Priors in Whole-Genome Regression: The Bayesian Alphabet Returns

Daniel Gianola¹

Department of Animal Sciences, Department of Biostatistics and Medical Informatics, and Department of Dairy Science, University of Wisconsin, Madison, Wisconsin 53706

ABSTRACT Whole-genome enabled prediction of complex traits has received enormous attention in animal and plant breeding and is making inroads into human and even *Drosophila* genetics. The term “Bayesian alphabet” denotes a growing number of letters of the alphabet used to denote various Bayesian linear regressions that differ in the priors adopted, while sharing the same sampling model. We explore the role of the prior distribution in whole-genome regression models for dissecting complex traits in what is now a standard situation with genomic data where the number of unknown parameters (p) typically exceeds sample size (n). Members of the alphabet aim to confront this overparameterization in various manners, but it is shown here that the prior is always influential, unless $n \gg p$. This happens because parameters are not likelihood identified, so Bayesian learning is imperfect. Since inferences are not devoid of the influence of the prior, claims about genetic architecture from these methods should be taken with caution. However, all such procedures may deliver reasonable predictions of complex traits, provided that some parameters (“tuning knobs”) are assessed via a properly conducted cross-validation. It is concluded that members of the alphabet have a room in whole-genome prediction of phenotypes, but have somewhat doubtful inferential value, at least when sample size is such that $n \ll p$.

WHOLE-genome enabled prediction of complex traits has received much attention in animal and plant breeding (e.g., Meuwissen *et al.* 2001; Heffner *et al.* 2009; Lorenz *et al.* 2011; de los Campos *et al.* 2012a; Heslot *et al.* 2012) and is making inroads into human and even *Drosophila* genetics (e.g., de los Campos *et al.* 2010, 2012b; Makowsky *et al.* 2011; Ober *et al.* 2012; Vázquez *et al.* 2012). This approach is known as “genomic selection” in breeding of agricultural species. The term “Bayesian alphabet” was coined by Gianola *et al.* (2009) to refer to a (growing) number of letters of the alphabet used to denote various Bayesian linear regressions used in genomic selection that differ in the priors adopted while sharing the same sampling model: a Gaussian distribution with mean vector represented by a regression on p markers, typically SNPs, and a residual variance, σ_e^2 . A recent review of some of these methods is in de los Campos *et al.* (2012a). In addition to prediction, this whole-

genome approach lends itself to investigation of “genetic architecture,” often defined as the number of genes affecting a quantitative trait, the allelic effects on phenotypes, and the frequency distribution spectrum of alleles at these genes (e.g., Hill 2012). If epistasis and pleiotropy are brought into the picture, this definition of genetic architecture needs to be expanded significantly.

Most researchers in genomic selection are familiar with most letters of the alphabet, but we provide a brief review of its ontogeny. The alphabet started with Bayes A and B (Meuwissen *et al.* 2001), but there has been rapid expansion since, as illustrated by the Bayes C π and D π methods (Habier *et al.* 2011). Apart from between-letter variation, there is also variation within letters, such as fast EM-Bayes A (Sun *et al.* 2012), fast Bayes B (Meuwissen *et al.* 2009), and BRR (Bayesian ridge regression on markers), which is equivalent to G-BLUP (Van Raden 2008) but with variance parameters estimated Bayesianly; the equivalence between G-BLUP and ridge regression is given, for example, in de los Campos *et al.* (2009a,b). The letter D has several variants: Bayes D0, D1, D2, and D3 (Wellman and Bennewitz 2012).

Here, L is used to denote the Bayesian Lasso (Park and Casella 2008; de los Campos *et al.* 2009a,b) while L1 and L2 can be used to refer to variants due to Legarra *et al.* (2011).

Copyright © 2013 by the Genetics Society of America
doi: 10.1534/genetics.113.151753

Manuscript received February 16, 2013; accepted for publication April 13, 2013

This article is dedicated to the late Professor George Casella (1951–2012), a great (and friendly) Bayesian statistician who made incursions into quantitative genetics at several points of his career.

¹Address for correspondence: 1675 Observatory Dr., University of Wisconsin, Madison, WI 53706. E-mail: gianola@ansci.wisc.edu

There is also the EL Bayesian Lasso of Mutshinda and Sillanpää (2010), with EL standing for “extended Lasso.” An almost empty hiatus spans from letters D to R (Erbe *et al.* 2012) with Bayes RS emerging even more recently (Brondum *et al.* 2012). Wang *et al.* (2013) presented Bayes TA, TB, and TC π , extensions of the corresponding letters to threshold models. The upper bound of the alphabet seems to have been defined by Larry Schaeffer (personal communication, Interbull Meeting, Guelph, 2011) when he threatened attendants of this conference with Bayes Z- Δ , although full details have not been published yet. The preceding review may not be comprehensive, as there may be other members of the alphabet that are unknown to the author. It is tempting to conjecture that there may be issues with individual members of the alphabet, as this continued growth is suggestive of a state of lack of satisfaction with any given letter.

This article explores the role of the prior distribution in whole-genome regression models for predicting or dissecting complex traits. In particular, we address a standard situation encountered in genomic selection: with genomic data, the number of unknown parameters exceeds sample size. Section *General Setting* presents the regression model and reminds readers that, for the preceding situation, regression coefficients on marker genotypes are not identified in the likelihood function so that the data do not contain information for inference that is uncontaminated from the effects of the prior, except in a subspace. Bayesian methods for confronting the blatant overparameterization of genomic selection models are reviewed in this section, where it is shown that the prior is always influential in this setting. The section *Bayesian Shrinkage* discusses how ridge regression produces frequency-dependent shrinkage, while Bayes A, Bayes B, Bayes L, and Bayes R effect a type of shrinkage that is both allelic frequency and effect-size dependent. After establishing in the preceding sections, hopefully in a firm manner, that all members of the alphabet do not lead to inferences that are devoid of the influence of the prior, it is argued in the *Discussion* that all such methods may deliver reasonable predictions of complex traits, provided that some parameters (“tuning knobs”) are assessed properly. It is concluded that, while members of the alphabet cannot be construed as providing solid inferences about “genetic architecture,” they do have a room in whole-genome prediction of phenotypes.

General Setting

Let $\mathbf{y}$ be an $n \times 1$ vector of target responses (e.g., phenotypes, preprocessed data). Using molecular markers, all members of the alphabet pose the same linear regression of phenotypes on marker codes, that is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad (1)$$

where $\mathbf{X}$ is an $n \times p$ matrix of marker codes (e.g., $-1, 0, 1$ for aa, Aa , and AA genotypes, respectively); when additive action is assumed, $\boldsymbol{\beta} = \{\beta_j\}$ is a vector of allelic substitution

effects for each of p markers, and $\mathbf{e}$ is a vector of residuals typically assigned the normal distribution $\mathbf{e}|\sigma_e^2 \sim N(\mathbf{e}|0, \mathbf{I}\sigma_e^2)$ where σ_e^2 is the residual variance, defined earlier.

In the standard additive model of quantitative genetics (e.g., Falconer and Mackay 1996), the β_j are fixed parameters, while the elements x_{ij} of $\mathbf{X}$ are random variables; e.g., members of the j th column of $\mathbf{X}$ may be realizations from a Hardy–Weinberg distribution with corealizations in columns j and j' reflecting some linkage disequilibrium distribution. The maximum-likelihood estimator of $\boldsymbol{\beta}$ treats $\mathbf{X}$ as a fixed matrix and satisfies the system of equations

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^{(0)} = \mathbf{X}'\mathbf{y},$$

where $\boldsymbol{\beta}^{(0)}$ may not be a unique solution (Searle 1971). If $n < p$ $\mathbf{X}'\mathbf{X}$ is singular so the maximum-likelihood estimator is not unique, as there is an infinite number of solutions to the equations above. Letting $(\mathbf{X}'\mathbf{X})^-$ be a generalized inverse of $\mathbf{X}'\mathbf{X}$, one solution is $\boldsymbol{\beta}^{(0)} = (\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{y}$ with expectation $E(\boldsymbol{\beta}^{(0)}|\boldsymbol{\beta}) = (\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{X}\boldsymbol{\beta}$, producing a biased estimator of $\boldsymbol{\beta}$, with at least $p - n$ of its elements being equal to 0. On the other hand, $E(\mathbf{y}|\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta} = \mathbf{g}$ (the genetic signal captured by markers) is estimated uniquely because its estimator, $\mathbf{X}\boldsymbol{\beta}^{(0)}$ is unique, although this reproduces $\mathbf{y}$ exactly in the $n < p$ situation. Fisher’s information content about $\boldsymbol{\beta}$ in the sample is $\mathbf{X}'\mathbf{X}\sigma_e^{-2}$ and, because this matrix is singular, one cannot speak about information pertaining to individual marker effects in a strict sense. However, the information content about $\mathbf{g} = \{g_i\}$ is $\mathbf{I}\sigma_e^{-2}$, meaning that information about each genotypic value g_i is proportional to that conveyed by a sample of size 1. Hence, in an $n < p$ model, maximum-likelihood cannot be used either as an inferential or as a predictive machine. In the latter case, it does not generalize to new samples, because it copies both noise and signal contained in model training data.

In their proposals for employing whole-genome markers in a linear regression model, Meuwissen *et al.* (2001) were inspired by the fact that animal breeders had dealt with the $n \ll p$ problem successfully in the context of predicting random effects via best linear unbiased prediction (BLUP); see Henderson (1984) for a review, with a gentler treatment in Mrode (2005). BLUP assumes that marker effects are drawn from some distribution with known variance components; only knowledge of the covariance structure is needed and the form of the distribution is immaterial, although a linear model must hold. An alternative is provided by the Bayesian treatment but, here, the meaning of probability and the manner in which unknowns are inferred are different from their frequentist counterparts (Gianola and Fernando 1986; Robinson 1991). The distinctions between these two views are emphasized next, but the two approaches confront $n \ll p$ by bringing external information to the problem, as noted early in the game by Robertson (1955).

The BLUP approach to whole-genome prediction assumes that $\boldsymbol{\beta}$ has a null mean vector and some known covariance

matrix $\mathbf{V}_\beta$; then $E(\mathbf{y}) = 0$ and the best linear unbiased predictor of β is

$$\text{BLUP}(\beta) = \mathbf{V}_\beta \mathbf{X}' (\mathbf{X} \mathbf{V}_\beta \mathbf{X}' + \mathbf{I} \sigma_e^2)^{-1} \mathbf{y}.$$

Here, β is regarded as a random draw from the distribution indicated above so, on average $E_y[\text{BLUP}(\beta)] = 0$, meaning that $\text{BLUP}(\beta)$ is unbiased with respect to the mean of the random effects distribution, $E(\beta | \mathbf{V}_\beta)$. BLUP envisages a sampling scheme where one draws a different realization of marker effects in every repetition of the sampling, such that, over all repetitions, 0 is obtained, on average. BLUP estimates zero without bias! However, when one is interested about individual marker effects (or about the genetic values of a given individual), the inference to be made pertains to the specific item of interest, and not to the average of their distribution. If so, BLUP is biased with respect to specific marker effects (the classical fixed model of quantitative genetics) because

$$E[\text{BLUP}(\beta) | \beta] = \mathbf{V}_\beta \mathbf{X}' (\mathbf{X} \mathbf{V}_\beta \mathbf{X}' + \mathbf{I} \sigma_e^2)^{-1} \mathbf{X} \beta,$$

so that the bias is $[\mathbf{I}_p - \mathbf{V}_\beta \mathbf{X}' (\mathbf{X} \mathbf{V}_\beta \mathbf{X}' + \mathbf{I} \sigma_e^2)^{-1} \mathbf{X}] \beta$, where $\mathbf{I}_p$ is an identity matrix of order p . The random effects treatment results in that $\text{BLUP}(\beta)$ is unique whether $n \ll p$ or not, but it produces a biased estimator of marker effects; this bias never disappears when $n \ll p$. On the other hand, if $n \rightarrow \infty$ and p stays fixed, bias goes away, given that the model is true. A toy example of the bias of BLUP with respect to the true, fixed, substitution effects is shown in the *Appendix (Bias of BLUP with respect to marker effects)*.

For the $n \ll p$ situation, Fan and Li (2001) discuss estimators that induce sparsity. However, to meet their so-called “oracle properties” (e.g., asymptotic unbiasedness), Fisher’s information matrix must be nonsingular for p_0 nonzero parameters ($p_0 < n$), these being the “true” effects of some markers on quantitative traits. BLUP and most members of the Bayesian alphabet do not produce a sparse model automatically; rather, they produce shrinkage of regression coefficients. Consider a sequence of P models of increasing dimensionality fitted to the same data, with $p_0 < p_1 < p_2 \dots < p_P$. The size of the “true” signal is dictated by the “true” effects p_0 and the size of the models could be viewed as corresponding to the number of SNPs in platforms of increasing density applied to the same data set. As marker density increases while n and p_0 remain fixed, estimates of marker effects must become necessarily smaller. How can “true” effects be learned properly if the model forces estimates to become smaller as p grows? Given the rhythm of technology, it is unlikely that we will reach the situation where $n \gg p_P$. At this point there is not much hope for learning marker effects in a manner that is free from making additional untestable assumptions. A minor complication: for the oracle properties to hold the true model must be “hit.” This is probably an unrealistic proposal when dealing with complex traits where many difficulties arise; for example,

linkage disequilibrium creates ambiguity because many markers can act as proxies for others and complex forms of epistasis are bound to produce havoc in a naive linear model on additive effects.

One way of tackling the $n \ll p$ problem is by introducing restrictions on the size of the regression coefficients, i.e., shrinkage or “regularization.” In the machine-learning literature this is attained via ad hoc penalty functions that produce regularization (e.g., Bishop 2006; Hastie *et al.* 2009). Bayesian methods with proper priors produce regularization automatically, to an extent that depends on the prior adopted. The various members of the Bayesian alphabet effect shrinkage in different manners, an issue explored subsequently in this article. Let all unknown parameters of a model be represented by $\theta = (\theta_1, \theta_2)$, where θ_1 and θ_2 denote distinct parameters, e.g., marker effects and their apparent (the reason for these terms is made clear later) variances, respectively, in Bayes A. The posterior distribution of θ (assume, for simplicity, that the residual variance is known) is

$$p(\theta_1, \theta_2 | \mathbf{y}, \sigma_e^2, H) \propto p(\mathbf{y} | \theta_1, \theta_2, \sigma_e^2, H) p(\theta_1, \theta_2 | H) \quad (2)$$

$$\propto p(\mathbf{y} | \theta_1, \sigma_e^2) p(\theta_1, \theta_2 | H). \quad (3)$$

Above, H is a set of more or less arbitrarily specified hyperparameters. Expression (3) results from the assumption that, given location effects θ_1 (e.g., allelic substitution effects), the data are conditionally independent of θ_2 . Further,

$$p(\theta_1, \theta_2 | H) = p(\theta_1 | \theta_2, H) p(\theta_2 | H) = p(\theta_1 | \theta_2) p(\theta_2 | H),$$

with the expression on the right side resulting because, given θ_2 , location effects θ_1 do not depend on H (e.g., Sorensen and Gianola 2002). Note that $p(\theta_1 | \theta_2)$ is a conditional prior distribution, while the marginal prior distribution actually assigned to θ_1 is

$$p(\theta_1 | H) = \int p(\theta_1 | \theta_2, H) p(\theta_2 | H) d\theta_2. \quad (4)$$

Likewise, $p(\theta_1 | \mathbf{y}, H)$ and $p(g(\theta_1) | \mathbf{y}, H)$ denote the marginal posterior distributions of θ_1 and $g(\theta_1)$, the latter being any function of θ_1 . For example, if θ_1 is the vector β , one may be interested in the posterior distribution of $\mathbf{X}\beta$, the marked signal. The results of a Bayesian analysis should not be interpreted from a frequentist perspective, as the meaning of probability is different in the two camps (Bernardo and Smith 1994; O’Hagan 1994; Sorensen and Gianola 2002). For example, BLUP is an unbiased predictor in conceptual repeated sampling, but corresponds to the posterior mean of marker effects in a Bayesian Gaussian model with known covariance structure. In the latter, the data are fixed; in the BLUP setting, the data vary at random.

