Joint Thoughts

Sarah Urbut

July 17, 2014

1 My questions

- Why is the single-tissue analysis sometimes more powerful?
- Why is the joint analysis more powerful even when only interested in the posterior probability of association in tissue s?
- In general, why does the joint analysis identify more eQTLs than pooling the single tissue analyses?
- Why is the method more powerful for eQTL that have modest) but consistent effects among tissues?

1.1 Why is the single-tissue analysis sometimes more powerful

In Tim's method, he notes that methods fall short in their ability to *jointly analyse data on all tissue to maximize power, while simulateously allowing for differences among eQTLs present in each tissue*. However, we jointly analyse data on all tissues *and all eQTLs*. We use all eQTLs to account for the uncertainty in (presumably shared) hyperparameters (such as the configuration weights η) but we also consider all tissues jointly to maximize the information we have towards begin an eQTL in even one tissue. Because we share information across genes to estimate the hyper parameters and some genes do not possess an eQTL active in more than one tissue, there will be some non-zero weight η on configurations that are not supported by the data (and consequently have a low BF_γ and thus the tissue by tissue analysis would perform best. i.e., the support in the data for each possible value of γ, relative to the null H_0, is quantified by the likelihood ratio, or *Bayes Factor* (BF)

$$\text{BF}_\gamma = \frac{\mathsf{P}(\text{data} \mid \text{true configuration is } \gamma)}{\mathsf{P}(\text{data} \mid H_0)}. \tag{1}$$

1.2 Why is the joint analysis more powerful even when only interested in the posterior probability of association in tissue s?

However, in all other cases, where a SNP is active in more than one tissue, the joint analysis is more powerful because the posterior probability of being an eQTL in even one tissue is:

$$\mathsf{P}(\text{eQTL in tissue } s \mid \text{data}, H_0 \text{ false}) = \sum_{\gamma:\gamma_s=1} \mathsf{P}(\text{true configuration is } \gamma \mid \text{data}, H_0 \text{ false}). \quad (2)$$

and thus if the evidence in favor a particular configuration, i.e., (1) is particularly strong because an eQTL is a strong eQTL in one tissue, the support for being an eQTL in the other tissue is strengthened by the BF for all configurations γ in which they occur together. Tim writes:

> For example, consider a SNP showing modest association with expression in Tissue 1. If this SNP also shows strong association in the other tissues, then it will be assigned a higher probability of being an active eQTL in Tissue 1 than it would if it showed no association in the other tissues.

This is because (2 will be inflated by the evidence in the data for a SNP being active in a particular configuration containing both tissues. This evidence will be greater than the evidence in the data for association between the SNP-gene pair in the tissue of interest.

1.3 In general, why does the joint analysis identify more eQTLs than pooling the single tissue analyses

Lastly, we identify more true eQTLs overall using the joint analysis when 2 or more tissues share an eQTL. When we pool single tissue analyses, not only are we considering more tests (and thus subject to greater Multiple Hypothesis Testing burden) we are effectively considering only one alternative model (i.e., $|\gamma| = 1$) for each tissue type while with BMA we are averaging over a range of alternative models. I do still have one question though: **when we consider the single tissue analyses, doesn't each single tissue analysis also encompass the universe of possibilities (i.e., the possibility active in A and B while not explicitly considering B?)**

The advantage of Bayesian Model Averaging is that basing inferences on one model alone is risky; presumably, ambiguity about model selection should dilute information about effect sizes and predictions, since part of the evidence is spent to specify the model (Leamer, 1978, page 91), Draper et al. (1987) and Hodges (1987).

The joint analysis is more powerful than ANOVA in these data, where simulations involve eQTLs that have similar effects among tissues. This is because ANOVA makes no presumptions about the shared effect, while our prior $p(\beta|\gamma, \theta)$ considers only a limited number of variance configurations and thus limits the heterogeneity in upweighting the probability of similar effects between tissues which share an eQTL.

