1 Two Models Walk into a Bar

Suppose that some phenotype of interest, z, can be described in individual i by the following relation:

$$z_i = E(z_i) + \epsilon_z \tag{1}$$

$$E(z_i) = \sum_{j=1}^{n} \beta_j y_{j,i} + \sum_{k=1}^{m} \gamma_k x_{k,i} \tag{2}$$

$$\epsilon_z \sim Normal(0, \sigma_z^2) \tag{3}$$

Where $y_{j,i}$ corresponds to gene expression in gene $j \in \{1, 2, ..., n\}$ in individual i, β_j corresponds to the effect of that j^{th}'s gene's expression on the expected value of z. Similarly, $x_{k,i}$ corresponds to the genotype at locus $k \in \{1, 2, ..., m\}$ in individual i, γ_k corresponds to the effect of that genotype on the expected value of z.

Further suppose that the expression of gene j in individual i can be written as:

$$y_{j,i} = E(y_{j,i}) + \epsilon_{y_j} \tag{4}$$

$$E(y_{j,i}) = \sum_{k=1}^{m} \eta_{j,k} x_{k,i} \tag{5}$$

$$\epsilon_{y_j} \sim Normal(0, \sigma_{y_j}^2) \tag{6}$$

That is, the expected value of expression in gene j corresponds to the sum of effects $\eta_{j,k}$ multiplied by the genotypes of individual i across loci $1 : m$.

We wish to make inference of the effects β_j, which represent the influence of gene j's expression on phenotype z. However, both gene expression and phenotypic data are not available from the same individuals, and often not available at all, so we are unable to simply fit the multiple regression model in (1-3). Rather, we may fit (or receive estimates from; see Section 4) two separate models, the first implied by (4) and a second of the form:

$$E(z_i) = \sum_{k=1}^{m} \theta_k x_{k,i} \tag{7}$$

Obtaining estimates $\hat{\eta}_{j,k}$ and $\hat{\theta}_k$, the former of which are centered on $\eta_{j,k}$ and the latter of which are centered on $\sum_{j=1}^{n} \eta_{j,k}\beta_j + \gamma_k$. This can be seen by substituting equation (4) into equation (2):

$$E(z_i) = \sum_{j=1}^{n} \beta_j \sum_{k=1}^{m} \eta_{j,k} x_{j,i} + \sum_{k=1}^{m} \gamma_k x_{k,i} \tag{8}$$

$$\epsilon_z \sim Normal(0, \sigma_z^2 + \sum_{j=1}^{n} \beta_j^2 \sigma_{y_j}^2) \tag{9}$$

followed by minor rearrangement of (8):

$$E(z_i) = \sum_{k=1}^{m} \sum_{j=1}^{n} \beta_j \eta_{j,k} x_{j,i} + \sum_{k=1}^{m} \gamma_k x_{k,i} \tag{10}$$

$$E(z_i) = \sum_{k=1}^{m} (\sum_{j=1}^{n} \beta_j \eta_{j,k} x_{j,i} + \gamma_k x_{k,i}) \tag{11}$$

Furthermore, $\sum_{j=1}^{n} \beta_j^2 \sigma_{y_j}^2$ in (9) corresponds to the weighted sum of a set of normally distributed random variables and is thus itself normal, enabling us to express it as a single normally distributed random variable whose variance is the sum of the variances of its individual components. For ease and compactness of expression, we may rewrite (9) as

$$\epsilon_z \sim Normal(0, \sigma_{z_{total}}^2) \tag{12}$$

As this resembles the model actually fitted during eQTL and GWAS association mapping (omitting covariates corresponding to the confounding effects of shared ancestry or other latent factors, effectively having regressed their confounding effects out to estimate within-group associations), we obtain estimates $\hat{\theta_k} = \sum_{j=1}^{n} \eta_{j,k} \beta_j + \gamma_k + \epsilon_{\theta_k}$ and $\hat{\eta}_{j,k} = \eta_{j,k} + \epsilon_{\eta_{j,k}}$, where the '$\epsilon$'s represent normally distributed random variables.

