DOI: 10.1534/genetics.105.049510

Genomic-Assisted Prediction of Genetic Value With Semiparametric Procedures

Daniel Gianola,*,†,‡,[1] Rohan L. Fernando[§] and Alessandra Stella[†]

*Department of Animal Sciences, University of Wisconsin, Madison, Wisconsin 53706, [†]Parco Tecnologico Padano, 26900 Lodi, Italy, [‡]Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, N-1432 Ås, Norway and [§]Department of Animal Science, Iowa State University, Ames, Iowa 50011

Manuscript received August 14, 2005
Accepted for publication April 18, 2006

ABSTRACT

Semiparametric procedures for prediction of total genetic value for quantitative traits, which make use of phenotypic and genomic data simultaneously, are presented. The methods focus on the treatment of massive information provided by, *e.g.*, single-nucleotide polymorphisms. It is argued that standard parametric methods for quantitative genetic analysis cannot handle the multiplicity of potential interactions arising in models with, *e.g.*, hundreds of thousands of markers, and that most of the assumptions required for an orthogonal decomposition of variance are violated in artificial and natural populations. This makes nonparametric procedures attractive. Kernel regression and reproducing kernel Hilbert spaces regression procedures are embedded into standard mixed-effects linear models, retaining additive genetic effects under multivariate normality for operational reasons. Inferential procedures are presented, and some extensions are suggested. An example is presented, illustrating the potential of the methodology. Implementations can be carried out after modification of standard software developed by animal breeders for likelihood-based or Bayesian analysis.

MASSIVE quantities of genomic data are now available, with potential for enhancing accuracy of prediction of genetic value of, *e.g.*, candidates for selection in animal and plant breeding programs or for molecular classification of disease status in subjects (Golub *et al.* 1999). For instance, Wong *et al.* (2004) reported a genetic variation map of the chicken genome containing 2.8 million single-nucleotide polymorphisms (SNPs) and demonstrated how the information can be used for targeting specific genomic regions. Likewise, Hayes *et al.* (2004) found 2507 putative SNPs in the salmon genome that could be valuable for marker-assisted selection in this species.

The use of molecular markers as aids in genetic selection programs has been discussed extensively. Important early articles are Soller and Beckmann (1982) and Fernando and Grossman (1989), with the latter focusing on best linear unbiased prediction of genetic value when marker information is used. Most of the literature on marker-assisted selection deals with the problem of locating one or few quantitative trait loci (QTL) using flanking markers. However, in the light of current knowledge about genomics, the widely used single-QTL search approach is naive, since there is evidence of abundant QTL affecting complex traits, as discussed, *e.g.*, by Dekkers and Hospital (2002). This would support the infinitesimal model of Fisher (1918) as a sensible statistical specification for many quantitative traits, with complications being the accommodation of nonadditivity and of feedbacks (Gianola and Sorensen 2004). Dekkers and Hospital (2002) observe that existing statistical methods for marker-assisted selection do not deal well with complexity posed by quantitative traits. Some difficulties are: specification of "statistical significance" thresholds for multiple testing, strong dependence of inferences on model chosen (*e.g.*, number of QTL fitted, distributional forms), inadequate handling of nonadditivity, and ambiguous interpretation of effects in multiple-marker analysis, due to collinearity.

Here, we discuss how large-scale molecular information, such as that conveyed by SNPs, can be employed for marker-assisted prediction of genetic value for quantitative traits in the sense of, *e.g.*, Meuwissen *et al.* (2001), Gianola *et al.* (2003), and Xu (2003). The focus is on inference of genetic value, rather than detection of quantitative trait loci. A main challenge is that of positing a functional form relating phenotypes to SNP genotypes (viewed as thousands of possibly highly colinear covariates), to polygenic additive genetic values, and to other nuisance effects, such as sex or age of an individual, simultaneously.

Standard quantitative genetics theory gives a mechanistic basis to the mixed-effects linear model, treated either from classical (Sorensen and Kennedy 1983; Henderson 1984) or from Bayesian (Gianola and

[1]*Corresponding author:* Department of Animal Sciences, 1675 Observatory Dr., Madison, WI 53706. E-mail: gianola@ansci.wisc.edu

Fernando 1986) perspectives. Meuwissen *et al.* (2001) and Gianola *et al.* (2003) exploit this connection and suggest highly parametric structures for modeling relationships between phenotypes and effects of hundreds or thousands of molecular markers. A first concern is the strength of their assumptions (*e.g.*, linearity, multivariate normality, proportion of segregating loci, spatial within-chromosome effects); it is unknown if their procedures are robust. Second, colinearity between SNP or marker genotypes is bound to exist, because of the sheer massiveness of molecular data plus cosegregation of alleles. While adverse effects of colinearity can be tempered when marker effects are treated as random variables, statistical redundancy is undesirable (Lindley and Smith 1972).

The genome seems to be much more highly interactive than what standard quantitative genetic models can accommodate (*e.g.*, D'Haeseleer *et al.* 2000). In theory, genetic variance can be partitioned into orthogonal additive, dominance, additive × additive, additive × dominance, dominance × dominance, etc., components, only under highly idealized conditions. These include linkage equilibrium, absence of natural or artificial selection, and no inbreeding or assortative mating (Cockerham 1954; Kempthorne 1954). Arguably, these conditions are violated in nature and in breeding programs. Actually, marker-assisted selection exploits existence of linkage disequilibrium, and even chance creates disequilibrium. Further, estimation of nonadditive components of variance is notoriously difficult, even under standard assumptions (Chang 1988). Therefore, it is doubtful whether or not standard quantitative genetic approaches can model fine-structure relationships between genotypes and phenotypes adequately, unless either departures from assumptions have mild effects or statistical constructs based on multivariate normality turn out to be more robust than what is expected on theoretical grounds. These considerations suggest that a nonparametric treatment of the data could be valuable.

On the other hand, application of the additive genetic model in selective breeding of livestock has produced remarkable dividends, as shown in Dekkers and Hospital (2002). Hence, a combination of nonparametric modeling of effects of molecular variables (*e.g.*, SNPs) with features of the additive polygenic mode of inheritance is appealing.

Our objective is to present semiparametric methods for prediction of genetic value for complex traits that make use of phenotypic and genomic data simultaneously. This article is organized as follows. kernel regression on snp markers introduces nonparametric regression, kernel functions, and smoothing parameters and proposes a nonparametric approximation to additive genetic value. Next, semiparametric kernel mixed model combines features of kernel regression with the mixed-effects linear model and describes classical and Bayesian implementations. reproducing kernel hilbert spaces mixed model uses established calculus of variations results and hybridizes the mixed-effects linear model with a regression on kernel basis functions. Estimation procedures are presented, the problem of incomplete genotyping is addressed, and a simulated example is given, to illustrate feasibility and potential. The article concludes with a discussion and with suggestions for additional research.

KERNEL REGRESSION ON SNP MARKERS

The regression function: Consider a stylized situation in which each of a series of individuals possesses a measurement for some quantitative trait denoted as y, as well as information on a possibly massive number of genomic variables, such as SNP "genotypes," represented by a vector $\mathbf{x}$. In the main, $\mathbf{x}$ is treated as a continuously valued vector of covariates, even though SNP genotypes are discrete (coding is done via dummy variates). Also, $\mathbf{x}$ could represent gene expression measurements from microarray experiments; here, it would be legitimate to regard this vector as continuous. Although gene expression measurements are typically regarded as response variables, there are contexts in which this type of information could be used in an explanatory role (Mallick *et al.* 2005).

Let the relationship between y and $\mathbf{x}$ be represented as

$$y_i = g(\mathbf{x}_i) + e_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where y_i is a measurement, such as plant height or body weight, taken on individual i, $\mathbf{x}_i$ is a $p \times 1$ vector of dummy SNP or microsatellite covariates observed on i, and $g(.)$ is some unknown function relating these genotypes to phenotypes. Define $g(\mathbf{x}_i) = E(y_i|\mathbf{x}_i)$ as the conditional expectation function, that is, the mean phenotypic value of an infinite number of individuals, all possessing the p-dimensional genotype $\mathbf{x}_i$. $e_i \sim (0, \sigma^2)$ is a random residual, distributed independently of $\mathbf{x}_i$ and with variance σ^2.

The conditional expectation function is

$$g(\mathbf{x}) = \frac{\int y p(\mathbf{x}, y)\, dy}{p(\mathbf{x})}. \tag{2}$$

Following Silverman (1986), consider a nonparametric kernel estimator of the p-dimensional density of the covariates,

$$\hat{p}(\mathbf{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right), \tag{3}$$

where $K((\mathbf{x}_i - \mathbf{x})/h)$ is a kernel function and h is a window width or smoothing parameter. In (3), $\mathbf{x}$ is the value ("focal point") at which the density is evaluated and $\mathbf{x}_i$ ($i = 1, 2, \ldots, n$) is the observed p-dimensional SNP genotype of individual i in the sample. Hence, (3)

estimates population densities (or frequencies). If $\hat{p}(\mathbf{x})$ is to behave as a multivariate probability density function, then it must be true that the kernel function is positive and that the condition

$$\int_{-\infty}^{\infty} \hat{p}(\mathbf{x})\mathbf{dx} = \frac{1}{nh^p}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big)\mathbf{dx} = 1$$

is satisfied. This implies that

$$\int_{-\infty}^{\infty} \frac{1}{h^p}K\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big)\mathbf{dx} = 1.$$

Similarly, and assuming that a single h parameter suffices, one can estimate the joint density of phenotype and genotypes at point $(y, \mathbf{x})$ as

$$\hat{p}(\mathbf{x}, y) = \frac{1}{nh^{p+1}}\sum_{i=1}^{n} K\Big(\frac{y_i - y}{h}\Big)K\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big),$$

where $K((y_i - y)/h)$ is also a kernel function; again, y_i is the observed sample value of variable y in individual i. The numerator of (2) can be estimated as

$$\begin{aligned}\int y\hat{p}(\mathbf{x}, y)\,dy &= \int y\frac{1}{nh^{p+1}}\sum_{i=1}^{n} K\Big(\frac{y_i - y}{h}\Big)K\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big)\,dy \\ &= \frac{1}{nh^p}\sum_{i=1}^{n}\Big[\frac{1}{h}\int yK\Big(\frac{y_i - y}{h}\Big)\,dy\Big]K\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big). \quad (4)\end{aligned}$$

In (4), let $z = (y - y_i)/h$, so that $dy = hdz$ and

$$\frac{1}{h}\int yK\Big(\frac{y_i - y}{h}\Big)\,dy = y_i\int K(z)\,dz + hE(z).$$

The kernel function is typically a proper probability density function chosen such that $\int K(z)\,dz = 1$ and $E(z) = \int zK(z)\,dz = 0$. If so, the preceding expression is equal to y_i, so that (4) becomes

$$\int y\hat{p}(\mathbf{x}, y)\,dy = \frac{1}{nh^p}\sum_{i=1}^{n} y_i K\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big). \quad (5)$$

Returning to (2), one can form the nonparametric estimator

$$\hat{E}(y\,|\,\mathbf{x}) = \hat{g}(\mathbf{x}) = \frac{\int y\hat{p}(\mathbf{x}, y)\,dy}{\hat{p}(\mathbf{x})},$$

which, upon replacing the numerator and denominator by (5) and (3), respectively, takes the form

$$\hat{g}(\mathbf{x}) = \sum_{i=1}^{n} w_i(\mathbf{x})y_i, \quad (6)$$

where

$$w_i(\mathbf{x}) = \frac{K((\mathbf{x}_i - \mathbf{x})/h)}{\sum_{i=1}^{n} K((\mathbf{x}_i - \mathbf{x})/h)}$$

is a weight that depends on the kernel function and window width chosen and on the $\mathbf{x}_i$ (*i.e.*, genotypes) observed in the sample. The linear combination of the observations (6) is called the Nadaraya–Watson estimator of the regression function (Nadaraya 1964; Watson 1964). As seen in (6), the fitted value at coordinate $\mathbf{x}$ is a weighted average of all data points, with the value of the weight depending on the "proximity" of $\mathbf{x}_i$ to $\mathbf{x}$ and on the value of the smoothing parameter h. For instance, if the kernel function has the Gaussian form

$$K\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big) = \frac{1}{(2\pi)^{p/2}}\exp\Big[-\frac{1}{2}\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big)'\Big(\frac{\mathbf{x}_i - \mathbf{x}}{h}\Big)\Big],$$

this has a maximum value of $(2\pi)^{-(p/2)}$ when $\mathbf{x} = \mathbf{x}_i$ and tails off to 0 as the distance between $\mathbf{x}$ and $\mathbf{x}_i$ increases. The values of $w_i(\mathbf{x})$ decrease more abruptly as $h \to 0$. This is the reason why this type of estimator is called "local," in the sense that observations with $\mathbf{x}_i$ coordinates closer to the focal point $\mathbf{x}$ are weighted more strongly in the computation of the fitted value $\hat{E}(y\,|\,\mathbf{x})$.