An important issue is the influence of priors on inference. Theory on Bayesian asymptotics dictates that, as sample size

grows, the influence of the prior vanishes gradually. In the limit, the posterior distribution becomes normal, centered at the maximum-likelihood estimator and with covariance matrix given by the inverse of Fisher's information measure, so the prior matters little in large samples (Bernardo and Smith 1994). However, this result holds for parameters that are likelihood identifiable, *i.e.*, when their maximum-likelihood estimator exists, but it must be kept in mind that markers are not QTL, so the marker-based model is arguably wrong. In an $n \ll p$ setting, true Bayesian learning can take place for at most n parameters or functions thereof, since $p - n$ parameters are unidentified. Gelfand and Sahu (1999) show that one can learn about at most n linearly independent functions of marker effects, such as $\mathbf{x}_i'\boldsymbol{\beta}$. Carlin and Louis (1996) and Sorensen and Gianola (2002) give an example where the marginal posterior distributions of unidentified parameters exist if these are assigned proper priors; however, the priors will always matter and their influence will never vanish asymptotically. In the $n \ll p$ setting, inferences about marker effects (often referred to as learning genetic architecture, *e.g.*, inferring effects of some QTL) are always influenced by the priors adopted, apart from the fact that the model is wrong, as argued above. This means that stories that can be made from posterior distributions will depend on stories that are made *a priori*. For example, Lehermeier *et al.* (2013) demonstrated the influence of priors on predictive ability from various Bayesian models (Bayes A, B, L) with simulated and empirical data. Also, Gianola *et al.* (2009) showed that the priors in Bayes A and B drive inferences on variances of marker-specific effects.

A formal verification that individual marker effects are not identified from a Bayesian perspective using a definition by Dawid (1979) is presented in the *Appendix (Marker effects are not identified from a Bayesian perspective in the $n < p$ setting)*; this holds for any model, linear or nonlinear. A proof that is specific to the linear regression model on p markers with sample size n is given in the *Appendix* as well (*Inferences in a linear model with unidentified parameters*); there, it is shown that what is learned about $\boldsymbol{\beta}$ is a function of what is learned about $\mathbf{X}\boldsymbol{\beta}$. In other words, Bayesian learning occurs for n items but then this knowledge is “distributed” into p pieces via the relationship between $\boldsymbol{\beta}$ and $\mathbf{X}\boldsymbol{\beta}$ induced by the prior.

In summary, proper Bayesian learning from data in a linear regression model with $n < p$ takes place only for linear combinations of marker effects that are identified in the likelihood, that is, estimable. Any other marker effects or linear combinations thereof are redundant in the sampling model, but their posterior distributions exist and the posterior mean will differ from the prior mean. It follows that mechanistic conjectures about genetic architecture in the $n < p$ situation are, to a large extent, driven by prior assumptions and not by data. This observation has been corroborated empirically (*e.g.*, Heslot *et al.* 2012; Ober *et al.* 2012; Lehermeier *et al.* 2013): Bayesian models differing in their prior produce different inferences about individual marker effects, but most

often deliver similar predictive abilities if tuned properly. Not surprisingly, the posterior distributions of $\mathbf{x}_i'\boldsymbol{\beta}$ (the signal to be predicted for datum i) from varying models are more similar to each other than the corresponding priors, as this function is likelihood identifiable. In short, extant theory says that given that a model is “true” (oracle principle 1, Fan and Li 2001), the posterior mean of an identifiable parameter or of a likelihood-identified combination of parameters will converge to its true value, including any “true zero,” as sample size goes to infinity (oracle principle 2). This works for $n > p$ or for some estimators where sparsity is built in automatically, but the model must be true; Fan and Li (2001) describe several such estimators.

A situation in which proper Bayesian learning can take place is presented in the *Appendix (An example of proper Bayesian learning)*.

Bayesian Shrinkage

Given that learning about genetic architecture without contamination from effects of the prior does not take place whenever $n \ll p$, a question is what the various different members of the alphabet do. We examine ridge regression (BLUP), Bayes A, Bayes B, Bayes L, and Bayes R and also give a warning about a commonly used description of a specific prior; these procedures are prototypical so there is no need to consider other letters of the alphabet. All these methods have been reported in the genomic selection literature. Since the marginal posterior distribution of marker effects (with the exception of that of BLUP under normality) cannot be arrived at analytically, the methods are appraised from a heuristic perspective.

BLUP (ridge regression)

The vector of marker effects $\boldsymbol{\beta}$ is assigned the normal prior $N(\boldsymbol{\beta}|0, \mathbf{I}\sigma_\beta^2)$. The structure of the problem is well known and the mixed model equations leading to BLUP satisfy

$$(\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda)\tilde{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\boldsymbol{\beta}^{(0)}, \quad (5)$$

where $\tilde{\boldsymbol{\beta}} = \text{BLUP}(\boldsymbol{\beta})$, $\lambda = \sigma_e^2/\sigma_\beta^2$ is the variance ratio, and $\boldsymbol{\beta}^{(0)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is as before. One can write

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda)^{-1}(\mathbf{X}'\mathbf{X}\boldsymbol{\beta}^{(0)} + \lambda \times 0), \quad (6)$$

so $\tilde{\boldsymbol{\beta}}$ (unique) is a matrix weighted average of solution $\boldsymbol{\beta}^{(0)}$ (not unique) and of the prior mean 0, where the weights are $\mathbf{X}'\mathbf{X}$ and λ , respectively. For fixed p , as n increases, the rank of $\mathbf{X}'\mathbf{X}$ will increase, eventually exceeding p and, in the limit, the posterior distribution will be centered at the unique maximum-likelihood estimator (by consistency, this will converge to the “true” value of $\boldsymbol{\beta}$, given the model).

Representation (5) suggests that the same amount of shrinkage is effected to all p markers (because the same variance ratio λ is added to every diagonal element of $\mathbf{X}'\mathbf{X}$), but

this is not the case. This is clear from Equation 6 where, for each marker effect, the contributions from the data vary over markers; this is more transparent from inspection of the solutions in scalar form. For marker 1, as an example, the estimator of substitution effect is

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n x_{i1} (y_i - x_{i2}\tilde{\beta}_2 - \dots - x_{ip}\tilde{\beta}_p)}{\sum_{i=1}^n x_{i1}^2 + \lambda} = \frac{\sum_{i=1}^n x_{i1}^2 \tilde{\beta}_1 + \lambda \times 0}{\sum_{i=1}^n x_{i1}^2 + \lambda},$$

where

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n x_{i1} (y_i - x_{i2}\tilde{\beta}_2 - \dots - x_{ip}\tilde{\beta}_p)}{\sum_{i=1}^n x_{i1}^2}.$$

Then, the BLUP $\tilde{\beta}_1$ of the allele substitution effect can be viewed, heuristically, as a weighted average of a “data driven” estimate ($\tilde{\beta}_1$) and of the mean of the prior distribution (0) where the respective weights are $\sum_{i=1}^n x_{i1}^2 / (\sum_{i=1}^n x_{i1}^2 + \lambda)$ and $\lambda / (\sum_{i=1}^n x_{i1}^2 + \lambda)$, respectively. This suggests less shrinkage toward zero for markers (j , say) having larger values of $\sum_{i=1}^n x_{ij}^2$. Now, if for any marker genotypes are coded as $-1, 0, 1$ for aa, Aa , and AA , respectively, it follows (assuming Hardy–Weinberg proportions and centered marker codes) that $E(x_{ij}^2) = \text{Var}(x_{ij}) = 2p_j(1 - p_j)$ so $E(\sum_{i=1}^n x_{ij}^2) = 2p_j(1 - p_j)n$, where p_j is the frequency of the A -type allele at that locus. Hence, at fixed sample size n , BLUP effects less shrinkage toward zero of markers that have intermediate allelic frequencies, simply because $2p_j(1 - p_j)$ is maximum at $p_j = \frac{1}{2}$. To illustrate, we use this Hardy–Weinberg approximation and plot (Figure 1) the weight “assigned to the data” for marker j

$$W_j = \frac{\sum_{i=1}^n x_{ij}^2}{\sum_{i=1}^n x_{ij}^2 + \lambda} \approx \frac{2p_j(1 - p_j)}{2p_j(1 - p_j) + \lambda/n},$$

against allelic frequency at $\lambda/n = 1, 0.1, 0.01$ and 0.001 , respectively. As depicted in Figure 1, the extent of shrinkage is frequency and sample size dependent, with some differential shrinkage (bottom two curves) taking place at large values of λ/n , that is, at small sample sizes, but with little or no differential shrinkage otherwise, unless alleles are rare. Then, the often-made statement that BLUP or ridge regression perform an homogeneous shrinkage of marker effects is not correct. In short, shrinkage is frequency and sample size dependent but effect-size independent.

Bayes A

Bayes A (Meuwissen *et al.* 2001) consists of a three-stage hierarchical model. The first stage is the normal regression (1); the second stage assigns a normal conditional prior to each of the marker effects, all possessing a null mean but with a variance that is specific to each marker; the third stage assigns the same scaled inverted chi-square distribution with known scale

Figure 1 Approximate weight (W) assigned to the data (the weight assigned to prior information is $1 - W$) as a function of allelic frequency at a marker locus. From top to bottom the lines give the trajectory for λ/n values of 0.001 (solid line), 0.01 (short dashes), 0.1 (long dashes), and 1 (dots).

(S_{β}^2) and degrees of freedom (ν) parameters to each of the marker variances. The mechanistic argument for the Bayes A prior was that markers may contribute differentially to genetic variance (they do, to an extent depending on their effects, allelic frequencies, and linkage disequilibrium relationships with causal variants), so it seemed a good idea to “estimate” such variances. There are two difficulties: the first one is that the marginal prior for the markers effects, resulting from deconditioning the second stage over the third stage as done in (4), is the same for all markers. Second, there is scant Bayesian learning for marker-specific variances. This was pointed out by Gianola *et al.* (2009), who showed that all markers have the same prior distribution: a $t(\beta_j|0, \nu, S_{\beta}^2)$ process with null mean and variance $\nu S_{\beta}^2 / (\nu - 2)$. Given that this prior is homoscedastic over markers, why is it that it behaves differently from ridge regression BLUP, where all individual marker effects are assigned the prior $N(\beta_j|0, \sigma_{\beta}^2)$?

In Bayes A, the marginal posterior distribution of marker effects cannot be arrived at in closed form, but insight can be obtained from inspection of the joint mode of the posterior distribution of β , assuming that the residual variance is known; recall that S_{β}^2 and ν are known hyperparameters in Bayes A. The hierarchical model is then

$$y_i|\beta, \sigma_e^2 \sim N(y_i|\mathbf{x}_i'\beta, \sigma_e^2), i = 1, 2, \dots, n; \quad \beta_j|S_{\beta}^2, \nu \sim \text{IID } t(\beta_j|0, S_{\beta}^2, \nu), j = 1, 2, \dots, p,$$

where $\mathbf{x}_i'$ is the i th row of $\mathbf{X}$. Conditionally on σ_e^2 , S_{β}^2 and ν , the joint posterior density is

$$p(\beta|S_{\beta}^2, \nu, \sigma_e^2, \mathbf{y}) \propto \prod_{i=1}^n \exp\left[-\frac{1}{2\sigma_e^2}(y_i - \mathbf{x}_i'\beta)^2\right] \prod_{j=1}^p \left[1 + \frac{\beta_j^2}{S_{\beta}^2\nu}\right]^{-(1+\nu)/2}. \quad (7)$$

Using results presented in the Appendix (Mode of the conditional posterior distribution in Bayes A), an iterative scheme for locating a mode of (7) is given by

$$\beta^{[t+1]} = (\mathbf{X}'\mathbf{X} + \mathbf{W}_\beta^{[t]})^{-1} \mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X} + \mathbf{W}_\beta^{[t]})^{-1} \mathbf{X}'\mathbf{X}\beta^{(0)} \quad (8)$$

with successive updating; here,

$$\mathbf{W}_\beta^{[t]} = \text{Diag} \left\{ \frac{\sigma_e^2}{S_\beta^2} \frac{(1 + 1/\nu)}{(1 + \beta_j^{2[t]}/S_\beta^2 \nu)} \right\}$$

is a diagonal matrix. If this converges, it will do so to one of perhaps many stationary points, as it is known that t -regression models may produce multimodal log-posterior surfaces, especially if ν is small (McLachlan and Krishnan 1997). Hence, iteration (8) may lead to a point receiving little posterior plausibility.

The role of $\mathbf{W}_\beta = \{w_{jj,\beta_j}\}$ in (8) parallels that of the inverse of the genetic variance–covariance matrix (times σ_e^2) in standard BLUP (Henderson 1984), so that the larger w_{jj,β_j} is, the stronger the shrinkage toward 0 (mean of the prior distribution). However, while the variance ratio $\lambda = \sigma_e^2/\sigma_\beta^2$ is constant in ridge regression BLUP, here it varies over markers, as it takes the form w_{jj,β_j} . As $\nu \rightarrow \infty$ (the t -distribution approaches a normal one), $\lambda_j \rightarrow \sigma_e^2/S_\beta^2$, resembling λ of BLUP. On the other hand, if the t prior has a finite number of degrees of freedom, markers whose effects are closer to 0 are shrunk more strongly than those with larger absolute values, simply because λ_j is larger for the former. To illustrate, let $\sigma_e^2 = S_\beta^2 = 1$ so that the “variance ratio” is $\lambda_j = (1 + 1/\nu)/(1 + \beta_j^2/\nu)$. Figure 2 illustrates the impact of the marker effect on the “variance ratio” for $\nu = 4, 6, 10$, and 1000. It is seen that λ_j becomes smaller (less shrinkage toward zero) as the absolute value of the effect of the marker increases; also, shrinkage increases as the degrees of freedom of the distribution increase, at any given marker effect. Eventually, when $\nu \rightarrow \infty$ (so that the prior is normal) the variance ratio takes the same value for all markers (thick line in Figure 2, almost horizontal, corresponding to $\nu = 1000$). For markers with effects near zero, the t -distribution shrinks effects more strongly than the normal process, but it does not severely penalize markers having strong effects on the phenotype. Hence, in Bayes A shrinkage is marker-effect specific, with this specificity becoming milder as the value of ν increases. Note that (7) also induces frequency-specific shrinkage, due to the Bayesian compromise between the prior and $\mathbf{X}'\mathbf{X}$, as in the case of BLUP. Hence, apart from the effects of sample size, there are two sources of shrinkage in Bayes A, contrary to a single one in BLUP. This seems to confer Bayes A more flexibility than BLUP, but this is not necessarily good because the extra parameters ν and S_β^2 (this one playing the role of σ_β^2) are influential, and may affect “inference” of marker effects adversely (Lehermeier *et al.* 2013). Naturally, these parameters can be assigned priors and inferred from the resulting Bayesian model, but this was not suggested by Meuwissen *et al.* (2001).