Thus, to retrieve our focal parameters β_j, we may fit a model of the form:

$$E(\hat{\theta_k}) = \sum_{j=1}^{n} \hat{\eta}_{j,k} \beta_j + \gamma_k \tag{13}$$

$$\epsilon_{\hat{\theta_k}} \sim Normal(0, \sigma_{\hat{\theta_k}}^2) \tag{14}$$

Horizontal pleiotropic effects corresponding to γ_k are not identifiable in this context without imposing strong structural constraints, as there would need be a unique effect for every locus $1 : k$. If they may be said to have $\mathrm{E}(\gamma_k) = \alpha$, their variance about α may be moved into the error term, $\epsilon_{\hat{\theta_k}}$, and their average contribution, α, can be written as an intercept to yield:

$$E(\hat{\theta_k}) = \alpha + \sum_{j=1}^{n} \hat{\eta}_{j,k} \beta_j \tag{15}$$

This model can be fitted via least squares or else using some robust alternative. More flexibly, we may perform inference in the hierarchical Bayesian framework, described below, which would smoothly permit incorporation of prior information (to e.g. regularize estimates of β_j), as well as propagate uncertainty in $\hat{\eta_j}$ and $\hat{\theta_k}$ and accommodate non-identifiability in β_j for the purposes of posterior predictive simulation.

2 Multiple Regression Madness

Additional difficulties arise, however, when we lack access to coefficient estimates from the multiple regression models in (5) and (7), and instead possess only those from regressions over each of m loci unconditional on states at all other $(m-1)$ loci. But they may be straightfowardly surmounted. Specifically, limited recombination may induce linkage disequilibrium between spatially proximate loci, confounding unconditional estimation of locus-specific effects and inducing spurious associations between states at loci and outcomes of interest. We can retrieve estimates and their variance-covariance matrix from the corresponding multiple regression by exploiting a few basic properties of covariances, namely that:

$$Cov(A, b \cdot C) = b \cdot Cov(A, C) \tag{16}$$

$$Cov(A, B + C) = Cov(A, B) + Cov(A, C) \tag{17}$$

and that the least-squares estimate for a slope coefficient, $\hat{\beta}$, in the model $y = a + \beta \cdot x + \epsilon$ can be expressed as:

$$\hat{\beta} = \frac{Cov(x, y)}{Var(x)} \tag{18}$$

where $Cov(x, y)$ and $Var(x)$ represent the sample covariance and variance, respectively. Thus, the unconditional covariance can be written as

$$Cov(x, y) = \hat{\beta} \cdot Var(x) \tag{19}$$

At Hardy-Weinberg equilibrium, the genotype of a diploid organism at some locus is a $binomial(2, p_x)$ random variable, where p_x is the allele frequency of the alternate allele, and the variance of this genotype can as such be written as

$$Var(x) = 2 \cdot p_x \cdot (1 - p_x) \tag{20}$$

Multiplying our unconditional estimate of $\hat{\beta}$ by this variance therefore yields the overall covariance between some outcome variable y and some predictor variable x_i.

Suppose, then, we have a model of the form:

$$y = \alpha + \sum_{i=1}^{n} \beta_i x_i + \epsilon \tag{21}$$

The total covariance between y and a specific x_j can be written as:

$$Cov(x_j, y) = Cov(x_j, \alpha + \sum_{i=1}^{n} \beta_i x_i + \epsilon) \tag{22}$$

$$= Cov(x_j, \alpha) + \sum_{i=1}^{n} Cov(x_j, \beta_i x_i) + Cov(x_j, \epsilon) \tag{23}$$

$$= 0 + \sum_{i=1}^{n} \beta_i Cov(x_j, x_i) + 0 \tag{24}$$

For each x_j we can obtain the same such equation, giving us a system of n equations with n unknowns, which can be solved in the usual way:

$$\begin{bmatrix} Var(x_1) & Cov(x_1,x_2) & \dots & Cov(x_1,x_n) \\ Cov(x_2,x_1) & Var(x_2) & \dots & Cov(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(x_n,x_1) & Cov(x_n,x_2) & \dots & Var(x_n) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} = \begin{bmatrix} Cov(x_1,y) \\ Cov(x_2,y) \\ \vdots \\ Cov(x_n,y) \end{bmatrix} \tag{25}$$

$$\Sigma_x \beta = Cov(x,y) \tag{26}$$

$$\beta = \Sigma_x^{-1} Cov(x,y) \tag{27}$$

to retrieve our desired multiple regression coefficients, given knowledge of the covariances of x (obtainable from a sufficiently similar population) and unconditional covariances between each x_i and y, thereby adjusting for any spurious covariance in the latter due to non-independence between x_i and x_j. The presence of additional, unreported regressors, such as those corresponding to sex, batch, or ancestry, should not be too damaging in this regard, so long as they don't covary with x_i too strongly. This operation has historically been referred to as a 'sweep', able to retrieve arbitrary sets of conditional associations from their marginal associations.