A specification that is more restrictive than (1) is the additive regression model

$$g(\mathbf{x}_i) = \sum_{j=1}^{p} E(y_i\,|\,x_{ij}) = \sum_{j=1}^{p} g_j(x_{ij}) \quad (7)$$

(Hastie and Tibshirani 1990; Fox 2005), where x_{ij} is the genotype for SNP j in individual i. In this model each of the "partial regression" functions is two-dimensional, thus allowing exploration of effects of individual SNPs on phenotypes, albeit at the expense of ignoring, *e.g.*, epistatic interactions between genotypes. A preliminary examination of relationships can be done via calculation of Nadaraya–Watson estimators of each of the $g_j(.)$ as

$$\hat{g}_j(x) = \frac{\sum_{i=1}^{n} K((x_{ij} - x)/h)y_i}{\sum_{i=1}^{n} K((x_{ij} - x)/h)}, \quad j = 1, 2, \ldots, p,$$

where the kernels are unidimensional. Naturally, one may wish to account for interactions between SNPs, so this type of analysis would be merely exploratory. A specification that is intermediate between (7) and (1) could include sums of single, pairwise, tripletwise, etc., SNP regression functions.

Impact of window width: The scatter plot shown in Figure 1, from Chu and Marron (1991), consists of data on log-income and age of 205 people. The solid line in Figure 1A is a moving weighted average of the points, with weights proportional to the curve at the bottom. Chu and Marron (1991) regard the dip in average income in the middle ages as "unexpected." Figure 1B gives results from three different smooths at window widths of 1, 3, and 9. When $h = 1$, the curve displays considerable sampling variation or roughness. On the other hand, when $h = 9$, local features disappear, because points that are far apart receive considerable

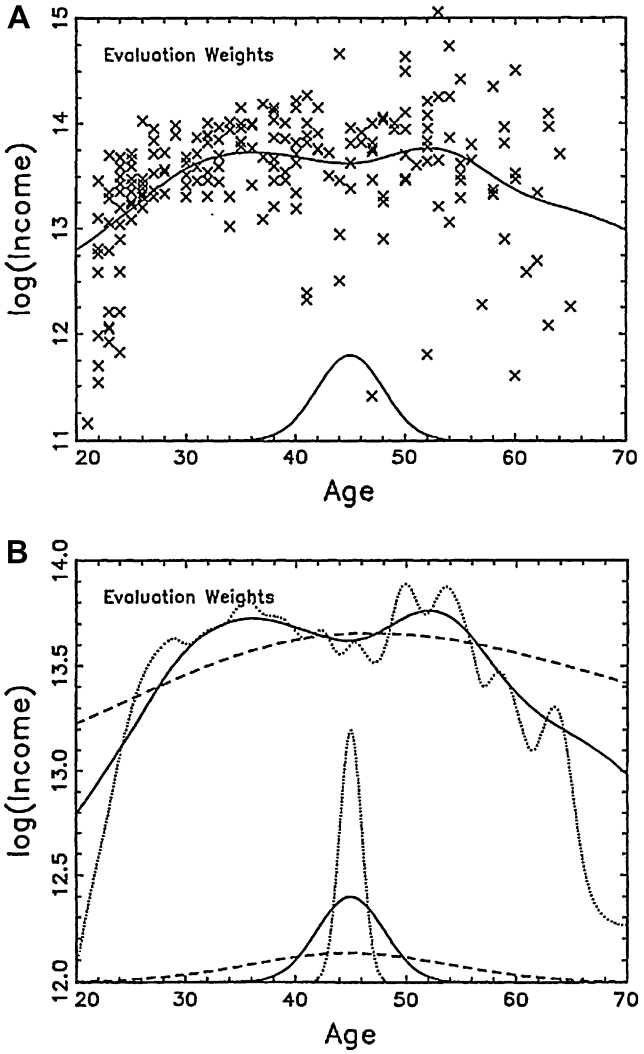

FIGURE 1.—Impact of window width on the regression function of log-income on age (CHU and MARRON 1991). (A) Scatter plot and (B) smooths for earning power data. Kernel is $N(0, 1)$. Window widths are represented by curves: solid curves, $h = 3$; dotted curves, $h = 1$; dashed curves, $h = 9$.

weight in the fitting procedure. If the dip is not an artifact, the oversmoothing results in a "bias." Hence, h must be gauged carefully.

MARRON (1988) and CHU and MARRON (1991) discuss data-driven procedures for assessing h. SILVERMAN (1986) gives a discussion in the context of density estimation, whereas MALLICK *et al.* (2005) consider Hilbert spaces kernel regression, with h treated as an unknown parameter. A conceptually simple and intuitively appealing procedure is cross-validation (*e.g.*, SCHUCANY 2004). For instance, in the "leave-one-out" procedure, the datum for case i, that is $(y_i, \mathbf{x}_i)$, is deleted, and a fit is carried out on the basis of the other $n - 1$ cases. Then, the prediction $\hat{g}_{i,-i}(\mathbf{x}_i \mid h)$ of y_i is formed, where the notation $-i$ indicates that all data other than that for case i are used for estimating the regression function. This process is repeated for all n data points. Subsequently, the cross-validation criterion

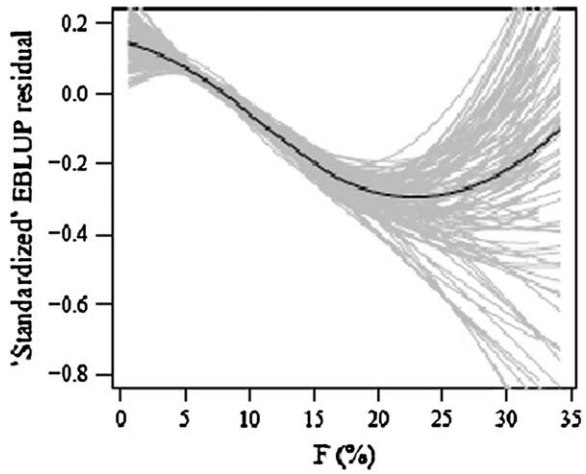

FIGURE 2.—LOESS curves of protein yield deviations (EBLUP) in Jersey cows against their inbreeding coefficients (F, %). The thick curve is the fitted regression (second-degree local polynomial, spanning parameter = 0.9); the other curves are 100 bootstrap replicates.

$$\mathrm{CV}(h) = \frac{\sum_{i=1}^{n} [y_i - \hat{g}_{i,-i}(\mathbf{x}_i \mid h)]^2}{n}$$

is formed. A cross-validation estimate of h is the minimizer of $\mathrm{CV}(h)$; this is found by carrying out the computations over a grid of h-values. HART and LEE (2005), however, present evidence of large variability of the leave-one-out estimates of h. HASTIE *et al.* (2001) discuss alternatives based on leaving out 10–20% of the sample values. RUPPERT *et al.* (2003) present procedures for simple calculation of cross-validation statistics.

D. GULISIJA, D. GIANOLA and K. A. WEIGEL (unpublished results) used another nonparametric procedure, LOESS, to study the relationship between performance and inbreeding in Jersey cows. There is some relationship between LOESS and kernel regression. In LOESS, the number of points contributing to a focal fitted value is fixed (contrary to kernel regression, where the actual number depends on the gentleness of the kernel chosen) and governed by a spanning parameter. This parameter (ranging between 0 and 1) is equivalent to h and dictates the fraction of all data points that contribute toward a fitted value. Figure 2 gives a LOESS fit for protein yield (actually, residuals from a parametric model) and 100 bootstrap replicates, illustrating uncertainty about the regression surface. Without getting into details, note that yield decreases gently at low values of inbreeding, followed by a faster linear decline, and then by an apparent increase. Irrespective of the variability (due to that few animals were either noninbred or highly inbred), neither the change of rate at low inbreeding nor the increase in yield at high consanguinity would be predicted by standard quantitative genetics theory. This is another illustration of how "irregularities" can be discovered nonparametrically, which would remain hidden otherwise.

Estimation of linear approximation: If the kernel function is a probability density function, the nonparametric density estimator (3) will be a density function as well, retaining differentiability properties of the kernel (SILVERMAN 1986). Consider $E(y \mid \mathbf{x}) = g(\mathbf{x})$ and suppose that one is interested in inferring a linear approximation to $g(\mathbf{x})$ near some fixed point $\mathbf{x}^*$ such as the mean value of the covariates; this leads to a nonparametric counterpart of additive genetic value. From a plant and animal breeding point of view, the concept of breeding value is essential in parametric models, so it seems important to develop a nonparametric counterpart as well. The linear function is

$$E^{\text{appr}}(y \mid \mathbf{x}) = g(\mathbf{x}^*) + \dot{\mathbf{g}}(\mathbf{x}^*)'(\mathbf{x} - \mathbf{x}^*), \tag{8}$$

where

$$\dot{\mathbf{g}}(\mathbf{x}^*) = \frac{\partial}{\partial \mathbf{x}} g(\mathbf{x})\bigg|_{\mathbf{x}=\mathbf{x}^*}$$

is the gradient of $g(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}^*$. This suggests the pseudomodel

$$y_i \approx g(\mathbf{x}^*) + \dot{\mathbf{g}}(\mathbf{x}^*)'(\mathbf{x} - \mathbf{x}^*) + e_i, \quad i = 1, 2, \ldots, n \tag{9}$$

and the approximate variance decomposition

$$\text{Var}(y) \approx \dot{\mathbf{g}}(\mathbf{x}^*)'\text{Var}(\mathbf{x})\dot{\mathbf{g}}(\mathbf{x}^*) + \sigma^2. \tag{10}$$

The first term in the expression above can be interpreted as the variance contributed by the linear effect of $\mathbf{x}$ in the neighborhood of $\mathbf{x}^*$. Further, if $\mathbf{x}$ is a vector of genotypes for molecular markers, this would be, roughly, the additive variance "due to" markers.