Bayes B

A formulation of Bayes B as a mixture at the level of effects, but not of their variances, as in Meuwissen *et al.* (2001), is

Figure 2 Impact of marker effect and of the degrees of freedom (d.f.) parameter on the extent of shrinkage toward the prior distribution in Bayes A: the larger the value in the y-axis, the stronger the shrinkage toward 0. d.f. = 4, solid line; d.f. = 6, long-dashed line; d.f. = 10, short-dashed line; d.f. = 1000, gray circles, almost horizontal.

in Gianola *et al.* (2009) and Habier *et al.* (2011). The hierarchical model is

$$y_i | \beta, \sigma_e^2 \sim N(\mathbf{x}_i' \beta, \sigma_e^2),$$

$$\beta_j | S_\beta^2, \nu, \pi \sim \text{IID} \begin{cases} 0 & \text{with probability } \pi \\ t(0, S_\beta^2, \nu) & \text{with probability } 1 - \pi \end{cases}, \quad j = 1, 2, \dots, p.$$

The prior is a mixture of a “0-state” (a point mass at 0) with a t -distribution, the mixing probabilities being π and $1 - \pi$, respectively, where π is assumed known and specified arbitrarily. Recall (in informal notation) that

$$t(\beta_j | 0, S_\beta^2, \nu) = \int N(\beta_j | 0, \sigma_\beta^2) \chi^{-2}(\sigma_\beta^2 | S_\beta^2, \nu) d\sigma_\beta^2, \quad j = 1, 2, \dots, p,$$

where $\chi^{-2}(\sigma_\beta^2 | S_\beta^2, \nu)$ is a scaled-inverted chi-square distribution assigned as prior to the variance of the j th marker effect, σ_β^2 . Meuwissen *et al.* (2001) formulated the mixture at the level of these variances, arguing as follows: “the distribution of genetic variances across loci is that there are many loci with no genetic variance (not segregating) and a few with genetic variance.” Gianola *et al.* (2009) were critical of this formulation, both from statistical and genetic points of view.

The hierarchical prior is deceiving because, in fact, Bayes B ends up assigning the same marginal prior to every marker. This follows from consideration of the mean and variance of a mixture, *e.g.*, Gianola *et al.* (2006). The mean of a mixture is the weighted average of the means of the components (the weights being the mixing probabilities π and $1 - \pi$) and the variance is the weighted average of the component variances, plus a term that can be interpreted as “variance” among component means. One has

$$E(\beta_j | \pi) = (1 - \pi)E[t(\beta_j | 0, S_\beta^2, \nu)] = 0, \quad j = 1, 2, \dots, p,$$

where $E[t(\beta_j | 0, S_\beta^2, \nu)] = 0$ is the mean of the t -distribution, and

$$\text{Var}(\beta_j|\pi) = (1 - \pi) \frac{S_\beta^2 \nu}{\nu - 2}, \quad j = 1, 2, \dots, p.$$

Above, $\text{Var}[t(\beta_j|0, S_\beta^2, \nu)] = S_\beta^2 \nu / (\nu - 2)$ is the variance of the t -distribution. It follows that Bayes B assigns, *a priori*, the same mean and variance to all marker effects and that it uses a prior that is even more precise than the prior in Bayes A (the prior variance is reduced by a fraction π in Bayes B, relative to that of Bayes A). This makes effective Bayesian learning even more difficult in Bayes B than in Bayes A, as it takes more information from the data to “neutralize” the prior of Bayes B than that of Bayes A. At any rate, none of these two regression models allows for proper learning about marker effects or genetic architecture in the $n \ll p$ setting, as argued earlier in this article.

As for Bayes A, no closed forms for the marginal posterior distributions of marker effects exist for Bayes B. The posterior expectation of β is

$$E_{\text{Bayes B}}(\beta|\pi, S_\beta^2, \nu, \sigma_e^2, \mathbf{y}) = (1 - \pi) E_{\text{Bayes A}}(\beta|\pi, S_\beta^2, \nu, \sigma_e^2, \mathbf{y}). \quad (9)$$

This indicates that shrinkage toward 0 is stronger than in Bayes A since posterior means are smaller in Bayes B by a fraction π . Coupled with the arbitrary assignment of a value to π , the implication is that the prior is even more influential in Bayes B than in Bayes A. This could have been expected intuitively, but the point has not been made before, at least in this manner.

Wimmer *et al.* (2012) noted that methods such as Bayes B have yielded better predictive abilities than BLUP in many simulation studies reported in the literature, but that this has not been observed with real data (e.g., Ober *et al.* 2012). Wimmer *et al.* (2012) investigated predictive abilities of these two methods in maize and in *Arabidopsis*. The target populations differed in effective population size and in extent of linkage disequilibrium. Despite expected differences in genetic architecture among populations and traits, predictive abilities delivered by BLUP and Bayes B did not differ significantly for their target traits. Further, they found via simulation (personal communication) that Bayes B was effective for learning genetic architecture in the $n \ll p$ setting only when the number of true nonzero marker effects (s) is such that $s \ll n$, given the true model. Otherwise, the error of estimation of marker effects was as poor as that of BLUP, the latter found to be more robust over a wide range of situations (this is ironic, because BLUP or G-BLUP were not tailored for learning genetic architecture). In short, they confirmed that, provided one “hits” the true model (thus fulfilling oracle property 1 of Fan and Li 2001), effective learning of “true genetic architecture” is possible only if the model is very sparse relative to sample size. The condition $s \ll n$ would lead to oracle property 2, as anticipated by standard asymptotic Bayesian theory under regularity conditions.

BayesSSVS was proposed by Verbyla *et al.* (2009), but it is not discussed here because it is similar to Bayes B. Bayes C π of Habier *et al.* (2011) provides a more sensible formulation of the mixture, but it is similar in spirit and shares the same limitations of Bayes B, since parameter identification is not attained for most of the unknowns. An interesting example of consequences of overparameterization in Bayes C π is provided by Duchein *et al.* (2012); these authors noted that as values of π went up in the sampling process, realizations of marker effect variances went down. Hints about genetic architecture from Bayes B or Bayes C π or from other members of the alphabet should be taken very cautiously, at least when $n \ll p$.

Bayes L

Lasso regression (Tibshirani 1996) inspired the Bayesian Lasso (Bayes L here) of Park and Casella (2008), a method with followers such as Vázquez *et al.* (2010) and Crossa *et al.* (2010) and with an implementation available in the software R described by Pérez *et al.* (2010). The linear regression model is given in (1), but the prior assigned to marker effects is a Laplace (double exponential, DE) distribution. All marker effects are assumed to be independently and identically distributed as DE with the prior density being

$$p(\beta|\lambda) = \frac{\lambda}{2} \exp(-\lambda|\beta|). \quad (10)$$

Here, $E(\beta|\lambda) = 0$ and $\text{Var}(\beta|\lambda) = 2/\lambda^2$ for all markers; as λ increases the variance of the DE distribution decreases and the density becomes sharper. This prior assigns the same variance or prior uncertainty to all marker effects, but it possesses thicker tails than the normal prior. A comparative discussion of the DE prior is in de los Campos *et al.* (2012a). Even though Bayes L bears a parallel with the Lasso, it does not “kill” or remove markers from the model, contrary to what happens in variable selection approaches. Bayes L poses a leptokurtic prior, so it is expected to shrink effects more strongly toward zero than the Gaussian prior, as opposed to inducing sparsity in the strict sense of the Lasso.

Bayes L shrinks strongly: To appraise how Bayes L shrinks marker effects, we examine the mode(s) of the joint posterior distribution of β using the DE prior (10), assuming that λ and the residual variance are known. As in Tibshirani (1996), write $|\beta_j| = \beta_j^2 / |\beta_j|$; with this representation $\sum_{j=1}^p |\beta_j| = \beta' \mathbf{W}_\beta^{-1} \beta$, where $\mathbf{W}_\beta^{-1} = \text{Diag}\{1/|\beta_j|\}$. Using this, the log-posterior (apart from an additive constant) is

$$L(\beta|\mathbf{y}, \lambda, \sigma_e^2) = -\frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \sigma_e^2 \lambda \beta' \mathbf{W}_\beta^{-1} \beta}{2\sigma_e^2}, \quad (11)$$

If, as in Tibshirani (1996), it is ignored that $\mathbf{W}_\beta^{-1}$ is a random matrix (because it is a function of $|\beta_j|$) this takes the form of a standard BLUP representation, so the mode of the conditional posterior distribution of β satisfies

$$\tilde{\beta} = (\mathbf{X}'\mathbf{X} + \sigma_e^2 \lambda \mathbf{W}_\beta^{-1})^{-1} \mathbf{X}'\mathbf{y}. \quad (12)$$

Contrary to BLUP–ridge regression where shrinkage factors are marker effect independent, these factors take the form $\sigma_e^2 \lambda / |\beta_j|$ in Bayes L, implying that markers with tiny effects are shrunk more strongly toward zero, as a larger number is added to the diagonal elements of the coefficient matrix leading to solution $\tilde{\beta}$. Note, however, that (12) is not an explicit system, so it would make sense to iterate; details on an iterative scheme are in the Appendix (*Mode of the conditional posterior distribution in Bayes L*).

The preceding implies that Bayes L produces a more “effectively sparse” model. This can be seen from inspection of an “effective number of parameters” measure (e.g., Tibshirani 1996; Ruppert *et al.* 2003) given by

$$\begin{aligned} \text{d.f.}_{\text{ridge}} &= \text{tr} \left[\mathbf{X} \left(\mathbf{X}'\mathbf{X} + \mathbf{I} \frac{\sigma_e^2}{\sigma_\beta^2} \right)^{-1} \mathbf{X}' \right] \\ &= \text{tr} \left[\left(\mathbf{X}'\mathbf{X} + \mathbf{I} \frac{\sigma_e^2}{\sigma_\beta^2} \right)^{-1} \mathbf{X}'\mathbf{X} \right], \end{aligned}$$

and

$$\text{d.f.}_{\text{Bayes L}} = \text{tr} \left[(\mathbf{X}'\mathbf{X} + \sigma_e^2 \lambda \mathbf{W}_\beta^{-1})^{-1} \mathbf{X}'\mathbf{X} \right].$$

If $\mathbf{X}$ is orthonormalized, so that $\mathbf{X}'\mathbf{X} = \mathbf{I}$ (with dispersion parameters scaled accordingly)

$$\text{d.f.}_{\text{ridge}} = \text{tr} \left[\left(\mathbf{I} + \frac{\sigma_e^2}{\sigma_\beta^2} \right)^{-1} \right] = p \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_e^2}, \quad (13)$$

and

$$\text{d.f.}_{\text{Bayes L}} = \text{tr} \left[\left(\mathbf{I} + \sigma_e^2 \lambda \mathbf{W}_\beta^{-1} \right)^{-1} \right] = \sum_{j=1}^p \frac{|\beta_j|}{|\beta_j| + \sigma_e^2 \lambda}. \quad (14)$$

This enables us to see that, in ridge regression, every degree of freedom (contributor to model complexity) represented by a column of the orthonormalized marker matrix is attenuated by the same factor, $\sigma_\beta^2 / (\sigma_\beta^2 + \sigma_e^2)$. On the other hand, in Bayes L markers having tiny effects are effectively, but not physically, wiped out of the model. Also, markers with strong effects receive a heavier weight in this overall measure of complexity.

We simulated $P = 100,000$ marker effects from DE distributions with mean 0 and variances 10^{-16} , 10^{-8} , or 10^{-4} ; setting $\sigma_e^2 = 1$, the preceding three values can be interpreted as the contribution of an individual marker to variance relative to residual variability. When a marker effect had a large variance (10^{-4}), the entire battery of markers, assuming *a priori* independence of effects, represented $\frac{10}{11}$ of the total variance; on the other hand, when markers were

assigned a variance of 10^{-16} , markers accounted for about $10^{-11} / (10^{-11} + 1)$ of the total variability. Since the variance of the DE distribution is $2/\lambda^2$ the settings led to λ values of $\sqrt{2} \times 10^8$, $\sqrt{2} \times 10^4$, and $\sqrt{2} \times 10^2$, respectively; larger values of λ produce stronger shrinkage toward 0. The shrinkage factor is $\sigma_e^2 \lambda / |\beta_j|$ for marker j in Bayes L vs. $\sigma_e^2 / \sigma_\beta^2$ in ridge regression–BLUP. The contribution of a marker to the model was assessed as follows: from (13), in ridge regression each marker contributes the same amount, $\sigma_\beta^2 / (\sigma_\beta^2 + \sigma_e^2)$, to model complexity, whereas in Bayes L the corresponding metric is $|\beta_j| / (|\beta_j| + \sigma_e^2 \lambda)$, as given in (14). For ridge regression, the effective number of parameters was approximately 10^{-11} , 10^{-3} , and 10, for $\sigma_\beta^2 = 10^{-16}$, 10^{-8} , and 10^{-4} , respectively. For Bayes L, the corresponding effective number of parameters was 4.96×10^{-12} , 4.98×10^{-4} , and 4.98, respectively. Clearly, Bayes L produced a model that was more sparse than ridge regression–BLUP. Each of the markers made a tiny contribution to model complexity; for instance, when the variance of the double exponential of marker effects was 10^{-16} , the relative contributions to the model of individual markers ranged from 0 to 10^{-16} ; when the variance was 10^{-8} , these ranged from 0 to 10^{-8} , while the range was 0 – 10^{-4} for $\sigma_\beta^2 = 10^{-4}$. A plot of the density of the effective contributions to the model of each of the 100,000 markers is in Figure 3, for the case $\sigma_\beta^2 = 10^{-4}$; >95% of the markers contributed $< 2 \times 10^{-4}$ effective degrees of freedom to the model. Hence, when a marker contributes to variance in a tiny manner, shrinkage of their individual effects toward 0 is very strong. Then, if a marker effect conveys the meaning of a fraction equal to 10^{-8} , say, of some physical parameter, what can this tell us about the state of nature (*i.e.*, genetic architecture) in the absence of effective Bayesian learning, as argued earlier in the article? Probably not much unless $n \gg p$ and the model fitted is the “true” one, the latter requiring the extraordinarily strong assumption that a complex trait is well represented by a (multiple) linear regression.

Bayes L With Gamma prior for λ^2 : The DE density (10) is indexed by a single positive parameter λ , and if this is treated as unknown, the marginal prior density of a marker effect is

$$p(\beta) = \int_0^\infty p(\beta|\lambda)p(\lambda)d\lambda,$$

where $p(\lambda)$ is the prior density of λ . Clearly $E(\beta) = E_\lambda E(\beta|\lambda) = 0$, but the prior variance of β will depend on the distribution assigned to λ . Typically, a $\Gamma(r, \delta)$ prior is placed on λ^2 with the density being

$$p(\lambda^2|r, \delta) = \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{r-1} \exp(-\delta \lambda^2), \quad (15)$$

and $E(\lambda^2|r, \delta) = r/\delta$ and $\text{Var}(\lambda^2|r, \delta) = r/\delta^2$. Since λ is positive, $p(\beta|\lambda) = p(\beta|\lambda^2)$, so that

Density of the effective contribution of markers to the model.

