Using the MVP Project to estimate Global Lipid Sharing

Sarah Urbut

Contents

1 Introduction

Variation in human lipids underlies many cardiovascular traits. While the heritability for any lipid trait is 40-60% ([1] , the proportion we can explain by specific variants is comparatively small. The novel approach we offer here allows groups of snps to be classified not only by their presence or absence of an effect n a particular lipid phenotype, but by their *relationships in continuous effects between lipid classes*, e.g., consistently larger effects in some lipids than other. Critically, quantifying the effect sizes of the snp pair across phenotypes considering the evidence contained in all lipid classes jointly thus reveals new patterns of activity across lipid classes, which differ in their relationship in sign and magnitude within and between conditions. The novel framework described here allows that these effects be 'shared' but not necessarily 'consistent' across lipid classes, thus capturing continuous variation heretofore missed. We hope to identify novel 'patterns' of sharing across lipid class, as well as to improve the power to detect a SNP that is 'modestly' associated with each phenotype class and thus exploit the boost in statistical power gained by joint analysis. By using an empirical bayes approach which learns the relative frequency of each pattern from the data to appropriately 'shrink' or nudge the posterior estimates of the effect, these shrunken estimates can then be used as updated effect size estimates in polygenic risk scoring for each lipid trait individuals, thus leveraging joint power in both the phenotypic space (across lipid classes) and later, in the genotypic space, but combining shrunken estimates to improve polygenic risk scoring.

These patterns of sharing can then be used to group like SNPs into classes by phenotypic pattern, which can ultimately help to to group mutation 'genetic risk' classes, as individuals with given mutation signatures might be expected to develop predictable patterns of lipid alterations across categories

1.1 Overview of Method

We assume we have a matrix of estimates of many 'effects' in multiple condition. For our purposes, we have a matrix of approximately 2 million SNPs with estimated (univariate) summary effects in each of four traits (lipids).

Our goal is to combine information across effects and conditions to produce improved estimates of the effects (and corresponding measures of significance) in each condition. In doing this we allow that some effects may be 'shared', being similar (though not identical) among conditions, while others may be 'specific' to only a subset of conditions.

Critically, we allow for subtle differences among those effects that are 'similar'. Here, we aim to learn about patterns of sharing of effects across lipids. Because these patterns are shared among SNPs, we can use the information contained in the larger data-set to help us better understand the global and SNP-specific patterns of effects of genetics on Lipid level. This allows us to make comparisons among phenotypes in which the SNP is called active, and among snps pairs with a similar degree of activity in a given phenotype. We will use **mash**to do so.

1.2 Model Description

We briefly summarize the **mash**model here [2]

Let b_{jr} $(j = 1, \ldots, J; r = 1, \ldots, R)$ denote the true value of effect j in condition (here lipid level) r. Further let $\hat{b}_{jr}$ denote the (observed) estimate of this effect, and $\hat{s}_{jr}$ the standard error of this estimate (so $\hat{b}_{jr}/\hat{s}_{jr}$ is the usual z statistic for testing whether b_{jr} is zero). We use $\boldsymbol{b}_j$ to denote the vector of effects $(b_{j1}, \ldots, b_{jR})$, and $\hat{\boldsymbol{b}}_j$ for the corresponding vector of estimates.

uses an Empirical Bayes (EB) approach. That is, we assume that the effects $\boldsymbol{b}_j$ come from some common ("prior") distribution g,

$$\boldsymbol{b}_j \sim g(\cdot), \tag{1}$$

and that the estimates $\hat{b}$ are normally distributed about the true effects:

$$\hat{b}_{jr}|b_{jr}, \hat{s}_{jr} \sim N(b_{jr}, \hat{s}_{jr}). \tag{2}$$

The EB approach first estimates g from the data, combining information across *all* effects, to yield an estimate $\hat{g}$ say. It then uses Bayes theorem to compute the posterior distribution for b_{jr} given the prior $\hat{g}$ and the data $(\hat{b}_{jr}, \hat{s}_{jr})$.

A key point is that the distribution g captures the similarity, or otherwise, of effects among conditions. Thus, in estimating g from the data, the EB approach adapts to the specific patterns of effects in the data at hand. If effects tend to be similar among conditions then the estimated $\hat{g}$ will reflect this, and the resulting posterior distributions will be "shrunk" (or "nudged") towards a common value, producing estimates that tend to be similar among conditions. Conversely, if effects tend to be specific to one or a few conditions then the estimated $\hat{g}$ and resulting estimates will reflect this.