We may also wish to retrieve $Cov(\hat{\beta})$, either to perform hypothesis-testing, or else to propagate inferential uncertainty to downstream procedures, such as that described in the first section. As before, the relevant information is contained in the precision matrix of X, represented above as Σ_x^{-1}, inversely weighted by the residual variance of y and one less than the sample size, n_{obs}. We can obtain that residual variance from the relations:

$$Var(\epsilon) = Var(y) \cdot (1 - r^2_{adj}) \tag{28}$$

where

$$r^2_{adj} = 1 - ((1 - r^2)(n_{obs} - 1)/(n_{obs} - n_p - 1)) \tag{29}$$

with n_p corresponding to the number of regressors in the model, and

$$r^2 = c^T R_x^{-1} c \tag{30}$$

where R is the correlation matrix of x and c is the vector of correlations $(r_{x_1,y}, r_{x_2,y}, ..., r_{x_{n_{obs}},y})$, obtainable from the covariances of x and y (19) and the variances of x (20) and y, the latter of which has usually been rescaled to unity in the original association mapping.

As before, effect sizes corresponding to unreported regressors stymie our ability to adequately retrieve $Cov(\hat{B})$, but their influence is minimal absent strong covariation between themselves and x.

3 Hierarchical Bayesian Regression

Having described our motivations, we now turn to the actual inference problem, fitting the model

$$\hat{\theta_k} = E(\hat{\theta_k}) + \epsilon_{\hat{\theta_k}} \tag{31}$$

where both components right-hand side are given in equations (13) and (14). Joint inference of all the effects of gene expression across all measured body tissues, aggregating signal across all loci, both *cis* and *trans* to the genes in question, would be preferable, though also quite intractable. In the interests of tractability, then, we may wish to restrict ourselves to only evaluating the effects of the same genes across different body tissues, as these should feature the strongest multicollinearity, as well as harnessing signal contained in only *cis*-SNPs (e.g. 1MB upstream and downstream from the gene under consideration), that our estimate of $Cov(x)$ be non-singular given finite sample sizes in reference panels.

A joint probability model requires that we specify priors over all parameters we'd care to make inference of, and several may suit our needs here. To enforce sparsity and ease interpretation of focal effects β_j, we can assign them a regularized horseshoe prior, where

$$\beta_j \sim Normal(0, \tau \cdot \widetilde{\lambda}_j) \tag{32}$$

$$\widetilde{\lambda}_j = \frac{c\lambda_m}{\sqrt{c^2 + \tau^2\lambda_j^2}} \tag{33}$$

$$\lambda_j \sim Half\text{-}Cauchy(0, 1) \tag{34}$$

$$c^2 \sim Inverse\text{-}Gamma(\frac{v}{2}, \frac{v}{2}s^2) \tag{35}$$

$$\tau \sim Half\text{-}Cauchy(0, 1) \tag{36}$$

$$v \sim Exponential(0.2) \tag{37}$$

$$s \sim Exponential(1) \tag{38}$$

$$\tag{39}$$

We may also wish that our model encode the possibility of both positively and negatively correlated effects across the same genes in different tissues, given that tissues often act in concert to manifest phenotypic variation. Thus, we may swap out (32) for something more of the flavor:

$$\vec{\beta} \sim Multivariate\text{-}Normal(\vec{0}, \Sigma_{\beta_j}) \tag{40}$$

$$\Sigma_\beta = \vec{S} R_\beta \vec{S} \tag{41}$$

$$\vec{S} = \tau(\widetilde{\lambda}_1, \widetilde{\lambda}_2, ..., \widetilde{\lambda}_n) \tag{42}$$

$$R_\beta \sim LKJ(1) \tag{43}$$

Additional distributions need also be specified in the probability model described by (31, 14, and 15). Here we might use something weakly regularizing, such as:

$$\alpha \sim Normal(0, 5) \tag{44}$$

$$ln(\sigma^2_{\hat{\theta}_k}) \sim Normal(0, 2) \tag{45}$$

$$ln(\sigma^2_{\hat{\theta}_{kj}}) \sim Normal(\mu, \phi^2) \tag{46}$$

$$(\mu, \phi^2) \sim Exponential(1) \tag{47}$$

potentially re-parameterizing components of the model to their non-centered form to improve sampling performance.

