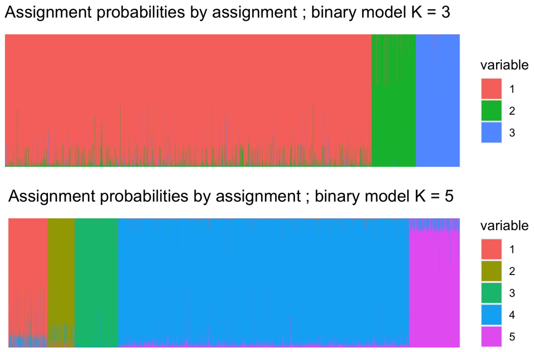

Figure 2: cluster assignment probabilities under the binary model

For $K = 5$ in the binary model, we found that around half the variants were assigned to the same cluster based off of their trait profiles (see figure 2). Most of the variants had a high probability of being assigned to the same cluster, indicating that the groupings are relatively distinct. Variants within a cluster that had a substantial probability of being assigned to other clusters shared their secondary assignments. Based off of preliminary exploration, these were usually variants only having one trait association so they could not be definitely clustered or ones that had downstream effects (e.g. a missense variant in the APOB gene also had effects on other cardiovascular biomarkers).

We chose $K = 5$ empirically based on which number of clusters had the most even distribution of variants. Additionally, we applied the method described in the STRUCTURE paper using Bayesian deviance for inference on K [1], which supported $K = 4$ or $K = 5$.

Based on the variant cluster assignments, we examined how traits were related to each other using their associated variant composition (see figure 3). As one would expect, traits known to have shared variants had similar proportions of variants across clusters; for example, HDL and Apolipoprotein A (ApoA), LDL and Apolipoprotein B (ApoB), eGFR and Creatinine clustered together along every dimension. Interestingly, traits did not completely group along their pre-defined trait categories. Creatinine and eGFR were grouped separately from most of the other traits labeled as Renal, such as Urate and Cystatin C. The traits labeled as cardiovascular were also separated on some dimensions, with HDL and ApoA forming one group, LDL and ApoB forming another, and triglycerides and cholesterol somewhere in between. Our distinctions are close enough to the pre-defined trait categories to be biologically plausible and suggest substantial heterogeneity in the genetic makeup of the trait categories.

3.2 Multivariate Normal Model

Based on the results of the binary model, we ran the continuous model with K defined to be either 3 or 5. Whereas the binary model assigned over half of the variants into one cluster, the continuous model produced much more balanced cluster sizes. This suggests that the clusters found in the continuous models are quite distinct from the clusters found in the binary model.

We can get a better sense of the identified clusters by visualizing the posterior mean of the multivariate normal mean parameter μ. Recall that μ is a 35 dimensional vector parameterizing the mean effect sizes for each trait. As such, we can visualize the mean parameters for each cluster by graphing (j, μ_{lj}) where j ranges across all 35 traits and l ranges across the k clusters. This particular visualization is shown in figure 6.

If the data truly followed our mixture model, we would expect the posterior means of the different clusters to be relatively different. In our case, we see a pretty large overall difference in magnitude between the estimate of the mean of cluster 1 compared to those of clusters 2 and 3. Combined with the assignment probabilities seen above, this seems to indicate that our model does capture some sort of structure in the effect sizes.

Comparing the model with 5 components against the model with 3 components, we see a mapping between clusters 1, 2, and 3 of the 3 component model and clusters 5, 1, and 3 of the 5 component model. In other words, the estimates of μ for clus-

