Predicting complex traits using a diffusion kernel on SNP data with an application to dairy cattle and wheat breeding

Gota Morota*[1], Masanori Koyama[2], Guilherme J. M. Rosa[1,3], Kent A. Weigel[4], and Daniel Gianola[1,3,4]

[1]Department of Animal Sciences, University of Wisconsin-Madison, WI, USA

[2]Department of Mathematics, University of Wisconsin-Madison, WI, USA

[3]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, WI, USA

[4]Department of Dairy Science, University of Wisconsin-Madison, WI, USA

October 1, 2012

1 Absract

Arguably, genotypes and phenotypes may be linked in functional forms that are not well addressed by the linear and additive models that are standard in quantitative genetics. Therefore, developing statistical learning models for predicting phenotypic values from all available molecular information that are capable of capturing complex genetic network architectures is of great importance. Bayesian kernel ridge regression is a non-parametric prediction model proposed for this purpose. Its essence is to create a spatial distance-based relationship matrix called kernel. Although, the set of all SNP genotype configurations on which a model is built is finite, past research has mainly used a Gaussian

*morota@ansci.wisc.edu

kernel. We sought to investigate the performance of a diffusion kernel, which was specifically developed to model discrete marker inputs, using Holstein cattle and wheat data. The predictive ability of the diffusion kernel was similar to that of non-spatial distance-based additive genomic relationship kernels in the Holstein data, but outperforming the latter in the wheat data. However, the difference in performance between the diffusion and the Gaussian kernels was negligible. It is concluded that the ability of a diffusion kernel of capturing total genetic variance is not better than that of a Gaussian kernel, at least for these data. Although, the diffusion kernel as a choice of basis function may have potential for use in whole-genome prediction, our results imply that embedding genetic markers into a non-Euclidean metric space has very small impact on prediction.

2 Introduction

Prediction of yet-to-be observed phenotypes for complex quantitative traits in agricultural species [1, 2] or for disease status in medicine [3] exploits connections between phenotypes, genealogies and DNA variations potentially representing functional diversity of organisms. Systems biology approaches have uncovered abundant epistasis in model organisms including the mouse and the rat [4], *Drosophila melanogaster* [5], and *Saccharomyces cerevisiae* [6], and Loewe [7] proposed an evolutionary systems biology framework to arrive at a better understanding of molecular interactions, given that epistatic interactions between mutations are commonly observed. Therefore, it seems reasonable to argue that genotypes and phenotypes may be connected in forms that are not well addressed by the linear and additive models that are standard in quantitative genetics. Bayesian regularized parametric linear additive smoothers, e.g., [8, 9] may not be fully adequate for capturing genetic signals under epistatic scenarios [10, 11]. Further, attempts to account for epistasis by including interactions in a linear model produces a highly parameterized model structure, possibly yielding a poor predictive ability in cross-validation, which does not scale well if high-order interactions are included in the model.

Genetic risk prediction in medicine relies on using genomic information for predicting the chance of contracting a disease, e.g., in personalized medicine for preventive treatment and clinical health care. Prediction of genetic risk derived from pre-selected marker variants is mainstream in this domain, as opposed to prediction based on fitting whole genome markers simultaneously, as done with great success in animal and plant breeding [8, 9, 10, 11]. However, the variants so detected are typically not useful for genetic risk prediction because they explain only a small fraction of the total genetic variance as estimated from covariances between relatives, e.g., using twin and family studies. Moreover, it has been shown that a large number of variants that do not reach genome-wide statistical significance contribute to the total additive genetic variance [12].

Development of statistical models for predicting phenotypic outcomes from all available molecular information that are capable of capturing complex genetic network architectures is therefore important. Arguably, a good predictive model should account for most of the genetic variability, as well as reflecting underlying genetic architecture properly. Also, a predictive model should be flexible with respect to type of input data, e.g., high-throughput chip-based genotypes or whole genome sequences, and mode of gene action.

An appealing alternative is provided by a kernel-based parametric method known as BLUP (Best Linear Unbiased Prediction) of genetic effects, developed in the 40's-50's by C. R. Henderson, an animal breeder [13]. BLUP can also be viewed as a regression of a phenotype on a pedigree-based relationships matrix **A** (when the model is additive) and it has been used for genetic improvement of livestock species for decades. This method was recently extended to incorporate SNPs (Single Nucleotide Polymorphisms) by replacing **A** by some genomic relationship matrix **G** [14], although there is no impediment to using **A** and **G** together [15]. BLUP is suited for handling a massive amount of genetic information because the computational burden can be proportional to the number of data points rather than to the number of predictor variables (e.g., markers), and this is particularly so if a common weight is assigned to a each marker. Recently, kernel-based non-parametric models e.g., [15, 16, 17, 18] have been proposed. A non-parametric treatment can accommodate nonlinear dependence of phenotypes on predictor variables without explicit modeling. This suggests that these procedures can potentially pick up various forms of gene action without posing richly parametrized structures that require making strong distribution and genetic architecture assumptions a priori [10, 15]. For example, Long et al. [16] used a computer simulation and found that the predictive ability of a non-parametric smoother was superior to that of a parametric linear counterpart when non-additive effects were strong. These authors also gave evidence that non-parametric smoothing is competitive to linear smoothing even when additivity accounts for most of the total genetic variability.

