DOI: 10.1534/genetics.107.077487

Note

Derivation of the Shrinkage Estimates of Quantitative Trait Locus Effects

Shizhong Xu[1]

Department of Botany and Plant Sciences, University of California, Riverside, California 92521

Manuscript received June 12, 2007
Accepted for publication August 3, 2007

ABSTRACT

The shrinkage estimate of a quantitative trait locus (QTL) effect is the posterior mean of the QTL effect when a normal prior distribution is assigned to the QTL. This note gives the derivation of the shrinkage estimate under the multivariate linear model. An important lemma regarding the posterior mean of a normal likelihood combined with a normal prior is introduced. The lemma is then used to derive the Bayesian shrinkage estimates of the QTL effects.

THE Bayesian shrinkage estimation of quantitative trait locus (QTL) effects was first introduced by Xu (2003) and later formalized by Wang *et al.* (2005). The multivariate version of the shrinkage estimation of QTL effects was recently developed by Yang and Xu (2007). The main purpose of the shrinkage estimation is to avoid variable selection for mapping multiple QTL. Once a normal prior distribution for each regression coefficient is incorporated into the QTL mapping program, the method can handle substantially more QTL effects than the classical maximum-likelihood (ML) method. In addition, the shrinkage method produces much clearer signals of QTL on the genome than the ML method. As a result, shrinkage mapping appears to have pointed to a new direction for future research in QTL mapping.

The key issue of shrinkage estimation is the normal prior distribution assigned to the regression coefficient (QTL effect). More importantly, different regression coefficients are assigned different normal priors. Because the variances in the prior distributions determine the degrees of shrinkage, assigning different prior variances to different regression coefficients allows the method to differentially shrink regression coefficients. A smaller prior variance will cause the regression coefficient to shrink more while a larger prior variance will lead to less shrinkage. This phenomenon is called selective shrinkage.

After incorporating the normal prior distribution into the likelihood function, we can derive the posterior distribution of the regression coefficient, which remains normal due to the conjugate nature of the normal prior. The posterior mean and posterior variance are used to generate a posterior sample of the regression coefficient. Formulas for the posterior mean and posterior variance are mathematically attractive (see Xu 2003; Wang *et al.* 2005; Yang and Xu 2007). However, due to page limitations of these publications, derivation of the formulas was not provided in these articles.

Derivation of the univariate shrinkage estimation closely followed Box and Tiao's (1973, Appendix A1.1) combination of a univariate normal likelihood and a univariate normal prior. Derivation of the multivariate shrinkage estimation followed the general Bayesian linear model of Lindley and Smith (1972) and the best linear unbiased prediction (BLUP) of Robinson (1991). The derivations presented by these authors were particularly targeted to statisticians and often difficult to understand by the audience of the genetics community. I have been regularly receiving e-mails and calls from readers asking for the derivation. These readers (almost all genetics professionals and students) are often interested in extending the shrinkage method to handle QTL mapping in different mapping populations. Understanding the derivation of these formulas is crucial to the development of new shrinkage methods. Simply pointing them to the above references often does not help too much because intermediate steps are needed to lead to the shrinkage estimate presented by Xu (2003). By doing this, I often give them an impression of irresponsibility. Therefore, I prepared a short note for the derivation and distributed the note to these interested readers. The note briefly summarizes the derivation using a language that is easy to understand by geneticists with basic statistical training. Given the increasing interest of the derivation from the QTL mapping community, it is more efficient to publish the note in Genetics where the very first shrinkage method (Xu 2003) was published.

[1]*Address for correspondence:* 900 University Ave., University of California, Riverside, CA 92521-0124. E-mail: xu@genetics.ucr.edu

THEORY AND MODEL

Shrinkage estimates: Let Y_j be an $m \times 1$ vector for the phenotypic values of m traits collected from the jth individual for $j = 1, \ldots, n$, where n is the sample size. This vector is described by the following linear model,

$$Y_j = b_0 + \sum_{k=1}^{p} X_{jk} b_k + e_j \; (j = 1, \ldots, n), \tag{1}$$

where b_0 is an $m \times 1$ vector for the population means (or intercept), X_{jk} is an $m \times q$ design matrix (determined by the genotypes of the jth individual at the kth locus), b_k is a $q \times 1$ vector for the regression coefficients (QTL effects) for locus k ($k = 1, \ldots, p$), e_j is an $m \times 1$ vector of residual errors with an assumed $N(0, D)$ distribution, and D is an $m \times m$ positive definite covariance matrix. When the kth regression coefficient is considered, all other regression coefficients are treated as constants and thus model (1) can be rewritten as