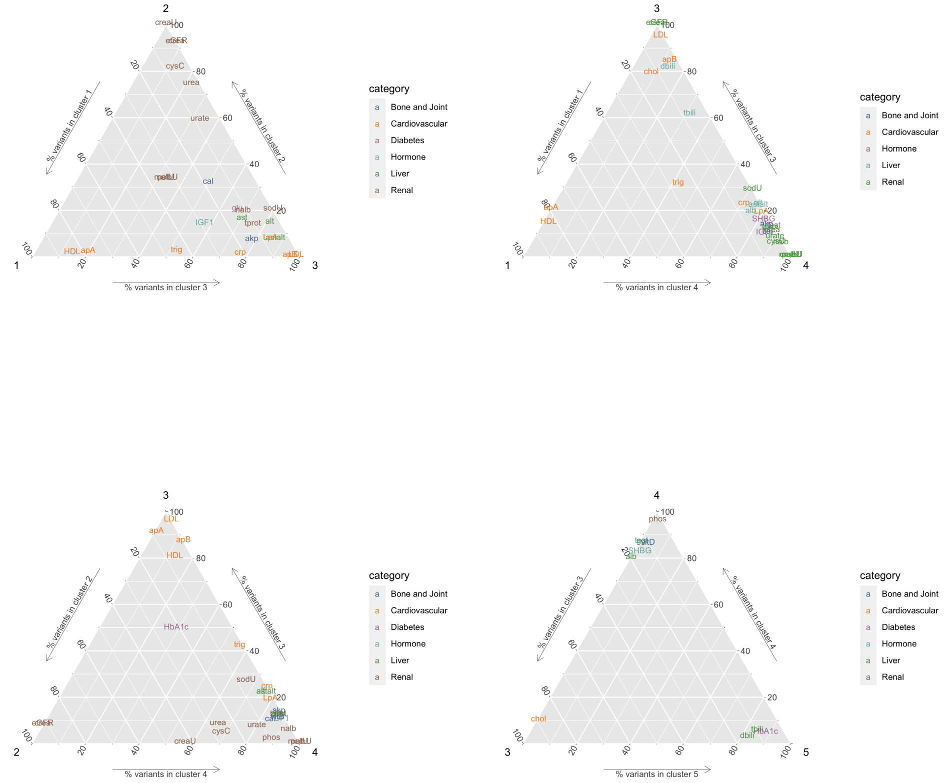

Figure 3: trait projections under the binary model

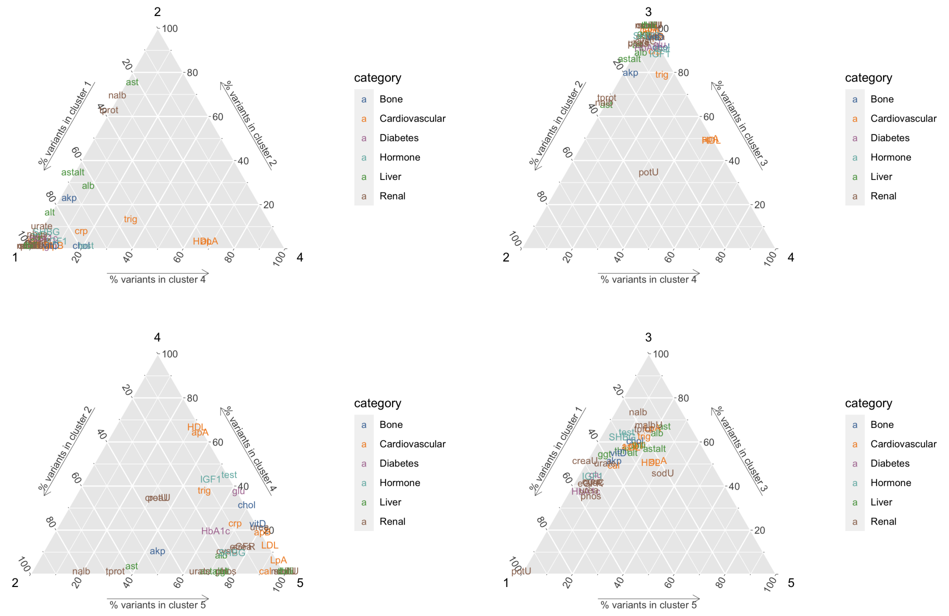

Figure 4: trait projections under the continuous model

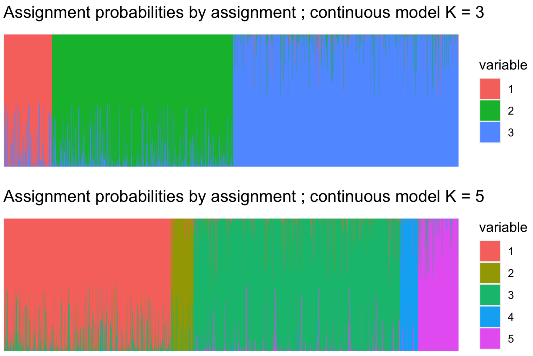

Figure 5: cluster assignment probabilities under the normal model

ters 1, 2, and 3 of the 3 component model match well with the estimates of μ for clusters 5, 1, and 3 of the 5 component model. The persistence of these three clusters further supports the idea that our model is finding underlying structure in the effect size summary statistics.