Kernel ridge regression [19], a kernel generalization of standard ridge regression [20], is also a non-parametric smoothing method. Ridge regression has received some attention in quantitative genetics in the context of mixed linear models [10, 15, 21, 22, 23, 24], and the non-parametric version is carried out by constructing a spatial distance-based relationship matrix called kernel, as opposed to using additive genomic relationship kernels **A** or **G**, which only embed correlations due to additive genetic effects of individuals. The choice of a kernel is equivalent to modeling covariance structure among individuals, and phenotypes are regressed on this kernel to obtain estimates of non-parametric regression coefficients.

A simulation study [18] found that in the presence of non-additive effects, a spatial distance-based kernel can outperform an additive genomic relationship kernel in predictive performance, but this has not been explored enough with real data. Further, while the set of all SNP genotype configurations on which a model is built is finite, past research has employed spatial distance-based kernels with infinite, unbounded domains, such as the Gaussian. Our first objective in this study is to compare a spatial distance kernel with a non-spatial distance kernel. Secondly, we assess the performance of a non-Gaussian spatial distance kernel by deploying kernels on graphs as the choice of a basis function, a procedure that is suitable for discrete input data structures. Instead of encoding SNP data in a continuous Euclidean space, as it is the case of the Gaussian kernel, we investigated kernels on a non-Euclidean space. We examined a diffusion kernel proposed by Kondor and Lafferty [25], Smola and Kondor [26] and Lafferty and Lebanon [27], which is a kernel defined for functions on discrete spaces, such as a graph. A brief review on 'kernels on graphs' is given by [28] and "graph kernels" are discussed in [29]. As it is shown later, the diffusion kernel can be viewed as a discretization of the Gaussian kernel. We also tested the sensitivity of applying the same bandwidth parameter to autosomes and allosomes in the spatial distance kernels.

This paper investigates the use of several kinds of kernels in a kernel ridge regression framework for genome-assisted prediction of quantitative traits. Two data sets representing Holstein cattle and wheat were employed for this purpose. The paper is organized as follows. In section 2, we describe the data and introduce basic notions of kernel ridge regression. We then apply the diffusion kernel to strings of dummy variable marker sequences; the motivation of the non-Euclidean metric space is followed by an introduction of the diffusion kernel. In section 3, main results are presented. In section 4, we address the implication of results obtained and make concluding remarks.

3 Materials and Methods

3.1 Data

Dairy cattle and wheat data were used. The dairy data was provided by the USDA-ARS Animal Improvement Programs Laboratory (Beltsville, MD) and represented 7,902 Holstein bulls each with 43,134 SNPs (MAF $>$ 0.025) spanning across the whole genome. The target response variable analyzed was progeny test predicted transmitting ability (PTA) of productive life (PL). PL is a measure of the observed length of time that a cow stays in the herd, from first calving to culling, and PTA is an estimate of half of the breeding value of a bull, which is a smoothed average assuming additive inheritance. PL is lowly heritable, with heritability estimated at 0.1 [30]. The genotype for each of 42,438 loci on autosomes was coded as 0 (homozygous for allele "a"), 1 (heterozygous), and 2 (homozygous for allele "A"), according to the number of copies of the "A" allele. The remaining 696 loci on the X chromosome were coded as either 0 or 2, representing absence or presence of the "A" allele respectively. Missing genotypes, due to either low call rates for some SNPs or poor DNA quality, were imputed via random sampling of genotypes with probabilities corresponding to observed genotype frequencies at each locus.

The wheat data included 599 inbred lines collected by the International Maize and Wheat improvement Center in Mexico (CIMMYT). Each line was genotyped with 1279 Diversity Array Technology (DArt) markers generated by Triticarte Pty. Ltd. These Binary markers take the form of presence or absence of one of the two possible alleles. The phenotype here was average grain yield of each line in the first out of 4 environments represented in the data set, scaled to have zero mean and variance one. Missing genotypes were imputed as for the Holstein data above. This data set has been also analyzed with support vector regression and neural network methods [17, 31].