A nonparametric estimator of (8) is given by

$$\hat{E}^{\text{appr}}(y \mid \mathbf{x}) = \hat{g}(\mathbf{x}^*) + \hat{\dot{\mathbf{g}}}(\mathbf{x}^*)'(\mathbf{x} - \mathbf{x}^*), \tag{11}$$

where

$$\hat{g}(\mathbf{x}^*) = \frac{\sum_{i=1}^n y_i K((\mathbf{x}_i - \mathbf{x}^*)/h)}{\sum_{i=1}^n K((\mathbf{x}_i - \mathbf{x}^*)/h)} = \sum_{i=1}^n w_i(\mathbf{x}^*) y_i, \tag{12}$$

and $\hat{\dot{\mathbf{g}}}(\mathbf{x}^*)$ is an estimate of the vector of first derivatives of $g(\mathbf{x})$ with respect to $\mathbf{x}$, evaluated at $\mathbf{x}^*$. The estimate could be the gradient of (6),

$$\begin{aligned}\hat{\dot{\mathbf{g}}}(\mathbf{x}^*) &= \sum_{i=1}^n y_i \frac{\partial}{\partial \mathbf{x}}\left[\frac{K((\mathbf{x}_i - \mathbf{x})/h)}{\sum_{i=1}^n K((\mathbf{x}_i - \mathbf{x})/h)}\right]\bigg|_{\mathbf{x}=\mathbf{x}^*} \\ &= \sum_{i=1}^n y_i[\mathbf{d}_i(\mathbf{x}^*) - w_i(\mathbf{x}^*)\mathbf{r}(\mathbf{x}^*)],\end{aligned} \tag{13}$$

where

$$\mathbf{d}_i(\mathbf{x}^*) = \frac{\dot{\mathbf{K}}((\mathbf{x}_i - \mathbf{x}^*)/h)}{\sum_{i=1}^n K((\mathbf{x}_i - \mathbf{x}^*)/h)},$$

$$\dot{\mathbf{K}}\left(\frac{\mathbf{x}_i - \mathbf{x}^*}{h}\right) = \frac{\partial}{\partial \mathbf{x}} K\left(\frac{\mathbf{x}_i - \mathbf{x}^*}{h}\right)\bigg|_{\mathbf{x}=\mathbf{x}^*},$$

and

$$\mathbf{r}(\mathbf{x}^*) = \frac{\sum_{i=1}^n \dot{\mathbf{K}}((\mathbf{x}_i - \mathbf{x}^*)/h)}{\sum_{i=1}^n K((\mathbf{x}_i - \mathbf{x}^*)/h)} = \sum_{i=1}^n \mathbf{d}_i(\mathbf{x}^*)$$

are vectors whose form depends on the kernel function used. Hence, (11) becomes

$$\begin{aligned}&\hat{E}^{\text{appr}}(y \mid \mathbf{x}) \\ &\quad = \sum_{i=1}^n w_i(\mathbf{x}^*) y_i + \left\{\sum_{i=1}^n y_i[\mathbf{d}_i(\mathbf{x}^*) - w_i(\mathbf{x}^*)\mathbf{r}(\mathbf{x}^*)]\right\}'(\mathbf{x} - \mathbf{x}^*).\end{aligned} \tag{14}$$

Likewise, an estimator of the variance contributed to (10) by the linear effect of $\mathbf{x}$ (nonparametric additive genetic variance from the SNPs) is given by $\hat{\dot{\mathbf{g}}}'(\mathbf{x}^*) \cdot \widehat{\text{Var}}(\mathbf{x})\hat{\dot{\mathbf{g}}}'(\mathbf{x}^*)$, where $\widehat{\text{Var}}(\mathbf{x})$ is some estimate of the covariance matrix of $\mathbf{x}$. Finally, the relative contribution to variance made by the linear effect of $\mathbf{x}(\tau^2)$ can be assessed as

$$\hat{\tau}^2 = \frac{\hat{\dot{\mathbf{g}}}'(\mathbf{x}^*)\widehat{\text{Var}}(\mathbf{x})\hat{\dot{\mathbf{g}}}'(\mathbf{x}^*)}{\hat{\dot{\mathbf{g}}}'(\mathbf{x}^*)\widehat{\text{Var}}(\mathbf{x})\hat{\dot{\mathbf{g}}}'(\mathbf{x}^*) + \hat{\sigma}^2}, \tag{15}$$

where $\hat{\sigma}^2$ is an estimate of σ^2, such as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \left[y_i - \hat{g}(\mathbf{x}^*) - \hat{\dot{\mathbf{g}}}(\mathbf{x}^*)'(\mathbf{x}_i - \mathbf{x}^*)\right]^2}{n}. \tag{16}$$

Gaussian kernel: A p-dimensional Gaussian kernel with a single band width parameter h has the form

$$K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) = \frac{1}{(2\pi)^{p/2}} \exp\left[-\frac{1}{2h^2}(\mathbf{x}_i - \mathbf{x})'(\mathbf{x}_i - \mathbf{x})\right],$$

so that the estimator of the density at $\mathbf{x}$ is the finite mixture of Gaussians,

$$\begin{aligned}\hat{p}(\mathbf{x}) &= \frac{1}{n}\sum_{i=1}^n \frac{1}{h^p} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) \\ &= \frac{1}{nh^p}\sum_{i=1}^n \frac{1}{(2h^2\pi)^{p/2}} \exp\left[-\frac{S_i(\mathbf{x})}{2h^2}\right],\end{aligned}$$

where, for $\mathbf{x}_i = [x_{i1}\ x_{i2}\ .\ x_{ip}]'$, at focal $\mathbf{x} = [x_1\ x_2\ .\ x_p]'$

$$\frac{S_i(\mathbf{x})}{h^2} = \left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right)'\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) = \frac{\sum_{k=1}^p (x_{ik} - x_k)^2}{h^2}.$$

The weights entering into estimator (6) are then

$$w_i(\mathbf{x}) = \frac{\exp\left[-\left(S_i(\mathbf{x})/2h^2\right)\right]}{\sum_{i=1}^n \exp\left[-\left(S_i(\mathbf{x})/2h^2\right)\right]}.$$

The fitted surface is given by

$$\hat{E}(y \mid \mathbf{x}) = \frac{\sum_{i=1}^n \exp\left[-\left(S_i(\mathbf{x})/2h^2\right)\right] y_i}{\sum_{i=1}^n \exp\left[-\left(S_i(\mathbf{x})/2h^2\right)\right]} = \frac{\mathbf{v}'(\mathbf{x})\mathbf{y}}{\mathbf{v}'(\mathbf{x})\mathbf{1}}, \tag{17}$$

where $\mathbf{y}$ is the $n \times 1$ vector of data points, $\mathbf{1}$ is an $n \times 1$ vector of ones, and $\mathbf{v}'(\mathbf{x})$ is an $n \times 1$ vector with typical element

$$v_i(\mathbf{x}) = \exp\left[-\frac{S_i(\mathbf{x})}{2h^2}\right]; \quad i = 1, 2, \ldots, n.$$

Consider next the linear approximation in (14). Using (17), write

$$\begin{aligned}\hat{E}^{\text{appr}}_{\text{Gauss}}(y\,|\,\mathbf{x}) &= \hat{g}(\mathbf{x}^*) + \hat{\dot{\mathbf{g}}}(\mathbf{x}^*)'(\mathbf{x} - \mathbf{x}^*) \\ &= \frac{\mathbf{v}'(\mathbf{x}^*)\mathbf{y}}{\mathbf{v}'(\mathbf{x}^*)\mathbf{1}} + \hat{\dot{\mathbf{g}}}(\mathbf{x}^*)'(\mathbf{x} - \mathbf{x}^*).\end{aligned}$$

Now, for a Gaussian kernel,

$$\dot{\mathbf{K}}\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) = \frac{\partial}{\partial \mathbf{x}} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right) = \frac{(\mathbf{x}_i - \mathbf{x})}{h^2} K\left(\frac{\mathbf{x}_i - \mathbf{x}}{h}\right).$$

Further,

$$\mathbf{d}_i(\mathbf{x}) = \frac{\dot{\mathbf{K}}((\mathbf{x}_i - \mathbf{x})/h)}{\sum_{i=1}^n K((\mathbf{x}_i - \mathbf{x})/h)} = \frac{(\mathbf{x}_i - \mathbf{x})w_i(\mathbf{x})}{h^2},$$

and

$$\mathbf{r}(\mathbf{x}) = \sum_{i=1}^n \mathbf{d}_i(\mathbf{x}^*) = \sum_{i=1}^n \frac{(\mathbf{x}_i - \mathbf{x})w_i(\mathbf{x})}{h^2}.$$

Hence

$$\begin{aligned}\hat{\dot{\mathbf{g}}}(\mathbf{x}^*) &= \sum_{i=1}^n y_i[\mathbf{d}_i(\mathbf{x}^*) - w_i(\mathbf{x}^*)\mathbf{r}(\mathbf{x}^*)] \\ &= \sum_{i=1}^n y_i\left[\frac{(\mathbf{x}_i - \mathbf{x}^*)w_i(\mathbf{x}^*)}{h^2}\right. \\ &\qquad \left. - w_i(\mathbf{x}^*)\sum_{i=1}^n \frac{(\mathbf{x}_i - \mathbf{x}^*)w_i(\mathbf{x}^*)}{h^2}\right] \\ &= \frac{\sum_{i=1}^n w_i(\mathbf{x}^*)y_i(\mathbf{x}_i - \mathbf{x}^*) - \hat{\dot{\mathbf{g}}}(\mathbf{x}^*)\sum_{i=1}^n (\mathbf{x}_i - \mathbf{x}^*)w_i(\mathbf{x}^*)}{h^2},\end{aligned}$$

and

$$\begin{aligned}\hat{E}^{\text{appr}}(y\,|\,\mathbf{x}) = &\frac{\mathbf{v}'(\mathbf{x}^*)\mathbf{y}}{\mathbf{v}'(\mathbf{x}^*)\mathbf{1}} \\ &+ \frac{\sum_{i=1}^n w_i(\mathbf{x}^*)y_i(\mathbf{x}_i - \mathbf{x}^*) - \hat{\dot{\mathbf{g}}}(\mathbf{x}^*)\sum_{i=1}^n w_i(\mathbf{x}^*)(\mathbf{x}_i - \mathbf{x}^*)}{h^2} \\ &\times (\mathbf{x} - \mathbf{x}^*).\end{aligned}$$

Kernels for discrete covariates: For a biallelic SNP, there are three possible genotypes at each "locus," as in stylized Mendelian situations. In a standard (parametric) analysis of variance representation, incidence situations (or additive and dominance effects at each of the loci) are described via two dummy binary variables per locus, and all corresponding epistatic interactions can be assessed from effects of cross-products of these variables. This leads to a highly parameterized structure and to formidable model selection problems.

Consider now the nonparametric approach. For an $\mathbf{x}$ vector with p coordinates, its statistical distribution is given by the probabilities of each of the 3^p combinations of binary outcomes. With SNPs, p can be very large (possibly much larger than n), so it is hopeless to estimate the probability distribution of genotypes accurately from observed relative frequencies, and smoothing is required (Silverman 1986). Kernel estimation extends as follows: for binary covariates the number of disagreements between a focal $\mathbf{x}$ and the observed $\mathbf{x}_i$ in subject i is given by

$$d(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x}_i - \mathbf{x})'(\mathbf{x}_i - \mathbf{x}),$$

where $d(.)$ takes values between 0 and p. For illustration, suppose that one has genotype *AABbccDd* as focal point and that individual i in the sample has been genotyped as *aaBbccDD*. The vectors of binary covariates for the focal and observed cases (save for an intercept) are given in the matrix

Genotype	*AA*	*Aa*	*aa*	*BB*	*Bb*	*bb*	*CC*	*Cc*	*cc*	*DD*	*Dd*	*dd*
Observed	0	0	1	0	1	0	0	0	1	1	0	0
Focal	1	0	0	0	1	0	0	0	1	0	1	0
Agree	N	Y	N	Y	Y	Y	Y	Y	Y	N	N	Y,

where N (Y) stands for disagreement (agreement). Then, $d(\mathbf{x}, \mathbf{x}_i) = 4$, which is twice the number of disagreements in genotypes because there are only 2 "d.f." per locus. In practice, one should work with a representation of incidences that is free of redundancies. For binary covariates, Silverman (1986) suggests the "binomial" kernel

$$K(\mathbf{x}, \mathbf{x}_i, h) = h^{p - d(\mathbf{x}, \mathbf{x}_i)}(1 - h)^{d(\mathbf{x}, \mathbf{x}_i)},$$

with $\frac{1}{2} \leq h \leq 1$; alternative forms of the kernel function are discussed by Aitchison and Aitken (1976) and Racine and Li (2004). It follows that the kernel estimate of the probability of observing the focal value $\mathbf{x}$ is

$$\hat{p}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n h^{p - d(\mathbf{x}, \mathbf{x}_i)}(1 - h)^{d(\mathbf{x}, \mathbf{x}_i)}. \tag{18}$$

If $h = 1$, the estimate is just the proportion of cases for which $\mathbf{x}_i = \mathbf{x}$; if $h = \frac{1}{2}$, every focal point gets an estimate equal to $(1/2)^p$, irrespective of the observed values $\mathbf{x}_1$, $\mathbf{x}_2, \ldots, \mathbf{x}_n$.

The nonparametric estimator of the regression function is

$$\hat{g}(\mathbf{x}) = \frac{\sum_{i=1}^n h^{p - d(\mathbf{x}, \mathbf{x}_i)}(1 - h)^{d(\mathbf{x}, \mathbf{x}_i)} y_i}{\sum_{i=1}^n h^{p - d(\mathbf{x}, \mathbf{x}_i)}(1 - h)^{d(\mathbf{x}, \mathbf{x}_i)}}.$$

Since a discrete distribution does not possess derivatives, the additive genetic value must be defined in the classical sense (*e.g.*, Falconer 1960), that is, by defining contrasts between expected values of individuals having

appropriate genotypes. Here, one can use either the vectorial representation $g(\mathbf{x})$ or the concept of the additive regression model in (7). Hereinafter, it is assumed that the distribution of $\mathbf{x}$ is continuous, and a continuous kernel function is employed throughout, as an approximation.