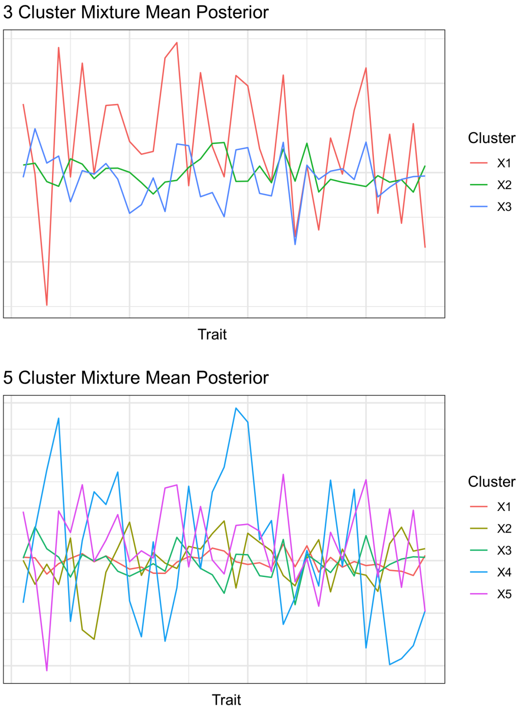

Figure 6: posterior mean for μ

As with the binary model, we projected traits based on the cluster assignements of their variants to estimate trait categories (see figure 4). In this case, the cluster proportions for each trait were not as distinct as in the binary model, with most traits falling into a single group.

4 Discussion

We used both a binary and continuous model for inference about distinctions in variant functionality and trait categorization. The former gave variants relatively definitive cluster assignments and resulted in trait similarities that were biologically interpretable. The latter incorporated information about association effect sizes for a more comprehensive representation of variant similarity, resulting in more diverse clustering, but less interpretable trait categories.

Variants that clustered into the same group have similar trait profiles, suggesting related functionality. The presence of variants with substantial probabilities of being in multiple clusters indicates that variants in different clusters are not completely distinct. Nevertheless, we were able to use the differences in variant effects between clusters to classify traits into reasonable groups.

There were multiple limitations to both of our approaches due to the simplifying assumptions we made. The binary model assumes independence between biomarkers and does not account for downstream affects. The continuous model incorporates many non-significant effect sizes, since we include all variants that reach genome-wide significance for at least one trait. This adds a lot of noise to the data and may be responsible for the poor distinguishability between trait cluster proportions. To improve our approach, we could incorporate the non-significant effect sizes though inverse-variance weights.

There are several possible extensions to our project. In our analysis, we combined protein-altering and non-coding variants, but it would be interesting to examine them separately. Additionally, we used the variant cluster assignments to determine trait categories, so it would be interesting to compare these with Gibbs sampling to cluster traits based on shared variant associations. Finally, we would want to compare our approach with other methods for estimating components of genetic associations on bio-bank scale summary statistics such as DeGAs [2].

5 Author Contributions

JF and DT were involved in conceptualization, method development, method implementation, and write-up. DT developed multivariate normal sampler and implementation. JF implemented binary sampler and plotting functionality.

6 Code Availability

Code and analysis available at https://github.com/JFreud/sumstats-cluster

References

[1] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, 06 2000.

[2] S. Sakaue, M. Kanai, Y. Tanigawa, J. Karjalainen, M. Kurki, S. Koshiba, A. Narita, T. Konuma, K. Yamamoto, M. Akiyama, K. Ishigaki, A. Suzuki, K. Suzuki, W. Obara, K. Yamaji, K. Takahashi, S. Asai, Y. Takahashi, T. Suzuki, N. Sinozaki, H. Yamaguchi, S. Minami, S. Murayama, K. Yoshimori, S. Nagayama, D. Obata, M. Higashiyama, A. Masumoto, Y. Koretsune, F. Gen, K. Ito, C. Terao, T. Yamauchi, I. Komuro, T. Kadowaki, G. Tamiya, M. Yamamoto, Y. Nakamura, M. Kubo, Y. Murakami, K. Yamamoto, Y. Kamatani, A. Palotie, M. A. Rivas, M. Daly, K. Matsuda, and Y. Okada. A global atlas of genetic associations of 220 deep phenotypes. *medRxiv*, 03 2021.

[3] N. Sinnott-Armstrong, Y. Tanigawa, D. Amar, N. Mars, C. Benner, M. Aguirre, G. Venkataraman, M. Wainberg, H. Ollila, T. Kiiskinen, A. Havulinna, J. Pirruccello, J. Qian, A. Shcherbina, F. Rodriguez, T. Assimes, V. Agarwala, R. Tibshirani, T. Hastie, and M. Rivas. Genetics of 35 blood and urine biomarkers in the uk biobank. *Nature Genetics*, 53, 02 2021.