Furthermore, our estimates $\hat{\theta_k}$ and $\hat{\eta_j}$ are not known without error, with uncertainty about each estimate retrievable under the procedure described in Section 2 and equivalent to a quadratic approximation of the joint posterior of θ_k and $\eta_{j,k}$, updated from a flat prior. We can propagate this uncertainty into our proability model by modifying (15) that

$$E(\theta_k) = \alpha + \sum_{j=1}^{n} \eta_j \beta_j \tag{48}$$

$$\theta_k \sim Multivariate\text{-}Normal(\vec{\hat{\theta}}, Cov(\hat{\theta})) \tag{49}$$

$$\eta_j \sim Multivariate\text{-}Normal(\vec{\hat{\eta}}, Cov(\hat{\eta})) \tag{50}$$

augmenting over the unobserved states, at least if n and m are not too great as to render sampling especially cumbersome.

Alternatively, a procedure similar to that described in Section 2 may be used to retrieve the joint distribution of all β_j, indexed not only across tissues but across genes as well, which may covary in their expression due to shared regulatory pathways. However, sample sizes for human expression data (as in GTEx) are far smaller than the total number of genes of interest, and it is unclear how using a Moore–Penrose inverse or coercing a singular sample covariance matrix to a positive definite one using e.g. Higham's (2002) method would affect the retrieval of focal parameter values. Additionally, the influence of *cis*-SNPs is expected to greatly exceed that of *trans*-SNPs in eQTL mapping, so their contribution to spurious covariance in GWAS is expected to be relatively small, the signal shared between eQTL and GWAS dominated by the pathway flowing through the local gene.

4 Munging Sumstats

It is often the case that received model fits (e.g. eQTL / GWAS sumstats) are not in directly comparable formats on a common scale, as required by (13) or for the manipulations described in Section 2. For convenience, we may wish to first munge everything to estimates and standard errors in units of, say, standard deviations outcome per (the same) alternate allele (relative to the same reference allele), which would further simplify calculation of LD covariance matrices. If reference and alternate alleles are inverted, one of the corresponding coefficients needs only be multiplied by -1, and if neither they nor their inverse match, that locus would merit exclusion from further analysis.

Suppose that we receive per-SNP sumstats corresponding to a standardized regression coefficient, $\hat{\beta}_{Std.}$ i.e. where both predictor and outcome have been rescaled to unit variance. To return these to units per SNP, one needs only divide by $sd(x)$, obtainable by taking the square root of eq. (20). Multiplying by sd(y) puts us back on the natural scale of the outcome variable, but otherwise these may be more interpretable when passed through the procedure in Section 1 when left in their standardized units. Standard errors may be likewise rescaled in this manner.

Alternatively, one may be presented with only a z-score, perhaps also accompanied by a 2-tailed a p-value. In this case, munging is also straightforward – the p-value, p, can be converted to an unsigned doubly standardized coefficient in one of two ways: approximately, with Fisher's z-transformation, $\tanh(Q_{normal}(1-\frac{p}{2})/\sqrt{n-3})$, where n is the sample size and $Q_{normal}()$ represents the quantile function of the standard normal distribution, or exactly, by evaluating $Q_t(1-\frac{p}{2}, n-2)/\sqrt{Q_t(1-\frac{p}{2}, n-2)^2+1}$, where Q_t is the quantile function of the Student's t distribution. Having obtained these unsigned coefficients, they can be signed appropriately and unstandardized, as above. Standard errors, meanwhile, can be obtained by dividing the coefficient by the corresponding t-value, which can be obtained by evaluating $|Q_t(p/2, n-2))|$. Note, that these assume that only two parameters features in the marginal association model; for example, an intercept and a slope. This is unlikely to be the case, as often many other regressors, such as 'Ancestry PCs', 'batch effects', or 'hidden factors'. But in most association mapping studies, these are often dwarfed by the sample size, and so a difference of a few degrees of freedom is unlikely to change their corresponding Student's t probability much at all, or even differ meaningfully from a normal approximation.