Figure 3 Density (over 100,000 markers) of the “effective degrees of freedom” contributed by a marker to the model for a double exponential prior distribution with variance 10^{-4} .

$$\begin{aligned}
 p(\beta|r, \delta) &= \int_0^\infty p(\beta|\lambda^2) p(\lambda^2|r, \delta) d\lambda^2 \\
 &\propto \int_0^\infty (\lambda^2)^{r+\frac{1}{2}-1} \exp\left[-\left(|\beta|\sqrt{\lambda^2} + \delta\lambda^2\right)\right] d\lambda^2.
 \end{aligned} \tag{16}$$

Using in the above expression an approximation given in the *Appendix (Approximation of an integral in Bayes L)*, Equation 16 gives

$$p(\beta|r, \delta) \propto_{\text{approx.}} e^{-|\beta|\sqrt{r/\delta}} \left\{ 1 - \frac{1}{2} \sqrt{\frac{\delta}{r}} |\beta| \frac{1}{2\delta} + \frac{\delta}{8r} \left(|\beta|^2 + \sqrt{\frac{\delta}{r}} |\beta| \right) \frac{4r+3}{4\delta^2} \right\}, \tag{17}$$

where $\propto_{\text{approx.}}$ means “approximately proportional to.” If only the first term of the approximation is used, after normalization one gets

$$p_1(\beta|r, \delta) \approx \frac{e^{-|\beta|\sqrt{r/\delta}}}{\int_{-\infty}^{\infty} e^{-|\beta|\sqrt{r/\delta}} d\beta}, \tag{18}$$

and this is a DE density with parameter $\lambda = \sqrt{r/\delta}$. If both the first and second terms of the approximation are employed, one gets

$$p_2(\beta|r, \delta) \approx \frac{e^{-|\beta|\sqrt{r/\delta}} \left(1 - \frac{1}{2} \sqrt{\delta/r} |\beta| (1/2\delta) \right)}{\int_{-\infty}^{\infty} e^{-|\beta|\sqrt{r/\delta}} \left(1 - \frac{1}{2} \sqrt{\delta/r} |\beta| (1/2\delta) \right) d\beta}. \tag{19}$$

Next, we examine the shape of the unnormalized density (19) for two different $\Gamma(r, \delta)$ prior distributions of λ^2 . Set-

ting $r = \delta$ gives Gamma distributions with expected value 1 and variance $\frac{1}{\delta}$; use of $r = \delta = 4$ and $r = \delta = 16$ produces prior distributions with variances $\frac{1}{4}$ and $\frac{1}{16}$, respectively, and the corresponding densities are shown in Figure 4, top left. Taking into account that the prior distributions of marker effects have null means, the variance of approximation (19) to the marginal prior of β was evaluated by numerical integration between -9 and 9 as

$$\text{Var}_2(\beta|r = \delta) = \frac{\int_{-9}^9 \beta^2 e^{-|\beta|} \left(1 - \frac{1}{2} |\beta| \frac{1}{2\delta} \right) d\beta}{\int_{-9}^9 e^{-|\beta|} \left(1 - \frac{1}{2} |\beta| \frac{1}{2\delta} \right) d\beta},$$

yielding 1.73 ($\delta = 4$) and 1.93 ($\delta = 16$). This produces a seemingly paradoxical situation, where the more uncertain prior for λ^2 ($\delta = 4$) gives a marginal prior for the marker effect that is more precise (as measured by the variance) than that for $\delta = 16$. The densities, shown in Figure 4, top right, seem indistinguishable. However, if the plot is zoomed in the middle and right tails of the distribution (left and right bottom, respectively) the prior with $\delta = 16$ turns out to be less sharp and with thicker tails, thus explaining its larger variance. Also, the prior probability that a marker has an effect ranging from -0.3 to 0.3 is 0.274 for $\delta = 4$ (more variable prior for λ^2) and 0.263 for $\delta = 16$; the probabilities that a marker has an effect from 2 to 7 are 0.058 ($\delta = 4$) and 0.065 ($\delta = 16$), respectively.

Bayes L with uniform prior on λ : In an attempt to make the prior in a Bayesian analysis less aggressive, one may naively think that Bayes’s “principle of insufficient reason” (the uniform prior) may render the analysis objective. Let the uniform prior on λ be $\lambda|L, U \sim \text{Uniform}(L, U)$, where L and U are the lower and upper bounds, respectively, of the prior distribution. Mixing the DE distribution with parameter λ over this prior gives as marginal density

$$p(\beta|L, U) = \frac{1}{U-L} \int_L^U \frac{\lambda}{2} \exp(-\lambda|\beta|) d\lambda.$$

As before, we employ a Taylor series to approximate $\exp(-\lambda|\beta|)$, but now around the expectation $m = (U+L)/2$ of the uniform distribution, giving

$$\exp(-\lambda|\beta|) \approx e^{-m|\beta|} \left[1 - |\beta|(\lambda - m) + \frac{1}{2} |\beta|^2 (\lambda - m)^2 \right].$$

Then

$$p(\beta|L, U) \propto_{\text{approx.}} \frac{e^{-m|\beta|}}{U-L} \int_L^U \left[1 - |\beta|(\lambda - m) + \frac{1}{2} |\beta|^2 (\lambda - m)^2 \right] \frac{\lambda}{2} d\lambda. \tag{20}$$

If the constant and the linear terms of the expansion are retained this produces

Figure 4 (A) Gamma prior density of lambda square for $r = \delta = 4$ (dot-dash) and $r = \delta = 16$ (solid). (B) Marginal prior densities of marker effects for $r = \delta = 4$ (dot-dash) and $r = \delta = 16$ (solid). (C) and (D) focus on the middle and right tails of the densities displayed in B, respectively.

$$p_{\text{unif},1}(\beta|L, U) \propto_{\text{approx.}} \frac{1}{U-L} e^{-m|\beta|} \int_L^U [1 - |\beta|(\lambda - m)] \frac{\lambda}{2} d\lambda.$$

Since λ is positive, one can set $L = 0$ and $m = U/2$ yielding

$$p_{\text{unif},1}(\beta_j|L = 0, U) \propto_{\text{approx.}} \frac{e^{-m|\beta|}}{U} \left(\frac{1}{4}U^2 + \frac{|\beta|U^2}{2} \left(\frac{m}{2} - \frac{U}{3} \right) \right) = \frac{U}{4} e^{-U|\beta|/2} \left(1 - |\beta| \frac{U}{6} \right).$$

A plot of $p_{\text{unif},1}(\beta_j|L = 0, U)$ is shown in Figure 5. As U increases, the prior distribution of the marker effect gets increasingly concentrated near 0, reaching a point mass in the limit. This implies that the regression model becomes effectively very simple if U is assigned large values, as most regression coefficients take values close to 0. In theory, this should produce underfitting and out of sample predictions that do not generalize well. It is thus intriguing why Legarra *et al.* (2011) obtained reasonable predictive accuracies when placing a uniform prior on λ , with $L = 0$ and $U = 10^6$. This theoretical

excursion suggests that a big warning should be inserted in documentation of software implementing DE regression models with a flat prior on the regularization parameter λ .

On parameterizations of Bayes L: How any Bayesian or “classical” model is parameterized depends on mechanistic (e.g., interpretation with respect to some theory) or computing considerations, but alternative parameterizations must be equivalent in terms of the inference attained. For example, a parameterization of the classical infinitesimal model (e.g., Hill 2012) in terms of additive genetic and environmental variances (V_A, V_E) must be equivalent to parameterization ($V - V_E, V_E$), where V is the phenotypic variance, or to parameterization ($Vh^2, (1 - h^2)V$), where h^2 is heritability. The second and third parameterizations do not imply causally that the genetic variance depends on the environmental variance or that the environmental variance depends on heritability. In likelihood-based inference there is invariance of parameters under transformation. However, care must be exercised in Bayesian analysis because parameters are random, so any rotation of coordinates (some transformations involve nonlinear

Figure 5 Prior density of a marker effect when a uniform (0, U) prior is adopted for λ at varying values of upper bound U of the uniform distribution: 1 (solid line), 5 (dashes), 10 (dots-dashes).

rotations) require intervention of the Jacobian of the transformation. One can go back and forth between parameterizations, provided that probability volumes are preserved properly. For instance, if one assigns independent priors to h^2 and V in a $(Vh^2, (1 - h^2)V)$ parameterization, those used in a V_A, V_E parameterization should be probabilistically consistent with the preceding, such that samples from the joint posterior of h^2 and V produce the same distribution as that obtained by sampling from the joint posterior of V_A, V_E . Further, conditioning and deconditioning may be necessary due to computing issues, *e. g.*, the Gibbs sampler works with conditional distributions, but the algorithm automates the deconditioning. It is precisely in this context that Legarra *et al.* (2011) misinterpreted the parameterization of Bayes L in Park and Casella (2008), de los Campos *et al.* (2009b), Weigel *et al.* (2009), and Vázquez *et al.* (2010) who, instead of working directly with prior (10) adopted a conditional prior discussed further below. All these authors have applied this parameterization successfully using data from animals and plants.

For reasons related to the behavior of Markov chain Monte Carlo algorithms for Bayes L, Park and Casella (2008) introduced a conditional DE distribution, with density This distribution has mean $E(\beta|\lambda, \sigma_e^2) = 0$ and variance $\text{Var}(\beta|\lambda, \sigma_e^2) = 2(\sigma_e^2/\lambda^2)$; this, of course, is not the variance of β_j . Legarra *et al.* (2011) incorrectly wrote $\text{Var}(\beta) = 2(\sigma_e^2/\lambda^2)$, and made the statement: “we do expect the distribution of SNP effects not to be related to unobservable, unaccounted (residual) effects that can, for example, vary from site to site for the same individuals.” It is fairly obvious that $\text{Var}(\beta|\lambda, \sigma_e^2)$ cannot be $\text{Var}(\beta_j)$ since

$$\text{Var}(\beta|\lambda) = E_{\sigma_e^2}[\text{Var}(\beta|\lambda, \sigma_e^2)] + \text{Var}_{\sigma_e^2}[E(\beta|\lambda, \sigma_e^2)] = E_{\sigma_e^2}\left[2\frac{\sigma_e^2}{\lambda^2}\right],$$

with the term $\text{Var}_{\sigma_e^2}[E(\beta|\lambda, \sigma_e^2)]$ dropping because it is null. Hence, $\text{Var}(\beta|\lambda)$ depends on the prior adopted for σ_e^2 . If σ_e^2 is assigned a scaled inverted chi-square distribution on ν_e

degrees of freedom and with scale S_e^2 , with density as in (29),

$$\begin{aligned} \text{Var}(\beta|\lambda, \nu_e, S_e^2) &= E_{\sigma_e^2}\left(2\frac{\sigma_e^2}{\lambda^2}|\lambda\right) = \frac{2}{\lambda^2} \int_0^\infty \sigma_e^2 p(\sigma_e^2|\nu_e, S_e^2) d\sigma_e^2 \\ &= \frac{2\nu_e S_e^2}{\lambda^2(\nu_e - 2)}, \quad \nu_e > 2. \end{aligned} \quad (21)$$

$$f(\beta/\lambda, \sigma_e^2) = \frac{\lambda}{2\sqrt{\sigma_e^2}} \exp\left(-\frac{\lambda}{2\sqrt{\sigma_e^2}}|\beta|\right).$$

Therefore, the variance of the prior distribution of marker effects does not depend on σ_e^2 but, rather, on λ^2 and on the parameters of the prior distribution of σ_e^2 . There is the additional complication that (21) does not take into account uncertainty associated with λ , and this is examined next.

Since λ must be positive, conditioning on λ is equivalent to conditioning on λ^2 , so that $E(\beta_j) = E_{\lambda^2}E(\beta_j|\lambda^2) = 0$, and

$$\text{Var}(\beta) = E_{\lambda^2}\text{Var}(\beta|\lambda^2) + \text{Var}_{\lambda^2}E(\beta|\lambda^2) = E_{\lambda^2}\text{Var}(\beta|\lambda^2).$$

Hence, unconditionally, use of (21) in $E_{\lambda^2}\text{Var}(\beta|\lambda^2)$ produces

$$\text{Var}(\beta|\nu_e, S_e^2) = \frac{2\nu_e S_e^2}{\nu_e - 2} \int \frac{1}{\lambda^2} p(\lambda^2) d\lambda.$$

If a $\Gamma(r, \delta)$ prior is placed on λ^2 , with density (13)

$$\text{Var}(\beta|\nu_e, S_e^2, r, \delta) = \frac{2\nu_e S_e^2}{\nu_e - 2} \int \frac{1}{\lambda^2} \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{r-1} \exp(-\delta\lambda^2) d\lambda^2.$$

Changing variables to $\theta = 1/\lambda^2$ gives

$$\begin{aligned} \text{Var}(\beta|\nu_e, S_e^2, r, \delta) &= \frac{2\nu_e S_e^2}{\nu_e - 2} \int \theta \frac{\delta^r}{\Gamma(r)} (\theta)^{-r+1} \exp\left(-\frac{\delta}{\theta}\right) \frac{1}{\theta^2} d\theta \\ &= \frac{2\nu_e S_e^2}{\nu_e - 2} \int \theta \frac{\delta^r}{\Gamma(r)} (\theta)^{-r-1} \exp\left(-\frac{\delta}{\theta}\right) d\theta. \end{aligned}$$

The integral is the expected value of a random variable (θ) following an inverted Gamma distribution with parameters r and δ , which is $\delta/(r - 1)$ ($r > 1$), so

$$\text{Var}(\beta|\nu_e, S_e^2, r, \delta) = \frac{2\nu_e S_e^2 \delta}{(\nu_e - 2)(r - 1)}. \quad (22)$$

As argued in Gianola *et al.* (2009), the connection between the variance of the prior distribution of marker effects and additive genetic variance is subtle and elusive. If $\text{Var}(\beta|\nu_e, S_e^2, r, \delta)$ were to be viewed as the variance of an additive effect in some infinitesimal model, how are its different components interpreted? If the standard infinitesimal model is parameterized in terms of (V_E, h^2) one can write

$$V_A = V_{E_1} \frac{h^2}{1-h^2}.$$

In (22) $\nu_e S_e^2 / (\nu_e - 2)$ is the counterpart of V_E , since this is the expected value of the prior distribution assigned to the residual variance, σ_e^2 . Then, $2\delta/(r-1)$ plays the role of $h^2/(1-h^2)$; since $\delta/(r-1)$ is the prior expectation of $1/\lambda^2$, it would turn out that $\lambda^2/2$ would be the counterpart of $(1-h^2)/h^2$.

The statements made in Legarra *et al.* (2011) are misleading due to an incorrect interpretation of the parameterization of Bayes L proposed by Park and Casella (2008), used to address a multimodality problem that seems to arise in nonhierarchical implementations of Bayes L in the sense of Kärkkäinen and Sillanpää (2012). These authors reported that hierarchical and nonhierarchical versions of the Bayesian Lasso led to different posterior inferences, but could not find clear reasons for this discrepancy. It might be related to lack of convergence of the Markov chain Monte Carlo scheme in the nonhierarchical parameterization or perhaps to some impropriety. Additional basic research is needed to explain this paradox, but Kärkkäinen and Sillanpää (2012) recommended the hierarchical implementation, possibly because of easier computation.

Bayes R

Erbe *et al.* (2012) presented this method as follows. Bayes R starts the hierarchical model with (1) and poses a mixture of four zero-mean normal distributions as a conditional prior for a specific SNP effect:

$$\begin{aligned} p(\beta | \sigma_{\beta_1}^2 = 0, \sigma_{\beta_2}^2 = 10^{-4}\sigma_g^2, \sigma_{\beta_3}^2 = 10^{-3}\sigma_g^2, \sigma_{\beta_4}^2 = 10^{-2}\sigma_g^2, \pi_1, \pi_2, \pi_3, \pi_4) \\ = \pi_1 \times 0 + \pi_2 N(\beta | 0, 10^{-4}\sigma_g^2) \\ + \pi_3 N(\beta | 0, 10^{-3}\sigma_g^2) + \pi_4 N(\beta | 0, 10^{-2}\sigma_g^2). \end{aligned} \quad (23)$$

Here, if the SNP effect is generated from the first component of the mixture (with probability π_1) it will be 0 with complete certainty; if drawn from the second component it will have a normal distribution with null mean and variance $\sigma_{\beta_2}^2 = 10^{-4}\sigma_g^2$, and so on. In Bayes R, $\sigma_g^2 = r^2\sigma^2$ is the assumed genetic variance, r^2 is the assumed reliability, and σ^2 is the variance of the target trait. Presumably, the assumption about r^2 is either model derived or based on prior cross-validation information, which is good Bayesian behavior, normatively. Makowsky *et al.* (2011) gave evidence that what one assumes about genetic variance from inference in training data is not recovered in cross-validation.