3.2 Kernel ridge regression

Our goal is to predict an unobserved response y, e.g., PL in $\mathbb{R}$ from a vector genotypes $\mathbf{x}$ at a large number of SNP loci; when p SNPs are considered, $\mathbf{x}$ is in $\mathbb{Z}_3^p$. To this end, we would like to establish a function $g : \mathbb{Z}_3^p \to \mathbb{R}$ mapping sequences of SNP genotypes onto the real line. A general setting is

$$y_i = g(\mathbf{x}_i) + \epsilon_i,$$

where y_i is a response variable on case $i(i = 1, 2, \cdots, n)$, $\mathbf{x}_i$ is $p \times 1$ vector of genotypes obtained on i, $g(\mathbf{x}_i)$ is a genetic effect interpretable as the conditional expectation function $g(\mathbf{x}_i) = E(\mathbf{y}_i | \mathbf{x} = \mathbf{x}_i)$, and ϵ_i is a residual.

We use kernel ridge regression to infer the unknown function g, and select an appropriate kernel $\mathcal{K}$ via a reproducing kernel Hilbert space $\mathcal{H}$ of functions on $\mathbb{Z}_3^p$, and optimize

$$\|\mathbf{y} - \mathbf{g}\|^2 + \lambda \|g\|^2_{\mathcal{H}}, \tag{3.1}$$

with respect to $\mathbf{g}$, where the first term is the residual sum of squares, and $\|g\|^2_{\mathcal{H}}$ is the squared norm of g under a Hilbert space; λ is a regularization parameter. The representer theorem [32] is used to find the optimal g.

In a non-parametric regression, the search space is infinite, but the representer theorem allows confining the search to a specific set of functions. It has been shown [10, 15, 24, 32] that the optimizer will be in the span of the functions indexed by the observed covariates, and that the problem simplifies to optimization of

$$\ell(\boldsymbol{\alpha} | \lambda) = \|\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}\|^2 + \lambda \|\mathbf{K}\boldsymbol{\alpha}\|^2_{\mathcal{H}}$$

where $\mathbf{K} = \{K(i,j) = K(x_i, x_j)\}$ is a $n \times n$ symmetric positive (semi) definite matrix; $\boldsymbol{\alpha}$ is an unknown $n \times 1$ vector of non-parametric regression coefficients; and $\mathbf{g} = \mathbf{K}\boldsymbol{\alpha}$, is the function that minimizes (3.1). By properties of a reproducing kernel, $\|\mathbf{K}\boldsymbol{\alpha}\|^2_{\mathcal{H}} = \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}$, so that the function to be minimized with respect to $\boldsymbol{\alpha}$ is

$$\ell(\boldsymbol{\alpha} | \lambda) = (\mathbf{y} - \mathbf{K}\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}. \tag{3.2}$$

This is equivalent to writing

$$\mathbf{y} = \mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

and then maximizing a penalized likelihood. This penalized likelihood is obtained by assuming that $y|\boldsymbol{\alpha}, \sigma_e^2$ and that $\boldsymbol{\alpha}$ follows $N(0, \mathbf{K}^{-1}\sigma_\alpha^2)$, where σ_e^2 is the variance of the residuals, and σ_α^2 is a variance component.

We review next additive genomic relationship kernels and the Gaussian kernel, and then present how one can build a kernel on a graph with discrete inputs. Hereafter, we denote $\mathbf{K}$ as the kernel matrix indexed by the observed covariate; and $K(i, j)$ indicates particular elements of $\mathbf{K}$; $\mathcal{K}$ is the infinite dimensional Gaussian kernel, or the $3^p \times 3^p$ dimensional kernel for the diffusion kernel.

3.3 Additive genomic relationship kernels

Two types of additive genomic relationship kernels were tested in this study. First, an additive genomic relationship matrix (**G1**) was constructed following VanRaden (2008)[14] as

$$\mathbf{G1} = \frac{\mathbf{ZZ}'}{2\sum p_j(1-p_j)},$$

where $\mathbf{Z} = Z_{ij}$ is a $n \times p$ matrix of centered SNP marker codes, with the entry for ith individual and the jth marker taking the form

$$Z_{ij} = \begin{cases} 0 - 2p_j & \text{if homozygous for “a”} \\ 1 - 2p_j & \text{if heterozygous} \\ 2 - 2p_j & \text{if homozygous for “A”}. \end{cases}$$

Here p_j is the frequency of allele “A” computed from a base population. The denominator of **G1** is a scaling parameter. In practice, the allele frequencies are estimated from the data at hand.

