Sherolock - Summary of Compound and Correlation Approaches

Jean Morrison

March 28, 2017

1 Goal

The goal of this project is to use summary statistics from GWAS to identify pairs of traits with association patterns that are consistent with a causal relationship. Our motivating problem is identifying genes for which expression levels are causally effect a phenotype of interest.

We consider there to be two elements to of a causal signature:

1. All eQTLs are GWAS SNPs
2. eQTL and GWAS effect sizes are correlated

Below I describe two approaches which I have called the "compound" approach and the "correlation" approach. In the "compound" approach we try to model both elements of the causal signature while in the "correlation" approach we focus on the second criterion. We construct the model in the "correlation" approach so that the first criterion is always present when the second criterion is true but not vice-versa. Both approaches have limitations discussed in Section 4. We will explore the performance of both approaches through simulations.

2 Correalation Approach

2.1 Version 1

Let $\hat{\beta}_{1,i}$ and $\hat{b}_{1,i}$, $i \in 1, \ldots, p$ be effect size and standard error estimates from an eQTL study. Let $\hat{\beta}_{2,i}$ and $\hat{b}_{2,i}$ be effect size and standard error estimates from GWAS. Let $\beta_{1,i}$ and $\beta_{2,i}$ be the true effects of SNP i on gene expression and phenotype respectively. We assume that the relationship of the estimated to true effect sizes is given by

$$\hat{\beta}_{s,i} \sim N(\beta_{s,i}, \hat{b}^2_{s,i}) \tag{1}$$

for $s = 1, 2$. That is, we assume that the SNPs are independent, the sample size of both studies is large enough for the normal approximation to hold and that $\hat{b}_{s,i}$ are very close to the true standard error of $\hat{\beta}_{s,i}$.

We assume that $\beta_{1,i}$ has distribution function $g_{1,i}$. We estimate $g_{1,i}$ as the posterior distribution of $\beta_{1,i}$ using the ASH model. That is

$$\beta_{1,i} \sim \pi^{(1)}_{0,i}\delta_0 + \sum_{k=1}^{K} \pi^{(1)}_{k,i} N(\mu_{k,i}, \sigma^2_k) \equiv g_{1,i}, \tag{2}$$

where σ_k for $k = 1, \ldots, K$ are a fixed grid. We model the distribution of $\beta_{2,i}$ conditional on $\beta_{1,i}$ as

$$\beta_{2,i} = \lambda\beta_{1,i} + u_i \tag{3}$$

where u_i is a random variable that represents the effect of SNP i in the phenotype through other mechanisms than the expression level of the gene. We allow the distribution of u_i to (potentially) depend on whether or not $\beta_{1,i}$ is an eQTL:

$$u_i|\beta_{1,i} = 0 \sim g_{2,0}$$
$$u_i|\beta_{1,i} \neq 0 \sim g_{2,1}$$

We assume that both $g_{2,0}$ and $g_{2,1}$ have the form of an ASH distribution so

$$g_{2,0} = \pi_0^{(2,0)}\delta_0 + \sum_{l=1}^{L} \pi_k^{(2,0)} N(0, \tau_l^2) \tag{4}$$

$$g_{2,1} = \pi_0^{(2,1)}\delta_0 + \sum_{l=1}^{L} \pi_k^{(2,1)} N(0, \tau_l^2) \tag{5}$$

where $\tau_1, \ldots, \tau_L$ are also a fixed grid. We can estimate $\pi_0^{(2,0)}, \ldots, pi_L^{(2,0)}$ and $\pi_0^{(2,1)}, \ldots, pi_L^{(2,1)}$ from representative sets of SNPs. One simple option is to set $g_{2,0} = g_{2,1}$ and estimate the parameters using ASH applied to all SNPs using the GWAS summary statistics. Another option is to estimate the parameters in $g_{2,0}$ using all SNPs that are not eQTLs for any gene and the parameters in $g_{2,1}$ using SNPs that are eQTLS for at least one gene.

Note that in this model, if λ is non-zero then all eQTLs also have non-zero effect in the GWAS. Therefore, rejecting $\lambda = 0$ gives a test for the presence of a causal signature.