$$Y_j^* = b_0 + X_{jk} b_k + e_j \; (j = 1, \ldots, n), \tag{2}$$

where

$$Y_j^* = Y_j - \sum_{k' \neq k}^{p} X_{jk'} b_{k'} \tag{3}$$

is the phenotypic value adjusted by all other regression coefficients that are not currently under consideration. Let us describe b_k by the following normal prior $b_k \sim N(\eta_k, \Gamma_k)$, where η_k is a $q \times 1$ vector for the means and Γ_k is a $q \times q$ prior variance–covariance matrix. The posterior distribution of b_k is multivariate normal with mean

$$E(b_k \mid Y^*, b_0, D, \eta_k, \Gamma_k) = \left[\sum_{j=1}^{n} X_{jk}^T D^{-1} X_{jk} + \Gamma_k^{-1}\right]^{-1} \left[\sum_{j=1}^{n} X_{jk}^T D^{-1} (Y_j^* - b_0) + \Gamma_k^{-1} \eta_k\right] \tag{4}$$

and variance–covariance matrix

$$\operatorname{var}(b_k \mid Y^*, b_0, D, \eta_k, \Gamma_k) = \left[\sum_{j=1}^{n} X_{jk}^T D^{-1} X_{jk} + \Gamma_k^{-1}\right]^{-1}. \tag{5}$$

In shrinkage analysis, we often set $\eta_k = 0$ for $k = 1, \ldots, p$; as such the posterior mean becomes

$$E(b_k \mid Y^*, b_0, D, \Gamma_k) = \left[\sum_{j=1}^{n} X_{jk}^T D^{-1} X_{jk} + \Gamma_k^{-1}\right]^{-1} \left[\sum_{j=1}^{n} X_{jk}^T D^{-1} (Y_j^* - b_0)\right]. \tag{6}$$

This posterior mean is called the shrinkage estimate of the regression coefficient b_k. When $\Gamma_k \to \infty$, the prior is flat, leading to the usual least-squares estimate,

$$E(b_k \mid Y^*, b_0, D) = \left[\sum_{j=1}^{n} X_{jk}^T D^{-1} X_{jk}\right]^{-1} \left[\sum_{j=1}^{n} X_{jk}^T D^{-1} (Y_j^* - b_0)\right]. \tag{7}$$

When $\Gamma_k \to 0$, we have $\Gamma_k^{-1} \to \infty$, which leads to $\left[\sum_{j=1}^{n} X_{jk}^T D^{-1} X_{jk} + \Gamma_k^{-1}\right]^{-1} \to 0$ and thus $E(b_k \mid Y^*, b_0, D, \eta_k, \Gamma_k) \to 0$, an estimate shrunken to zero. Therefore, matrix Γ_k serves as a factor to determine the degree of shrinkage for the estimate of b_k. Because Γ_k varies, the degree of shrinkage also varies across k. To prove the shrinkage estimate, I first introduce the following lemma:

LEMMA. *Assume that parameter b can be inferred from two independent sources of information. Let $b \mid I_1 \sim N(\beta_1, \Sigma_1)$ and $b \mid I_2 \sim N(\beta_2, \Sigma_2)$ be the distributions of the two sources of information. When we combine I_1 and I_2, the distribution of b remains multivariate normal $b \mid I_1, I_2 \sim N(\beta, \Sigma)$ with mean $\beta = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\beta_1 + \Sigma_2^{-1}\beta_2)$ and variance–covariance matrix $\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$.*

Proof of the lemma. The distribution of b given the two sources of information is described by

$$\begin{aligned} p(b \mid I_1, I_2) &= p(b \mid I_1) p(b \mid I_2) \\ &= C \exp\left\{-\frac{1}{2}\left[(b - \beta_1)^T \Sigma_1^{-1} (b - \beta_1) + (b - \beta_2)^T \Sigma_2^{-1} (b - \beta_2)\right]\right\}, \end{aligned} \tag{8}$$

where C is a constant with respect to b. When deriving a distribution, we are interested only in the kernel of the distribution. A kernel of a distribution is the central part of the distribution function, the part that remains when constants are disregarded. In the above distribution, the logarithm of the kernel is

$$K(b) = -\frac{1}{2}\left[(b - \beta_1)^T \Sigma_1^{-1} (b - \beta_1) + (b - \beta_2)^T \Sigma_2^{-1} (b - \beta_2)\right], \tag{9}$$

which is further expressed by

$$\begin{aligned} K(b) = &-\frac{1}{2}\left[b^T (\Sigma_1^{-1} + \Sigma_2^{-1}) b - 2 b^T (\Sigma_1^{-1}\beta_1 + \Sigma_2^{-1}\beta_2)\right] \\ &- \frac{1}{2}(\beta_1^T \Sigma_1^{-1} \beta_1 + \beta_2^T \Sigma_2^{-1} \beta_2). \end{aligned} \tag{10}$$