The mean of the mixture is obviously 0. Since the four components of the mixture have null means, the variance, given $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$, is

$$\text{Var}(\beta | \boldsymbol{\pi}) = (\pi_2 \times 10^{-4} + \pi_3 \times 10^{-3} + \pi_4 \times 10^{-2})\sigma_g^2.$$

Further,

$$\text{Var}(\beta) = E_{\boldsymbol{\pi}}[\text{Var}(\beta | \boldsymbol{\pi})] + \text{Var}_{\boldsymbol{\pi}}[E(\beta | \boldsymbol{\pi})] = E_{\boldsymbol{\pi}}[\text{Var}(\beta | \boldsymbol{\pi})].$$

Erbe *et al.* (2012) used a Dirichlet distribution with parameter vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)'$ as prior for the elements of $\boldsymbol{\pi}$, so that

$$\text{Var}(\beta | \boldsymbol{\alpha}) = E_{\boldsymbol{\pi}}[\text{Var}(\beta | \boldsymbol{\pi})] = \frac{(10^{-4}\alpha_2 + 10^{-3}\alpha_3 + 10^{-2}\alpha_4)}{\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4} \sigma_g^2. \quad (24)$$

In particular, Erbe *et al.* (2012) took $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$, producing a uniform distribution on $\boldsymbol{\pi}$. It follows that all SNPs have the same marginal prior distribution, with null mean, and variance

$$\text{Var}(\beta_j | \boldsymbol{\alpha}) = \frac{r^2\sigma^2}{400} \left(1 + \frac{1}{10} + \frac{1}{100}\right) = \frac{111}{4 \times 10^4} r^2\sigma^2.$$

This suggests that a simple ridge regression-BLUP obtained by solving

$$\left[\mathbf{X}'\mathbf{X} + \frac{\sigma_e^2(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)}{r^2\sigma^2(10^{-4}\alpha_2 + 10^{-3}\alpha_3 + 10^{-2}\alpha_4)} \right] \hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

may deliver predictive abilities that are similar to those of Bayes R, except that it would differ with respect to Bayes R on how marker effects are shrunk.

Insight on how shrinkage takes place in Bayes R is gained by inspecting the joint posterior density of all marker effects, given r^2 , σ^2 , and $\boldsymbol{\pi}$. Here

$$\begin{aligned} p(\beta | \mathbf{y}, \boldsymbol{\pi}, r^2, \sigma^2) &\propto \exp\left(-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma_e^2}\right) \\ &\times \prod_{j=1}^p \left[\pi_1 \times 0 + \pi_2 N(\beta_j | 0, \sigma_2^2) + \pi_3 N(\beta_j | 0, \sigma_3^2) \right. \\ &\left. + \pi_4 N(\beta_j | 0, \sigma_4^2) \right], \end{aligned} \quad (25)$$

where $\sigma_2^2 = r^2\sigma^2 10^{-4}$, $\sigma_3^2 = r^2\sigma^2 10^{-3}$, and $\sigma_4^2 = r^2\sigma^2 10^{-2}$ (these values can be modified *a piacere*). Taking derivatives of the log-posterior with respect to $\boldsymbol{\beta}$ gives (apart from an additive constant)

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} \log[p(\beta | \mathbf{y}, \boldsymbol{\pi}, r^2, \sigma^2)] &= \frac{1}{\sigma_e^2} (\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}) \\ &+ \left\{ \frac{\sum_{i=2}^4 \pi_i \frac{d}{d\beta_j} \phi_i(\beta_j | 0, \sigma_i^2)}{\pi_2 \phi_2(\beta_j | 0, \sigma_2^2) + \pi_3 \phi_3(\beta_j | 0, \sigma_3^2) + \pi_4 \phi_4(\beta_j | 0, \sigma_4^2)} \right\}, \end{aligned} \quad (26)$$

where $\{\cdot\}$ denotes a $p \times 1$ vector. Above, $\phi_i(\beta_j | 0, \sigma_i^2)$ ($i = 2, 3, 4$) is the density of β_j under the normal distribution corresponding to component i of the mixture, with

$$\frac{d}{d\beta_j} \phi_i(\beta_j|0, \sigma_i^2) = -\frac{\phi_i(\beta_j|0, \sigma_i^2)}{\sigma_i^2} \beta_j.$$

Employing the preceding expression in Equation 26 yields

$$\frac{\partial}{\partial \beta} \log[p(\beta|\pi, r^2, \sigma^2)] = \left(\frac{1}{\sigma_e^2}\right)(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\beta) - \left\{ \frac{\sum_{i=2}^4 \pi_i \phi_i(\beta_j|0, \sigma_i^2)}{\sum_{i=2}^4 \pi_i \phi_i(\beta_j|0, \sigma_i^2)} \beta_j \right\}. \quad (27)$$

Setting this to zero and rearranging leads to iteration

$$\beta^{[t+1]} = (\mathbf{X}'\mathbf{X} + \Omega_{\beta^{[t]}})^{-1} \mathbf{X}'\mathbf{y},$$

where $\Omega_{\beta^{[t]}}$ is a $p \times p$ diagonal matrix with typical element

$$\Omega_{jj, \beta}^{[t]} = \sigma_e^2 \frac{\sum_{i=2}^4 \pi_i \phi_i(\beta_j^{[t]}|0, \sigma_i^2) (1/\sigma_i^2)}{\sum_{i=2}^4 \pi_i \phi_i(\beta_j^{[t]}|0, \sigma_i^2)} = \sum_{i=2}^4 \pi_{ij}^{[t]} \frac{\sigma_e^2}{\sigma_i^2}, \quad (28)$$

where

$$\pi_{ij}^{[t]}(\beta_j) = \frac{\pi_i \phi_i(\beta_j^{[t]}|0, \sigma_i^2)}{\sum_{i=2}^4 \pi_i \phi_i(\beta_j^{[t]}|0, \sigma_i^2)}, \quad i = 1, 2, \dots, 4 \text{ and } j = 1, 2, \dots, p.$$

This is interpretable as the probability that a value β_j in the course of iteration comes from the i th component of the mixture, as the value of β_j changes iteratively. Note that $\Omega_{jj, \beta}$ is a weighted average of the shrinkage factors σ_e^2/σ_i^2 corresponding to those that would be employed if the variance parameter of the i th component of the mixture were to be used in ridge regression-BLUP. If σ_i^2 is taken as constant over the three “slab” components, Bayes R reduces to BLUP. On the other hand, when σ_i^2 varies over components, the ratio σ_e^2/σ_i^2 will be larger for components having the smallest variance. Observe that π_1 does not play a role in this posterior mode interpretation of how Bayes R effects shrinkage.

In summary, Bayes R assigns the same prior distribution to all markers in the battery of SNPs, one with null mean and variance (for a mixture of K components)

$$\text{Var}(\beta|\alpha) = \sum_{k=1}^K \frac{\alpha_k}{\sum_{k=1}^K \alpha_k} \sigma_k^2,$$

where the α 's are the parameters of the prior distribution of the mixing probabilities π . Bayes R takes $\sigma_1^2 = 0$.

The superior performance of Bayes R over other methods found by Erbe *et al.* (2012) probably results from using prior empirical knowledge about r^2 , the assumed reliability. Bayes R has been extended to Bayes RS (Brondum *et al.* 2012). This is a minor variant of Bayes R in which the mixture (23) is expanded by a factor S , so that there are now S mixtures of four normal distributions each. The letter S denotes a number of chromosome segments constructed in some

manner that reflects prior knowledge that some such segments contribute more variance than others. Using the arguments outlined above, it is easy to see that Bayes RS leads to a shrinkage that, instead of being component specific, is now region-component specific.

An incorrect prior often used in the Bayesian alphabet

The following statement is found at high frequency in the genomic selection literature: “The prior distribution of the residual variance is $\chi^{-2}(\sigma_e^2|\nu_e = -2, S_e^2 = 0)$, meaning that the degrees of freedom of the prior is -2 and that the scale parameter is null.” Examples are Meuwissen *et al.* (2001) and Jia and Jannink (2012), respectively. Note that Bayes's theorem returns with null posterior density or probability any parameter value that is assigned 0 density or mass *a priori*. If the prior density (or probability) of parameter θ is such that $p(\theta|\text{hyperparameters}) = 0$, it must be that

$$p(\theta|\text{hyperparameters}, \mathbf{y}) = \frac{p(\mathbf{y}|\theta, \text{hyperparameters}) \times 0}{p(\mathbf{y}|\text{hyperparameters})} = 0,$$

as well. The prior $\chi^{-2}(\sigma_e^2|\nu_e = -2, S_e^2 = 0)$ is absurd for two reasons. First, a scaled inverted chi-square distribution exists only if both ν_e and S_e^2 are >0 . To see the second reason, we write the prior density explicitly, that is,

$$p(\sigma_e^2|\nu_e, S_e^2) = \frac{(\nu_e S_e^2/2)^{\frac{\nu_e}{2}}}{\Gamma(\nu_e/2)} \times (\sigma_e^2)^{-\left(\frac{\nu_e}{2}+1\right)} \exp\left(-\frac{\nu_e S_e^2}{2\sigma_e^2}\right), \quad (29)$$

so for $S_e^2 = 0$ and any “legal” value of ν_e , $p(\sigma_e^2|\nu_e, S_e^2 = 0) = 0\nu\sigma_e^2$. Then, it must be that $p(\sigma_e^2|\nu_e, S_e^2, \mathbf{y}) = 0\nu\sigma_e^2$ as well. Hence, a scaled inverted chi-square with a null scale parameter is not a probability model at all, as it does not assign appreciable density to any value of the unknown residual variance. It does not convey uncertainty whatsoever: any value of the residual variance is assigned a density of zero, prior and posterior to observing data.

A possible reason for this mistake is as follows: Sorensen and Gianola (2002), as many other Bayesians often do, write the prior as being proportional to the kernel of the scaled inverted chi-square density, that is, as

$$p(\sigma_e^2|\nu_e, S_e^2) \propto (\sigma_e^2)^{-\left(\frac{\nu_e}{2}+1\right)} \exp\left(-\frac{\nu_e S_e^2}{2\sigma_e^2}\right),$$

and note that this kernel reduces to a uniform distribution by taking $\nu_e = -2$ and $S_e^2 = 0$, yielding $p(\sigma_e^2|\nu_e, S_e^2) \propto 1$. However, it takes more than a kernel to make a density, as multiplication of 1 times the integration constant $(\nu_e S_e^2/2)^{\nu_e/2}/\Gamma(\nu_e/2)$ produces zero.

Discussion

The main message from this article is that it is not clear how one can learn about genetic architecture from data in $n \ll p$

situations. This is because individual marker effects are not estimable from the likelihood, apart from the fact that it is unlikely that a multiple linear regression provides a sensible description of biological complexity. On the other hand, it is feasible to learn about the signal $\mathbf{X}\boldsymbol{\beta}$ because there is information about this unknown vector in the data, although equivalent to that conveyed by a sample of size 1. Unfortunately, the Bayesian alphabet (Gianola *et al.* 2009) continues to grow under the incorrect perception that different specifications stemming from various choices of prior inform about genetic architecture; Erbe *et al.* (2012) and Brondum *et al.* (2012) provide good examples of this. It is difficult to defend such claims unless $n \gg p$ and provided that the model is “true” and effectively sparse (Wimmer *et al.* 2012). Otherwise, the prior always matters whenever $n \ll p$ and different priors lead to different claims about the state of nature, merely because their shrinkage behavior in finite samples varies. All members of the alphabet produce unique point and interval Bayesian estimates of marker effects, but the driver is the prior and not the data.

Mixtures of Gaussian distributions are widely used in nonparametric density estimation (Wasserman 2010) because most distributions can be approximated well. Mixtures can capture vagaries from cryptic distributions but at the expense of parsimony, thus posing the risk of copying noise, as opposed to signal, especially if the mixture model has too many parameters. McLachlan and Peel (2000) give a warning: estimation of the parameters of a mixture (Ψ) on the basis of data are meaningful only if Ψ is likelihood identifiable. In Bayes RS (apart from nuisance effects and the residual variance) the number of unknown parameters is $2p + 4S$. Here, $2p$ comes from the fact that each marker is assigned a distinct variance; the $4S$ comes from the fact that there are S segments each having four segment-specific mixing probabilities $\boldsymbol{\pi}_s (s = 1, 2, \dots, S)$. Unfortunately, $n \ll \ll 2p + 4S$, and this creates a huge identification deficit relative to the information content in a sample of size n . In a Bayesian context, there is the additional issue (occurring even when $n > p$) called label switching, leading Celeux *et al.* (2000) to write: “Although somewhat presumptuous, we consider that almost the entirety of Markov chain Monte Carlo samplers for mixture models has failed to converge!” In view of these pitfalls, one wonders what meaningful mechanistic sense can be extracted from these richly parameterized specifications intended to inform about genetic architecture.

Although their inferential outcomes may be misleading, one should not dismiss the potential value of Bayes B, C, R, RS, or of any of the mixture models proposed so far as prediction machines. Predictive distributions stemming from the various members of the alphabet may be analytically distinct from each other, but such differences are seldom revealed in cross-validation (e.g., Heslot *et al.* 2012); an exception is Lehermeier *et al.* (2013). Below we review how the alphabet can be interpreted from a predictive perspective.

A pioneer of Bayesian predictive inference (Geisser 1993) wrote:

Clearly hypothesis testing and estimation as stressed in almost all statistics books involve parameters...this presumes the truth of the model and imparts an inappropriate existential meaning to an index or parameter...inferring about observables is more pertinent since they can occur and be validated to a degree that is not possible for parameters.

Bayesian methods play an important role in machine learning (e.g., Bishop 2006; Barber 2012; Dehmer and Basak 2012; Rogers and Girolami 2012). A reason is that Bayes’s theorem provides a predictive distribution automatically, something that has not been appreciated in full yet in the whole-genome prediction literature.

The problem of prediction can be cast as one of making statements about future data $\mathbf{y}_f$, given past data $\mathbf{y}$. A model M (e.g., Bayes L) with parameter vector $\boldsymbol{\theta}$ is fitted (trained) to $\mathbf{y}$, leading to the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}, H, M)$, where H denotes hyperparameters. If $\mathbf{y}_f$ is treated as an unknown, the prior becomes $p(\boldsymbol{\theta}, \mathbf{y}_f|H, M) = p(\mathbf{y}_f|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|H, M)$ so that

$$p(\boldsymbol{\theta}, \mathbf{y}_f|\mathbf{y}, H, M) \propto p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{y}_f, M)p(\mathbf{y}_f|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|H, M).$$

Since past observations do not depend on future observations, given the parameters, $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{y}_f, M) = p(\mathbf{y}|\boldsymbol{\theta}, M)$, so that

$$p(\mathbf{y}_f|\mathbf{y}, H, M) \propto \int p(\mathbf{y}_f|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|\mathbf{y}, H, M)d\boldsymbol{\theta}. \quad (30)$$

This is the predictive distribution, where parameters $\boldsymbol{\theta}$ do not necessarily play an “existential role” in the sense of Geisser (1993); rather, they are tools enabling one to go from past to future observations. Note that

$$p(\mathbf{y}_f|\mathbf{y}, H, M) = E_{\boldsymbol{\theta}|\mathbf{y}, H, M} [p(\mathbf{y}_f|\boldsymbol{\theta}, M)],$$

meaning that the predictive distribution weights an infinite number of predictions made at a specific values of $\boldsymbol{\theta}$, with the averaging distributions being $p(\boldsymbol{\theta}|\mathbf{y}, H, M)$; this posterior conveys the plausibility assigned to a specific value of $\boldsymbol{\theta}$, posterior to the observed data $\mathbf{y}$. For example, for ridge regression–BLUP, the posterior distribution of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}|\mathbf{y}$, variances $\sim N(\tilde{\boldsymbol{\beta}}, (\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda)^{-1}\sigma_e^2)$, where $\tilde{\boldsymbol{\beta}}$ is the solution to Equation 5. It follows that the posterior distribution of the signal is $\mathbf{X}\boldsymbol{\beta}|\mathbf{y}$, variances $\sim N(\mathbf{X}\tilde{\boldsymbol{\beta}}, \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda)^{-1}\mathbf{X}'\sigma_e^2)$. This implies that the predictive distribution of a future vector of data $\mathbf{y}_f = \mathbf{X}_f\boldsymbol{\beta} + \mathbf{e}_f$ would also be normal

$$\mathbf{y}_f|\mathbf{y}, \text{variances} \sim N\left(\mathbf{X}_f\tilde{\boldsymbol{\beta}}, \mathbf{X}_f(\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda)^{-1}\mathbf{X}_f'\sigma_e^2 + \mathbf{I}_f\sigma_{e_f}^2\right).$$

Here, the strong assumption that the stochastic process generating current and future data are the same is made;

typically, it is assumed that $\sigma_{e_f}^2 = \sigma_e^2$, but this may not be realistic. While the different priors of the alphabet lead to different predictive distributions, it is to be expected that at least the point predictions will be fairly similar. This is because $\mathbf{X}\boldsymbol{\beta}$ is identified in the likelihood, so some Bayesian learning about the signal will take place, especially when $\mathbf{y}$ is a vector of preprocessed means (e.g., means of daughter yield deviations for a battery of dairy cattle bulls with a large number of progeny records). In the latter case, the various members of the alphabet are expected to differ minimally in predicting ability.