To flexibly model the effect distribution g we use a mixture of multivariate normal (MVN) distributions:

$$g(\cdot; \boldsymbol{\pi}, \boldsymbol{U}) = \sum_{k=1}^{K} \sum_{l=1}^{L} \pi_{k,l} \; N_R(\cdot; \mathbf{0}, \omega_l U_k).$$

(3)

1.3 Interpretation

Here each covariance matrix U_k corresponds to a different "class" of effect, which allow for different patterns of sharing across subgroups of effects, whereas each scaling coefficient ω_l

corresponds to a different "size" of effect. The mixture proportions $\pi_{k,l}$ determine the relative frequency of each size-class combination. We allow the scaling coefficients ω_l to vary on a fixed dense grid of values spanning "very small" to "very large", to capture the full range of effects that could reasonably occur (as in [3]). Estimating g then involves estimating the parameters $\pi and \mathrm{U}. With \mathrm{U} fixed, it is a simply convex optimization problem$

The estimates $\hat{b}_{jr}$ are assumed to be independent and normally distributed about the true effects, and the true effects are assumed to follow (1.2), yielding

$$p(\hat{\boldsymbol{b}}_j|\boldsymbol{b}_j, V_j) = N_R(\hat{\boldsymbol{b}}_j; \boldsymbol{b}_j, V_j), \tag{4}$$

$$p(\boldsymbol{b}_j|,) = \sum_{k=1}^{K}\sum_{l=1}^{L} \pi_{k,l}\ N_R(\boldsymbol{b}_j; \mathbf{0}, \omega_l U_k). \tag{5}$$

where $N_R(\cdot; \boldsymbol{\mu},)$ denotes the density of the R-dimensional multivariate normal (MVN) distribution with mean $\boldsymbol{\mu}$ and covariance matrix , and the scaling parameters $\omega_1, \ldots, \omega_L$ are fixed on a dense grid. Here V_j denotes the $R \times R$ diagonal matrix with diagonal elements $s_{j1}^2, \ldots, s_{jR}^2$.

Assuming V_j to be diagonal in (4) is not strictly necessary: the methods implemented here apply for any user-supplied values for V_j, which represents the covariance matrix of the estimates $\hat{\boldsymbol{b}}_j$. Here we recognize that the variance of $V(\boldsymbol{\epsilon}) = V_j$ is likely not in fact diagonal, given the correlated errors among lipid phenotypes and thus it is necessary to use a covariance matrix which recognizes this correlation in the error structure so as not to erroneously estimate the effects. This is a new feature of the **mash** package, and will be essential to the analysis of the lipid data given the expected correlation in errors from a phenotype that is a linear combination of other phenotypes. However, because different individuals contribute summary statistics to each phenotype class, the expected errors are night as tightly correlated as otherwise might be expected.

1.4 Application

In our application to the Lipid data from the MVP project $R = 4$, so each U_k is a 4 by 4 covariance matrix, and each component of the mixture (1.2) is a distribution in 4 dimensions. Visualizing such a distribution is challenging, but we can get some insight from the first eigenvector of U_k, v_k say, which captures the principal direction of the effects in component k. If U_k is dominated by this principal direction (which is the case for many k in our application) then we can think of effects from that component as being of the form λv_k for some scalar λ. For example, if the elements of the vector v_k are approximately equal then component k captures effects that are approximately equal in all conditions. Or, if v_k has one large element, with other elements close to 0, then component k corresponds to an effect that is strong in only one condition. Though not evident in the gene expression case but possible in the more general case, these effects may be inversely correlated among subgroups, and our model also allows for this by estimating these patterns from the data.

2 Aim 1: Simulations

We can use the matrix of summary statistic data from the MVP project to create a J snp x 4 matrix of SNPS. It is crucial to recognize that many of these SNPS represent reproduction of one LD block, in which we can assume at most one causal SNP per LD Block using existing methods to partition the genome into appropriate blocks ([4]). These potentially 'causal' SNPs will then represent the 'max' set, and the correlated SNPs will represent the 'null' set accordingly, in line with the **mash** framework ([2] et al). From this, we select the top 5% as defined as maximum Z statistic across conditions to initialize our estimates of the variance matrix of the effects.

We can then simulate a data set with true effects reflecting these empirical patterns, as well as 90% null effects. Our aim will be to correctly characterize true associations as true without

erroneously concluding that the null associations are also real. We can compare to univariate shrinkage methods, as well as methods that provide fixed (configuration) approaches to joint estimation [5].

We can compare RRMSE and ROC curves under the setting of mostly shared, mostly single condition, correlated and uncorrelated errors, using $\hat{V}$ matrices of varying degrees of correlation.

3 Aim 2: Use the MVP Summary Statistic Data

While in simulated data, we can compare among methods in terms of accuracy and power, I can borrow from machine learning approaches to employ a 'testing-training' framework to find the model of best fit when varying the number and nature of components. That is, I will divide the approximately 2 million SNPS into a testing and training set, and build the covariance matrices as well as the learn the empirical bayes weights from the training data, and then compute likelihood and posterior on the test data. The model fit which maximizes the likelihood on the test data is the best fit.