A second additive genomic relationship matrix ($\mathbf{G2}$) was as in Yang et al. (2010) [12]

$$\mathbf{G2} = \frac{\mathbf{W}\mathbf{W}'}{p}$$

where $\mathbf{W}$ is a matrix of standardized genotypes [33] with its jth column being

$$\mathbf{w}_{.j} = \frac{\mathbf{z}_{.j}}{\sqrt{2p_j(1-p_j)}},$$

where $\mathbf{z}_{.j}$ is the jth column of $\mathbf{Z}$ and p represents the number of SNPs.

Since the Holstein data set led to non-positive $\mathbf{G1}$ and $\mathbf{G2}$ matrices, as suggested by Strandén and Christensen [34], $\mathbf{G}(i = 1, 2)$ was modified to $\mathbf{G}_i^* = 0.95\mathbf{G}_i + 0.05\mathbf{I}$, yielding $\mathbf{G}^*$ matrices that provided valid kernels. The wheat data produced positive definite genomic relationship kernel matrices.

3.4 Gaussian Kernel

In a Gaussian kernel, the distance between a pair (i, j) of genotypes is represented as a squared Euclidean norm. Given a positive bandwidth parameter θ, the kernel takes the form

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}_j) &= \exp(-\theta d_{ij}^2) \\ &= \prod_{k=1}^{p} \exp(-\theta(x_{ik} - x_{jk})^2) \end{aligned}$$

where $d_{ij}^2 = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ik} - x_{jk})^2 + \cdots + (x_{ip} - x_{jp})^2}$, and x_{ik} $(i, j = 1, \cdots, n, k = 1, \cdots, p)$ is the SNP genotype for individual i at SNP k. A small Euclidean distance between two individuals reflects a strong similarity in state between their genotypes. As θ increases, the kernel evaluation approaches $K(\mathbf{x}_i, \mathbf{x}_j) = 0$, producing a “sharp” or “local” kernel. On the other hand, as $\theta \to 0$, the kernel approaches 1, i.e., a situation where the two individuals “match” perfectly, providing a “global” kernel.

3.5 Non-Euclidean metric space

The SNP data on p loci on some individual often comes as $\mathbf{x} = (x_1, x_2, \ldots, x_p) \in \mathbb{Z}_3^p$, which is clearly a discrete space, as there are 3^p possible configurations of genotypes (not all of which are observable). Before defining the diffusion kernel consider the meaning of 'diffusion on a graph'. Suppose $p = 1$, and consider a function k_x that measures the spread of 'influence' of the genotype at this locus over the other possible genotypes by assuming that the 'influence' diffuses like heat does. Let $k_{\tilde{x}}(0, x) = 1_{x=\tilde{x}}(x)$, be the indicator function for genotype $\tilde{x}$ on $\mathbb{Z}_3$. We call this the time 0 diffusion, since in this case $\tilde{x}$ has absolutely no influence on other genotypes; that is, the influence of $\tilde{x}$ does not diffuse out to its neighbors. Now, define the time t diffusion of the 'influence' of genotype $\tilde{x}$ on genotype x to be

$$k_{\tilde{x}}(t, x) = k_{\tilde{x}}(t-1, x) + \sum_{|x-x'|=1} \alpha[k_{\tilde{x}}(t-1, x') - k_{\tilde{x}}(t-1, x)] \tag{3.3}$$

where α is constant rate of diffusion and each summand is the differential gradient of the 'influence' between genotypes x and x'. This is illustrated in Table 1. As stated above, there is no diffusion at $t = 0$. Subsequently, the time 1 diffusion with $\alpha = 0.1$ when $\tilde{x} = 1$ is computed as:

$$\begin{aligned}
k_1(1, x=0) &= k_1(0, x=0) + \alpha[k_1(0, x^{'}=1) - k_1(0, x=0)] \\
&= 0 + 0.1[1-0] \\
&= 0.1 \\
k_1(1, x=1) &= k_1(0, x=1) + \alpha[k_1(0, x^{'}=0) - k_1(0, x=1)] + \alpha[k_1(0, x^{'}=2) - k_1(0, x=1)] \\
&= 1 + 0.1[0-1] + 0.1[0-1] \\
&= 0.8 \\
k_1(1, x=2) &= k_1(0, x=2) + \alpha[k_1(0, x^{'}=1) - k_1(0, x=2)] \\
&= 0 + 0.1[1-0] \\
&= 0.1
\end{aligned}$$

As shown in Table 1, as t increases the 'influence' spreads over all genotypes more evenly; also, the larger α is, the faster the diffusion is with respect to time t.