> This is presumably because our simulations involved eQTLs that have similar effects in each tissue, and our prior distribution $p(\beta|\gamma, \theta)$ explicitly up-weights eQTLs with this feature.

1.4 Why is the method more powerful for eQTL that have modest) but consistent effects among tissues?

Tim writes that

> In many cases, the eQTLs detected by BF_{BMA} but not by the tissue-by-tissue analysis have modest effects that are consistent across tissues. Because their effects are modest in each tissue, they fail to reach the threshold for statistical significance in any single tissue, and so the tissue-by-tissue analysis misses them. But because their effects are consistent across tissues, the joint analysis is able to detect them.

It isn't as simple as the sum of the evidence in the data in support of being active in each tissue because (7) is a *weighted average* of the evidence in favor of any particular configuration, and thus presumably it takes 1/3 of the evidence in support of being an eQTL in any particular tissue. **or maybe not? Maybe it really is pooling all the evidence in favor of the SNP being an eQTL – for config [1 1 1], there is three times the data to support the QTL's activity. Also if it tends to occur altogether ever, both the likelihood and the prior will incorporate 1) the extra data and 2) the correlation between tissues 3) upweight the shared effect** Note that the computation of the (1) involves the **vector** of gene expression values across tissue types, and thus aggregates the expression data across subgroups in favor of a particular configuration. It isn't directly clear to me how the BF directly takes advantage of the correlation structure in the data - I would think that in measuring evidence for the configuration [1 1 1] for example, we actually lose some degrees of freedom in estimating each coefficient. *Doesn't this pose a variable selection problem?? Tabachnick and Fidell (2007) argue that there is little sense in using MANOVA on dependent variables that effectively measure the same concept.*

But Matthew's 2013 paper shows that when there exists correlation in the data, $\mathrm{BF}_{all} or BF_{BMA} is always more powerful tha$
$so somehow, the correlation structure in the data is incorporated to reinforce the evidence that a SNP is a ne QTL in one tissue. M$

$$\mathrm{BF}_{\mathrm{BMA}} = \frac{\mathsf{P}(\text{data} \mid H_0 \text{ false})}{\mathsf{P}(\text{data} \mid H_0 \text{ true})} = \sum_{\gamma \neq (0,\ldots,0)} \eta_\gamma \, \mathrm{BF}_\gamma. \tag{3}$$

Furthermore, the prior up weights shared effects - 'because their effects are consistent across tissues, the joint analysis is able to detect them'. Is this similar to having increased resolution to detect SNPs that are associated with correlated phenotypes, or is this a function of a prior explicitly up weighting the probability of shared effects among tissues? How does this correlation structure increase BF?

In comparing his analysis with the tissue specific analysis in figure 5, he notes that tissue specific analyses "fail to take account of incomplete power to detect eQTLs at any given threshold" but is this different that simply stating that the joint analysis sums the evidence for being a QTL over all tissues while single tissue analyses are really asking a different question? I think Ding (2010) is referring to the sample size of the study – "Unfortunately, this method will underestimate the overlap percentage whenever either of the two studies is underpowered (in that case, many true eQTLs might be detected in one study but missing from the list of eQTLs detected in the second study)" and not the fact that he fails to 'double' the evidence in favor of an eQTL. In figure 4 and 5, he shows that calling an eQTL in a 'tissue' specific analysis is like considering the [0 1 0] or [1 0 0] configuration, but this would imply that a univariate analysis only considers the case where the QTL is active in one tissue (and none of the others) while I feel that a univariate analysis, by the law of total probability, also considers the cases where it is active in other tissues though it may not be incorporated directly into the likelihood. *Isn't the univariate analysis actually:*

$$\mathsf{P}(\text{eQTL in tissue of interest}) = \sum_{i \ldots K} \mathsf{P}(\text{eQTL in tissue of interest and eQTL in tissue } i). \tag{4}$$

