Modeling Beyond the Configuration Model

Sarah Urbut

February 24, 2015

Contents

1 Beyond the Config Model

Recall that the Config Model made the following assumptions:

2 Understanding the model

Recall that for each tissue, we model the potential genetic association between a target SNP and the expression levels of a target gene by the simple linear regression model (**??**). In vector form, this model is represented by

$$\boldsymbol{y}_r = \mu_r \mathbf{1} + \beta_{\mathbf{r}} \boldsymbol{g}_{\mathbf{r}} + \boldsymbol{e}_{\mathbf{r}}, \ \boldsymbol{e}_{\mathbf{r}} \sim \mathcal{N}(\mathbf{0}, \sigma^{\mathbf{2}}_{\mathbf{r}} \mathbf{I}), \tag{1}$$

where r indexes one of the S tissue types examined and the vectors $\boldsymbol{y}_r, \boldsymbol{g}_r$ and $\boldsymbol{e}_r$ denote the expression levels, the genotypes of the samples and the residual errors respectively for the r^{th} tissue type. The intercept term, μ_r, and the residual error variance, σ_r^2 are allowed to vary with tissue type. The regression coefficient β_r denotes the effect of the eQTL in tissue r.

When the tissue samples are taken from the same individuals we allow that the observations on the same individual may be correlated with one another, but here I work with summary statistics and thus we assume that the residual errors among tissues.

Specifically, let $E := (\boldsymbol{e}_1 \cdots \boldsymbol{e}_s)$ denote the $N \times R$ matrix of residual errors, the we assume it to follow a matrix-variate normal (MN) distribution, i.e.,

$$E \sim \text{MN}(0, I, \Sigma). \tag{2}$$

That is, the vectors $(\epsilon_{1i}, \ldots, \epsilon_{Ri})$ are independent and identically distributed as $\mathcal{N}(0, \Sigma)$. The (unknown) $S \times S$ covariance matrix Σ quantifies the correlations in the error between the R tissues; it can vary from gene to gene and is estimated from the data (see below). In practice, we find that this matrix is close to diagonal, and again, when we work with summary statistics, we assume that the matrix of standard errors of $\hat{\beta}$, $\text{V}_{gp} as approximated by is diagonal$.

But critically, the value of our method comes in considering the tissues jointly, and thus makes use of the multivariate vector of expression levels across all tissues.

Now, the likelihood for gene g and SNP p is:

$$Y_g | X_p, B_{gp}, X_c, B_{gc}, \Sigma_{gp} \sim \mathcal{N}_{N \times R}(X_p B_{gp} + X_c B_{gc}, I_N, \Sigma_{gp}) \tag{3}$$

where:

- Y_g is the $N \times R$ matrix of expression levels;
- X_p is the $N \times 1$ matrix of genotypes (assuming the same individuals in all tissues);
- B_{gp} is the unknown $1 \times R$ matrix of genotype effect sizes;
- X_c is the $N \times (1 + Q)$ matrix of known covariates (including a column of 1's for the intercepts);
- B_{gc} is the unknown $(1 + Q) \times R$ matrix of covariate effect sizes (including the μ_s);
- $\mathcal{N}_{N \times R}$ is the matrix Normal distribution;
- Σ_{gp} is the unknown $R \times R$ covariance matrix of the errors.

For mathematical convenience (especially in the case of multiple SNPs), we vectorize the rows of B_{gp} into $\boldsymbol{\beta}_{gp}$. Here, as we focus on one SNP at a time, we directly have $\boldsymbol{\beta}_{gp} = B_{gp}^T$.