Writing (3.3) in vector form, with $\mathbf{k}_{\tilde{x}}(t, x) = \mathbf{k}_{\tilde{x}}(t)$, we get

$$\begin{aligned}\mathbf{k}_{\tilde{x}}(t) &= \mathbf{k}_{\tilde{x}}(t-1) + \alpha\mathbf{H}\mathbf{k}_{\tilde{x}}(t-1) \\ &= (\mathbf{I} + \alpha\mathbf{H})\mathbf{k}_{\tilde{x}}(t-1) \qquad (3.4) \\ &= (\mathbf{I} + \alpha\mathbf{H})^t\mathbf{k}_{\tilde{x}}(0)\end{aligned}$$

were $\mathbf{I}$ is a 3×3 identity matrix; $\mathbf{k}_{\tilde{x}}(0)$ is a constant 3×1 matrix of initial values, and

$$\mathbf{H} = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix} \qquad (3.5)$$

with the first, second, and third rows of the $\mathbf{H}$ matrix corresponding to k_0, k_1, and k_2 respectively. The negative of this matrix is called the Laplacian of a graph Γ given by

$$0 - 1 - 2. \qquad (3.6)$$

Let Γ be an undirected graph with vertex set $V(\Gamma)$. In general, the Laplacian of a graph Γ, $L(\Gamma)$ is a $V(\Gamma)$ dimensional square matrix given by

$$\begin{aligned}\mathbf{L}(\Gamma) &= -\mathbf{H}(\Gamma) \\ &= -\mathbf{A}(\Gamma) + \mathbf{\Lambda}\end{aligned}$$

where $\mathbf{A}$ is an adjacency matrix and $\mathbf{\Lambda}$ is a diagonal matrix with $\Lambda_{ii} = \sum_{j=1}^{n} A_{ij}$. We can therefore generalize this 'diffusion' for any graph Γ by using $\mathbf{H}(\Gamma) = -\mathbf{L}(\Gamma)$. Under this definition, given any $V(\Gamma)$ dimensional vector $\mathbf{w}$,

$$\mathbf{w}^t\mathbf{H}(\Gamma)\mathbf{w} = -\sum_{i \sim j}(w_i - w_j)^2 \leq 0$$

which shows that $\mathbf{H}(\Gamma)$ is a negative semi-definite matrix.

The most naive way of constructing a graph on $\mathbb{Z}_3^p$ is a Hamming graph. For the case $p = 1$, a Hamming graph is simply a complete graph of size 3, and has the form

$$\begin{array}{ccc} 0 & - & 1 \\ \backslash & & / \\ & 2 & \end{array} \tag{3.7}$$

On this graph, the distance from genotype 0 ('aa') to genotype 2 ('AA') is the same as that from 0 ('aa') to 1 ('Aa'). Since genotype 'aa' has no copies of the 'A' allele, it may be more reasonable to assume that genotype 'Aa' is closer to 'AA', which has two copies of the 'A' allele. This can be viewed from a mutational perspective as well. Genotype 0 ('aa') requires two mutations to become genotype 2 ('AA'), while genotype 1 ('Aa') requires only one mutation. Thus, the graph of interest would be given by (3.6). The latter is a path graph for SNP data, which will be taken as a minimal basis for our graph. In a path graph, all vertices are on a straight line, as in (3.6).

A SNP grid of p loci is a p dimensional grid with vertices in $\mathbb{Z}_3^p$, with two vertices $\mathbf{x}$ and $\mathbf{x}'$ being adjacent if and only if

$$\sum_{i=1}^{p} |x_i - x_i'| = 1.$$

For example, the graph below is the grid for 2 loci derived from the Cartesian graph product of two path graphs as in (3.6):

$$\begin{array}{ccccc} 02 & - & 12 & - & 22 \\ | & & | & & | \\ 01 & - & 11 & - & 21 \\ | & & | & & | \\ 00 & - & 10 & - & 20 \end{array} \tag{3.8}$$

The graph Laplacian for graph (3.8) is a square matrix of dimension $3^2 \times 3^2$:

$$
\begin{aligned}
\mathbf{L}(\Gamma) &= -\mathbf{H}(\Gamma) \\
&= \begin{bmatrix}
2_{00} & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
-1 & 3_{01} & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & -1 & 2_{02} & 0 & 0 & -1 & 0 & 0 & 0 \\
-1 & 0 & 0 & 3_{10} & -1 & 0 & -1 & 0 & 0 \\
0 & -1 & 0 & -1 & 4_{11} & -1 & 0 & -1 & 0 \\
0 & 0 & -1 & 0 & -1 & 3_{12} & 0 & 0 & -1 \\
0 & 0 & 0 & -1 & 0 & 0 & 2_{20} & -1 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & -1 & 3_{21} & -1 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 2_{22}
\end{bmatrix}
\end{aligned}
$$

where the subscripts denote the vertices of graph (3.8). When there are p loci, the p-dimensional grid graph has 3^p vertices corresponding to sequences of genotypes, such that two vertices are adjacent if and only if just one SNP locus differs by 1. Now, suppose $p = 3$. The Cartesian graph product of (3.6) and (3.8) yields a 3 dimensional grid graph with 3^3 vertices, as shown in Figure 1. The diffusion kernel computes a similarity between two vertices on this graph, and projects this information into a more interpretable space.