We can see that this kernel involves another constant, $-\frac{1}{2}(\beta_1^T \Sigma_1^{-1} \beta_1 + \beta_2^T \Sigma_2^{-1} \beta_2)$, which can be ignored also. Therefore, the actual kernel that contains only the linear and quadratic functions of b is

$$\begin{aligned} K(b) = -\frac{1}{2}\Big[&b^T (\Sigma_1^{-1} + \Sigma_2^{-1}) b - 2 b^T (\Sigma_1^{-1} + \Sigma_2^{-1})(\Sigma_1^{-1} + \Sigma_2^{-1})^{-1} \\ &\times (\Sigma_1^{-1}\beta_1 + \Sigma_2^{-1}\beta_2)\Big]. \end{aligned} \tag{11}$$

Let $\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}$ and $\beta = (\Sigma_1^{-1} + \Sigma_2^{-1})^{-1}(\Sigma_1^{-1}\beta_1 + \Sigma_2^{-1}\beta_2)$. The kernel is simplified into

$$K(b) = -\frac{1}{2}(b^T \Sigma^{-1} b - 2 b^T \Sigma^{-1} \beta), \tag{12}$$

which turns out to be the kernel of $N(\beta, \Sigma)$. Therefore, we conclude that $b \mid I_1, I_2 \sim N(\beta, \Sigma)$.

Derivation of the shrinkage estimates: We now use the above lemma to derive the shrinkage estimate of b_k. The two sources of information for b_k come from the data (Y^*) and the prior. Information from the data is used to infer b_k through the maximum-likelihood method. The log-likelihood function is

$$L(b_k) = -\frac{n}{2}\ln|D| - \frac{1}{2} \times \sum_{j=1}^{n}(Y_j^* - b_0 - X_{jk}b_k)^T D^{-1}(Y_j^* - b_0 - X_{jk}b_k). \quad (13)$$

The maximum-likelihood estimate of b_k is

$$\hat{b}_k = \left[\sum_{j=1}^{n} X_{jk}^T D^{-1} X_{jk}\right]^{-1}\left[\sum_{j=1}^{n} X_{jk}^T D^{-1}(Y_j^* - b_0)\right] \quad (14)$$

and the variance of this estimate is

$$\mathrm{var}(\hat{b}_k) = \left[\sum_{j=1}^{n} X_{jk}^T D^{-1} X_{jk}\right]^{-1}. \quad (15)$$

Let $\beta_1 = \hat{b}_k$ and $\Sigma_1 = \mathrm{var}(\hat{b}_k)$. After some algebraic manipulation on the likelihood function, we find that Equation 13 has the following normal kernel with respect to b_k,

$$K_1(b_k) = -\frac{1}{2}(b_k - \beta_1)^T\Sigma_1^{-1}(b_k - \beta_1). \quad (16)$$

Therefore, the distribution of b_k inferred from the data is $b_k \mid I_1 \sim N(\beta_1, \Sigma_1)$. The second source of information for b_k is the prior distribution $N(\eta_k, \Gamma_k)$. If we let $\beta_2 = \eta_k$ and $\Sigma_2 = \Gamma_k$, the distribution of b_k from the second source of information is $b_k \mid I_2 \sim N(\beta_2, \Sigma_2)$. According to the lemma, the posterior mean of b_k is

$$\begin{aligned} &E(b_k \mid Y^*, b_0, D, \eta_k, \Gamma_k) \\ &\quad = [\mathrm{var}^{-1}(\hat{b}_k) + \Gamma_k^{-1}]^{-1}[\mathrm{var}^{-1}(\hat{b}_k)\hat{b}_k + \Gamma_k^{-1}\eta_k] \\ &\quad = \left[\sum_{j=1}^{n} X_{jk}^T D^{-1} X_{jk} + \Gamma_k^{-1}\right]^{-1}\left[\sum_{j=1}^{n} X_{jk}^T D^{-1}(Y_j^* - b_0) + \Gamma_k^{-1}\eta_k\right] \end{aligned} \quad (17)$$

and the posterior variance is

$$\begin{aligned} &\mathrm{var}(b_k \mid Y^*, b_0, D, \eta_k, \Gamma_k) \\ &\quad = [\mathrm{var}^{-1}(\hat{b}_k) + \Gamma_k^{-1}]^{-1} = \left[\sum_{j=1}^{n} X_{jk}^T D^{-1} X_{jk} + \Gamma_k^{-1}\right]^{-1}. \end{aligned} \quad (18)$$

This concludes the derivation of the shrinkage estimate of b_k.