We use the notion of *configuration*, a latent indicator R-dimensional vector $\boldsymbol{\gamma}_{gp}$ such that $\gamma_{gpr} = 1$ means the eQTL is active in tissue r, i.e. $b_{gpr} \neq 0$, whereas $\gamma_{gpr} = 0$ means the eQTL is inactive in tissue r, i.e. $b_{gpr} = 0$. Moreover, we introduce an unknown mean $\bar{b}_{gp}$, to finally get the following "spike-and-slab" prior allowing to borrow information across tissues in which the eQTL is active:

$$b_{gpr} | \gamma_{gpr}, \bar{b}_{gp}, \phi \sim \gamma_{gpr} \mathcal{N}(\bar{b}_{gp}, \phi^2) + (1 - \gamma_{gpr}) \delta_0 \tag{4}$$

where δ_0 is a point mass at 0, and

$$\bar{b}_{gp} | \omega \sim \mathcal{N}(0, \omega^2) \tag{5}$$

Thus if $\gamma_{gpr} = 0$, then β_{gpr} is by definition 0. Otherwise, the tissue specific variance is ϕ.

Whereas the configuration handles qualitative heterogeneity (having an effect or not), the hyperparameters ϕ and ω handles quantitative heterogeneity (having possibly different, non-null effects). By integrating out $\bar{b}_{gp}$, we can see that $\phi^2 + \omega^2$ controls the average magnitude of the effect in any tissue and $\phi^2/(\omega^2 + \phi^2)$ controls the amount of heterogeneity.

Equivalently, we can write this prior as a multivariate Normal:

$$\boldsymbol{b}_{gp} | U_0 \sim \mathcal{N}_R(\mathbf{0}, U_0) \tag{6}$$

where, following Wen (2014), U_0 is parametrized as $(\Gamma_{gp}, \Delta_{gp})$:

$$p(U_0) = \mathsf{P}(\Gamma_{gp})p(\Delta_{gp}|\Gamma_{gp}) \tag{7}$$

so that Γ_{gp} is a binary matrix consisting of entry-wise non-zero indicators and is identical in size and layout to U_0, and Δ_{gp} is an indexed set of numerical values quantifying each non-zero entry in Γ_{gp}. The skeleton Γ_{gp} has $\boldsymbol{\gamma}_{gp}$ on the diagonal. Each off-diagonal entry $\Gamma_{gp,ij}$ is equal to 1 has long as diagonal elements $\Gamma_{gp,ii}$ and $\Gamma_{gp,jj}$ are both equal to 1. Thus, for a configuration 1-0-1 for example, all the entries containing tissue 2 (i.e., the prior on the effect size covariances with tissue 2 and tissue 2's effect size variance) will be 0. Otherwise,

In the current application, we choose:

$$U_0 \,|\, \boldsymbol{\gamma}_{gp} = \mathbf{1}, \phi, \omega \;\; = \begin{pmatrix} \phi^2 + \omega^2 & \cdots & \omega^2 \\ \vdots & \ddots & \vdots \\ \omega^2 & \cdots & \phi^2 + \omega^2 \end{pmatrix} \tag{8}$$

In terms of notation, the 0 in U_0 indicates the prior.

In practice, we use a known grid for prior variances, i.e. L pairs of values (ϕ_l, ω_l) leading to a mixture of multivariate Normals:

$$\boldsymbol{b}_{gp}|\boldsymbol{\lambda}, \boldsymbol{U}_0(\boldsymbol{\gamma}_{gp}) \sim \sum_{l=1}^{L} \lambda_l \, \mathcal{N}_R(\mathbf{0}, U_{0l}(\boldsymbol{\gamma}_{gp})) \tag{9}$$

where $U_{0l}(\boldsymbol{\gamma}_{gp})$ indicates that the U_{0l} matrix is a function of the configuration $\boldsymbol{\gamma}_{gp}$.