The predictive distribution can be used to check whether observed data are consistent with what a model would lead one to expect. Sorensen and Waagepetersen (2003) used this idea to examine goodness of fit of model for litter size in pigs. However, the predictive approach outlined above does not take uncertainty about the model into account, and this may understate variability seriously. Bayesians address this via model averaging, where the predictive distribution is averaged over models, that is,

$$p(\mathbf{y}_f|\mathbf{y}) = \int p(\mathbf{y}_f|\mathbf{y}, \mathbf{H}_M, \mathbf{M}) d\mu(\mathbf{M}|\mathbf{y}).$$

This integral represents both the situation where the number of models is finite and countable or infinite. In the first case the integral is a sum and the measure $\mu(\mathbf{M}|\mathbf{y})$ is the posterior probability of model $\mathbf{M}$. In the second case the number of possible models may be huge, e.g., in variable selection approaches for linear models aiming to include or exclude p markers, there are 2^p possible specifications. If p is very large, the number of models is practically infinite, so the measure $\mu(\mathbf{M}|\mathbf{y})$ is the posterior density assigned to a specific model.

Although $p(\mathbf{y}_f|\mathbf{y})$ provides a more sensible assessment of predictive uncertainty, in practice one proceeds by constructing cross-validation distributions, with respect to one or several competing models. Each prediction generates an error, and this error will have a cross-validation distribution. The relevance of cross-validation is another important contribution of Meuwissen *et al.* (2001) to whole-genome prediction. Here, hyperparameters of genomic selection models (e.g., π in Bayes C π) can be viewed as “tuning knobs” and evaluated over a grid. Unfortunately, the reality is that Manhattan plots tend to overwhelm cross-validation graphs in genome-wide association studies.

Also, differences in predictive ability are often masked by the variation conveyed by a properly constructed cross-validation distribution (e.g., González-Camacho *et al.* 2012). On the other hand, the various Bayesian predictive machines resulting from different priors may possess differential robustness in finite samples. For instance, some priors may be less sensitive with respect to differences in true genetic architecture (Wimmer *et al.* 2012).

Given that the data do not contain information about individual marker effects, variation in inference is an artifact

caused by the various priors. This leads to the question: How much does one prior differ from another one? Information on this can be obtained by use of some notion of statistical distance between distributions, such as the Kullback–Leibler (KL) metric. For example, Gianola *et al.* (2009) used KL for debunking the notion that marker-specific-effect variances in Bayes A tell us something about genetic variability of chromosomal regions. Recently, Lehermeier *et al.* (2013) used a metric that is easier to interpret than KL, the Hellinger distance or HD (e.g., Roos and Held 2011). They found that Bayesian learning in Bayes A and Bayes B was more limited than with Bayes L or Bayesian ridge regression. In our context, the HD between prior $N(\boldsymbol{\beta}|\mathbf{0}, \sigma_\beta^2)$ assigned to a marker effect in ridge regression and prior $t(\boldsymbol{\beta}|\mathbf{0}, S_\beta^2, \nu)$ of Bayes A is

$$\text{HD}(N, t) = \sqrt{1 - \int \sqrt{N(\boldsymbol{\beta}|\mathbf{0}, \sigma_\beta^2) t(\boldsymbol{\beta}|\mathbf{0}, S_\beta^2, \nu)} d\boldsymbol{\beta}}.$$

HD takes values between 0 and 1, with 1 corresponding to the situation where, say, any realization from $t(\boldsymbol{\beta}|\mathbf{0}, S_\beta^2, \nu)$ is assigned 0 density under $N(\boldsymbol{\beta}|\mathbf{0}, \sigma_\beta^2)$, and vice versa. Similar expressions hold for HD(N , DE), where DE($\boldsymbol{\beta}|\mathbf{0}, \lambda$) is the zero-mean double-exponential distribution with parameter λ that is used in Bayes L and for HD(t , DE). To compare these three priors, we took $\sigma_\beta^2 = 1$, $S_\beta^2\nu/(\nu - 2) = 1$, and $2/\lambda^2 = 1$, so that the three priors had the same variance; for the t -distribution we assigned $\nu = 6$, to produce sufficiently thick tails. With these assignments $S_\beta^2 = \frac{2}{3}$ and $\lambda = \sqrt{2}$, so that, using numerical integration between -10 and 10 , $\text{HD}(N, t) = 0.0690$. Further, $\text{HD}(N, \text{DE}) = 0.122$, and

$$\text{HD}(t, \text{DE}) = \sqrt{1 - \int \sqrt{\frac{\Gamma[3.5] [1 + \beta^2/4]^{-(3.5)} \exp(-\sqrt{2}|\beta|)}{\Gamma[3] \sqrt{4\pi}} \frac{1}{\sqrt{2}}}} d\beta = 0.06.$$

This shows, at least when variances are matched, that these three priors are not too different from each other, so differences in inference would stem from difference in the type and extent of shrinkage effected. However, if priors are not matched, these distances would be expected to increase. Since ridge regression–BLUP, Bayes A, and Bayes L postulate the same sampling model, whenever $n \ll p$ differences in posterior inferences between these three members of the Bayesian alphabet must be due to the fact that the priors are very different and influential.

To conclude, whole-genome prediction can be useful for providing locally valid predictions of complex traits. However, the additive regression models employed therein should not be taken at face value from an inferential perspective unless an additive model with many 0 coefficients turns out to hold as approximately true (oracle property 1 met), and $n \gg p_0$, where p_0 is the number of nonzero coefficients (oracle property 2 met). If these two conditions are (ever) fulfilled, it may be that the genetic

architecture of the very elusive additive QTL (on whose existence the statistical abstraction of marker-assisted inference is based) will be unraveled by statistical means.

The question of the extent to which an additive genetic model is a good representer of complexity is another issue yet to be sorted out. The Bayesian alphabet may expand further on this matter, e.g., Bayes A may grow into Bayes AAA if additive $\times$ additive $\times$ additive epistasis is included in a model. Additional expansions of the Bayesian alphabet to accommodate epistatic interactions will further exacerbate the inferential problems, because of a vast increase in number of regression coefficients. It is far from obvious how genetic architecture of complex traits can be learned via highly dimensional statistical models.

Acknowledgments

A big note of thanks goes to Christos Dadousis, Christina Lehermeier, Valentin Wimmer, and Chris-Carolin Schön (Technische Universität München, TUM, Germany) and William G. Hill (University of Edinburgh) for providing a thorough external review of the manuscript. Eduardo Manfredi (Institut National de la Recherche Agronomique, Toulouse, France) is acknowledged for pointing out the article of Duchemin *et al.* (2012), who detected overparameterization problems of Bayes C π . Heather Adams, Juan Manuel González Camacho, Gota Morota, and Francisco Peñagaricano, all from Wisconsin, and Brad Carlin (Minnesota) are thanked for their comments on an earlier draft of this article. The author is indebted to Chiara Sabatti, the Associate Editor handling the review, and to two anonymous reviewers for their constructive criticism leading to a more succinct manuscript, albeit a much less humorous one than the original submission. Work was partially supported by the Wisconsin Agriculture Experiment Station.

Literature Cited

- Barber, D., 2012 *Bayesian Reasoning and Machine Learning*. Cambridge University Press, Cambridge, UK.
- Bernardo, J. M., and A. F. M. Smith, 1994 *Bayesian Theory*. Wiley, Chichester, UK.
- Bishop, C. M., 2006 *Pattern Recognition and Machine Learning*. Springer, New York.
- Brondum, R. F., G. Su, M. S. Lund, P. J. Bowman, M. E. Goddard *et al.*, 2012 Genome specific priors for genomic prediction. *BMC Genomics* 10.1186/1471-2164-13-543.
- Carlin, B. P., and T. A. Louis, 1996 *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.
- Celeux, G., M. Hurn, and C. Robert, 2000 Computational and inferential difficulties with mixture posterior distributions. *J. Am. Stat. Assoc.* 95: 957–979.
- Crossa, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño *et al.*, 2010 Prediction of genetic value of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186: 713–724.
- Dawid, A. P., 1979 Conditional independence in statistical theory (with discussion). *J. R. Stat. Soc. B* 41: 1–31.
- Dehmer, M., and S. C. Basak, 2012 *Statistical and Machine Learning Approaches for Network Analysis*. Wiley, Hoboken, NJ.
- de los Campos, G., D. Gianola, and G. J. M. Rosa, 2009a Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *J. Anim. Sci.* 87: 1883–1887.
- de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009b Predicting quantitative traits with regression models for dense molecular markers and pedigrees. *Genetics* 182: 375–385.
- de los Campos, G., D. Gianola, and D. B. Allison, 2010 Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Rev. Genet.* 11: 880–886.
- de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus, 2012a Whole genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193: 327–345.
- de los Campos, G., Y. C. Klimentidis, A. I. Vázquez, and D. B. Allison, 2012b Prediction of expected years of life using whole-genome markers. *PLoS ONE* 7: 1–7.
- Duchemin, S. I., C. Colombani, A. Legarra, G. Baloche, H. Larroque *et al.*, 2012 Genomic selection in the French Lacaune dairy sheep breed. *J. Dairy Sci.* 95: 2723–2733.
- Erbe, M., B. J. Hayes, L. K. Matukumali, S. Goswami, P. J. Bowman *et al.*, 2012 Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95: 4114–4129.
- Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*, Ed. 4. Longmans Green, Harlow, UK.
- Fan, J., and R. Li, 2001 Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* 96: 1348–1360.
- Geisser, S., 1993 *Predictive Inference: An Introduction*. Chapman & Hall, New York.
- Gelfand, A. E., and S. K. Sahu, 1999 Identifiability, improper priors, and Gibbs sampling for generalized linear models. *J. Am. Stat. Assoc.* 94: 247–253.
- Gianola, D., and R. L. Fernando, 1986 Bayesian methods in animal breeding theory. *J. Anim. Sci.* 63: 217–244.
- Gianola, D., B. Heringstad, and J. Ødegård, 2006 On the quantitative genetics of mixture characters. *Genetics* 173: 2247–2255.
- Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. L. Fernando, 2009 Additive genetic variability and the Bayesian alphabet. *Genetics* 183: 347–363.
- González-Camacho, J. M., G. de los Campos, P. Pérez, D. Gianola, J. E. Cairns *et al.*, 2012 Genome-enabled prediction of genetic values using radial basis function neural networks. *Theor. Appl. Genet.* 125: 759–771.
- Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*. Available at: <http://www.biomedcentral.com/1471-2105/12/186>
- Hastie, T., R. Tibshirani, and J. Friedman, 2009 *The Elements of Statistical Learning*, Ed. 2. Springer, New York.
- Heffner, E. L., M. E. Sorrells, and J. L. Jannink, 2009 Genomic selection for crop improvement. *Crop Sci.* 49: 1–12.
- Henderson, C. R., 1977 Best linear unbiased prediction of breeding values not in the model for records. *J. Dairy Sci.* 60: 783–787.
- Henderson, C. R., 1984 *Applications of Linear Models in Animal Breeding*. University of Guelph, Ontario, Canada.
- Heslot, N., M. E. Sorrells, J. L. Jannink, and H. P. Yang, 2012 Genomic selection in plant breeding: a comparison of models. *Crop Sci.* 52: 146–160.
- Hill, W. G., 2012 Quantitative genetics in the genomics era. *Curr. Genomics* 13: 196–206.

- Janss, L., G. de los Campos, N. Sheehan, and D. Sorensen, 2012 Inferences from genomic models in stratified populations. *Genetics* 92: 693–704.
- Jia, Y., and J.-L. Jannink, 2012 Multiple trait genomic selection methods increase genetic value prediction accuracy. *Genetics* 192: 1513–1522.
- Kärkkäinen, H. P., and M. K. Sillanpää, 2012 Back to basis for Bayesian model building in genomic selection. *Genetics* 191: 969–987.
- Legarra, A., C. Robert-Granié, P. Croiseau, F. Guillaume, and S. Fritz, 2011 Improved Lasso for genomic selection. *Genet. Res.* 93: 77–87.
- Lehermeier, C., V. Wimmer, T. Albrecht, H. Auinger, D. Gianola *et al.*, 2013 Sensitivity to prior specification in Bayesian genome-based prediction models. *Stat. Appl. Genet. Mol. Biol.* DOI: 10.1515/sagmb-2012-0042.
- Lorenz, A. J., S. Chao, F. G. Asoro, E. L. Heffner, T. Hayashi *et al.*, 2011 Genomic selection in plant breeding: knowledge and prospects. *Adv. Agron.* 110: 77–123.
- Makowsky, R., N. M. Pajewski, Y. C. Klimentidis, A. I. Vázquez, C. W. Duarte *et al.*, 2011 Beyond missing heritability: prediction of complex traits. *PLoS Genet.* 7(4): 10.1371/journal.pgen.100205.
- McLachlan, G., and T. Krishnan, 1997 *The EM Algorithm and Extensions*. Wiley, New York.
- McLachlan, G., and D. Peel, 2000 *Finite Mixture Models*. Wiley, New York.
- Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157: 1819–1829.
- Meuwissen, T. H. E., T. R. Solberg, R. Shepherd, and J. A. Woolliams, 2009 A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value. *Genet. Sel. Evol.* 41: (2) 1–10.
- Mrode, R., 2005 *Linear Models for the Prediction of Animal Breeding Values*. Ed. 2. CABI, Wallingford, UK.
- Mutshinda, C.M., and M. J. Sillanpää, 2010 Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics* 86: 1067–1075.
- Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu *et al.*, 2012 Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genet.* 8 (5): e1002685 10.1371/journal.pgen.1002685.
- O'Hagan, A., 1994 *The Advanced Theory of Statistics: Vol. 2B. Bayesian Inference*. Arnold, Cambridge, UK.
- Park, T., and G. Casella, 2008 The Bayesian Lasso. *J. Am. Stat. Assoc.* 103: 681–686.
- Pérez, P., G. de los Campos, J. Crossa, and D. Gianola, 2010 Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian Linear Regression Package in R. *Plant Genome* 3: 106–116.
- Robertson, A., 1955 Prediction equations in quantitative genetics. *Biometrics* 11: 95–98.
- Robinson, G. K., 1991 That BLUP is a good thing: the estimation of random effects. *Stat. Sci.* 6: 15–32.
- Rogers, S., and M. Girolami, 2012 *A First Course in Machine Learning*. CRC Press, Boca Raton, FL.
- Roos, M., and L. Held, 2011 Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Anal.* 6: 259–278.
- Ruppert, D., M. P. Wand, and R. J. Carroll, 2003 *Semipara metric Regression*. Cambridge University Press, New York.
- Searle, S. R., 1971 *Linear models*. Wiley, New York.
- Searle, S. R., 1966 *Matrix Algebra for the Statistical Sciences*. Wiley, New York.
- Sillanpää, M., 2012 *Bayesian Lasso-Related Methods for Genomic Prediction Methods for Genomic Predictions and QTL Analysis Using SNP Data*, p. 20. Eucarpia, Programme, Information, Abstracts, T4, Hohenheim University, Stuttgart, Germany.
- Sorensen, D., and D. Gianola, 2002 *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer, New York.
- Sorensen, D., and R. Waagepetersen, 2003 Normal linear models with genetically structured residual variance heterogeneity: a case study. *Genet. Res.* 82: 207–222.
- Sun, X., L. Qu, D. J. Garrick, J. C. M. Dekkers, and R. L. Fernando, 2012 A fast EM algorithm for Bayes A-like prediction of genomic breeding values. *PLoS ONE* 7(11): 1–9. e49157 10.1371/journal.pone.0049157.
- Tibshirani, R., 1996 Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. A Stat. Soc.* 58: 267–288.
- Van Raden, P. M., 2008 Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91: 4414–4423.
- Vázquez, A. I., G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola *et al.*, 2010 Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. *J. Dairy Sci.* 93: 5942–5949.
- Vázquez, A. I., G. de los Campos, Y. C. Klimentidis, G. J. M. Rosa, D. Gianola *et al.*, 2012 A comprehensive genetic approach for improving prediction of skin cancer risk in humans. *Genetics* 192: 1493–1502.
- Verbyla, K. L., P. J. Bowman, B. J. Hayes, and M. E. Goddard, 2009 Sensitivity of genomic selection to using different prior distributions. *BMC Proc.* 03/2010; 4 (Suppl) 1:S5. 1–4 (doi: 10.1186/1753-6561-4-S1-S5).
- Wang, C.-L., X.-D. Ding, J.-Y. Wang, J.-F. Liu, W.-X. Fu *et al.*, 2013 Bayesian methods for estimating GEBVs of threshold traits. *Heredity* 110: 213–219.
- Wasserman, L., 2010 *All of Nonparametric Statistics*. Springer, New York.
- Weigel, K. A., G. de los Campos, O. González-Recio, H. Naya, X. L. Wu *et al.*, 2009 Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J. Dairy Sci.* 92: 5248–5257.
- Wellmann, R., and J. Bennewitz, 2012 Bayesian models with dominance effects for genomic evaluation of quantitative traits. *Genet. Res.* 94: 21–37.
- Wimmer, V., T. Albrecht, C. Lehermeier, H.-J. Auinger, Y. Wang *et al.*, 2012 *Eucarpia: Programme, Information, Abstracts*. T7, p. 30. Hohenheim University, Stuttgart, Germany.