SEMIPARAMETRIC KERNEL MIXED MODEL

General considerations: Consider now a situation for which there might be an operational or mechanistic basis for specifying at least part of a model. For instance, suppose that y is a measure on some quantitative trait, such as milk production of a cow. Animal breeders have exploited to advantage the infinitesimal model of quantitative genetics (Fisher 1918). Vectorial representations of this model are given by Quaas and Pollak (1980), and applications to natural populations are discussed by Kruuk (2004). In this section, we combine features of the infinitesimal model with a nonparametric treatment of genomic data and present semiparametric implementations.

Statistical specification: Model (1) is expanded as

$$y_i = \mathbf{w}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u} + g(\mathbf{x}_i) + e_i, \quad i = 1, 2, \ldots, n, \tag{19}$$

where $\boldsymbol{\beta}$ is a vector of nuisance location effects and $\mathbf{u}$ is a $q \times 1$ vector containing additive genetic effects of q individuals (these effects are assumed to be independent of those of the markers), some of which may lack a phenotypic record; $\mathbf{w}_i'$ and $\mathbf{z}_i'$ are known nonstochastic incidence vectors. As before, $g(\mathbf{x}_i)$ is some unknown function of the SNP data. It is assumed that $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$, where σ_u^2 is the "unmarked" additive genetic variance and $\mathbf{A}$ is the additive relationship matrix, whose entries are twice the coefficients of coancestry between individuals. Let $\mathbf{e} = \{e_i\}$ be the $n \times 1$ vector of residuals, and assume that $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, where σ_e^2 is the residual variance. Note that the model implies that

$$y_i - g(\mathbf{x}_i) = \mathbf{w}_i'\boldsymbol{\beta} + \mathbf{z}_i'\mathbf{u} + e_i \tag{20}$$

and

$$y_i - \mathbf{w}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{u} = g(\mathbf{x}_i) + e_i. \tag{21}$$

The preceding means that: (1) the offset $y_i - g(\mathbf{x}_i)$ follows a standard mixed-effects model, and (2) if $\boldsymbol{\beta}$ and $\mathbf{u}$ were known, one could use (6) to estimate $g(\mathbf{x}_i)$ employing $y_i - \mathbf{w}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{u}$ as "observations." This suggests two strategies for analysis of data, discussed below.

Strategy 1—mixed model analysis: This follows from representation (20). First, estimate $g(\mathbf{x}_i)$, for $i = 1, 2, \ldots, n$, via $\hat{g}(\mathbf{x}_i)$, as in (6) or, if a Gaussian kernel is adopted, as in (17). Then, carry out a mixed model analysis using the "corrected" data vector and pseudomodel

$$\mathbf{y}^* = \{y_i - \hat{g}(\mathbf{x}_i)\} = \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where $\mathbf{W} = \{\mathbf{w}_i'\}$ and $\mathbf{Z} = \{\mathbf{z}_i'\}$ are incidence matrices of appropriate order. The pseudomodel ignores uncertainty about $g(\mathbf{x})$, since $\hat{g}(\mathbf{x}_i)$ is treated as if it were the true regression (on SNPs) surface.

Under the standard multivariate normality assumptions of the infinitesimal model, one can estimate the variance components σ_u^2 and σ_e^2 from $\mathbf{y}^*$ via restricted maximum likelihood (REML) (Patterson and Thompson 1971) and form empirical best linear unbiased estimators and predictors of $\boldsymbol{\beta}$ and $\mathbf{u}$, respectively, by solving the Henderson mixed model equations

$$\begin{bmatrix} \mathbf{W}'\mathbf{W} & \mathbf{W}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{W} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}(\sigma_e^2/\sigma_u^2) \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{W}'\mathbf{y}^* \\ \mathbf{Z}'\mathbf{y}^* \end{bmatrix} \tag{22}$$

(Henderson 1973). The ratio σ_e^2/σ_u^2 is evaluated at REML estimates of the variance components. Solving system (22) is a standard problem in animal breeding even for very large q, since $\mathbf{A}^{-1}$ is easy to compute. The two-stage procedure could be iterated several times, *i.e.*, use the solutions to (22), to obtain a new estimate of $g(\mathbf{x})$ using $y_i - \mathbf{w}_i'\hat{\boldsymbol{\beta}} - \mathbf{z}_i'\hat{\mathbf{u}}$ as "data," and then update the pseudodata $\mathbf{y}^*$, etc.

The "total" additive genetic effect of an individual possessing a vector of SNP covariates with focal value $\mathbf{x}$ can be defined as the sum of the additive genetic effect of the SNPs, in the sense of (8), plus the polygenic effect u_i, that is,

$$T_i(\mathbf{x}) = g(\mathbf{x}^*) + \dot{\mathbf{g}}(\mathbf{x}^*)'(\mathbf{x} - \mathbf{x}^*) + u_i.$$

An empirical predictor of T_i can be formed by adding (14) to the ith component of $\hat{\mathbf{u}}$ in the mixed model equations. It is not obvious how a measure of uncertainty about $T_i(\mathbf{x})$ can be constructed using this procedure.

A Bayesian approach can be used instead, using the corrected data $y_i - \hat{g}(\mathbf{x}_i)$ as observations. Under standard assumptions made for the prior and the likelihood (*e.g.*, Wang *et al.* 1993, 1994; Sorensen and Gianola 2002), one can draw samples $j = 1, 2, \ldots, m$ from the pseudoposterior distribution $[\boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_e^2 \,|\, \mathbf{y}^*]$ via Gibbs sampling and then form "semiparametric" draws of the total genetic value as

$$\begin{aligned} T_i^{(j)}(\mathbf{x}) = &\sum_{k=1}^{n} w_k(\mathbf{x}^*) y_k \\ &+ \left\{ \sum_{k=1}^{n} y_k [\mathbf{d}_k(\mathbf{x}^*) - w_k(\mathbf{x}^*)\mathbf{r}(\mathbf{x}^*)] \right\}' \\ &\times (\mathbf{x} - \mathbf{x}^*) + u_i^{(j)}, \end{aligned}$$

for $j = 1, 2, \ldots, m$. These $T_i^{(j)}(\mathbf{x})$ can be construed as draws from a semiparametric pseudoposterior distribution; one can readily calculate the means, median, percentiles, and variance of this distribution and produce posterior summaries, leading to an approximation to uncertainty about total genetic value.

Irrespective of whether classical or Bayesian viewpoints are adopted, this approach ignores the error of estimation of $g(\mathbf{x})$, as noted earlier.

Strategy 2—random *g*(.)function: The estimate of $g(\mathbf{x})$ can be improved as follows. Consider (21) and suppose, temporarily, that $\boldsymbol{\beta}$ and $\mathbf{u}$ are known. Then, form the offset $y_i - \mathbf{w}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{u}$ and the alternative Nadaraya–Watson estimator of $g(\mathbf{x}_i)$:

$$\begin{aligned}\hat{g}(\mathbf{x}\,|\,\boldsymbol{\beta},\mathbf{u},\mathbf{y},h) &= \hat{E}(y_i - \mathbf{w}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{u}\,|\,\mathbf{x}) \\ &= \sum_{k=1}^{n} w_k(\mathbf{x})(y_k - \mathbf{w}_k'\boldsymbol{\beta} - \mathbf{z}_k'\mathbf{u}). \qquad (23)\end{aligned}$$

The problem is that the location vectors $\boldsymbol{\beta}$ and $\mathbf{u}$ are not observable, so they must be inferred somehow.

Regard now $\boldsymbol{\beta},\mathbf{u}$ as unknown quantities possessing some prior distribution, which is modified via Bayesian learning into the pseudoposterior process $[\boldsymbol{\beta},\mathbf{u}\,|\,\mathbf{y}^*]$, after observation of the pseudodata $\mathbf{y}^*$ and of the $n \times p$ matrix of observed SNP covariates $\mathbf{X}$. Then, for some band width h, $\hat{g}(\mathbf{x}\,|\,\boldsymbol{\beta},\mathbf{u},\mathbf{y},h)$ would also be a random variable possessing some pseudoposterior distribution. Let $\boldsymbol{\beta}^{(j)}$, $\mathbf{u}^{(j)}(j = 1, 2, \ldots, m)$ be a draw from the pseudoposterior process $\left[\boldsymbol{\beta},\mathbf{u},\sigma_u^2,\sigma_e^2\,|\,\mathbf{y}^*,h\right]$, as in the preceding section. Further, regard $\hat{g}(\mathbf{x}\,|\,\boldsymbol{\beta}^{(j)},\mathbf{u}^{(j)},\mathbf{y},h)$ as a draw from the pseudoposterior distribution of $\hat{g}(\mathbf{x}\,|\,\boldsymbol{\beta},\mathbf{u},\mathbf{y},h)$, given $\mathbf{y}^*$. Subsequently, estimate features of the pseudoposterior distribution of $\hat{g}(\mathbf{x}\,|\,\boldsymbol{\beta},\mathbf{u},h)$ by ergodic averaging. For example, an estimate of its posterior expectation would be

$$\hat{E}_{\boldsymbol{\beta},\mathbf{u}|\mathbf{y}^*,h}[\hat{g}(\mathbf{x}\,|\,\boldsymbol{\beta},\mathbf{u},h)] = \frac{1}{m}\sum_{j=1}^{m}\hat{g}(\mathbf{x}\,|\,\boldsymbol{\beta}^{(j)},\mathbf{u}^{(j)},h),$$

where the samples $\boldsymbol{\beta}^{(j)},\mathbf{u}^{(j)}$ are obtained via a Markov chain Monte Carlo (MCMC) procedure from the pseudoposterior distribution $\left[\boldsymbol{\beta},\mathbf{u},\sigma_u^2,\sigma_e^2\,|\,\mathbf{y}^*,h\right]$.

The MCMC procedure would consist of sequential, iterative, sampling from all conditional pseudoposterior distributions, as follows:

Sample $\boldsymbol{\beta}$ from $[\boldsymbol{\beta}\,|\,\text{ELSE}]$, where ELSE denotes all other parameters, $\mathbf{y}^*$, $\mathbf{X}$, and h. Using standard results, this conditional distribution has the form

$$\boldsymbol{\beta}\,|\,\text{ELSE} \sim N(\tilde{\boldsymbol{\beta}},\mathbf{V}_\beta),$$

where

$$\tilde{\boldsymbol{\beta}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'(\mathbf{y}^* - \mathbf{Z}\mathbf{u}),$$

and

$$\mathbf{V}_\beta = (\mathbf{W}'\mathbf{W})^{-1}\sigma_e^2.$$

Sample $\mathbf{u}$ from $[\mathbf{u}\,|\,\text{ELSE}]$. Using similar standard results, the distribution to sample from is

$$\mathbf{u}\,|\,\text{ELSE} \sim N(\tilde{\mathbf{u}},\mathbf{V}_u),$$

where

$$\tilde{\mathbf{u}} = \left(\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2}\right)^{-1}\mathbf{Z}'(\mathbf{y}^* - \mathbf{W}\boldsymbol{\beta})$$

and

$$\mathbf{V}_u = \left(\mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2}\right)^{-1}\sigma_e^2.$$

Sample the two variance components from the scaled inverse chi-square distributions

$$\sigma_u^2 \sim (\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + \nu_u S_u^2)\chi^{-2}_{q+\nu_u}$$

and

$$\sigma_e^2 \sim [(\mathbf{y}^* - \mathbf{W}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{y}^* - \mathbf{W}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \nu_e S_e^2]\chi^{-2}_{n+\nu_e}.$$

Above, ν_u, S_u^2 and ν_e, S_e^2 are known hyperparameters of independent scaled inverse chi-square priors assigned to σ_u^2 and σ_e^2, respectively.