Given the estimate of $g_{1,i}$, $g_{2,0}$ and $g_{2,1}$ we can write a likelihood for $\hat{\beta}_{2,i}$

$$\begin{aligned} P(\hat{\beta}_{2,i}|\hat{g}_{1,i}, \hat{g}_{2,0}, \hat{g}_{2,1}, \hat{b}_{2,i}) =& \pi_{0,i}^{(1)} P(\hat{\beta}_2|\beta_{1,i} = 0, \hat{g}_{2,0}, \hat{b}_{2,i}) + \left(1 - \pi_{0,i}^{(1)}\right) P(\hat{\beta}_2|\beta_{1,i} \neq 0, \hat{g}_{2,1}, \hat{b}_{2,i}) \\ & \pi_{0,i}^{(1)} \sum_{l=0}^{L} \pi_l^{(2,0)} N\left(\hat{\beta}_{2,i}; 0, \tau_l^2 + \hat{b}_{2,i}^2\right) + \\ & \left(1 - \pi_{0,i}^{(1)}\right) \sum_{k=1}^{K} \frac{\pi_{k,i}^{(1)}}{1 - \pi_{0,i}^{(1)}} \sum_{l=0}^{L} \pi_l^{(2,1)} N(\hat{\beta}_{2,i}; \lambda\mu_{k,i}^{(1)}, \lambda^2\sigma_k^2 + \hat{b}_{2,i}^2 + \tau_l^2) \end{aligned} \tag{6}$$

where $\sigma_0 = 0$. We can then write the likelihood as a function of λ as

$$l(\lambda) = \prod_{i=1}^{p} P(\hat{\beta}_{2,i}|\hat{g}_{1,i}, \hat{g}_{2,0}, \hat{g}_{2,1}, \hat{b}_{2,i}).$$

We can then maximize the likelihood to obtain an estimate of λ.

2.2 Version 0

Version 0 is my first attempt at this model and is like Version 1 except that $g_{2,1} = N(0, \omega^2)$. This means that every eQTL also effects the phenotype through some mechanism other than through gene expression levels so is less satisfying than the mixture. It does make for a simpler likelihood function:

$$\begin{aligned} P(\hat{\beta}_{2,i}|\hat{g}_{1,i}, \hat{g}_{2,0}, \hat{g}_{2,1}, \hat{b}_{2,i}) =& \pi_{0,i}^{(1)} P(\hat{\beta}_2|\beta_{1,i} = 0, \hat{g}_{2,0}, \hat{b}_{2,i}) + \left(1 - \pi_{0,i}^{(1)}\right) P(\hat{\beta}_2|\beta_{1,i} \neq 0, \hat{g}_{2,1}, \hat{b}_{2,i}) \\ & \pi_{0,i}^{(1)} \sum_{l=0}^{L} \pi_l^{(2,0)} N\left(\hat{\beta}_{2,i}; 0, \tau_l^2 + \hat{b}_{2,i}^2\right) + \\ & \left(1 - \pi_{0,i}^{(1)}\right) \sum_{k=1}^{K} \frac{\pi_{k,i}^{(1)}}{1 - \pi_{0,i}^{(1)}} N(\hat{\beta}_{2,i}; \lambda\mu_{k,i}^{(1)}, \lambda^2\sigma_k^2 + \hat{b}_{2,i}^2 + \omega^2) \end{aligned} \tag{7}$$

This is more efficient to evaluate. If $L = K$ then the version in (6) has about K^2 terms while the version in (7) has only about $2K$ terms.

3 Compound Approach

3.1 Model

The raw data for this approach is tha same as in Section 2 and we use the same model for the relationship of the estimated to true effect sizes,

$$\hat{\beta}_{s,i} \sim N(\beta_{s,i}, \hat{b}_{s,i}^2) \tag{8}$$

for $s = 1, 2$.

We model SNPs as belonging to one of four categories defined by whether or not $\beta_{1,i}$ and $\beta_{2,i}$ are equal to zero. We let α_1, α_2, α_3 and α_4 be the overall proportions of SNPs belonging to each category in the table:

	$\beta_{1,i} = 0$	$\beta_{1,i} \neq 0$
$\beta_{2,i} = 0$	α_1	α_2
$\beta_{2,i} \neq 0$	α_3	α_4

Let $\mathbf{Z}$ be a latent vector of states indicating which quadrant of the table each SNP falls in to so $Z_i = 1$ indicates that $\beta_{1,i} = 0$ and $\beta_{2,i=0}$, $Z_i = 2$ indicates that $\beta_{1,i} \neq 0$ and $\beta_{2,i=0}$, etc. If SNP i is in category 1, 2, or 3, then $\beta_{1,i}$ and $\beta_{2,i}$ are independent because one is fixed at 0. For SNPs in category 2, we model the eQTL effect as being drawn from a $N(0, \sigma_1^2)$ distribution. For SNPs in category 3, we model the GWAS effect as being drawn from a $N(0, \sigma_2^2)$ distribution. For SNPs in category 4, the effects are not necessarily independent and dependence of the effects can indicate a causal signature. For these SNPs we model