In other words, suppose we had the following situation for a gene SNP pair across tissues in 10 individuals (a through j):

$$\begin{pmatrix} a & 1 & 0 & 1 \\ b & 1 & 1 & 0 \\ c & 1 & 1 & 1 \\ d & 0 & 1 & 1 \\ e & 0 & 0 & 1 \\ f & 1 & 0 & 0 \\ g & 0 & 0 & 1 \\ h & 1 & 0 & 0 \\ i & 1 & 1 & 0 \\ j & 1 & 0 & 0 \end{pmatrix}$$

Does a univariate analysis only consider the gene expression of individuals f and j (because these are [1 0 0] configurations)? Doesn't it count the levels of gene expression for all individuals, even when there is expression in other tissues as well? I recognize that we will ignore the evidence that the other tissues might suggest about the probability of being an eQTL in one tissue, but the univariate analysis would be really limited if it ONLY considered the $|\gamma| = 1$ case. *I think this might be a BMA issue again, where the strength of the method lies in the hierarchical weight it puts on the multi-tissue configuration where as the univariate analysis implicitly treats every configuration equally.* Does the univariate analysis put all weight on [1 0 0]? Yes – it does, again strength of Bayesian Model Averaging.

1.5 In summary, we are really asking several different questions.

1. How does joint analysis help our posterior probability of activity in a particular tissue?
2. How does joint analysis help our overall posterior probability of being an eQTL at all?
3. How does joint analysis help our support for a particular configuration?

For the first question, I think the answer is that for a SNP that has 'modest' activity in one tissue, the Posterior Probability will consider the evidence in favor of its (potentially stronger) activity in additional tissues in the configuration which includes tissue s to increase the posterior probability of being the a QTL in the modest tissue s.

For the second question, even if the evidence is modest in each tissue, the $\mathrm{BF}_{\mathrm{BMA}}$ will be elevated (if the averaging doesn't dilute each tissues' effects) by the aggregate evidence in all tissues. There are two helpful features of our model: the hierarchical weighting will be greater for the shared configuration because this tends to occur more frequently among genes and thus matches the 'true alternative' (see Madigan and Raftery 1994). Secondly, we aggregate evidence across correlated tissues in favor of the SNP being active at all, strengthening our finding of support in the data for the activity of the SNP. Let's reconsider the computation of each BF_γ :

BF_γ in (1) using

$$\mathrm{BF}_\gamma = \sum_{j=1}^{M} w_j \mathrm{BF}_\gamma(\phi_j, \omega_j) \tag{5}$$

where M is the total number of grid points and $\mathrm{BF}_\gamma(\phi_j, \omega_j)$ is given by

$$\mathrm{BF}_\gamma(\phi, \omega) = \frac{p(Y|G, \phi, \omega, \gamma)}{p(Y|G, H_0)} = \frac{\int p(Y \mid G, \mu, b, \Sigma) p(\mu, \Sigma) p(b|\gamma, \phi, \omega)\, db\, d\mu\, d\Sigma}{\int p(Y \mid G, \mu, b = 0, \Sigma) p(\mu, \Sigma)\, d\mu\, d\Sigma} \tag{6}$$

Thus the value of the joint analysis for detecting the probability of being an eQTL at all occurs in two ways:

When averaging over configurations in which it is active in more than one tissue, we consider twice the evidence which could help to strengthen our case. Call this the 'sample size' effect. This is similar to the benefit of a meta-anlysis in which the data are pooled to draw more accurate conclusions about

Our choice of a prior specifically up weights those situations in which the effects are similar across tissues, thus giving our method an extra boost. This is what is implied by 'sharing information' across tissues and is similar to the fixed effect

For three, this refers to joint as in considering all genes simultaneously, and seems more straightforward to me: infer configuration weights that maximize overall patterns observed in the data.