Form draws from the pseudoposterior distribution of $\hat{g}(\mathbf{x}\,|\,\boldsymbol{\beta},\mathbf{u},\mathbf{y},h)$ as

$$\hat{g}(\mathbf{x}\,|\,\boldsymbol{\beta}^{(j)},\mathbf{u}^{(j)},h).$$

As usual, early draws in the MCMC procedure are discarded as burn-in and, subsequently, m samples are collected to infer features of interest. Note that the procedure termed as strategy 1 gives $\hat{g}(\mathbf{x})$ as an estimate of the relationship between phenotypes and SNP data. In strategy 2, one can use $\hat{g}(\mathbf{x}\,|\,\boldsymbol{\beta} = \overline{\boldsymbol{\beta}},\mathbf{u} = \overline{\mathbf{u}},h)$ instead, where $\overline{\boldsymbol{\beta}}$ and $\overline{\mathbf{u}}$ are means of the pseudoposterior distributions $[\boldsymbol{\beta}\,|\,\mathbf{y}^*,h]$ and $[\mathbf{u}\,|\,\mathbf{y}^*,h]$, respectively. Under strategy 2, a point predictor of total additive genetic value could be

$$\begin{aligned}\overline{T}_i(\mathbf{x}) = &\sum_{i=1}^{n} w_i(\mathbf{x}^*)(y_i - \mathbf{w}_i'\overline{\boldsymbol{\beta}} - \mathbf{z}_i'\overline{\mathbf{u}}) \\ &+ \left\{\sum_{i=1}^{n}(y_i - \mathbf{w}_i'\overline{\boldsymbol{\beta}} - \mathbf{z}_i'\overline{\mathbf{u}})[\mathbf{d}_i(\mathbf{x}^*) - w_i(\mathbf{x}^*)\mathbf{r}(\mathbf{x}^*)]\right\}' \\ &\times(\mathbf{x} - \mathbf{x}^*) + \overline{u}_i,\end{aligned}$$

where $\overline{u}_i$ is the posterior mean of u_i.

REPRODUCING KERNEL HILBERT SPACES MIXED MODEL

General: What follows is motivated by developments in Mallick *et al.* (2005) for classification of tumors using microarray data. The underlying theory is outside the scope of this article. Only essentials are given here, and foundations are in Wahba (1990, 1999).

Using the structure of (19), consider the penalized sum of squares

$$\mathrm{SS}[g(\mathbf{x}),h]=\sum_{i=1}^{n}[y_i-\mathbf{w}_i'\boldsymbol{\beta}-\mathbf{z}_i'\mathbf{u}-g(\mathbf{x}_i)]^2+h\|g(\mathbf{x})\|, \tag{24}$$

where, as before, h is a smoothing parameter (possibly unknown) and $\|g(\mathbf{x})\|$ is some norm or "stabilizer." For instance, in smoothing splines, $\|g(\mathbf{x})\|$ is a function of the second derivatives of $g(\mathbf{x})$ integrated between end points that compose the data. The second term in (24) acts as a penalty: if the unknown function $g(\mathbf{x})$ is rough, in the sense of having slopes that change rapidly, the penalty increases. The main problem here is that of finding the function $g(\mathbf{x})$ that minimizes (24). Since $\mathrm{SS}[g(\mathbf{x}),h]$ is a functional on $g(\mathbf{x})$, this is a variational or calculus of variations problem over a space of smooth curves. The solution was given by Kimeldorf and Wahba (1971) and Wahba (1999), and the minimizer admits the representation

$$g(.)=\alpha_0+\sum_{j=1}^{n}\alpha_j K(.,\mathbf{x}_j),$$

where $K(.,.)$ is called a reproducing kernel. A possible choice for the kernel (Mallick *et al.* 2005) is the single smoothing parameter Gaussian function

$$K_h(\mathbf{x},\mathbf{x}_j)=\exp\left[-\frac{(\mathbf{x}-\mathbf{x}_j)'(\mathbf{x}-\mathbf{x}_j)}{h}\right].$$

Mixed model representation: We embed these results into (19), leading to the specification

$$y_i=\mathbf{w}_i'\boldsymbol{\beta}+\mathbf{z}_i'\mathbf{u}+\sum_{j=1}^{n}\exp\left[-\frac{(\mathbf{x}_i-\mathbf{x}_j)'(\mathbf{x}_i-\mathbf{x}_j)}{h}\right]\alpha_j+e_i, \tag{25}$$

with the intercept parameter α_0 included as part of $\boldsymbol{\beta}$. Note that there are as many regressions α_j as there are data points. However, the roughness penalty in the variational problem leads to a reduction in the effective number of parameters in reproducible kernel Hilbert spaces (RKHS) regression, as it occurs in smoothing splines (Fox 2002).

Define the $1\times n$ row vector

$$\mathbf{t}_i'(h)=\left\{\exp\left[-\frac{(\mathbf{x}_i-\mathbf{x}_j)'(\mathbf{x}_i-\mathbf{x}_j)}{h}\right]\right\},\quad j=1,2,\ldots,n;$$

the $n\times 1$ column vector $\boldsymbol{\alpha}=\{\alpha_j\}$, $j=1,2,\ldots,n$; and the $n\times n$ matrix

$$\mathbf{T}(h)=\begin{bmatrix}\mathbf{t}_1'(h)\\ \mathbf{t}_2'(h)\\ .\\ \mathbf{t}_n'(h)\end{bmatrix}.$$

Then (25) can be written in matrix form as

$$\mathbf{y}=\mathbf{W}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}+\mathbf{T}(h)\boldsymbol{\alpha}+\mathbf{e}.$$

Suppose, further, that the α_j coefficients are exchangeable according to the distribution $\alpha_j\sim N(0,\sigma_\alpha^2)$. Hence, for a given smoothing parameter h, we are in the setting of a mixed-effects linear model.

Given h, σ_u^2, σ_e^2, and σ_α^2 (at a given h, the three variance components may be estimated by, *e.g.*, REML) one can obtain predictions of the polygenic breeding values $\mathbf{u}$ and of the coefficients $\boldsymbol{\alpha}$ from the solutions to the system

$$\begin{bmatrix}\mathbf{W}'\mathbf{W} & \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{T}(h)\\ \mathbf{Z}'\mathbf{W} & \mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}(\sigma_e^2/\sigma_u^2) & \mathbf{Z}'\mathbf{T}(h)\\ \mathbf{T}'(h)\mathbf{W} & \mathbf{T}'(h)\mathbf{Z} & \mathbf{T}'(h)\mathbf{T}(h)+\mathbf{I}(\sigma_e^2/\sigma_\alpha^2)\end{bmatrix}\times\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\\ \hat{\boldsymbol{\alpha}}\end{bmatrix}=\begin{bmatrix}\mathbf{W}'\mathbf{y}\\ \mathbf{Z}'\mathbf{y}\\ \mathbf{T}'(h)\mathbf{y}\end{bmatrix}. \tag{26}$$

The total additive genetic value of an individual possessing a vector of SNP covariates $\mathbf{x}_i$ can be defined as

$$\begin{aligned}T_i&=u_i+\left\{\frac{\partial}{\partial\mathbf{x}}\sum_{j=1}^{n}\exp\left[-\frac{(\mathbf{x}-\mathbf{x}_j)'(\mathbf{x}-\mathbf{x}_j)}{h}\right]\alpha_j\right\}'_{\mathbf{x}=\mathbf{x}^*}(\mathbf{x}_i-\mathbf{x}^*)\\ &=u_i+\left\{\sum_{j=1}^{n}\exp\left[-\frac{(\mathbf{x}^*-\mathbf{x}_j)'(\mathbf{x}^*-\mathbf{x}_j)}{h}\right]\boldsymbol{\eta}_j^*\right\}'(\mathbf{x}_i-\mathbf{x}^*),\end{aligned} \tag{27}$$

where

$$\boldsymbol{\eta}_j^*=-\frac{2(\mathbf{x}^*-\mathbf{x}_j)'\alpha_j}{h}.$$

Since T_i is a linear function of the unknown u_i and α_j effects, its empirical best linear unbiased predictor, assuming known h, can be obtained by replacing these effects by the corresponding solutions from (26).

Incomplete genotyping: At least in animal breeding, it is not feasible to have all individuals genotyped for SNPs. On the other hand, the number of animals with phenotypic information available is typically in the order of hundreds of thousands, and genotyping is selective, *e.g.*, young bulls that are candidates for progeny testing in dairy cattle production. Animals lacking molecular data are not a random sample from the population, and ignoring this issue may lead to biased inferences. Unless missingness of molecular data is ignorable, in the sense of, *e.g.*, Henderson (1975), Rubin (1976), Gianola and Fernando (1986), or Im *et al.* (1989), the procedures given below require modeling of the missing data process, which is difficult and may lack robustness. Here, it is assumed that missingness is ignorable, enabling use of likelihood-based or Bayesian procedures as if selection had not taken place (Sorensen *et al.* 2001). Two *ad hoc* procedures are discussed, and an

alternative approach, suitable for kernel regression, is presented in the CONCLUSION.

Let the vector of phenotypic data be partitioned as $\mathbf{y} = [\mathbf{y}_1' \mathbf{y}_2']'$, where $\mathbf{y}_1$ ($n_1 \times 1$) consists of records of individuals lacking SNP data, whereas $\mathbf{y}_2$ ($n_2 \times 1$) includes phenotypic data of genotyped individuals. Often, it will be the case that $n_1 > p \gg n_2$. We adopt the model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix} \mathbf{u} + \begin{bmatrix} \mathbf{0} \\ \mathbf{T}_2(h) \end{bmatrix} \boldsymbol{\alpha} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}. \quad (28)$$

For the sake of flexibility, assume that $\mathbf{e}_1 \sim N(\mathbf{0}, \mathbf{I}_{n_1}\sigma^2_{e_1})$ and $\mathbf{e}_2 \sim N(\mathbf{0}, \mathbf{I}_{n_2}\sigma^2_{e_2})$ are mutually independent but heteroscedastic vectors. In short, the key assumption made here is that the random effect $\boldsymbol{\alpha}$ affects $\mathbf{y}_2$ but not $\mathbf{y}_1$ or, equivalently, that it gets absorbed into $\mathbf{e}_1$. With this representation, the mixed model equations take the form

$$\begin{bmatrix} \sum_{i=1}^{2} \frac{1}{\sigma^2_{e_i}}\mathbf{W}_i'\mathbf{W}_i & \sum_{i=1}^{2} \frac{1}{\sigma^2_{e_i}}\mathbf{W}_i'\mathbf{Z}_i & \frac{1}{\sigma^2_{e_2}}\mathbf{W}_2'\mathbf{T}_2(h) \\ \sum_{i=1}^{2} \frac{1}{\sigma^2_{e_i}}\mathbf{Z}_i'\mathbf{W}_i & \sum_{i=1}^{2} \frac{1}{\sigma^2_{e_i}}\mathbf{Z}_i'\mathbf{Z}_i + \mathbf{A}^{-1}\frac{1}{\sigma^2_u} & \frac{1}{\sigma^2_{e_2}}\mathbf{Z}_2'\mathbf{T}_2(h) \\ \frac{1}{\sigma^2_{e_2}}\mathbf{T}_2'(h)\mathbf{W}_2 & \mathbf{T}_2'(h)\mathbf{Z} & \frac{1}{\sigma^2_{e_2}}\mathbf{T}_2'(h)\mathbf{T}_2(h) + \mathbf{I}\frac{1}{\sigma^2_\alpha} \end{bmatrix} \times \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{2} \frac{1}{\sigma^2_{e_i}}\mathbf{W}_i'\mathbf{y}_i \\ \sum_{i=1}^{2} \frac{1}{\sigma^2_{e_i}}\mathbf{Z}_i'\mathbf{y}_i \\ \frac{1}{\sigma^2_{e_2}}\mathbf{T}_2'(h)\mathbf{y}_2 \end{bmatrix} \quad (29)$$

If SNP data are missing completely at random and h, σ^2_u, σ^2_e, and σ^2_α are treated as known, then $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$, and $\hat{\mathbf{u}}$ and $\hat{\boldsymbol{\alpha}}$ are unbiased predictors of $\mathbf{u}$ and $\boldsymbol{\alpha}$, respectively. They are not "best," in the sense of having minimum variance or minimum prediction error variance, because the smooth function $g(\mathbf{x})$ of the SNP markers is missing in the model for individuals that are not genotyped (HENDERSON 1974).