$$\begin{pmatrix}\beta_{1,i}\\ \beta_{2,i}\end{pmatrix} | Z_i = 4 \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\tilde{\sigma}_2\\ \rho\sigma_1\tilde{\sigma}_2 & \tilde{\sigma}_2^2\end{pmatrix}\right) \tag{9}$$

where $\tilde{\sigma}_2$ is distinct from σ_2. This model induces a four part mixture distribution for $\begin{pmatrix}\beta_{1,i}\\ \beta_{2,i}\end{pmatrix}$:

$$\begin{pmatrix}\beta_{1,i}\\ \beta_{2,i}\end{pmatrix} \sim \alpha_1 \begin{pmatrix}\delta_0\\ \delta_0\end{pmatrix} + \alpha_2 N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & 0\\ 0 & 0\end{pmatrix}\right) + \alpha_3 N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0 & 0\\ 0 & \sigma_2^2\end{pmatrix}\right) + \tag{10}$$

$$\alpha_4 N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\tilde{\sigma}_2\\ \rho\sigma_1\tilde{\sigma}_2 & \tilde{\sigma}_2^2\end{pmatrix}\right). \tag{11}$$

Note that this model is asymmetric in that the GWAS effect size of SNPs that are also eQTLs has a different distribution from the GWAS effect size of SNPs that are not eQTLS (i.e. $\tilde{\sigma}_2$ is distinct from σ_2) while eQTL effect sizes for SNPs that also effect the phenotype have the same distribution as eQTLs that that don't (i.e. σ_1 in the fourth part of the mixture is the same as σ_1 in the second part of the mixture).

Integrating out β_1 and β_2 gives the joint distribution of $\hat{\beta}_1$ and $\hat{\beta}_2$:

$$\begin{aligned}\begin{pmatrix}\hat{\beta}_{1,i}\\ \hat{\beta}_{2,i}\end{pmatrix} \sim & \alpha_1 N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}b_{1,i}^2 & 0\\ 0 & b_{2,i}^2\end{pmatrix}\right) + \\ & \alpha_2 N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 + b_{1,i}^2 & 0\\ 0 & b_{21}^2\end{pmatrix}\right) + \\ & \alpha_3 N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}b_{1,i}^2 & 0\\ 0 & \sigma_2^2 + b_{2,i}^2\end{pmatrix}\right) + \\ & \alpha_4 N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 + b_{1,i}^2 & \rho\sigma_1\tilde{\sigma}_2\\ \rho\sigma_1\tilde{\sigma}_2 & \tilde{\sigma}_2^2 + b_{2,i}^2\end{pmatrix}\right).\end{aligned} \tag{12}$$

This model as an equivalent conditional formulation. We can model

$$\beta_{1,i} \sim \pi_0^{(1)}\delta_0 + \left(1 - \pi_0^{(1)}\right) N(0, \sigma_1^2). \tag{13}$$

We then model the distribution of $\beta_{2,i}$ conditional on whether or not $\beta_{1,i}$ is equal to zero.

$$\begin{aligned}\beta_{2,i}|\beta_{1,i} = 0 &\sim \pi_0^{(2,0)}\delta_0 + \left(1 - \pi_0^{(2,0)}\right) N(0, \sigma_2^2)\\ \beta_{2,i}|\beta_{1,i} \neq 0 &\sim \pi_0^{(2,1)}\delta_0 + \left(1 - \pi_0^{(2,1)}\right) N(\lambda\beta_1, \tau^2)\end{aligned} \tag{14}$$

Parameter	Complete independence		No Correlation		Complete causality	
	Cond	Joint	Cond	Joint	Cond	Joint
Probabilities	$\pi_0^{(2,1)} = \pi_0^{(2,0)}$	$\alpha_1 = \pi_0^{(1)}\pi_0^{(2,0)}$ $\alpha_2 = (1-\pi_0^{(1)})\pi_0^{(2,0)}$ $\alpha_3 = \pi_0^{(1)}(1-\pi_0^{(2,0)})$ $\alpha_4 = (1-\pi_0^{(1)})(1-\pi_0^{(2,0)})$	$\pi_0^{(2,1)} > 0^{**}$	$\alpha_2 > 0^{**}$	$\pi_0^{(2,1)} = 0$	$\alpha_2 = 0$
Correlation	$\lambda = 0$	$\rho = 0$	$\lambda = 0$	$\rho = 0$	$\lambda \neq 0^{**}$	$\rho \neq 0^{**}$
Variance	$\tau^2 = \sigma_2^2$	$\tilde{\sigma}_2^2 = \sigma_2^2$	$\tau^2 = \sigma_2^2$	$\tilde{\sigma}_2^2 = \sigma_2^2$	**	**

Table 1: Parameter restrictions for three different models. The ** symbol indicates that a parameter is un-constrained or effectively unconstrained if it is excluded from taking on a single value.