Communicating editor: C. Sabatti

Appendix

Bias of BLUP with Respect to Marker Effects

As a toy example, let $n = 3$ and $p = 4$. The model includes an intercept plus the effect of 3 markers, and the incidence matrix is

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & -1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & -1 \end{bmatrix}.$$

The first column contains the dummy variable for the intercept, and the remaining columns are the genotype codes for the markers at each of three loci. The first observation (row 1 of $\mathbf{X}$) pertains to an individual that is Aa (coded as 0), bb (coded as -1), CC (coded as 1), and so on. This matrix has rank 3, and a generalized inverse of $\mathbf{X}'\mathbf{X}$ is

$$(\mathbf{X}'\mathbf{X})^- = \begin{bmatrix} 3 & -4 & 2 & 0 \\ -4 & 6 & -3 & 0 \\ 2 & -3 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

If the true values of the intercept and of the marker effects are denoted as a , b , c , and d , respectively, the expected value of the maximum-likelihood estimator of the four parameters is

$$E(\beta^{(0)}|\beta) = \begin{bmatrix} a \\ b \\ c-d \\ 0 \end{bmatrix},$$

with the expected value of the effect of the third marker being 0 instead of d because of the rank deficiency. Now, we use BLUP with $\mathbf{V}_\beta = \mathbf{I}_3\sigma_\beta^2$ and variance ratio $\lambda = \sigma_e^2/\sigma_\beta^2$ (σ_β^2 is the variance of marker effects) and calculate it (Henderson 1984) as

$$\text{BLUP}(\beta) = (\mathbf{X}'\mathbf{X} + \mathbf{I})^{-1}\mathbf{X}'\mathbf{y}.$$

For this example,

$$(\mathbf{X}'\mathbf{X} + \mathbf{I})^{-1} = \begin{bmatrix} \frac{\lambda^2+6\lambda+6}{k} & -\frac{2\lambda+8}{k} & \frac{2}{k} & -\frac{2}{k} \\ -\frac{2\lambda+8}{k} & \frac{\lambda^2+7\lambda+12}{k} & -\frac{\lambda+3}{k} & \frac{\lambda+3}{k} \\ \frac{2}{k} & -\frac{\lambda+3}{k} & \frac{\lambda^3+7\lambda^2+11\lambda+1}{k\lambda} & \frac{2\lambda^2+9\lambda+1}{k\lambda} \\ -\frac{2}{k} & \frac{\lambda+3}{k} & \frac{2\lambda^2+9\lambda+1}{k\lambda} & \frac{\lambda^3+7\lambda^2+11\lambda+1}{k\lambda} \end{bmatrix},$$

where $k = \lambda^3 + 9\lambda^2 + 20\lambda + 2$. Then

$$E(\text{BLUP}(\beta)|\beta) = (\mathbf{X}'\mathbf{X} + \mathbf{I})^{-1}\mathbf{X}'\mathbf{X} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}.$$

After tedious algebra, one arrives at

$$E(\text{BLUP}(\beta)|\beta) = \begin{bmatrix} a\left(\frac{3}{k}q_4 - \frac{2}{k}q_6\right) + b\left(\frac{4}{k\lambda} + \frac{2}{k}q_4 - \frac{2}{k}q_6\right) - c\left(\frac{1}{k}q_6 - \frac{8}{k\lambda}\right) + d\left(\frac{1}{k}q_6 - \frac{8}{k\lambda}\right) \\ a\left(\frac{2}{k}c_2 - \frac{3}{k}q_6\right) - b\left(\frac{2}{k}q_6 - \frac{2}{k}q_2 + \frac{2}{k\lambda}q_5\right) + c\left(\frac{1}{k}q_2 - \frac{4}{k\lambda}q_5\right) - d\left(\frac{1}{k}q_2 - \frac{4}{k\lambda}q_5\right) \\ a\left(\frac{6}{k} - \frac{2}{k}q_5\right) + b\left(\frac{4}{k} - \frac{2}{k}q_5 - \frac{1}{k\lambda}q_3 + \frac{1}{k\lambda}q_1\right) - c\left(\frac{1}{k}q_5 + \frac{2}{k\lambda}q_3 - \frac{2}{k\lambda}q_1\right) + d\left(\frac{1}{k}q_5 + \frac{2}{k\lambda}q_3 - \frac{2}{k\lambda}q_1\right) \\ -a\left(\frac{6}{k} - \frac{2}{k}q_5\right) - b\left(\frac{4}{k} - \frac{2}{k}q_5 - \frac{1}{k\lambda}q_3 + \frac{1}{k\lambda}c_1\right) + c\left(\frac{1}{k}q_5 + \frac{2}{k\lambda}q_3 - \frac{2}{k\lambda}q_1\right) - d\left(\frac{1}{k}q_5 + \frac{2}{k\lambda}q_3 - \frac{2}{k\lambda}q_1\right) \end{bmatrix},$$

where

$$\begin{aligned} q_1 &= \lambda^3 + 7\lambda^2 + 11\lambda + 1, \\ q_2 &= \lambda^2 + 7\lambda + 12, \\ q_3 &= 2\lambda^2 + 9\lambda + 1, \\ q_4 &= \lambda^2 + 6\lambda + 6, \\ q_5 &= \lambda + 3, \\ q_6 &= 2\lambda + 8. \end{aligned}$$

Conditionally on β , all marker effects are estimated with a bias that involves all other markers (and the intercept as well). Since inferences on genetic architecture are primarily based on point estimates (it should be noted that the biased estimator is more precise), it is quite clear that such inferences are not “clean.”

Marker Effects Are Not Identified from a Bayesian Perspective in the $n < p$ Setting

Let a Bayesian linear model consist of location parameters θ_A and θ_B (this partition has a different meaning from the one given above), with likelihood $p(\mathbf{y}|\theta_A, \theta_B)$. If the conditional posterior density of θ_B is such that

$$p(\theta_B|\theta_A, \mathbf{y}) = p(\theta_B|\theta_A),$$

then θ_B is not identifiable, meaning that observation of data does not increase knowledge about θ_B beyond what is conveyed by the conditional prior $p(\theta_B|\theta_A)$ (Dawid 1979; Gelfand and Sahu 1999). For the model in (1), in the $n < p$ situation matrix $\mathbf{X}_{n \times p}$ has rank n , and one can reorganize its columns into

$$\mathbf{X}\beta = [\mathbf{X}_1 \quad \mathbf{X}_2] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix},$$

where $\mathbf{X}_1$ is $n \times n$ with rank n , and $\mathbf{X}_2$ is $n \times (p - n)$, with the vector of marker effects β partitioned accordingly. Changing variables as

$$\begin{bmatrix} \theta_A \\ \theta_B \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \\ 0 & \mathbf{I}_{(p-n) \times (p-n)} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

produces the inverse transformations $\beta_2 = \theta_B$ and $\beta_1 = \mathbf{X}_1^{-1}(\theta_A - \mathbf{X}_2\theta_B)$; because the transformation is linear, the Jacobian does not involve the parameters. Using the new parameterization model (1) can now be written as

$$\mathbf{y} = \theta_A + \mathbf{e},$$

implying that the data contain information about θ_A but not about θ_B (the latter can represent any marker effect, by construction). Then, irrespective of the joint prior distribution assigned to θ_A and θ_B , the posterior is

$$p(\theta_A, \theta_B|\mathbf{y}) \propto p(\mathbf{y}|\theta_A, \theta_B)p(\theta_B|\theta_A)p(\theta_A) \propto p(\mathbf{y}|\theta_A)p(\theta_B|\theta_A)p(\theta_A),$$

so

$$p(\theta_B|\theta_A, \mathbf{y}) = p(\theta_B|\theta_A),$$

verifying that the $p - n$ marker effects are not likelihood identified. As pointed out by Gelfand and Sahu (1999), this does not mean that there is not Bayesian learning about θ_B . It means, however, that data “speak” about θ_A and that what can be said about θ_B depends on what has been spoken about θ_A , with the pipe lining of knowledge done through the prior distribution. This can be seen more clearly by writing the posterior of θ_B as

$$p(\theta_B|\mathbf{y}) = \int p(\theta_B|\theta_A, \mathbf{y})p(\theta_A|\mathbf{y})d\theta_A = \int p(\theta_B|\theta_A)p(\theta_A|\mathbf{y})d\theta_A \\ = E_{p(\theta_A|\mathbf{y})}[p(\theta_B|\theta_A)].$$

This representation enables one to see that marginal inferences about individual marker effects are the weighted average of an infinite number of inferences made from the conditional prior $p(\theta_B|\theta_A)$, where the averaging distribution is the posterior of the signal $p(\theta_A|\mathbf{y})$. If θ_B is any marker effect, say β_j , the preceding becomes

$$p(\beta_j|\mathbf{y}) = \int [p(\beta_j|\mathbf{X}_1\beta_1)]p(\mathbf{X}_1\beta_1|\mathbf{y})d(\mathbf{X}_1\beta_1).$$

In conclusion, for any letter of the alphabet and for any prior distribution adopted, any inference made about genetic architecture always depends on the form of $p(\beta_j|\mathbf{X}_1\beta_1)$ or, more generally, of $p(\theta_B|\theta_A)$, and these densities depend on the prior adopted, but not on the data. Proper Bayesian learning takes place for $\mathbf{X}_1\beta_1$ only.

Inferences in a Linear Model with Unidentified Parameters

In the context of model (1), the likelihood function (assuming known σ_e^2) is

$$l(\beta|\mathbf{y}, \sigma_e^2) \propto \exp \left[-\frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma_e^2} \right].$$

For the $n < p$ situation, and with $\beta^{(0)}$ being a solution to the normal equations corresponding to generalized inverse $(\mathbf{X}'\mathbf{X})^-$, the (singular) likelihood is expressible as

$$l(\beta|\mathbf{y}, \sigma_e^2) \propto \exp \left(-\frac{(\beta - \beta^{(0)})'\mathbf{X}'\mathbf{X}(\beta - \beta^{(0)})}{2\sigma_e^2} \right).$$

Letting $r = \text{rank}(\mathbf{X})$ and using results from linear model theory (if $n \ll p$, then $r \leq n$), it follows that

$$\mathbf{X}_{n \times p} \beta_{p \times 1} = (\mathbf{X}_{n \times p} \mathbf{Q}_{1, p \times r})(\mathbf{L}_{r \times p} \beta_{p \times 1}) + (\mathbf{X}_{n \times p} \mathbf{Q}_{2, p \times (p-r)})(\mathbf{H}_{(p-r) \times p} \beta_{p \times 1}) = \mathbf{K}_1 \alpha_1 + \mathbf{K}_2 \alpha_2,$$

where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are partitions of a $p \times p$ matrix of rank-preserving elementary operators (Searle 1966); $\alpha_1 = \mathbf{L}\beta$ is an $r \times 1$ vector of likelihood-identified estimable functions and $\alpha_2 = \mathbf{H}\beta$ is a $(p - r) \times 1$ vector of pseudoparameters; $\mathbf{K}_1 = \mathbf{X}\mathbf{Q}_1$ and $\mathbf{K}_2 = \mathbf{X}\mathbf{Q}_2$ are incidence matrices, with $\mathbf{K}_2 = 0$ (α_2 is a pseudoparameter, because it is effectively wiped out of the model). The genetic signal is given by $\mathbf{K}_1\alpha_1$ but we include α_2 as well, to see what Bayesian inference does for something on which the data lack information.