An alternative consists of writing the bivariate model

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{T}_2(h) \end{bmatrix} \boldsymbol{\alpha} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}$$

and then assigning to the polygenic component, the multivariate normal distribution

$$\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{A}\sigma^2_{u_1} & \mathbf{A}\sigma_{u_{12}} \\ \mathbf{A}\sigma_{u_{12}} & \mathbf{A}\sigma^2_{u_2} \end{bmatrix} \right).$$

Here, $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$ are additive genetic variances in individuals without and with molecular information, respectively, and $\sigma_{u_{12}}$ is their additive genetic covariance. Computations would be those appropriate for a two-trait linear model analysis (HENDERSON 1984; SORENSEN and GIANOLA 2002).

Bayesian analysis: To illustrate, consider the first of the two options in the preceding section. Suppose a kernel has been chosen but that the value of h is uncertain, so that the model unknowns are

$$\boldsymbol{\theta} = \left[\boldsymbol{\beta}', \mathbf{u}', \boldsymbol{\alpha}', \sigma^2_u, \sigma^2_\alpha, \sigma^2_{e_1}, \sigma^2_{e_2}, h \right]'.$$

Let the prior density have the form

$$\begin{aligned} p(\boldsymbol{\theta} \mid H) = {} & p(\boldsymbol{\beta}) N\left(\mathbf{u} \mid \mathbf{0}, \mathbf{A}\sigma^2_u\right) N\left(\boldsymbol{\alpha} \mid \mathbf{0}, \mathbf{I}\sigma^2_\alpha\right) \\ & \times p\left(\sigma^2_u \mid \nu_u, S^2_u\right) p\left(\sigma^2_\alpha \mid \nu_\alpha, S^2_\alpha\right) p\left(\sigma^2_{e_1} \mid \nu_e, S^2_e\right) \\ & \times p\left(\sigma^2_{e_2} \mid \nu_e, S^2_e\right) p(h \mid h_{\min}, h_{\max}), \end{aligned} \quad (30)$$

where H denotes the set of all known hyperparameters (whose values are fixed *a priori*) and $N(. \mid ., .)$ indicates a multivariate normal distribution with appropriate mean vector and covariance matrix. The four variance components $\sigma^2_u, \sigma^2_\alpha, \sigma^2_{e_1}, \sigma^2_{e_2}$ are assigned independent scaled inverse chi-square prior distributions with degrees of freedom ν and scale parameters S^2, with appropriate subscripts. Assign an improper prior distribution to each of the elements of $\boldsymbol{\beta}$ and, as in MALLICK *et al.* (2005), adopt a uniform prior for h, with lower and upper boundaries $h_{\min}$ and $h_{\max}$, respectively.

Given the parameters, observations are assumed to be conditionally independent, and the distribution adopted for the sampling model is

$$N\left(\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \middle| \begin{bmatrix} \mathbf{W}_1\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u} \\ \mathbf{W}_2\boldsymbol{\beta} + \mathbf{Z}_2\mathbf{u} + \mathbf{T}_2(h)\boldsymbol{\alpha} \end{bmatrix}, \begin{bmatrix} \mathbf{I}_{n_1}\sigma^2_{e_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_2}\sigma^2_{e_2} \end{bmatrix} \right).$$

Given h, one is again in the setting of the Bayesian analysis of a mixed linear model, and Markov chain Monte Carlo procedures for this situation are well known. Under standard conjugate prior parameterizations, all conditional posterior distributions are known, save for that of h. Hence, one can construct a Gibbs–Metropolis sampling scheme in which conditional distributions are used for drawing $\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\alpha}, \sigma^2_u, \sigma^2_\alpha, \sigma^2_{e_1}, \sigma^2_{e_2}$ and a Metropolis update is employed for h.

Location effects $\boldsymbol{\beta}$ are drawn from a multivariate normal distribution with mean vector and covariance matrix

$$\begin{aligned} \overline{\boldsymbol{\beta}} = {} & \left(\sum_{i=1}^{2} \frac{1}{\sigma^2_{e_i}}\mathbf{W}_i'\mathbf{W}_i \right)^{-1} \\ & \times \left[\sum_{i=1}^{2} \frac{1}{\sigma^2_{e_i}}\mathbf{W}_i'(\mathbf{y}_i - \mathbf{Z}_i\mathbf{u}) - \frac{1}{\sigma^2_{e_2}}\mathbf{W}_2'\mathbf{T}_2(h)\boldsymbol{\alpha} \right] \end{aligned} \quad (31)$$

and

$$\mathbf{V}_\beta = \left(\sum_{i=1}^{2} \frac{1}{\sigma^2_{e_i}}\mathbf{W}_i'\mathbf{W}_i \right)^{-1}, \quad (32)$$

respectively. Likewise, the additive genetic effects can be sampled from a normal distribution centered at

$$\overline{\mathbf{u}} = \left(\sum_{i=1}^{2}\frac{1}{\sigma^2_{e_i}}\mathbf{Z}_i'\mathbf{Z}_i + \mathbf{A}^{-1}\frac{1}{\sigma^2_u}\right)^{-1} \times \left[\sum_{i=1}^{2}\frac{1}{\sigma^2_{e_i}}\mathbf{Z}_i'(\mathbf{y}_i - \mathbf{W}_i\boldsymbol{\beta}) - \frac{1}{\sigma^2_{e_2}}\mathbf{W}_2'\mathbf{T}_2(h)\boldsymbol{\alpha}\right] \quad (33)$$

and with covariance matrix

$$\mathbf{V}_u = \left(\sum_{i=1}^{2}\frac{1}{\sigma^2_{e_i}}\mathbf{Z}_i'\mathbf{Z}_i + \mathbf{A}^{-1}\frac{1}{\sigma^2_u}\right)^{-1}. \quad (34)$$

The conditional posterior distribution of the coefficients $\boldsymbol{\alpha}$ is multivariate normal as well, with mean vector

$$\overline{\boldsymbol{\alpha}} = \left(\frac{1}{\sigma^2_{e_2}}\mathbf{T}_2'(h)\mathbf{T}_2(h) + \mathbf{I}\frac{1}{\sigma^2_\alpha}\right)^{-1}\frac{1}{\sigma^2_{e_2}}\mathbf{T}_2'(h)(\mathbf{y}_2 - \mathbf{W}_2\boldsymbol{\beta} - \mathbf{Z}_2\mathbf{u}) \quad (35)$$

and variance–covariance matrix

$$\mathbf{V}_\alpha = \left(\frac{1}{\sigma^2_{e_2}}\mathbf{T}_2'(h)\mathbf{T}_2(h) + \mathbf{I}\frac{1}{\sigma^2_\alpha}\right)^{-1}. \quad (36)$$

All four variance components have scaled inverse chi-square conditional posterior distributions and are conditionally independent. The conditional posterior distributions to sample from are

$$\sigma^2_u \,|\, \text{ELSE} \sim (\mathbf{u}'\mathbf{A}^{-1}\mathbf{u} + \upsilon_u S^2_u)\chi^{-2}_{q+\upsilon_u}, \quad (37)$$

$$\sigma^2_\alpha \,|\, \text{ELSE} \sim (\boldsymbol{\alpha}'\boldsymbol{\alpha} + \nu_\alpha S^2_\alpha)\chi^{-2}_{n_2+\upsilon_\alpha}, \quad (38)$$

$$\sigma^2_{e_1} \,|\, \text{ELSE} \sim [(\mathbf{y}_1 - \mathbf{W}_1\boldsymbol{\beta} - \mathbf{Z}_1\mathbf{u})' \times (\mathbf{y}_1 - \mathbf{W}_1\boldsymbol{\beta} - \mathbf{Z}_1\mathbf{u}) + \nu_e S^2_e]\chi^{-2}_{n_1+\upsilon_e}, \quad (39)$$

and

$$\sigma^2_{e_2} \,|\, \text{ELSE} \sim [(\mathbf{y}_2 - \mathbf{W}_2\boldsymbol{\beta} - \mathbf{Z}_2\mathbf{u} - \mathbf{T}_2(h)\boldsymbol{\alpha})' \times (\mathbf{y}_2 - \mathbf{W}_2\boldsymbol{\beta} - \mathbf{Z}_2\mathbf{u} - \mathbf{T}_2(h)\boldsymbol{\alpha}) + \nu_e S^2_e]\chi^{-2}_{n_2+\upsilon_e}. \quad (40)$$

The most difficult parameter to sample is h. Its conditional posterior density can be represented as

$$p(h \,|\, \text{ELSE}) = \frac{F(h)}{\int_{h_{\min}}^{h_{\max}} F(h)dh}, \quad (41)$$

where

$$F(h) = \exp\left[-\frac{1}{2\sigma^2_{e_2}} \times \sum_{i=1}^{n_2}\left\{y_i - \mathbf{w}_i'\boldsymbol{\beta} - \mathbf{z}_i'\mathbf{u} - \sum_{j=1}^{n_2}\exp\left[-\frac{(\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)}{h}\right]\alpha_j\right\}^2\right].$$

Density (41) is not in a recognizable form. However, a Metropolis algorithm (Metropolis *et al.* 1953) can be tailored for obtaining samples from the distribution $[h \,|\, \text{ELSE}]$. Suppose that the Markov chain is at state $h^{[t]}$. A proposal value h^* is drawn from some symmetric candidate-generating distribution and accepted with probability

$$\gamma = \min\left[\frac{p(h^* \,|\, \text{ELSE})}{p(h^{[t]} \,|\, \text{ELSE})}, 1\right].$$

If the proposal is accepted, then set $h^{[t+1]} = h^*$; otherwise, stay at $h^{[t]}$. If a Hastings update is employed, instead, an adaptive proposal distribution could be, *e.g.*, a finite mixture of the prior density and of a scaled inverse chi-square density f with mean value $h^{[t]}$ and variance $\left(h_{\text{largest}} - h_{\text{smallest}}\right)^2/12$,

$$m(h) = w\frac{1}{h_{\max} - h_{\min}} + (1 - w)f,$$

where h_{largest} and h_{smallest} are the largest and smallest value of h, respectively, accepted up to time t. The weight w in $(0, 1)$ is assessed such that a reasonable acceptance rate for the Metropolis–Hastings algorithm is attained.

Illustrative example: Phenotypic and genotypic values were simulated (single replication) for a sample of N unrelated individuals, for each of two situations. The trait was determined either by five biallelic QTL, having additive gene action, or by five pairs of biallelic QTL, having additive-by-additive gene action. Under nonadditive gene action, the additive genetic variance was null, so that all genetic variance was of the additive-by-additive type. Heritability (ratio between genetic and phenotypic variance) in both settings was set to 0.5.

Genotypes were simulated for a total of 100 biallelic markers, including the "true" QTL; all loci were simulated to be in gametic-phase equilibrium. Since all individuals were unrelated and all loci were in gametic-phase equilibrium, only the QTL genotypes and the trait phenotypes would provide information on the genotypic value. In most real applications, the location of the QTL will not be known, and so many loci that are not QTL will be included in the analysis.

A RKHS mixed model was used to predict genotypic values, given phenotypic values and genotypes at the QTL and at all other loci. The model included a fixed intercept α_0 and a random RKHS regression coefficient α_i for each subject; additive effects were omitted from the model, as individuals were unrelated, precluding separation of additive effects from residuals, in the absence of knowledge of variance components. The genetic value g_i of a subject was predicted as

$$\hat{g}_i = \hat{\alpha}_0 + \mathbf{t}_i'(h)\hat{\boldsymbol{\alpha}},$$

where $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}$ were obtained by solving (26) for this model, using a Gaussian kernel at varying functions of h.

TABLE 1

Accuracy of predicting genotypic values using the RKHS mixed model

Gene action	N	k	h	$\sigma_e^2/\sigma_\alpha^2$	Accuracy
Additive	1000	100	8920	0.0005	0.95
Nonadditive	1000	10	47	0.01	0.96
Nonadditive	1000	25	54	0.01	0.81
Nonadditive	1000	50	98	0.01	0.50
Nonadditive	1000	100	436	0.0005	0.00
Nonadditive	5000	100	300	0.0005	0.85

N is the sample size, k is the number of loci fitted in the model, including the QTL, h is the bandwidth, and $\sigma_e^2/\sigma_\alpha^2$ is the ratio between environmental and RKHS variance.