3.6 Diffusion Kernel on a non-Euclidean metric space

Consider now the continuous analogue of the diffusion scheme above. This can be done by making 'time' or 'space' continuous, and 'time' will be made continuous first. Let $\alpha = \theta h$ ($\theta > 0$) and $t = 1/h$. By using a small h, we can achieve a discretization of the 'diffusion time' on a much finer scale, and the coefficient matrix is

$$(\mathbf{I} + \theta h \mathbf{H}(\Gamma))^{1/h} \tag{3.9}$$

If an infinitesimal scale is considered by taking $h \to 0$, (3.9) converges to

$$\begin{aligned}\lim_{h\to 0}\left(\mathbf{I}+\theta h\mathbf{H}(\Gamma)\right)^{1/h} &= \exp(\theta\mathbf{H}) \\ &= \sum_{k=0}^{\infty}\frac{\theta^k}{k!}\mathbf{H}^k = \mathbf{I}+\theta\mathbf{H}+\frac{\theta^2}{2}\mathbf{H}^2+\frac{\theta^3}{3!}\mathbf{H}^3+\cdots+\frac{\theta^n}{n!}\mathbf{H}^n+\cdots\end{aligned} \tag{3.10}$$

If a graph Γ with a Laplacian $\mathbf{L}(\Gamma)$ is considered, then $\exp(-\theta\mathbf{L}(\Gamma))$ is called the diffusion kernel or heat kernel for graph Γ, where θ is a rate of diffusion [25]. Here putting $\mathcal{K} = \exp(\theta\mathbf{H})$ and taking the derivative with respect to θ gives,

$$\frac{d}{d\theta}\mathcal{K} = \mathbf{H}\mathcal{K} \tag{3.11}$$

which is a discrete diffusion equation (heat equation) on a graph with $\mathbf{H} = -\mathbf{L}(\Gamma)$. Note that diffusion kernels always need to be associated with a graph.

A Gaussian kernel is obtained by making this diffusion kernel ”space” continuous. The connection between the two kernels is provided in Appendix A.

3.7 Diffusion Kernel indexed by observed covariates

When a graph Γ is large and asymmetric, the computation of the diffusion kernel $\mathcal{K}(\Gamma)$ can be forbiddingly hard. For instance, for a SNP grid with 43134 loci, the dimension of $\mathcal{K}$ is 3^{43134} by 3^{43134}. Symmetry helps, however. If a closed form of $\mathcal{K}$ can be arrived at, there is no need to compute $\mathcal{K}(\mathbf{x},\mathbf{x}')$ for all pairs of genotype sequences $\mathbf{x},\mathbf{x}'$. This is indeed the case for the Gaussian kernel, where the dimension of $\mathcal{K}$ is infinite. With Kondor and Lafferty's result given in [25], we may obtain the closed form of the diffusion kernel from the sample for our SNP grid.

First one needs to consider the Cartesian graph product for the diffusion kernel of a graph. Let $\mathcal{K}_1(\theta)$ and $\mathcal{K}_2(\theta)$ be the kernels for graphs Γ_1 and Γ_2, respectively. The diffusion kernel for $\Gamma = \Gamma_1 \square \Gamma_2$ is given by [25]

$$\mathcal{K}_1(\theta)\otimes\mathcal{K}_2(\theta). \tag{3.12}$$

were $\Box$ denotes the Cartesian graph product and $\otimes$ is the tensor product (infinite dimensional Kronecker product). Consider a graph with one locus, Γ_0, with form $0-1-2$. Then, we see that the diffusion kernel of the SNP grid on p loci with bandwidth parameter θ is given by

$$\mathcal{K}_{\theta}^{\otimes p} = \bigotimes_{i=1}^{p} \mathcal{K}_{\theta}(\Gamma_0).$$

To this end, we just need to compute $\mathcal{K}_{\theta}(\Gamma_0) = \exp(\theta \mathbf{H})$ with $\mathbf{H}$ in (3.5).

