File S2

Variance of Lasso Estimate of Bin Effect

The GlmNet/R program (Friedman et al. 2010) does not provide the standard error for the Lasso estimate $\hat{\gamma}_k$ (Tibshirani 1996). Here we present an approximate formula to calculate the standard error. Let $\text{var}(\hat{\gamma}_k)$ be the variance of the estimated bin effect. The standard error simply takes the square root of this variance. The Lasso estimate of a bin effect is a kind of shrinkage estimate. Therefore, $\hat{\gamma}_k$ and $\text{var}(\hat{\gamma}_k)$ are considered as the "posterior mean" and "posterior variance" of γ_k, respectively. We may find the "prior variance" retrospectively using $\hat{\gamma}_k$ and the data. If γ_k were estimated from a single bin model, the variance of that estimate would be

$$\text{var}(\tilde{\gamma}_k) = (Z_k^T Z_k)^{-1}\hat{\sigma}^2 \tag{11}$$

where

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta} - Z\hat{\gamma})^T(y - X\hat{\beta} - Z\hat{\gamma}) \tag{12}$$

is the estimated residual error variance. Let σ_k^2 be the "prior variance" of γ_k. Combining the prior variance σ_k^2 and the variance from the data $\text{var}(\tilde{\gamma}_k)$, we can generate the posterior variance expressed by

$$\text{var}(\hat{\gamma}_k) = \left(\frac{1}{\sigma_k^2} + \frac{1}{\text{var}(\tilde{\gamma}_k)}\right)^{-1} = \frac{\sigma_k^2\hat{\sigma}^2}{\hat{\sigma}^2 + \sigma_k^2 Z_k^T Z_k} \tag{13}$$

When the prior variance is estimated from the data, it can be interpreted as the expectation of γ_k^2,

$$\sigma_k^2 = E(\gamma_k^2) = \hat{\gamma}_k^2 + \text{var}(\hat{\gamma}_k) \tag{14}$$

Therefore,

$$\sigma_k^2 = \hat{\gamma}_k^2 + \frac{\sigma_k^2\hat{\sigma}^2}{\hat{\sigma}^2 + \sigma_k^2 Z_k^T Z_k} \tag{15}$$

The above equation has only one unknown quantity, σ_k^2, which has two solutions with the positive one being

$$\hat{\sigma}_k^2 = \frac{\hat{\gamma}_k^2 Z_k^T Z_k + \sqrt{(\hat{\gamma}_k^2 Z_k^T Z_k)^2 + 4\hat{\sigma}^2\hat{\gamma}_k^2 Z_k^T Z_k}}{2Z_k^T Z_k} \tag{16}$$

Therefore, the posterior variance $\text{var}(\hat{\gamma}_k)$ is obtained by substituting σ_k^2 appearing in equation (13) by the solution given in equation (16),

$$\text{var}(\hat{\gamma}_k) = \left(\frac{1}{\hat{\sigma}_k^2} + \frac{1}{\text{var}(\tilde{\gamma}_k)}\right)^{-1} = \frac{\hat{\sigma}_k^2\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{\sigma}_k^2 Z_k^T Z_k} \tag{17}$$

When $m \gg n$ and n is very large, the Lasso estimate of σ^2 is often near zero (perfect fit), leading to $\sigma_k^2 \approx \hat{\gamma}_k^2$. The corresponding Wald test under $\sigma_k^2 \approx \hat{\gamma}_k^2$ is then

$$W_k = \hat{\gamma}_k^2 \frac{\hat{\sigma}^2 + \hat{\sigma}_k^2 Z_k^T Z_k}{\hat{\sigma}_k^2 \hat{\sigma}^2} \approx \hat{\gamma}_k^2 \frac{\hat{\sigma}^2 + \hat{\gamma}_k^2 Z_k^T Z_k}{\hat{\gamma}_k^2 \hat{\sigma}^2} \approx \frac{\hat{\gamma}_k^2}{\hat{\sigma}^2} Z_k^T Z_k + 1 \qquad (18)$$

which was given by Hu et al. (2012) without providing a proof.