TABLE 2

Accuracy of predicting genotypic values using the MLR mixed model

Gene action	N	k	$\sigma_e^2/\sigma_\alpha^2$	Accuracy
Additive	1000	100	32	0.95
Nonadditive	1000	10	208	0.0
Nonadditive	1000	25	748	0.18
Nonadditive	1000	50	760	0.0
Nonadditive	1000	100	328	0.0
Nonadditive	5000	100	3400	0.0

N is the sample size, k is the number of loci fitted in the model, including the QTL, and $\sigma_e^2/\sigma_\alpha^2$ is the ratio between the environmental variance and that "due to" linear regression.

The mean squared error of prediction of genetic value (MSEP) was calculated as

$$\text{MSEP}\left(h, \sigma_e^2/\sigma_\alpha^2\right) = \sum_i (g_i - \hat{g}_i)^2, \qquad (42)$$

and a grid search was used to determine the values of h and $\sigma_e^2/\sigma_\alpha^2$ that minimized (42). To evaluate the performance of $\hat{g}_i$ as a predictor, another sample of 1000 individuals ("PRED") was simulated, including genotypes, genotypic values, and phenotypes. This was deemed preferable to doing prediction in the training sample, to reduce dependence between performance and (h, $\sigma_e^2/\sigma_\alpha^2$), whose values were assessed in the training sample The genotypic values of the subjects in PRED were predicted, given their genotypes, using $\hat{g}_i$. Genotypic values were also predicted using a multiple linear regression (MLR) mixed model with a fixed intercept and random regression coefficients on the linear effects of genotypes.

Results for the RKHS mixed model are in Table 1, and those for the MLR mixed model are in Table 2, where "accuracy" is the correlation between true and predicted genetic values. When gene action was strictly additive, the two methods (each fitting $k = 100$ loci) had the same accuracy, indicating that RKHS performed well even when the parametric assumptions were valid. On the other hand, when inheritance was purely additive by additive, the parametric MLR was clearly outperformed by RKHS, irrespective of the number of loci fitted. An exception occurred at $k = 100$ and $N = 1000$; here, the two methods were completely inaccurate. However, when N was increased from 1000 to 5000, the accuracy of RKHS jumped to 0.85 (Table 1), whereas MLR remained inaccurate (Table 2). Note that the accuracy of RKHS decreased (with N held at 1000) when k increased. We attribute this behavior to the use of a Gaussian kernel when, in fact, covariates are discrete.

CONCLUSION

This article discusses approaches for prediction of genetic value using markers for the entire genome, such as SNPs, and phenotypic measurements for complex traits. In particular, theory for nonparametric and semiparametric procedures, *i.e.*, kernel regression and reproducing kernel Hilbert spaces regression, is developed. The methods consist of a combination of features of the classical additive genetic model of quantitative genetics with an unknown function of SNP genotypes, which is inferred nonparametrically. Mixed-model and Bayesian implementations are presented. The procedures can be computed using software developed by animal breeders for likelihood-based and Bayesian analysis, after modifications.

Except for the parametric part of the model, that is, the standard normality assumption for additive genetic values and model residuals, the procedures attempt to circumvent potential difficulties posed by violation of assumptions required for an orthogonal decomposition of genetic variance stemming from SNP genotypes (Cockerham 1954; Kempthorne 1954). Our expectation is that the nonparametric function of marker genotypes, $g(\mathbf{x})$, captures all possible forms of interaction, but without explicit modeling. The procedures should be particularly useful for highly dimensional regression, including the situation in which the number of SNP variables (p) exceeds amply the number of data points (n). Instead of performing a selection of a few "significant" markers on the basis of some *ad hoc* method, information on all molecular polymorphisms is employed, irrespective of the degree of colinearity. This is because the procedures should be insensitive to difficulties caused by colinearity, given the forms of the estimators, *e.g.*, (6) or (23). It is assumed that the assignment of genotypes to individuals is unambiguous, *i.e.*, $\mathbf{x}$ gives the genotypes for SNP markers free of error.

Our methods share the spirit of those of Meuwissen *et al.* (2001), Gianola *et al.* (2003), Xu (2003), Yi *et al.* (2003), Ter Braak *et al.* (2005), Wang *et al.* (2005), and Zhang and Xu (2005), but without making strong

assumptions about the form of the marker–phenotype relationship, which is assumed linear by all these authors, and without invoking parametric distributions for pertinent effects.

A hypothetical example was presented, illustrating potential and computational feasibility of at least the RKHS procedure; a standard additive genetic model was outperformed by RKHS when additive-by-additive gene action was simulated. Comparisons between parametric and nonparametric procedures are needed. It is unlikely that computer simulations would shed much light in this respect. First, a huge number of simulation scenarios can be envisaged, resulting from limitless combinations of parameter values, numbers of markers, marker effects, residual distributions, etc. Second, simulations tend to have local value only, that is, conclusions are tentative only within the limited experimental region explored and are heavily dependent on the state of nature assumed. Third, end points and gauges of a simulation tend to be arbitrary. For instance, should frequentist procedures be assessed on Bayesian grounds and vice versa? We believe that studies based on predictive cross-validation for a range of traits and species are perhaps more fruitful. These studies will be conducted once adequate and reliable phenotypic-SNP data sets become more widely available.

There are at least two difficulties with the proposed methodology. As noted above, it is assumed that the SNP genotypes are constructed without error, which is seldom the case. To solve this problem, one would need to build an error-in-variables model, but at the expense of introducing additional assumptions. A second difficulty is that posed by the fact that many individuals will lack SNP information, at least in animal breeding. Earlier, we presented approximate procedures on the basis of the assumption that missingness of SNP data is ignorable, such that the effect of $g(\mathbf{x})$ can be absorbed into a residual that has a different variance from that peculiar to individuals possessing SNP data or into the additive genetic value in a two-trait implementation. A more appropriate treatment of missing data requires imputation of genotypes for individuals lacking SNP information. If the SNP data are missing completely at random or just at random, the solutions to system (29), after augmentation with the missing $\mathbf{T}_1(h)$, would give the means of the posterior distributions of $\boldsymbol{\beta}, \mathbf{u}$, and $\boldsymbol{\alpha}$, conditionally on the variance components and on h (Gianola and Fernando 1986; Sorensen *et al.* 2001). However, the sampling procedure should address the constraint that SNP genotypes of related individuals must be more alike than those of unrelated subjects. In our treatment, and for operational reasons, we adopted the simplifying assumption that SNP genotypes are independently distributed. This may be anticonservative and could lead to some bias if SNP data are not missing completely at random.

It is unknown to what extent our procedures are robust with respect to the choice of kernel function. Silverman (1986) discusses several options and, in the context of univariate density estimation, concludes that different kernels differ little in mean squared error. Also, an inadequate specification of the smoothing parameter h may affect inference adversely. In this respect, the procedures discussed in reproducing kernel hilbert spaces mixed model provide for automatic specification of the band-width parameter. Here, one could either make an analysis conditionally on the "most likely" value of h or, alternatively, average over its posterior distribution, to take uncertainty into account fully. Here, we have focused on using a continuous kernel, primarily to exploit differentiability properties. However, the vector of markers, $\mathbf{x}$, has a discrete distribution, and careful investigation is needed to assess the adequacy of such an approximation.

A procedure to accommodate missing genotypes in kernel regression is as follows. Replace $g(\mathbf{x}_i)$ in (19) by

$$\begin{aligned} g(\mathbf{x}_i) &= g(\hat{\mathbf{x}}_i) + g(\mathbf{x}_i) - g(\hat{\mathbf{x}}_i) \\ &= g(\hat{\mathbf{x}}_i) + q_i, \end{aligned} \tag{43}$$

where $\hat{\mathbf{x}}_i$ is the conditional expectation of $\mathbf{x}_i$ given all the observed genotypes in the pedigree, and $q_i = g(\mathbf{x}_i) - g(\hat{\mathbf{x}}_i)$. The conditional expectation in (43) is a function of the conditional probabilities of the missing genotypes given the observed genotypes. Much research has been devoted to computing such probabilities from complex pedigrees using either approximations (Van Arendonk *et al.* 1989; Fernando *et al.* 1993; Kerr and Kinghorn 1996) or MCMC samplers (Sheehan and Thomas 1993; Jensen *et al.* 1995; Jensen and Kong 1999; Fernández *et al.* 2002; Stricker *et al.* 2002).

When genotypes for individual i are observed, $\hat{\mathbf{x}}_i = \mathbf{x}_i$, and q_i is null. When genotypes for i are missing, $g(\hat{\mathbf{x}}_i)$ is an approximation to the conditional expectation of $g(\mathbf{x}_i)$ given the observed genotypes, and q_i is a random variable with approximately null expectation. Now, the genotypic value of an individual can be written as

$$g(\mathbf{x}_i) + u_i = g(\hat{\mathbf{x}}_i) + q_i + u_i, \tag{44}$$

where, from the linear approximation given by (9), the variance of q_i is

$$\mathrm{Var}(q_i) = \hat{\mathbf{g}}(\mathbf{x}^*)' \mathbf{V}_i \hat{\mathbf{g}}(\mathbf{x}^*), \tag{45}$$

with $\mathbf{V}_i$ being the conditional covariance matrix of $\mathbf{x}_i$ given the observed genotypes of relatives. The jth diagonal in $\mathbf{V}_i$ is a function of the conditional probabilities of the missing genotypes at locus j given the observed genotypes, and the jkth off-diagonal element is a function of the conditional probabilities of the missing genotypes at loci j and k given the observed genotypes. An

estimate of $\mathrm{Var}(q_i)$ can be obtained by replacing $\dot{\mathbf{g}}(\mathbf{x}^*)$ by its estimate in (45).

In the simplest approach to modeling the covariances of the q_i, the random effects u_i and q_i are combined as

$$a_i = u_i + q_i.$$

Now, the model for y_i is written as

$$y_i = \mathbf{w}_i'\boldsymbol{\beta} + g(\hat{\mathbf{x}}_i) + \mathbf{z}_i'\mathbf{a} + e_i. \tag{46}$$

An approximation to the covariance matrix Σ_a of $\mathbf{a}$ can be obtained by using the usual tabular algorithm that is based only on pedigree information, after setting the ith diagonal of Σ_a to $\mathrm{Var}(u_i) + \mathrm{Var}(q_i)$. The inverse of this approximate Σ_a is sparse and it can be computed efficiently (Henderson 1976).

Although q_i is not null only for individuals with missing genotypes, as discussed below, the observed genotypes on relatives can provide information on the segregation of alleles in individuals with missing genotypes. Thus, observed genotypes can provide information on covariances between the q_i. Let q_{ij} denote the additive genotypic value at locus j, which can be written as

$$q_{ij} = v_{ij}^{\mathrm{m}} + v_{ij}^{\mathrm{p}},$$

where v_{ij}^{m} and v_{ij}^{p} are the gametic values of haplotypes h_{ij}^{m} and h_{ij}^{p}. When genotype information is available at a locus, it is more convenient to work with the gametic values v_{ij}^{m} and v_{ij}^{p} rather than the genotypic values q_{ij}. For an individual with missing genotypes, the variance of v_{ij}^{x} can be written as

$$\mathrm{Var}(v_{ij}^{x}) = \sum_{h_{ij}^{x} \neq h_{ij}^{\prime x}} \Pr(h_{ij}^{x})\Pr(h_{ij}^{\prime x})(\alpha_{h_{ij}^{x}} - \alpha_{h_{ij}^{\prime x}})^2, \tag{47}$$

for $x = \mathrm{m}$ or p, where $\alpha_{h_{ij}^{\mathrm{m}}}$ is the additive effect of haplotype h_{ij}^{m}. These additive effects can be estimated from the linear function given by (8) as follows. Let $\mathbf{x}_{h_j}$ denote the value of $\mathbf{x}$ with all elements set to its mean value except at locus j where one of the haplotypes is set to h_j. Then,

$$\hat{\alpha}_{h_j} = \hat{g}(\mathbf{x}^*)'(\mathbf{x}_{h_j} - \mathbf{x}^*).$$

The covariance between v_{ij}^{x} and $v_{i'j}^{\mathrm{m}}$, for example, can be written as

$$\begin{aligned}\mathrm{Cov}(v_{ij}^{x}, v_{i'j}^{\mathrm{m}}) = {} & \mathrm{Cov}(v_{ij}^{x}, v_{dj}^{\mathrm{m}})\Pr(h_{i'j}^{\mathrm{m}} \leftarrow h_{dj}^{\mathrm{m}}) \\ & + \mathrm{Cov}(v_{ij}^{x}, v_{dj}^{\mathrm{p}})\Pr(h_{i'j}^{\mathrm{m}} \leftarrow h_{dj}^{\mathrm{p}}),\end{aligned} \tag{48}$$

where $\Pr(h_{i'j}^{\mathrm{m}} \leftarrow h_{dj}^{\mathrm{m}})$ is the probability that the maternal haplotype of i' is inherited from the maternal haplotype of d its dam. Suppose genotype information is available on the ancestors of d and on the descendants of i'. Then, the segregation probabilities in (48) may be different from 0.5, and thus the observed genotypes will contribute information for genetic covariances at this locus. However, even in this situation, the inverse of the gametic covariance matrix $\mathbf{G}_j$ for locus j that is constructed using (47) and (48) is sparse and it can be computed efficiently (Fernando and Grossman 1989).