With this result, one can create the $\mathbf{H}$ matrix for a SNP grid as follows. Let $\mathbf{x}$ and $\mathbf{x}'$ be SNP data for p loci; n_s be the number of loci for which $|\mathbf{x}_i - \mathbf{x}_i'| = s$, and m_{11} be the number of loci for which $\mathbf{x}_i = \mathbf{x}_i' = 1$. In other words, n_1 is the number of loci at which two individuals differ by 1, and m_{11} is the number of loci at which two individuals share heterozygous states. Then

$$K_{\theta}^{snpgrid}(\mathbf{x}, \mathbf{x}^{'}) \propto \left(\frac{-2e^{-3\theta}+2}{e^{-3\theta}+3e^{-\theta}+2}\right)^{n_1} \left(\frac{e^{-3\theta}-3e^{-\theta}+2}{e^{-3\theta}+3e^{-\theta}+2}\right)^{n_2} \left(\frac{4e^{-3\theta}+2}{e^{-3\theta}+3e^{-\theta}+2}\right)^{m_{11}} \tag{3.13}$$

with proportionality constant $(e^{-3\theta} + 3e^{-\theta} + 2)^q$, where $q = n_1 + n_2 + m_{11}$. The last term is a contribution from heterozygosity. We refer to this as SNP grid kernel, specifically developed to model SNP data in this study. A proof of this result is given in the Appendix B.

3.8 Diffusion kernel for binary genotypes

Another diffusion kernel tailored for binary genotypes is required for chromosome X of sires or for the wheat inbred lines. In this setting, instead of (3.6), the path graph for one locus ($p = 1$) is

$$0 - 2$$

and the corresponding graph Laplacian is given by

$$\begin{aligned}\mathbf{L}(\Gamma) &= -\mathbf{H}(\Gamma)\\ &= \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\end{aligned} \tag{3.14}$$

as opposed to (3.5). For two loci ($p = 2$), the Cartesian product of graphs $\Gamma_1(0 - 2)$ and $\Gamma_2(0 - 2)$ yields the graph

$$\begin{array}{ccc} 00 & - & 01 \\ | & & | \\ 10 & - & 11 \end{array} \tag{3.15}$$

where the first digits $\in V(\Gamma_1)$ and the second digits $\in V(\Gamma_2)$. Then, the associated graph Laplacian is

$$\begin{aligned} \mathbf{L}(\Gamma) &= -\mathbf{H}(\Gamma) \\ &= \begin{bmatrix} 2_{00} & -1 & -1 & 0 \\ -1 & 2_{01} & -1 & 0 \\ -1 & 0 & 2_{10} & -1 \\ 0 & -1 & -1 & 2_{11} \end{bmatrix} \end{aligned}$$

where the subscripts denote the rows and columns of vertices of graph (3.15). Specifically, we compute $\mathbf{K}_\theta = \exp(\theta\mathbf{H})$ with $\mathbf{H}$ defined in (3.14) and perform the tensor product p times. With this, the kernel is given by

$$K_\theta^{hypercube}(\mathbf{x}, \mathbf{x}') \propto \left(\frac{1 - \exp(-2\theta)}{1 + \exp(-2\theta)}\right)^{d(\mathbf{x},\mathbf{x}')} \tag{3.16}$$

where $d(\mathbf{x}, \mathbf{x}')$ is the Hamming distance, that is, number of coordinates at which $\mathbf{x}$ and $\mathbf{x}'$ differ [25]. Following Kondor and Lafferty [25], this diffusion kernel for binary markers will be referred to as the hypercube kernel.

3.9 Combining SNP grid kernels and hypercube kernels

In the Holstein data, we additionally combined the two kernels derived from autosomes and from chromosome X to see the influence of applying a same value of the bandwidth parameter to different types of chromosomes. This is giving by

$$\mathbf{K}^{all} = \mathbf{K}^{snpgrid} \# \mathbf{K}^{hypercube}. \tag{3.17}$$

where # is a Hadamard product of matrices. In general, given a set of n individuals, we may partition SNPs into several subsets, say $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_r)$. If $\mathbf{K}^i$ is the diffusion kernel corresponding to subset $\mathbf{x}_i$, then the diffusion kernel for all sets can be computed as

$$\mathbf{K}^{all} = \mathbf{K}^1 \# \mathbf{K}^2 \# \cdots \# \mathbf{K}^r.$$

This result also holds for the Gaussian kernel, but not necessarily so for every kernel, e.g., the exponential kernel defined with the Euclidean distance ($||\mathbf{x}_i - \mathbf{x}_j||$) does not hold this property.