To simplify the notation, as there are $J = 2^R - 1$ active configurations, we now write U_{0jl} for the prior covariance matrix for configuration j and grid values l. When combined with the prior on configurations, this leads to:

$$\boldsymbol{b}_{gp}|\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{U}_0 \sim \sum_{j=1}^{J} \eta_j \sum_{l=1}^{L} \lambda_l \, \mathcal{N}_R(\mathbf{0}, U_{0jl}) \tag{10}$$

Which leads us to a corresponding multivariate mixture posterior on $\boldsymbol{b}_{gp}$.

$$\begin{aligned} p(\boldsymbol{b}_{gp}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\eta}}, v_{gp} = 1) = \sum_{l=1}^{L} \sum_{J=1}^{J} & p(\boldsymbol{b}_{gpjl}|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, c_{gpj} = 1, d_{gpl} = 1, v_{gp} = 1) \\ & \times \mathsf{P}(d_{gpl} = 1, c_{gpj} = 1|\hat{\boldsymbol{b}}_{gp}, \hat{V}_{gp}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\eta}}, v_{gp} = 1) \end{aligned} \tag{11}$$

3 Updates

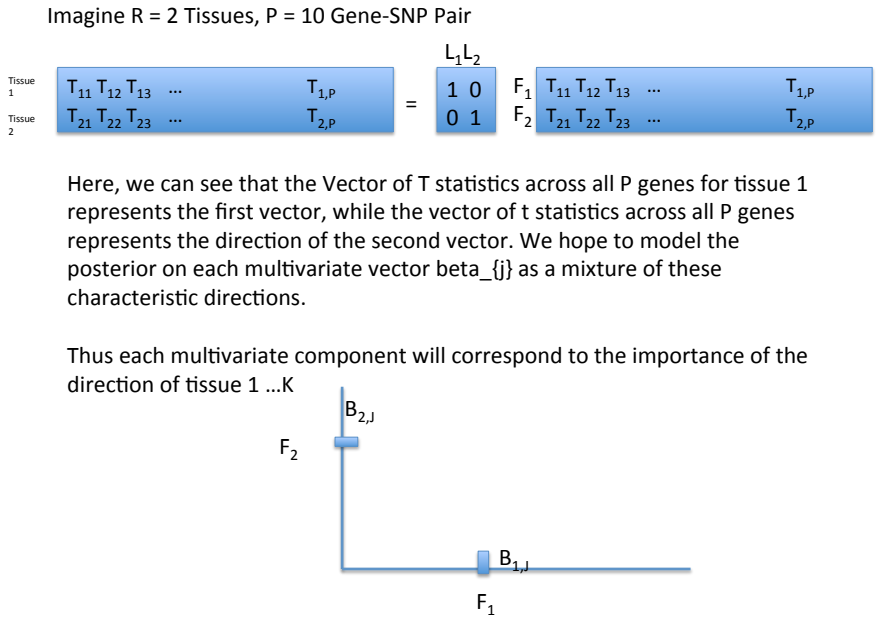
Now, suppose that instead of restricting ourselves to 'fixed' $U_0$of the form:

$$U_0 \,|\, \boldsymbol{\gamma}_{gp} = \mathbf{1}, \phi, \omega \;\; = \begin{pmatrix} \phi^2 + \omega^2 & \cdots & \omega^2 \\ \vdots & \ddots & \vdots \\ \omega^2 & \cdots & \phi^2 + \omega^2 \end{pmatrix} \tag{12}$$

We allow our estimate of U_0 to be informed by the data. Specifically, let us imagine modeling the prior on the covariance of effects as a linear combination of the effects in each tissue, such that we

can imagine a covariance matrix that reflects a linear combination of 'stretches' along axes defined by standardized effects in each tissue. In the univariate case, ash assumes that all β_j come from some shared mixture distribution where the proportion of each component is 'learned' from the data by maximizing the likelihood across all gene SNP pairs. Now, each $\boldsymbol{b}_j$ is actually a vector of effect sizes across tissues for a given gene-snp pair, and so we simply make the assumption that this shared distribution from which all $\boldsymbol{b}_j$ arise is a mixture of multivariate normals, with each of the K multivariate normals specified by the prior covariance matrix for $\boldsymbol{b}$, U_k.