The posterior variance $\text{var}(\hat{\gamma}_k)$ provides a convenient way to draw the Wald test statistic. The retrospectively inferred "prior variance" $\hat{\sigma}_k^2$ also gives us a chance to evaluate the "degree of confidence" for each bin effect and "the effective number of tests", which were proposed by MacKay (1992) and Tipping (2001). The degree of confidence (also called degree of freedom) for bin *k* is defined by

$$d_k = 1 - \frac{\text{var}(\hat{\gamma}_k)}{\hat{\sigma}_k^2} = \frac{\hat{\sigma}_k^2 - \text{var}(\hat{\gamma}_k)}{\hat{\sigma}_k^2} = \frac{\hat{\sigma}_k^2 Z_k^T Z_k}{\hat{\sigma}^2 + \hat{\sigma}_k^2 Z_k^T Z_k} \qquad (19)$$

Note that $0 \le d_k \le 1$, with 1 indicating perfect confidence and 0 indicating no confidence. In fact, d_k is the reduction of the posterior variance relative to the prior variance. In shrinkage analysis, not every single test is counted as a test. If a bin effect has a confidence $d_k = 0.5$, this test only counts as a "half test". For the entire genome, we have m bins and thus *m* Wald tests, but the "effective number of tests" is only

$$m_e = \sum_{k=1}^{m} d_k = \sum_{k=1}^{m} \frac{\hat{\sigma}_k^2 Z_k^T Z_k}{\hat{\sigma}^2 + \hat{\sigma}_k^2 Z_k^T Z_k} \qquad (20)$$

Recall that when $m \square\ n$ and *n* is very large, the Lasso estimate of σ^2 is often near zero and $\sigma_k^2 \approx \hat{\gamma}_k^2$. However, most bin effects will be estimated at exactly zero. In this case, the effective number of tests simply equals the number of non-zero effects. The effective number of tests implies a different method to correct multiple tests, if it is necessary to consider such a correction. Assume that the nominal *p*-value criterion is 0.05. The genome-wide *p*-value criterion after Bonferroni correction for multiple tests should be $0.05/m_e$, rather than $0.05/m$. The conventional Bonferroni correction using the number of bins to adjust for the *p*-value is too conservative (over-correction) and the effective number of tests will correct the over-correction. The effective number of test share a similar nature as the QTL intensity in the Bayesian QTL mapping (Sillanpää and Arjas 1998; Sillanpää and Arjas 1999).

Table S3 lists the effective numbers of tests for all the eight traits under the natural bin analysis. Taking the yield (YD) trait for example, the effective number of tests is 26.56, which leads to a critical genome-wide *p*-value of $0.05/26.56 = 0.00188$ after the Bonferroni correction for multiple tests. Whether a bin is associated or not with YD, we should use 0.00188 as the criterion, rather than the 0.05 criterion. Alternatively, the Chi-square one criterion should be $\chi_1^{-2}(1-0.00188) = \chi_1^{-2}(0.99812) = 9.63$, which is converted into to a LOD score criterion of $9.63/4.61 = 2.09$. In other words, a bin is declared as significant if its Wald test statistic is larger than 9.63 or its LOD score is greater than 2.09. When this criterion (effective number of tests) is used, no bins are associated with YD. If the original Bonferroni correction were conducted, the *p*-value criterion would be $0.05/1619 = 3.0883 \times 10^{-5}$, which is even more stringent. The LOD score tests and the thresholds using this effective number of tests criterion are illustrated

in Supplemental Figure S1 for traits YD, TP, GN, KGW and Figure S2 for traits GL, GW, HD and OsC1. Among the eight traits, the first six traits have relatively large effective numbers of tests. The last two traits have small effective numbers of tests, especially, the single-gene-controlled trait OsC1 has $m_e = 1.2$, which is virtually a single gene test. This is a known single-gene-controlled trait and, theoretically, the effective number of tests should be exactly one. The known gene is located in bin 868 and this bin completely co-segregates with the color trait, whose LOD score and confidence are $LOD_{868} = 46737$ and $d_{868} = 0.999995$, respectively. However, out of 210 lines, 209 lines co-segregate with the bin 867 (the one next to the bin containing the color gene). This bin (bin 867) has a LOD score 0.0609, a *p*-value 0.5964 and a degree of confidence 0.212056. As a result, the effective number of tests for OsC1 is $0.999995 + 0.212056 = 1.2121$. Interestingly, a single mismatch causes a LOD score drop from 46737 (perfect match) to 0.6 (one mismatch). This kind of result would not be possible if one bin were fit to the model at a time.

References

Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1): 1-22.

Hu Z, Wang Z, Xu S. 2012. An infinitesimal model for quantitative trait genomic value prediction. *PLoS One* **7**: e41336.

MacKay DJC. 1992. Bayesian interpolation. *Neural Computation* **4**: 415-447.

Sillanpää MJ, Arjas E. 1998. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**(3): 1373-1388.

Sillanpää MJ, Arjas E. 1999. Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* **151**: 1605-1619.

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* **58**: 267-288.

Tipping ME. 2001. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* **1**: 211-244.