I will then report posterior as above To specify the posterior distributions, recall the following standard result for Bayesian analysis of an R-dimensional MVN. If $\boldsymbol{b} \sim N_R(0, U)$, and $\hat{\boldsymbol{b}}|\boldsymbol{b} \sim N_R(\boldsymbol{b}, V)$ then

$$\boldsymbol{b}|\hat{\boldsymbol{b}} \sim N_R(\tilde{\boldsymbol{\mu}}, \tilde{U}), \tag{6}$$

where:

$$\tilde{U} = \tilde{U}(U, V) := (U^{-1} + V^{-1})^{-1}, \tag{7}$$

$$\tilde{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}}(U, V, \hat{\boldsymbol{b}}) := \tilde{U}(U, V)V^{-1}\hat{\boldsymbol{b}}. \tag{8}$$

This result is easily extended to the case where the prior on $\boldsymbol{b}$ is a mixture of MVNs (5). In this case the posterior distribution is simply a mixture of MVNs:

$$p(\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{V}_j, \hat{}) = \sum_p^P \tilde{\pi}_{jp} N_R(\boldsymbol{b}_j; \tilde{\boldsymbol{\mu}}_{jp}, \tilde{U}_{jp}) \tag{9}$$

where $\tilde{\boldsymbol{\mu}}_{jp} = \tilde{\boldsymbol{\mu}}(\Sigma_p, V_j, \hat{\boldsymbol{b}}_j)$ (equation (8)), $\tilde{U}_{jp} = \tilde{U}(\Sigma_p, V_j)$ (equation (7)), and

$$\tilde{\pi}_{jp} = \frac{\hat{\pi}_p N_R(\hat{\boldsymbol{b}}_j; \boldsymbol{0}, \Sigma_p + V_j)}{\sum_{p=1}^P \hat{\pi}_p N_R(\hat{\boldsymbol{b}}_j; \boldsymbol{0}, \Sigma_p + V_j)}. \tag{10}$$

From this is is straightforward to compute the posterior mean

$$E(\boldsymbol{b}_j|\hat{\boldsymbol{b}}_j, \hat{V}_j, \hat{}) = \sum_p^P \tilde{\pi}_{jp} \tilde{\boldsymbol{\mu}}_{jp} \tag{11}$$

and posterior variance

$$\mathrm{Var}(\boldsymbol{b}_{jr}|\hat{\boldsymbol{b}}_j, \hat{V}_j, \hat{}) = \sum_{p=1}^{\tilde{P}} (\tilde{U}_{jp,rr} + \tilde{\boldsymbol{\mu}}_{jp,r}^2) - [\sum_p^P \tilde{\pi}_{jp} \tilde{\boldsymbol{\mu}}_{jp,r}]^2 \tag{12}$$

as well as the local false sign rate.

- We can use the local false sign rate to assess significance in each subgroup, and compare to univariate approaches.
- We can categories SNPS as 'shared' if all the effects are of the same sign, and if the snps is considered significant (at an $lfsr \leq \alpha$) in all or most conditions.
- We can also find those SNPS with similar patterns of sign of effects across conditions.

4 Aim 2: Incorporate into PRS

Once we have posterior estimates for each SNPs effect size that has been 'jointly' shrunken thus exploiting any sharing of power and precision across classes, we can use these as the predictive covariates in PRS algorithm. Critically, such methods traditionally fit variants which satisfy a particular significance threshold in one disease, but using jointly shrunken estimates adds a layer of aggregate information. Down the line, we might wish to build a polygenic risk score for a multivariate phenotype as well, thus leveraging the additional heritability of multiple independent variables to predict multivariate traits.

5 Aim 3: Admixture Modeling to Re-characterize the Friedrickson classes

Existing terminology describe individuals as falling into one of 5 classes. However, at the individual characteristic (not summary statistic) level, individuals might have proportional membership in K classes, where each class is characterized by a pattern of effects across lipid classes. We can borrow from the admixture modeling literature ([6]) to use Grade of Membership models, where proportional ancestry is replaced by proportional membership in a phenotype pattern class.

References

1. Goode, E. L., Cherny, S. S., Christian, J. C., Jarvik, G. P. & de Andrade, M. Heritability of longitudinal measures of body mass index and lipid and lipoprotein levels in aging twins. *Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies* **10**, 703–711 (2007).

2. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics* **51**, 187–195 (2019). URL `http://www.nature.com/articles/s41588-018-0268-8`.

3. Stephens, M. False Discovery Rates: A New Deal. *bioRxiv* 038216 (2016). URL `http://biorxiv.org/content/early/2016/01/29/038216`.

4. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* btv546 (2015). URL `https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv546`.

5. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. *PLoS Genet* **9**, e1003486 (2013). URL `http://dx.doi.org/10.1371/journal.pgen.1003486`.

6. Dey, K. K., Hsiao, C. J. & Stephens, M. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLOS Genetics* **13**, e1006599 (2017). URL `https://dx.plos.org/10.1371/journal.pgen.1006599`.