In selecting these prior covariance matrices U_k, we can 'learn' these covariance matrices from the data, or specify a fixed set of covariance matrices from which we learn the relative weight, similar to the methods above. We will begin by modeling the prior on β_{gp} as descending from a mixture of multivariate normals, using a learned combination of weighted covariance matrix of the t statistics, and the first K factors of the Sparse Factor Representation. Importantly, learning across gene-snp pairs j also allows us to share information across tissues, if we learn the proportion descending from a component which puts heavy weight on a shared effect (as in the fixed effect case).



4 Understanding SFA

Both PCA and SFA attempt to approximate each tissues' $\beta_{i,..p}$ vector by a linear combination of $\beta_{i,..p}$ across tissues.

In SFA, at gene-snp pair j in tissue 1, tissue 1 will have loading on only the first factor, which represents the direction of effect in tissue 1, $\beta_{1,j}$. Similarly, we can see that F[1,1] results from L[,1] * G[,1] and so if L[,1] (the first loading, with coordinates for each tissue) is a vector of [1 0 0 .. R]

Then the corresponding Factor will receive only input from $\beta_{1,J}$ and thus can be thought of the direction corresponding to effect in tissue 1. The entire vector $F[1,]$ can then be thought of the pointing in the direction of 'true' effect over all gene SNP pairs for tissue 1, and F[2,] can be thought of the direction of effect across all genes in tissue 2, etc. But because tissue 1 β_j will have loading only on $F[1,]$, the predicted gene expression in tissue 1 will simply be the β_{j1} for tissue 1 at each of the J gene-snp pairs. By contrast, the first factor of PCA will be a weighted average of $\beta_{j1..R}$ across all tissues for a given gene-snp pair j, because each tissue β will receive equal loading since there is no requirement that the column sum of L be 1. Thus in PCA, each of the j elements of each factor $F[,k] = [1..j]$ represent a linear combination of the effect size of β_j across all tissues, while in sparse factor analysis they represent only the effect in β_k. Similarly, the loadings in PCA represent linear combinations of the j genes for each of the n individuals, while in factor analysis they can be more appropriately thought of as the relative importance of a particular factor (here, vector of effects in one tissue) on that sample. Since in practice we do not observe the true β_{jl} and are thus attempting to make an inference from the observed