For an improved approximation to the modeling of covariances between the q_i, the model for y_i is written as

$$y_i = \mathbf{w}_i'\boldsymbol{\beta} + g(\hat{\mathbf{x}}_i) + \sum_{j \in L} \mathbf{k}_j'\mathbf{v}_j + \mathbf{z}_i'\mathbf{a}^* + e_i, \tag{49}$$

where L is the set of k loci with the largest effects on the trait. The covariance matrix for the vector $\mathbf{v}_j$ of gametic values can be computed using (47) and (48). The usual tabular method based on pedigree information is used to compute Σ_{a^*}, after setting the ith diagonal to

$$\mathrm{Var}(a_i^*) = \mathrm{Var}(a_i) - \sum_{j \in L}[\mathrm{Var}\left(v_{ij}^{\mathrm{m}}\right) + \mathrm{Var}\left(v_{ij}^{\mathrm{p}}\right)]. \tag{50}$$

Our models extend directly to binary or ordered categorical responses when a threshold-liability model holds (Wright 1934; Falconer 1965; Gianola 1982; Gianola and Foulley 1983). Here, the $g(\mathbf{x})$ function would be viewed as affecting a latent variable; Mallick *et al.* (2005) used this idea in analysis of gene expression measurements.

Extension to multivariate responses is less straightforward. It is conceivable that each trait may require a different function of SNP genotypes. It is not obvious how this problem should be dealt with without making strong parametric assumptions.

In conclusion, we believe that this article gives a first description of nonparametric and semiparametric procedures that may be suitable for prediction of genetic value using dense marker data. However, considerable research is required for tuning, extending, and validating some of the ideas presented here.

Miguel A. Toro, Bani Mallick, and two anonymous reviewers are thanked for useful comments. Research was completed while the senior author was visiting Parco Tecnologico Padano, Lodi, Italy. Support by the Wisconsin Agriculture Experiment Station and by grants National Research Initiatives Competitive Grants Program/U.S. Department of Agriculture 2003-35205-12833, National Science Foundation (NSF) DEB-0089742, and NSF Division of Mathematical Sciences (DMS)–NSF DMS-044371 is acknowledged.

LITERATURE CITED

Aitchison, J., and C. G. G. Aitken, 1976 Multivariate binary discrimination by the kernel method. Biometrika **63:** 413–420.

Chang, H. L. A., 1988 Studies on estimation of genetic variances under nonadditive gene action. Ph.D. Thesis, University of Illinois, Urbana–Champaign, IL.

Chu, C. K., and J. S. Marron, 1991 Choosing a kernel regression estimator. Stat. Sci. **6:** 404–436.

Cockerham, C. C., 1954 An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics **39:** 859–882.

Dekkers, J. C. M., and F. Hospital, 2002 The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. **3:** 22–32.

D'Haeseleer, P., S. Liang and R. Somogyi, 2000 Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics **16:** 707–726.

Falconer, D. S., 1960 *Introduction to Quantitative Genetics*. Longman, London.

Falconer, D. S., 1965 The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann. Hum. Genet. **29:** 51–76.

Fernández, S. A., R. L Fernando, B. Gulbrandtsen, C. Stricker, M. Schelling *et al.*, 2002 Irreducibility and efficiency of ESIP to sample marker genotypes in large pedigrees with loops. Genet. Sel. Evol. **34:** 537–555.

Fernando, R. L., and M. Grossman, 1989 Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **21:** 467–477.

Fernando, R. L., C. Stricker and R. C. Elston, 1993 An efficient algorithm to compute the posterior genotypic distribution for every member of a pedigree without loops. Theor. Appl. Genet. **87:** 89–93.

Fisher, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. **52:** 399–433.

Fox, J., 2002 *An R and S-PLUS Companion to Applied Regression*. Sage, Thousand Oaks, CA.

Fox, J., 2005 *Introduction to Nonparametric Regression* (Lecture Notes) (http://socserv.mcmaster.ca/jfox/Courses/Oxford).

Gianola, D., 1982 Theory and analysis of threshold characters. J. Anim. Sci. **54:** 1079–1096.

Gianola, D., and R. L. Fernando, 1986 Bayesian methods in animal breeding theory. J. Anim. Sci. **63:** 217–244.

Gianola, D., and J. L. Foulley, 1983 Sire evaluation for ordered categorical data with a threshold model. Genet. Sel. Evol. **15:** 201–244.

Gianola, D., and D. Sorensen, 2004 Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. Genetics **167:** 1407–1424.

Gianola, D., M. Perez-Enciso and M. A. Toro, 2003 On marker-assisted prediction of genetic value: beyond the ridge. Genetics **163:** 347–365.

Golub, T. R., D. Slonim, P. Tamayo, C. Huard, M. Gasenbeek *et al.*, 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286:** 531–537.

Hart, J. D., and C. L. Lee, 2005 Robustness of one-sided cross-validation to autocorrelation. J. Multivar. Anal. **92:** 77–96.

Hastie, T. J., and R. J. Tibshirani, 1990 *Generalized Additive Models*. Chapman & Hall, London.

Hastie, T., R. Tibshirani and J. Friedman, 2001 *The Elements of Statistical Learning*. Springer, New York.

Hayes, B., J. Laerdahl, D. Lien, A. Adzhubei and B. Høyheim, 2004 Large scale discovery of single nucleotide polymorphism (SNP) markers in Atlantic Salmon (Salmo salar). AKVAFORSK, Institute of Aquaculture Research (www.mabit.no/pdf/hayes.pdf).

Henderson, C. R., 1973 Sire evaluation and genetic trends, pp. 10–41 in *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush*. American Society of Animal Science and American Dairy Science Association, Champaign, IL.

Henderson, C. R., 1974 General flexibility of linear model techniques for sire evaluation. J. Dairy Sci. **57:** 963–972.

Henderson, C. R., 1975 Best linear unbiased estimation and prediction under a selection model. Biometrics **31:** 423–447.

Henderson, C. R., 1976 A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics **32:** 69–83.

Henderson, C. R., 1984 *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, ON, Canada.

Im, S., R. L. Fernando and D. Gianola, 1989 Likelihood inferences in animal breeding under selection: a missing data theory viewpoint. Genet. Sel. Evol. **21:** 399–414.

Jensen, C. S., and A. Kong, 1999 Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops. Am. J. Hum. Genet. **65:** 885–901.

Jensen, C. S., A. Kong and U. Kjaerulff, 1995 Blocking Gibbs sampling in very large probabilistic expert systems. Int. J. Hum. Comp. Stud. **42:** 647–666.

Kempthorne, O., 1954 The correlation between relatives in a random mating population. Proc. R. Soc. Lond. Ser. B **143:** 103–113.

Kerr, R. J., and B. P. Kinghorn, 1996 An efficient algorithm for segregation analysis in large populations. J. Anim. Breed. Genet. **113:** 457–469.

Kimeldorf, G., and G. Wahba, 1971 Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. **33:** 82–95.

Kruuk, L. E. B., 2004 Estimating genetic parameters in natural populations using the 'animal model'. Philos. Trans. R. Soc. Lond. B **359:** 873–890.

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

Mallick, B. K., D. Ghosh and M. Ghosh, 2005 Bayesian classification of tumours by using gene expression data. J. R. Stat. Soc. B **67:** 219–234.

Marron, J. S., 1988 Automatic smoothing parameter selection: a survey. Emp. Econ. **13:** 187–208.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H Teller and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1091.

Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2001 Is it possible to predict the total genetic merit under a very dense marker map? Genetics **157:** 1819–1829.

Nadaraya, E. A., 1964 On estimating regression. Theor. Probab. Appl. **9:** 141–142.

Patterson, H. D., and R. Thompson, 1971 Recovery of interblock information when block sizes are unequal. Biometrika **58:** 545–554.

Quaas, R. L., and E. J. Pollak, 1980 Mixed model methodology for farm and ranch beef cattle testing programs. J. Anim. Sci. **51:** 1277–1287.

Racine, J., and Q. Li, 2004 Nonparametric estimation of regression functions with both categorical and continuous data. J. Econom. **119:** 99–130.

Rubin, D. B., 1976 Inference and missing data. Biometrika **63:** 581–582.

Ruppert, D., M. P. Wand and R. J. Carroll, 2003 *Semiparametric Regression*. Cambridge University Press, Cambridge, UK.

Schucany, W. R., 2004 Kernel smoothers: an overview of curve estimators for the first graduate course in nonparametric statistics. Stat. Sci. **4:** 663–675.

Sheehan, N., and A. Thomas, 1993 On the irreducibility of a Markov chain defined on a space of genotype configurations by a sample scheme. Biometrics **49:** 163–175.

Silverman, B. W., 1986 *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.

Soller, M., and J. Beckmann, 1982 Restriction fragment length polymorphisms and genetic improvement. Proc. 2nd World Congr. Genet. Appl. Livestock Prod. **6:** 396–404.

Sorensen, D., and D. Gianola, 2002 *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer-Verlag, New York.

Sorensen, D. A., and B. W. Kennedy, 1983 The use of the relationship matrix to account for genetic drift variance in the analysis of genetic experiments. Theor. Appl. Genet. **66:** 217–220.

Sorensen, D., R. L. Fernando and D. Gianola, 2001 Inferring the trajectory of genetic variance in the course of artificial selection. Genet. Res. **77:** 83–94.

Stricker, C., M. Schelling, F. Du, I. Hoeschele, S. A. Fernandez *et al.*, 2002 A comparison of efficient genotype samplers for complex pedigrees and multiple linked loci. 7th World Congress of Genetics Applied to Livestock Production. INRA, Castanet-Tolosan, France. CD-ROM communication no. 21–12.

Ter Braak, C. J. F., M. Boer and M. Bink, 2005 Extending Xu's (2003) Bayesian model for estimating polygenic effects using markers of the entire genome. Genetics **170:** 1435–1438.

Van Arendonk, J. A. M., C. Smith and B. W. Kennedy, 1989 Method to estimate genotype probabilities at individual loci in farm livestock. Theor. Appl. Genet. **78:** 735–740.

Wahba, G., 1990 *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia.

Wahba, G., 1999 Support vector machines, reproducing kernel Hilbert spaces and the randomized GAVC, pp. 68–88 in *Advances in Kernel Methods*, edited by B. Schölkopf, C. Burges and A. Smola. MIT Press, Cambridge, MA.

Wang, C. S., J. J. Rutledge and D. Gianola, 1993 Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet. Sel. Evol. **25:** 41–62.

Wang, C. S., J. J. Rutledge and D. Gianola, 1994 Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. Genet. Sel. Evol. **26:** 91–115.

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

Watson, G. S., 1964 Smooth regression analysis. Sankhyā A **26:** 359–372.

Wong, G. K., B. Liu, J. Wang, Y. Zhang, X. Yang *et al.*, 2004 A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature **432:** 717–722.

Wright, S., 1934 The results of crosses between inbred strains of guinea pigs, differing in number of digits. Genetics **19:** 537–551.

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

Yi, N., G. Varghese and D. A. Allison, 2003 Stochastic search variable selection for identifying multiple quantitative trait loci. Genetics **164:** 1129–1138.

Zhang, Y., and S. Xu, 2005 A penalized maximum-likelihood method for estimating epistatic effects of QTL. Heredity **95:** 96–104.

Communicating editor: R. W. Doerge