1.6 Summary

I've been thinking a lot about the power of joint analysis:

As to the multiphen paper I asked about, I think the situation is different because even if the SNP is only associated with one trait, two traits may be correlated and thus a joint analysis might be helpful (I.e., Matthews 2013 plos one paper) because it effectively doubles the evidence in favor if association in even one tissue. Multiple phenotype analysis then exploits this dependency. In our single tissue case, expression among tissues was not correlated so estimating hyper parameters that put non zero weight on the multi tissue configuration reduces the sensitivity of our model by supporting a hypothesis inconsistent with the evidence.

2 Literature Review: Evidence for Joint Support

So Why Does Joint Analysis work? It doesn't seem intuitively apparent (to me at least) that a joint analysis immediately increases your power. Why does considering other information (if that information is somewhat correlated with your question of interest) always increase your power?

Skol (2006) compares the benefits of a 2-stage design with that of replication. He compares a replication analysis, in which markers that exceed a Significance threshold in phase 1 are considered in phase 2 and a new statistic that ignores phase 1 results is constructed, with that of a joint analysis, in which t statistics from both phases are combined into a joint statistic. The joint method is always more powerful - i.e., a greater number of true associations are detected. He writes:

> This makes sense, as joint analysis makes full use of stage 1 data, including the strength of evidence for the observed stage 1 association, whereas replication-based analysis uses only the information that the stage 1 association exceeds the threshold for follow-up but otherwise ignores the strength of the stage 1 evidence.

I see that it's important to consider all information, but doesn't this also dilute the presumably stronger effect observed in phase 2? He writes that "unless the risk allele has a much larger effect in the stage 2 samples, joint analysis will remain more powerful than replication-based analysis." Presumably the general principle here is that considering all the information strengthens support in the data for a particular hypothesis.

Wen and Stephens (2014) show that the Bayes Factor assessing the evidence for a particular configuration can be written as the product of the evidence in each subgroup and the consistence of effects among subgroups. Thus in particular, if all subgroups show effects in the same direction, then the first term may be large and 'boost' the evidence of association. So: If there are modest (but shared) effects,

Plos (2013), Matthew writes

> The posterior probability that any *particular* coordinate of Y is associated with g will always be less that the overall posterior probability that at least one coordinate of Y is associated.

This is analogous to summing over all of the configurations where it is active in at least one tissue. *But it's not this simple is it? Don't we have to take some weighted average? Couldn't it actually weaken the evidence in the strongest tissue?*

Consider again:

$$\mathrm{BF}_{\mathrm{BMA}} = \frac{\mathsf{P}(\text{data} \mid H_0 \text{ false})}{\mathsf{P}(\text{data} \mid H_0 \text{ true})} = \sum_{\gamma \neq (0,\ldots,0)} \eta_\gamma \, \mathrm{BF}_\gamma. \tag{7}$$

We indeed add all of the evidence of a variety of univariate associations, but we dilute them accordingly. So if we observe the same modest effect in each tissue, will just take 1/S times that ... right? Indeed, in Matthew's paper, equation 11 shows the standard univariate analysis - which assigns 0 weight to being active in more than one tissue. But I thought that a univariate analysis considers all the setting where a SNP is active in a particular tissue, regardless of its activity in other tissues as well.

The univariate analyses still consider all the data correct? It's just they put punitive (or 0) priors on some of the tissue-specific effects? Yes.

Bayesian Model Averaging

So maybe this is actually a question of specifying the correct model, in which matching the correct model yields the highest evidence in favor of the alternative. But **bma** considers the weighted average of evidence against the global null, i.e., that a SNP is not associated with expression in any tissue, but averaging over all configurations. I think this requires that the null hypothesis in the denominator of the BF be the same in all cases, but it seems that the 'separate' analyses would imply a univariate null - i.e., that the SNP is inactive in a *tissue* - rather than the multivariate null. Maybe not. Consider:

$$\mathrm{BF}_{\mathrm{BMA}} = \frac{\mathsf{P}(D|M_1)}{\mathsf{P}(D|M_0)} = \sum_{\gamma \neq (0,\ldots,0)} \frac{\mathsf{P}(\gamma_1|M_1)\mathsf{P}(D|\gamma_1, M_1)\, d\gamma_1}{\mathsf{P}(\gamma_0|M_0)\mathsf{P}(D|\gamma_0, M_0)}. \tag{8}$$