Univariate version of the shrinkage estimate: The shrinkage estimate of the regression coefficient given by Xu (2003) is a special case of the general shrinkage estimate. The regression model of Xu (2003) is

$$y_j = b_0 + \sum_{k=1}^{p} x_{jk}b_k + e_j, \quad (19)$$

where every variable in the equation is a scalar rather than a matrix. When focused on the kth regression coefficient, the model is rewritten as

$$y_j^* = b_0 + x_{jk}b_k + e_j, \quad (20)$$

where $y_j^* = y_j - \sum_{k' \neq k}^{p} x_{jk'}b_{k'}$ is the adjusted data. Let us assume $e_j \sim N(0, \sigma_0^2)$, where σ_0^2 is the univariate version of matrix D. Assume that the prior distribution for b_k is $N(0, \sigma_k^2)$. Therefore, the univariate versions of η_k and Γ_k are $\eta_k = 0$ and $\Gamma_k = \sigma_k^2$, respectively. Substituting all the parameters of Equations 4 and 5 by their univariate counterparts, we have

$$\begin{aligned} &E(b_k \mid y^*, b_0, \sigma_0^2, \sigma_k^2) \\ &\quad = \left[\sum_{j=1}^{n} x_{jk}^2/\sigma_0^2 + \sigma_k^{-2}\right]^{-1}\left[\sum_{j=1}^{n} x_{jk}(y_j^* - b_0)/\sigma_0^2\right] \\ &\quad = \left[\sum_{j=1}^{n} x_{jk}^2 + \sigma_0^2/\sigma_k^2\right]^{-1}\left[\sum_{j=1}^{n} x_{jk}(y_j - b_0 - \textstyle\sum_{k' \neq k}^{p} x_{jk'}b_{k'})\right] \end{aligned} \quad (21)$$

and

$$\begin{aligned} &\mathrm{var}(b_k \mid y^*, b_0, \sigma_0^2, \sigma_k^2) \\ &\quad = \left[\sum_{j=1}^{n} x_{jk}^2/\sigma_0^2 + \sigma_k^{-2}\right]^{-1} = \left[\sum_{j=1}^{n} x_{jk}^2 + \sigma_0^2/\sigma_k^2\right]^{-1}\sigma_0^2. \end{aligned} \quad (22)$$

These equations are exactly the same as Equations 5 and 6 given by Xu (2003).

DISCUSSION

There are several alternative ways to prove the shrinkage estimation, such as the conditional distribution of multivariate normal variables (Giri 1996). The method presented in this note is a generalization of Box and Tiao's (1973, Appendix A1.1) combination of a univariate normal likelihood and a univariate normal prior. Using the method of Box and Tiao (1973), we can extend the lemma to the situation of inferring b from more than two independent sources of information. Let m be the number of sources of information (independent of each other) used to infer b and the distribution from the ith source is $N(\beta_i, \Sigma_i)$ for $i = 1, \ldots, m$. The posterior distribution of b combining all the sources of information is $b \mid I_1, \ldots, I_m \sim N(\beta, \Sigma)$, where

$$\beta = (\Sigma_1^{-1} + \Sigma_2^{-1} + \ \ldots\ + \Sigma_m^{-1})^{-1}(\Sigma_1^{-1}\beta_1 + \Sigma_2^{-1}\beta_2 + \ \ldots\ + \Sigma_m^{-1}\beta_m) \quad (23)$$

and

$$\Sigma = (\Sigma_1^{-1} + \Sigma_2^{-1} + \ \ldots\ + \ \Sigma_m^{-1})^{-1}. \quad (24)$$

One can use mathematical induction to prove Equations 23 and 24, starting from $m = 2$ (given in the lemma) and moving to $m + 1$ and so on.