4.1 Visualization

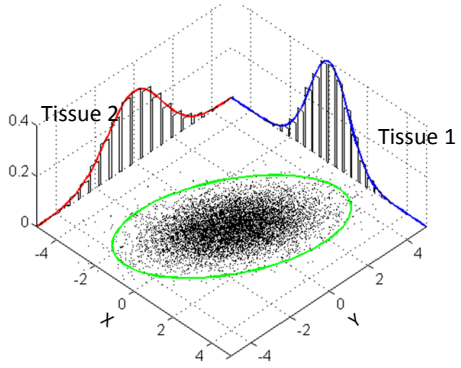
- Think of the 'Factors" as corresponding to the direction of effect for tissue 1...R across all p SNPs. Along the same lines of thinking as Ash, we can actually assume that there are 'shared' patterns of effect in tissue 1 across all j gene-snp pairs, such that this direction can probably be specified by a reduced set of gene-snp Pairs (in fact, the maximal number of linearly independent columns must be R, if R¡¡P).
- If we say that all vectors $_j$ 'share' information, we can estimate them from the same multivariate normal distribution. In fact, we use a mixture of multivariate normals, which is estimated hierarchically by finding the optimal combination of distributions to maximize the likelihood across all gene-snp pairs. Thus if some gene-snp pairs tend to exhibit stronger 'effects' in tissue 1, they will push the weight $\pi_{k,l}$ corresponding to a large prior variance for $\boldsymbol{b}_1$ up, while if most 'gene-snp' pairs show similar large effects in all tissues, then the $\pi_{k,l}$ might be very similar when ω_l is large, if we allow each of the $U_{k,l}$ matrices to represent the $L[,k]L[,k]^t$ matrix. That is, the matrix quantifying the loadings of 'eigentissue K' for each tissue, or equivalently, a weighted combination of all-gene SNP pairs' characteristic pattern K expression.
- We can thus form a mixture of multivariate normals, from which each unobserved $_j$ multivariate vectors is expected to lie. In the univariate case, the variance specifies the 'width' of this distribution along the 'tissue 1' axis, for example, where the tissue 1 axis can be thought of as a summary of the direction of behavior of all j end snp pairs in tissue 1. Now, we add an additional tissue 2 axis, which is similarly constructed as a summary of behavior of all j gene-snp pairs in tissue 2 (i.e., each of the j elements of this vector corresponds to the ideal activity in tissue k at gene snp pair j, thus the axis is itself j dimensional). The prior covariance term just specifies the width of the ellipse formed by the covariance between these two distribution for every mixture component - i.e., how well does activity in tissue 1 for β_{j1} correspond to activity in tissue 2 for β_{j2}.
- You can see that in SFA, the loadings result from a linear combination of the genes, and the factors result from a linear combination of the individuals. So for v[1,1] (or equivalently L[1,k] expresses how much of eigen tissue 'k' individual 1 contains across the 'optimal' combination of genes. It us the 'eigengen' because it represent a summary of 'k'ness across the genome.
- If we know that in SVD, the Us result from linear combination of the individuals while the Vs result from a linear Combination. In PCA, the first factor might represent the genomic profile of an 'average individual' at each gene, then in Factor Analysis it simply represents 'individual 1' type eigenarray (i.e., the expression of tissues most similar to group 1 in our example across the whole genome).

Thus the 'linear combination' would put no weight on the contribution of the genomic profile fro-
mother tissue types, and all weight on samples of tissue type 'k'. The Vs are linear combinations of
the columns of $U^t and similarly, each element of V[,k] represents a linear combination of the features of an individual.$
because each V^k is a genomic summary of every individual, while the true Covariance Matrix
$X^t X$ is compares individuals as well, but element by element. We can think of the loadings as how much eigentissue k an ind
e.g., how much that gene contributes to eigenpattern 1. Equivalently, the loadings in L[1] represent how much that individ

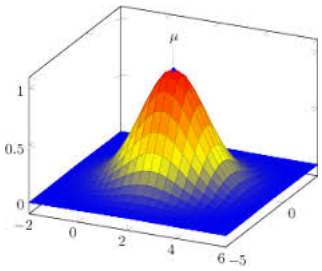
- The reason that we will use L[,k]*L[,k]' to approximate the covariance matrix of V is because

$$X_{rxp} = L_{rxk} F_{kxp} X_{rxp} X_{rxp}^t \approx \frac{1}{P} LFF'L' \approx L_k L_k'$$

In the two tissue case, all b_{j1} (i.e., the effect of each gene snp pairs in tissue 1) lie along the blue curve, while all b_{j2} lie along the red curve. The off-diagonal entries of the covariance represent the 'tightness of the ellipse', where a large correlation (or covariance) means ellipse is very narrow and expression in tissue 1 perfectly dictates expression in tissue 2.



A 'mixture' of multivariate normals means that we seek to find an optimal combination of these ellipse and marginal width plots to maximize the likelihood across all gene-snp pairs, assuming again that all b_j lie somewhere along the contour of the resulting function.



Shared mean effect

5 General Algorithm

For all j gene-snp pairs, beta $\boldsymbol{b}_j$ represent the unknown standardized effect of a snp 'p' on gene 'g'.

$$\boldsymbol{b}_j | \boldsymbol{\pi}, \boldsymbol{U}_0 \sim \sum_{k,l} \pi_{k,l} \, \mathcal{N}_R(\mathbf{0}, \omega_l^2 U_k) \tag{14}$$

As mentioned, we can 'learn' U_k from the data or specify each U_k and learn the relative proportion of each mixture component from the data via the EM algorithm.