Which implies that each configuration is a rejection of the global null, i.e., that the SNP is no effect in any tissue. Doesn't the result of Matthew's statement (that the strength of evidence against the multivariate null is always greater than the strength of evidence against a particular univariate null hypothesis.) Huberty and Morris (1989) found that it was almost never beneficial to do a multivariate analysis followed by univariate confirmations because they *failed to uncover the results uncovered in the first place by the multivariate analyses* - each one implicitly puts prior weight of 0 on alternative configurations.

Huberty writes that multivariate tests are appropriate when determining outcome variable subsets that account for group separation, determining the relative contribution to group separation of the outcome variables in the final subject, and identifying underlying factor associated with the obtained multivariate results. He describes considering some linear composite of the outcome variable (linear Discriminant Function – i.e., LDF) Correlations between each outcome variable can then be revealed. But the way that we choose the composites is to maximize the effects! I.e., the optimization of the compositions is based on a criterion that is internal to the outcome variables, namely, the maximization of variance accounted for in the variable set. Think of this as reversing the regression I guess - finding the set of tissues that best reveal the variation in genotype.

An objective of multivariate analysis is to *"increase the sensitivity of the analysis through the exploitation of the intercorrelation among the response variables so that indications that may not be noticed in separate univariate analyses stand out more clearly in the multivariate analyses"*(1984, Ketternring). So do we exploit this by considering correlation in effect size or of correlation in error? I think it's through correlation in effect size, because we impose shared effect size prior. Huberty and Morris argues that it is illogical to confirm using univariate F values to identify some constructs, which inherently depend on intercorreation of constituent variables - indeed such univariate tests may fail to reveal the very associations identified by the multivariate tests in the first place. If we think about the choice of as those that maximize the linear combination of dependent variables that reveal variation in g (because Matthew's note 2 shows that the regression of g Y is equivalent to Y g), then this is easy to see in the likelihood case. We can see that the method used for MANOVA takes into account the correlations between the dependent variables as well as the differences between their means.

1. The eigenvalues show the ratio of SSB/SSW for a univariate analysis in which the discriminant function (composite variable) corresponding to the eigenvalue is the dependent variable, and the
2. The eigenvectors show the coefficients which can be used to the create the discriminant function.
3. Eigen analysis is such that each discriminant function has the highest-possible SSB/SSW, given that each successive discriminant function is uncorrelated with the previous one(s).
4. There are as many eigenvalues (and eigenvectors), and therefore discriminant functions, as there are dependent variables, or the number of groups minus one, whichever is the smaller. (If the independent variable is a numeric variable, there is only one discriminant function.)

The value of our method (and that enumerated in Stephen 2013) is the applicability of Bayesian Model Averaging (see Hoteling and Kass and Raftery, 1995). There is more evidence against the global null, and yet interpretability is an issue, and so by enumerating the possibilities under the null and taking a 'learned' approach to approximating the true alternative (by applying data-sensitive weights to each alternative) we attempt to maximize evidence against global null – great! My question lies:

Are we taking full advantage of the fully-joint configuration (i.e., 1 1 1 or 1 1 0 1 1) because by imposing a diagonal prior on σ and integrating out σ, where do we maximize the correlations among tissues that reveal differences in the data? I think we could do beyond the sample size advantage (see figure 5 in Tim's paper) and beginenumerate

How does a Bayesian multivariate analysis take advantage of these intercorrelation if its not through the estimate of linear combination of dependent variable?? See Huberty and Kettenring

Since from our conversations, it seems that the errors E can assume to be diagonal and if we are grabbing $\hat{\beta}$ from a variety of studies, than how do we take advantage of the correlations in the Joint Configuration? I realize that we don't integrate out the hyper parameters (the priors on the heterogeneity) and so this accounts for heterogeneity, but in general – how do we account for the dependency in the BF as opposed to Hotleellins Lambda?