Bayesian shrinkage estimation refers to the biased estimation of a regression coefficient toward zero using

a prior variance as a factor to control the degree of shrinkage. A normal prior is often selected because it is a conjugate prior so that the posterior distribution remains normal. A normal posterior simplifies the MCMC sampling process because the Gibbs sampler can be used to draw the regression coefficient. Other prior distributions have been proposed, *e.g.*, the mixture prior of two normal distributions (George and Mcmulloch 1993; Yi *et al.* 2003) and the spike and slab model (Ishwaran and Rao 2005). A *t*-distribution may also be used as a prior for the regression coefficient. However, the posterior distribution using a nonnormal prior rarely has an explicit form of a distribution, making Gibbs sampling impossible and thus complicating the MCMC sampling process.

The shrinkage method for regression analysis may also be called the random model approach to regression analysis, or simply random regression, because each regression coefficient is treated as a random effect with a (prior) normal distribution. It is well known that there is no limit in the number of random effects that can be handled by a random model. The success of a random linear model analysis, however, depends on the variance components chosen for the random model. If a random model contains an excessively large number of regression coefficients, most of them will be zero or close to zero. The sparse nature of the regression coefficients cannot be characterized by the random linear model alone and it must be accompanied by an efficient method to choose the variance components. In QTL mapping, the number of variance components can be extremely large, making subjective selection of the variance components impossible. Therefore, the variance components must be estimated from the data.

The most convenient way to estimate the variance components is to use the maximum-likelihood method. The estimated variance components are used in place of the prior variances to estimate the regression coefficients. The method is called the empirical Bayes method as far as the estimation of regression coefficients is concerned (Xu 2007). To reflect the sparse nature of the regression coefficients, a prior distribution is often assigned to each variance component. This is called hierarchical modeling (Gelman 2005). Furthermore, the prior distribution should be highly concentrated around zero. Many different prior distributions can be chosen for the variance components, but the scaled inverse chi-square distribution is the most convenient and flexible prior with such a property (Lindley and Smith 1972). Exponential distribution (Tibshirani 1996) and half *t*-distribution (Gelman 2006) have also been used. The prior choice for variance components of the random regression analysis is a very active research area to explore. More efficient priors may be developed in the future.

In the random regression analysis, the variance of a regression coefficient is not the primary interest of the investigator; rather, it is used only for the purpose of controlling the magnitude of the shrinkage. If the regression coefficients are batched (clustered) so that regression coefficients in the same batches share the same prior distribution, the variance may be estimated accurately and the estimate of it may be meaningful (Gelman 2005). In this case, the primary interest has been shifted from the regression coefficients to the variances of the regression coefficients; the method is better called the analysis of variances (ANOVA) (Gelman 2005). In the usual shrinkage analysis, the regression coefficients are not batched; *i.e.*, every regression coefficient has its own prior variance, and the estimated variance for a regression coefficient may vary drastically across the posterior sample. This problem may look very bad, but will not seriously harm the Bayesian shrinkage estimates of the regression coefficients. One can minimize the variation of the sampled variance across the posterior sample by using some proper prior distribution for the variance (Gelman 2005).

LITERATURE CITED

Box, G. E. P., and G. C. Tiao, 1973 *Bayesian Inference in Statistical Analysis*. Wiley & Sons, New York.

Gelman, A., 2005 Analysis of variance–why it is more important than ever. Ann. Stat. **33:** 1–53.

Gelman, A., 2006 Prior distribution for variance parameters in hierarchical models. Bayesian Anal. **1:** 515–533.

George, E. I., and R. E. McMulloch, 1993 Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **91:** 883–904.

Giri, N. C., 1996 *Multivariate Statistical Analysis*. Marcel Dekker, New York.

Ishwaran, H., and J. S. Rao, 2005 Spike and slab variable selection: frequentist and Bayesian strategies. Ann. Stat. **33:** 730–773.

Lindley, D. V., and A. F. M. Smith, 1972 Bayes estimates for the linear model. J. R. Stat. Soc. Ser. B **34:** 1–41.

Robinson, G. K., 1991 That BLUP is a good thing: the estimation of random effects. Stat. Sci. **6:** 15–32.

Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B **58:** 267–288.

Wang, H., Y. M. Zhang, X. Li, G. L. Masinde, S. Mohan *et al.*, 2005 Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics **170:** 465–480.

Xu, S., 2003 Estimating polygenic effects using markers of the entire genome. Genetics **163:** 789–801.

Xu, S., 2007 An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics **63:** 513–521.

Yang, R., and S. Xu, 2007 Bayesian shrinkage analysis of quantitative trait loci for dynamic traits. Genetics **176:** 1169–1185.

Yi, N., V. George and D. B. Allison, 2003 Stochastic search variable selection for identifying quantitative trait loci. Genetics **164:** 1129–1138.

Communicating editor: B. J. Walsh