In the special case where K = 1, as a first proposal for U_k when R is small, we let $\boldsymbol{t}_r$ represent the vector of t statistics across all j gene-snp pairs in tissue r. and μ_r represent the mean t statistic across all j gene-snp pairs in tissue r.

$$U_0 \,|\omega \;\; = \begin{pmatrix} \omega^2 \;(\boldsymbol{t}_r - \mu_r)(\boldsymbol{t}_r - \mu_r)^t & \cdots & \omega^2(\boldsymbol{t}_r - \mu_r)(\boldsymbol{t}_{r'} - \mu_{r'})^t \\ \vdots & \ddots & \vdots \\ \omega^2(\boldsymbol{t}_r - \mu_r)(\boldsymbol{t}_{r'} - \mu_{r'})^t & \cdots & \omega(\boldsymbol{t}_{r'} - \mu_{r'})(\boldsymbol{t}_{r'} - \mu_{r'})^t \end{pmatrix} \tag{15}$$

Here, we pre-specify $\omega_l^2 \in [0, 0.01, 0.02...]$ where each ω^2 can be though of as a 'stretch' along the vector corresponding to each factor. (But what is the connection between the factors and the prior covariance matrix? Or the covariance matrix of t statistics).

Similarly, can let $U_k = \boldsymbol{\gamma}_k \boldsymbol{\gamma}_{\boldsymbol{k}}^t$ where $\boldsymbol{\gamma_k}$ represents the R by 1 vector corresponding to the 'kth' loading from the Sparse Factor representations of the covariance matrix of t statistics, which summarize behavior of all t_j in tissue 1..k*respectively*.

6 EM Algorithm Outline

Given these U_k, we can now obtain MLEs for π_{kl} by a simple EM algorithm, in which we seek to choose the 'optimal' combination of covariance matrices $_k$ from which to model the posterior distribution of β_j.

We will use the following algorithm:

1) For each prior covariance matrix U_k and 'stretch' factor ω_l compute

$$\begin{aligned} L(\beta_j; k, l) = \Pr(\hat{\boldsymbol{b}_j}|k, l) \\ = \Pr(\hat{\boldsymbol{b}}_j; 0, \omega_l^2 U_k + \hat{V}) \end{aligned} \tag{16}$$

Then, we can set

$$w_{jkl} = \frac{\pi_{k,l} L(\beta_j; k, l)}{\sum_{k,l} \pi_{k,l} L(\beta_j; k, l)} \tag{17}$$

We can then update our estimate of $\pi_{k,l}$ as

$$\pi_{kl}^{i+1} = \frac{\sum_j w_{jkl}}{\sum_{j,k,l} w_{jkl}} \tag{18}$$

Which amounts to iteratively updating the weights such that they maximize the likelihood of the data across all gene snp pairs, assuming all vectors β_j arise from some shared mixture distribution and thus inform our estimates of each β_j. We can begin with $U_0 = \mathrm{I}$, $U_1 = \hat{\Sigma}$ where $\hat{\Sigma}$ represents the estimated RxR covariance matrix of T statistics, and U_2 is $U_{ki,j}$ as $F_i \dot{F}_j^t$, which, if we assume the F_k are broadly capturing direction of effect across genes in tissue 1..R and there are at most k=R factors, then each element i,j of U_k is the factor corresponding to the covariance between tissue i and tissue 2 in the case of the approximated covariance matrix.

In order to determine the hierarchical weight that separates the relative importance of one tissue expression pattern, we will separate each of the U$_k$*ratherthanusingthesumofthefirstK*(*asyousuggestedinthe*VV't*example*).*Int* (approximated covariance matrix) will then be derived from the loadings on tissue-specific K expression profile for each tissue. Each element of U_k will then represent the covariance of tissue r_i and r_j based on their loadings on this expression profile.