I recognize the likelihood will increase with increased sample size of additional tissues (i.e., 1 01 with the third tissue being easier to apprehend increases the likelihood and thus the posterior corresponding to the first tissue – but is it just a sample size???

The assumption in Stephens (2013) about

```
lm(g~y)
```

being the same as

```
manova(lm(y~g))
```

only holds if the columns of Y are IID (which might be true - reference: personal conversation, generative model. Indeed, in Tim's simulations comparing with ANOVA,

```
m1 <- lm(y ~ xs)
m2 <- lm(y ~ xs * xg)
pval <- anova(m1, m2)[[6]][2]
```

he doesn't use the LDFs from MANOVA.

Multivariate more powerful then univariate assumption in Matthew's paper (see figure 5[2,2]: Does this hold because (incorrectly) counting $\mathbf{Y_U}$ as $\mathbf{Y_d}$ (see figure 5) is better than counting it as Y_I.

Suggestions: Could we performs some kind of eigendecomposition to search for the best linear disriminant function that 'clusters tissues' based on patterns of common association - e.g., kidney, liver and adipose reveal a particular pattern of separation for SNP A while brain, blood and heart reveal a different pattern? We could use sparse factor analysis, and wouldn't require the strong assumption of the tissues being IID... we could still use the type model.

3 Gaining a better intuition on Bayes Factors

We can understand Bayes factors as a ratio of marginal likelihood obtained by considering the conditional density of Y evaluated at a variety of values of β, each weighted accordingly by their prior distribution. I think it is easiest to first consider the normal case, in which

$$K = \frac{\mathsf{P}(D|M_1)}{\mathsf{P}(D|M_2)} = \frac{\int \mathsf{P}(\theta_1|M_1)\mathsf{P}(D|\theta_1, M_1)\, d\theta_1}{\int \mathsf{P}(\theta_2|M_2)\mathsf{P}(D|\theta_2, M_2)\, d\theta_2}$$

$$\int \mathsf{P}(\theta_1|M_1)\mathsf{P}(D|\theta_1, M_1)\, d\theta_1 = \int \mathsf{P}(\theta_1|D)\mathsf{P}(D|M_1)\, d\theta_1$$

$$= \mathbf{C}\int \mathsf{P}(\theta_1|D) d\theta_1$$

$$= \frac{\mathbf{C}}{\mathbf{K}}$$

Where $\mathbf{C}$ represents some constant equivalent to the marginal density of D. In the normal case with known variance, the marginal density of D is just $\mathbf{N}$ with mean equivalent to the prior mean, μ_0 and variance equal to the sum of the prior variance $\mathbf{W_0}$ and the data variance $\mathbf{V}$. Furthermore, $\mathbf{K}$ is some constant of proportionality used to normalize the posterior density of μ given θ. Thus we can view the Bayes factor as a ratio of the marginal density of data $\mathbf{D}$ to the constant of proportionality for the posterior distribution.

Questions: itemize

Using the LaPlace approximation, puts most of its mass around the MLE of the data. I see that under the alternative case. But then what happens in the null case (see Kass and Raftery 1995)

How do the Bayes factors relate to each of the inferences on β? The betas are still ID in the random effects case, correct?

How do the Bayes factors relate to the T staitscs - in the derivation in Wakefield, 2009,

How does this translate to fixed vs random effects? In the random effects case, it is more products in the [1 1 1] model, correct? See Wen (2014)

Can we think of both Bayes factors showing approximately the density of the data at its MLE, multiplied by the prior probability of the MLE specific to each model? This would make sense, and then in the RE case, each study or tissue would have its own MLE.