We can translate between parameters in the joint and conditional formulation as follows:

$$\begin{aligned}
\pi_0^{(1)} &= \alpha_1 + \alpha_3 \\
\pi_0^{(2,0)} &= \frac{\alpha_1}{\alpha_1 + \alpha_3} \\
\pi_0^{(2,1)} &= \frac{\alpha_2}{\alpha_2 + \alpha_4} \\
\lambda &= \rho\frac{\tilde{\sigma}_2}{\sigma_1} \\
\tau^2 &= (1-\rho^2)\tilde{\sigma}_2^2.
\end{aligned} \tag{15}$$

This model contains 7 parameters but only two tell us about the presence of a causal signature, λ and $\pi_0^{(2,1)}$ (conditional formulation) or ρ and α_2 (joint formulation). The causal signature is indicated by $\lambda \neq 0$ and $\pi_0^{(2,1)} = 0$

Note that in this approach $\pi_0^{(2,1)}$ has a slightly different interpretation than in the correlation approach. In this approach, $\pi_0^{(2,1)}$ is the proportion of eQTLs that do not influence the phenotype at all. In the correlation model $\pi_0^{(2,1)}$ is the proportion of eQTLs that do not have pleiotropic effects (effects not mediated by the expression levels) on the phenotype.

3.2 Causal and Non-Causal models

To reduce the complexity of the problem, we estimate 4 parameters from the data and fix them. These are $\pi_0^{(1)}$ and σ_1 which we estimate by applying BVSR to the eQTL summary statistics only and $\pi_0^{(2,0)}$ and σ_2 which we estimate by applying BVSR to the GWAS summary statistics only. The idea behind this strategy is that there are not enough eQTLs to substantially influence the estimation of $\pi_0^{(2,0)}$ and σ_2, even in the presence of a causal relationship. This leaves us with one of the mixture proportions, ρ and $\tilde{\sigma}_2$ (or λ and τ in the conditional formulation).

We set the following priors on the three remaining parameters

$$z \equiv arctanh(\rho) \sim N(0, 0.25) \tag{16}$$

$$k \equiv \frac{\tau}{\sigma_2} \sim \chi^2(1) \tag{17}$$

$$\pi_0^{(2,1)} \sim Beta\left(C\pi_0^{(2,0)}, C\left(1-\pi_0^{(2,0)}\right)\right) \tag{18}$$

where C is a constant

We define three models to compare shown in Table 1.

3.3 Comparing Models

To compare the three models in Table 1, we have used two different strategies:

1. Compute the *expected log pointwise posterior density* under each model based on samples from the posterior distribution. To compare two models we test that the difference in elpd is different from zero.

2. Compute the Bayes factor comparing pairs of models. For this computation we numerically estimate the marginal likelihood of each model given the data using importance sampling with samples from the posterior providing the importance distribution.

4 Comparison of Compound and Correlation Approaches

The compound and correlation approaches differ in the following ways:

- In the correlation approach (version 1), it is impossible to have a model in which $\lambda = 0$ but all eQTLs are GWAS SNPs. In the compound approach, this is technically achievable by having both λ and $\pi_0^{(2,1)}$ equal 0. However, this conformation is very close to the boundaries of both the NC and CC models. In the NC model it is possible to choose $\pi_0^{(2,1)}$ very close to 0 and in the CC model it is possible to choose λ very close to 0. If the truth were that both $\pi_0^{(2,1)}$ and λ were equal to 0, we might find slightly more evidence for the CC model because the prior on $\pi_0^{(2,1)}$ places lots of weight close to 1. However, it is also likely that the two models would be hard to distinguish.

- In the compound approach, all SNPs falling into category 4 (both eQTL and GWAS) have a pleiotropic effect on the phenotype (i.e. an effect that is not mediated by the expression level. This is similar to Version 0 of the correlation approach. This may lead to a reduction in power if many SNPs in category 4 have little pleiotropic effect.

- The compound approach uses a two component distribution for the effect sizes in given in equations (13) and (14) while the correlation approach uses the more flexible unimodal distribution from ASH. In principle, we could use ASH type unimodal distributions for the distribution of $\beta_{1,i}$ and of $\beta_{2,i}|\beta_{1,i} = 0$ in the compound approach. However, this would make the gibbs sampler and the joint distribution in (12) more complicated so I have not pursued this.

- In the correlation approach, we can test if $\lambda = 0$ using maximum likelihood (no need to use a sampler) and there is only one free parameter. The likelihood is complex which makes numerical maximization somewhat time consuming but it is still more efficient than the compound approach.