If β is assigned the normal prior $N(\beta|0, \mathbf{V}_\beta)$,

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} | \mathbf{V}_\beta \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{L}\mathbf{V}_\beta\mathbf{L}' & \mathbf{L}\mathbf{V}_\beta\mathbf{H}' \\ \mathbf{H}\mathbf{V}_\beta\mathbf{L}' & \mathbf{H}\mathbf{V}_\beta\mathbf{H}' \end{bmatrix} \right). \quad (\text{A.1})$$

The model is now $\mathbf{y} = \mathbf{K}_1\alpha_1 + \mathbf{K}_2\alpha_2 + \mathbf{e}$, and the likelihood under the new parameterization becomes

$$l(\alpha_1, \alpha_2|\mathbf{y}, \sigma_e^2) \propto \exp \left(-\frac{\begin{bmatrix} (\alpha_1 - \alpha_1^{(0)})' & (\alpha_2 - \alpha_2^{(0)})' \end{bmatrix} \begin{bmatrix} \mathbf{K}_1'\mathbf{K}_1 & \mathbf{K}_1'\mathbf{K}_2 \\ \mathbf{K}_2'\mathbf{K}_1 & \mathbf{K}_2'\mathbf{K}_2 \end{bmatrix} \begin{bmatrix} \alpha_1 - \alpha_1^{(0)} \\ \alpha_2 - \alpha_2^{(0)} \end{bmatrix}}{2\sigma_e^2} \right) \quad (\text{A.2})$$

$$= \exp\left(-\frac{(\alpha_1 - \alpha_1^{(0)})' \mathbf{K}_1' \mathbf{K}_1 (\alpha_1 - \alpha_1^{(0)})}{2\sigma_e^2}\right). \quad (\text{A.3})$$

Expression (A.3) indicates that at most, r parameters are likelihood identified, but (A.2) is retained to illustrate what the prior does. It is well known (e.g., Gianola and Fernando 1986; Sorensen and Gianola 2002) that combining (A.1) with (A.2) leads to the posterior distribution

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} | \mathbf{y}, \mathbf{V}_\beta, \sigma_e^2 \sim N\left(\begin{bmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \end{bmatrix}, \begin{bmatrix} \mathbf{K}_1' \mathbf{K}_1 + \mathbf{V}^{11} & \mathbf{K}_1' \mathbf{K}_2 + \mathbf{V}^{12} \\ \mathbf{K}_2' \mathbf{K}_1 + \mathbf{V}^{21} & \mathbf{K}_2' \mathbf{K}_2 + \mathbf{V}^{22} \end{bmatrix}^{-1} \sigma_e^2\right), \quad (\text{A.4})$$

where

$$\begin{bmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{K}_1' \mathbf{K}_1 + \sigma_e^2 \mathbf{V}^{11} & \mathbf{K}_1' \mathbf{K}_2 + \sigma_e^2 \mathbf{V}^{12} \\ \mathbf{K}_2' \mathbf{K}_1 + \sigma_e^2 \mathbf{V}^{21} & \mathbf{K}_2' \mathbf{K}_2 + \sigma_e^2 \mathbf{V}^{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{K}_1' \mathbf{y} \\ \mathbf{K}_2' \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_1' \mathbf{K}_1 + \sigma_e^2 \mathbf{V}^{11} & \sigma_e^2 \mathbf{V}^{12} \\ \sigma_e^2 \mathbf{V}^{21} & \sigma_e^2 \mathbf{V}^{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{K}_1' \mathbf{y} \\ 0 \end{bmatrix},$$

since $\mathbf{K}_2 = 0$, and where

$$\begin{bmatrix} \mathbf{V}^{11} & \mathbf{V}^{12} \\ \mathbf{V}^{21} & \mathbf{V}^{22} \end{bmatrix} = \begin{bmatrix} \mathbf{L} \mathbf{V}_\beta \mathbf{L}' & \mathbf{L} \mathbf{V}_\beta \mathbf{H}' \\ \mathbf{H} \mathbf{V}_\beta \mathbf{L}' & \mathbf{H} \mathbf{V}_\beta \mathbf{H}' \end{bmatrix}^{-1}.$$

The p -dimensional distribution (A.4) is nonsingular, but it is based on a likelihood that is defined in r dimensions only! Note that the posterior mean satisfies

$$\begin{bmatrix} \mathbf{K}_1' \mathbf{K}_1 + \sigma_e^2 \mathbf{V}^{11} & \sigma_e^2 \mathbf{V}^{12} \\ \sigma_e^2 \mathbf{V}^{21} & \sigma_e^2 \mathbf{V}^{22} \end{bmatrix} \begin{bmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{K}_1' \mathbf{y} \\ 0 \end{bmatrix}. \quad (\text{A.5})$$

The coefficient matrix in (A.5) is the counterpart of $\mathbf{X}'\mathbf{X}$ and it is proportional to the negative matrix of second derivatives of the log-posterior with respect to α_1 and α_2 . This shows that proper Bayesian learning takes place only for α_1 , as the information about α_2 and the co-information about α_1 and α_2 come from the prior only. Further, note the relationship

$$\tilde{\alpha}_2 = -(\mathbf{V}^{22})^{-1} \mathbf{V}^{21} \tilde{\alpha}_1, \quad (\text{A.6})$$

indicating that what is learned about α_2 is solely a function of what is learned about α_1 . This is verified by inserting relationship (A.6) in Equations A.5 above, giving

$$(\mathbf{K}_1' \mathbf{K}_1 + \sigma_e^2 \mathbf{V}^{11} - \sigma_e^2 \mathbf{V}^{12} (\mathbf{V}^{22})^{-1} \mathbf{V}^{21}) \tilde{\alpha}_1 = \mathbf{K}_1' \mathbf{y}.$$

Using properties of inverses of partitioned matrices, $\mathbf{V}_{11}^{-1} = \mathbf{V}^{11} - \mathbf{V}^{12} (\mathbf{V}^{22})^{-1} \mathbf{V}^{21}$, so that

$$\tilde{\alpha}_1 = (\mathbf{K}_1' \mathbf{K}_1 + \sigma_e^2 \mathbf{V}_{11}^{-1})^{-1} \mathbf{K}_1' \mathbf{y}. \quad (\text{A.7})$$

The preceding confirms that the data inform about α_1 but not about α_2 ; what is learned about the latter from phenotypes is done indirectly, through α_1 . Such an “indirect” inference parallels the concept of “prediction of breeding values of individuals without phenotypes” (Henderson 1977). In the molecular markers setting, n linear combinations of markers are learned from the data, but $p - n$ remain at the mercy of the prior. In other words, one does not clearly know what marker effects are being learned from the data, unless the model is parameterized deliberately. This is clearly shown in the first section of the *Appendix*.

An example of Proper Bayesian Learning

To illustrate a case of proper Bayesian learning where a connection with genomic BLUP arises, consider inferring the signal $\mathbf{g} = \mathbf{X}\beta$. This is likelihood identified (estimable) because $E(\mathbf{y} | \mathbf{X}\beta) = \mathbf{g}$, and the likelihood is

$$l(\mathbf{g}|\mathbf{y}) \propto \exp \left[-\frac{(\mathbf{g} - \mathbf{y})'(\mathbf{g} - \mathbf{y})}{2\sigma_e^2} \right],$$

with a maximum at $\hat{\mathbf{g}} = \mathbf{y}$. The information matrix is

$$E_{\mathbf{y}|\mathbf{g}, \sigma_e^2} \left[\frac{\partial^2}{\partial \mathbf{g} \partial \mathbf{g}'} \frac{(\mathbf{g} - \mathbf{y})'(\mathbf{g} - \mathbf{y})}{2\sigma_e^2} \right] = \frac{1}{\sigma_e^2} \mathbf{I}_{n \times n},$$

meaning that, for each individual signal, the information content is proportional to what is conveyed by a sample of size $n = 1$ (if the response variates are means of preprocessed data, the information content will be higher). For the prior $N(\boldsymbol{\beta}|\mathbf{0}, \mathbf{V}_\beta)$, the resulting prior for the signal is $\mathbf{g}|\mathbf{V}_\beta \sim N(\mathbf{0}, \mathbf{XV}_\beta\mathbf{X}')$ and standard results for Bayesian inference give $\mathbf{g}|\mathbf{y}, \mathbf{V}_\beta, \sigma_e^2 \sim N(\tilde{\mathbf{g}}, \mathbf{V}_g)$ as posterior distribution, where

$$\tilde{\mathbf{g}} = \left[\frac{1}{\sigma_e^2} \mathbf{I} + (\mathbf{XV}_\beta\mathbf{X}')^{-1} \right]^{-1} \mathbf{y}.$$

and

$$\mathbf{V}_g = \left[\frac{1}{\sigma_e^2} \mathbf{I} + (\mathbf{XV}_\beta\mathbf{X}')^{-1} \right]^{-1}.$$

A special case is when $\mathbf{V}_\beta = \mathbf{I}_p \sigma_\beta^2$, so that for $\lambda = \sigma_e^2 / \sigma_\beta^2$ being the variance ratio, $\tilde{\mathbf{g}} = [\mathbf{I} + (\mathbf{XX}')^{-1} \lambda]^{-1} \mathbf{y}$, and $\mathbf{V}_g = [\mathbf{I} + (\mathbf{XX}')^{-1} \lambda]^{-1} \sigma_e^2$. Using well-established results known from prediction of random variables dating back to Henderson (1977) but rediscovered recently (e.g., Janss *et al.* 2012) one can easily find the posterior distribution of $\boldsymbol{\beta}$ from that of $\mathbf{g}$, and vice versa. Here, take $\boldsymbol{\alpha}_1 = \mathbf{X}\boldsymbol{\beta}$ so that $\alpha_1 \sim N_n(0, \mathbf{XX}'\sigma_\beta^2)$ and $\tilde{\alpha}_1 = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} = \tilde{\mathbf{g}}$, which is known as genomic BLUP. Any marker effect can be learned indirectly from $\tilde{\mathbf{g}}$ using standard BLUP theory as $\tilde{\boldsymbol{\beta}} = \mathbf{X}'(\mathbf{XX}')^{-1} \tilde{\mathbf{g}}$.

Mode of the Conditional Posterior Distribution in Bayes A

Taking logs of (7) yields

$$L = \log \left[p(\boldsymbol{\beta} | S_\beta^2, \nu, \sigma_e^2, \mathbf{y}) \right] = -\frac{1}{2\sigma_e^2} \sum_{i=1}^n (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 - \left(\frac{1+\nu}{2} \right) \sum_{j=1}^p \log \left[1 + \frac{\beta_j^2}{S_\beta^2 \nu} \right]. \quad (\text{A.8})$$

The gradient vector is

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma_e^2} \sum_{i=1}^n \frac{\partial (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2}{\partial \boldsymbol{\beta}} - \left(\frac{1+\nu}{2} \right) \sum_{j=1}^p \frac{\partial}{\partial \boldsymbol{\beta}} \log \left[1 + \frac{\beta_j^2}{S_\beta^2 \nu} \right] = \frac{1}{\sigma_e^2} \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{x}_i' \boldsymbol{\beta}) - \frac{(1+\nu)}{S_\beta^2 \nu} \\ &\quad \begin{bmatrix} \frac{\beta_1}{1 + \frac{\beta_1^2}{S_\beta^2 \nu}} \\ \frac{\beta_2}{1 + \frac{\beta_2^2}{S_\beta^2 \nu}} \\ \vdots \\ \frac{\beta_p}{1 + \frac{\beta_p^2}{S_\beta^2 \nu}} \end{bmatrix} \\ &= \frac{1}{\sigma_e^2} \mathbf{X}' \mathbf{y} - \frac{1}{\sigma_e^2} (\mathbf{X}' \mathbf{X} + \mathbf{W}_\beta) \boldsymbol{\beta}, \end{aligned} \quad (\text{A.9})$$

where $\mathbf{X}' \mathbf{y} = \{\sum_{i=1}^n \mathbf{x}_i y_i\}$, $\mathbf{X}' \mathbf{X} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$, and

$$\mathbf{W}_\beta = \text{Diag} \left\{ \frac{\sigma_e^2}{S_\beta^2} \frac{(1 + 1/\nu)}{(1 + \beta_j^2/S_\beta^2 \nu)} \right\}.$$

Setting to zero, to satisfy the first-order condition, leads to

$$(\mathbf{X}'\mathbf{X} + \mathbf{W}_\beta)\hat{\beta} = \mathbf{X}'\mathbf{y}.$$

This system is not explicit in β (because marker effects appear nonlinearly in $\mathbf{W}_\beta$) but a functional iteration can be developed to locate stationary points.

Mode of the Conditional Posterior Distribution in Bayes L

As a side note, consider what happens if it is not ignored that $\mathbf{W}_\beta^{-1} = \text{Diag}\{1/|\beta_j|\}$ is a random matrix, contrary to what was done by Tibshirani (1996) in a modal representation of Bayes L. Recalling that $|\beta_j| = \beta_j^2/|\beta_j|$ and that $d|x|/dx = \text{sign}(x)$,

$$\frac{\partial}{\partial \beta_j} |\beta_j| = \frac{\partial}{\partial \beta_j} \left(\frac{\beta_j^2}{|\beta_j|} \right) = \frac{2\beta_j}{|\beta_j|} - \frac{\beta_j^2}{|\beta_j|^2} \text{sign}(\beta_j) = \frac{2\beta_j}{|\beta_j|} - \text{sign}(\beta_j).$$

Differentiating (11) with respect to β

$$\begin{aligned} \frac{\partial L(\beta|\mathbf{y}, \lambda, \sigma_e^2)}{\partial \beta} &= - \frac{\frac{\partial}{\partial \beta} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \sigma_e^2 \lambda \frac{\partial}{\partial \beta} \sum_{j=1}^p |\beta_j|}{2\sigma_e^2} \\ &= - \frac{1}{2\sigma_e^2} \left[-2\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) + 2\sigma_e^2 \lambda \mathbf{W}_\beta^{-1} \beta - \sigma_e^2 \lambda \mathbf{s}_\beta \right], \end{aligned}$$

where $\mathbf{s}_\beta$ is a vector containing the signs of the elements of β . Here, the first-order condition would lead to the iteration

$$(\mathbf{X}'\mathbf{X} + \sigma_e^2 \lambda \mathbf{W}_{\beta^{[t]}}^{-1})\beta^{[t+1]} = \mathbf{X}'\mathbf{y} + \frac{\sigma_e^2 \lambda}{2} \mathbf{s}_{\beta^{[t]}}.$$

Approximation of an Integral in Bayes L

The integral in (16) can be approximated using a second-order expansion around $\lambda^2 = r/\delta$ such that (ignoring the subscript in β_j)

$$\exp \left[-(|\beta| \sqrt{\lambda^2}) \right] \approx e^{-|\beta| \sqrt{r/\delta}} \left[1 - \frac{1}{2} \sqrt{\frac{\delta}{r}} |\beta| \left(\lambda^2 - \frac{r}{\delta} \right) + \frac{\delta}{8r} \left(|\beta|^2 + \sqrt{\frac{\delta}{r}} |\beta| \right) \left(\lambda^2 - \frac{r}{\delta} \right)^2 \right].$$

Use of this in (16) produces

$$\begin{aligned} \int_0^\infty (\lambda^2)^{r+\frac{1}{2}-1} \exp \left[-(|\beta| \sqrt{\lambda^2} + \delta \lambda^2) \right] d\lambda^2 &= \int_0^\infty \exp \left[-|\beta| \sqrt{\lambda^2} \right] (\lambda^2)^{r+\frac{1}{2}-1} \exp(-\delta \lambda^2) d\lambda^2 \\ &\approx e^{-|\beta| \sqrt{r/\delta}} \int_0^\infty (\lambda^2)^{r+\frac{1}{2}-1} \exp(\delta \lambda^2) d\lambda^2 \\ &\quad - e^{-|\beta| \sqrt{r/\delta}} \frac{1}{2} \sqrt{\frac{\delta}{r}} |\beta| \int_0^\infty \left(\lambda^2 - \frac{r}{\delta} \right) (\lambda^2)^{r+\frac{1}{2}-1} \exp(-\delta \lambda^2) d\lambda^2 \\ &\quad + e^{-|\beta| \sqrt{r/\delta}} \frac{\delta}{8r} \left(|\beta|^2 + \sqrt{\frac{\delta}{r}} |\beta| \right) \int_0^\infty \left(\lambda^2 - \frac{r}{\delta} \right)^2 (\lambda^2)^{r+\frac{1}{2}-1} \exp(-\delta \lambda^2) d\lambda^2. \end{aligned}$$

Note that $(\lambda^2)^{r+\frac{1}{2}-1} \exp(-\delta \lambda^2)$ is the kernel of a $\Gamma\left(r + \frac{1}{2}, \delta\right)$ distribution, so that

$$\int_0^\infty (\lambda^2)^{r+\frac{1}{2}-1} \exp(-\delta \lambda^2) d\lambda^2 = \left[\frac{\delta^{r+\frac{1}{2}}}{\Gamma(r+\frac{1}{2})} \right]^{-1},$$

$$\int_0^\infty \left(\lambda^2 - \frac{r}{\delta} \right) (\lambda^2)^{r+\frac{1}{2}-1} \exp(-\delta \lambda^2) d\lambda^2 = \left[\frac{\delta^{r+\frac{1}{2}}}{\Gamma(r+\frac{1}{2})} \right]^{-1} \frac{1}{2\delta},$$

and

$$\int_0^\infty \left(\lambda^2 - \frac{r}{\delta} \right)^2 (\lambda^2)^{r+\frac{1}{2}-1} \exp(-\delta \lambda^2) d\lambda^2 = \left[\frac{\delta^{r+\frac{1}{2}}}{\Gamma(r+\frac{1}{2})} \right]^{-1} \frac{4r+3}{4\delta^2}.$$