3.10 Bayesian treatment of kernel ridge regression

Once the choice of the kernel is determined, (3.2) can be maximized by taking the derivative of $\ell(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ to obtain

$$\hat{\boldsymbol{\alpha}} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y}$$

where λ is a regularization parameter. Here, implementation of kernel ridge regression was casted in a Bayesian framework with $\lambda = \frac{\sigma_\epsilon^2}{\sigma_\alpha^2}$, where σ_ϵ^2 and σ_α^2 are the residual variance and the variance attached to $\boldsymbol{\alpha}$ respectively. Then [35, 36], note that

$$\begin{aligned} \exp(-\frac{1}{2}\ell(\alpha)) &= \exp\left\{-\frac{1}{2}\left[(\mathbf{y} - \mathbf{K}\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}\right]\right\} \\ &\propto \exp\left(-\frac{1}{2\sigma_\epsilon^2}(\mathbf{y} - \mathbf{K}\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{K}\boldsymbol{\alpha})\right) \exp\left(-\frac{1}{2\sigma_\alpha^2}\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha}\right) \end{aligned}$$

This is proportional to $p(\boldsymbol{\alpha}|\mathbf{y}, \sigma_e^2, \sigma_\alpha^2) \propto p(\mathbf{y}|\boldsymbol{\alpha}, \sigma_e^2)p(\boldsymbol{\alpha}|\sigma_\alpha^2)$, i.e., the posterior density of $\boldsymbol{\alpha}$ (given σ_e^2 and σ_α^2) for the linear model

$$\mathbf{y} = \mathbf{K}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

with $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}\sigma_e^2)$ and with prior $\boldsymbol{\alpha} \sim N(0, \mathbf{K}^{-1}\sigma_\alpha^2)$. Minimizing $\ell(\boldsymbol{\alpha})$ will maximize $\exp(-\frac{1}{2}\ell(\boldsymbol{\alpha}))$, so $\hat{\boldsymbol{\alpha}}$ is the conditional posterior mode of $\boldsymbol{\alpha}$. One may change the basis $\mathbf{K}$ using the eigenvalue decomposition $\mathbf{K} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}'$, where $\boldsymbol{\Lambda}$ is the matrix of eigenvectors of $\mathbf{K}$ and $\boldsymbol{\Psi}$ is a diagonal matrix whose diagonals are the eigenvalues, as shown in de los Campos et al. [35], such that, for $\boldsymbol{\delta} = \boldsymbol{\Psi}\boldsymbol{\Lambda}'$ one gets, in a fully Bayesian model,

$$\begin{cases} \mathbf{y} = \boldsymbol{\Lambda}\boldsymbol{\delta} + \boldsymbol{\epsilon}, \\ p(\boldsymbol{\epsilon}, \boldsymbol{\delta}, \sigma_\epsilon^2, \sigma_\alpha^2) \propto N(\boldsymbol{\epsilon}|0, \mathbf{I}\sigma_\epsilon^2)N(\boldsymbol{\delta}|0, \boldsymbol{\Psi}\sigma_\alpha^2)p(\sigma_\epsilon^2, \sigma_\alpha^2) \end{cases}$$

Once a prior is assigned to σ_e^2 and σ_α^2, a MCMC scheme can be used to infer all unknown parameters, including λ. Scaled inverse chi-square prior distributions were assigned to σ_e^2 and σ_α^2, each with 3 degrees of freedom and a scale parameter equal to 1. Samples from posterior distributions were obtained by the Gibbs sampler in [35], and each of the analysis was based on 100,000 MCMC samples with the first 60,000 samples discarded as burn-in. After burn-in, samples were thinned at a rate of 10, resulting in 4000 mildly correlated samples for posterior inference. Convergence was monitored by inspecting trace plots of each of the two variance parameters. A bandwidth parameter θ yielding high predictive ability is needed as well. However, sampling of the bandwidth parameter in MCMC sampling requires computation of kernels at each iteration, which is very demanding given the number of individuals and SNPs considered in our study. For this reason, evaluation of the diffusion kernel was performed over a fixed grid of values of θ. The range of θ considered provides average values of $K(\mathbf{x}, \mathbf{x}^{'})$ that were evenly spaced, approximately, between 0.13 to 0.99. Computation of kernels and Gibbs sampling was carried out in Fortran and in R (http://www.R-project.org), respectively.

3.11 Assessment of predictive ability

The predictive abilities of RKHS models with either a diffusion kernel or a Gaussian kernel were assessed by cross-validation. A subset of 5403 bulls born from 1952 through 2003 was used as training set for the Holstein data. A testing set of 2499 bulls born from 2004 through 2006 was used to evaluate predictive ability. For the wheat data, a 10 fold cross-validation scheme was applied by assigning 599 lines randomly to one of 10 disjoint subsets. Each set was used for validation in turn, while the other 9 subsets were used to train the model. To illustrate, we estimated $\boldsymbol{\alpha}$ in the Holstein data using the training set $\mathbf{y} = (y_1, \cdots, y_{5403})'$ and their corresponding SNP genotypes $\mathbf{x}_1, \cdots, \mathbf{x}_{5403}$, and then predicted responses in the testing set as:

$$\hat{\mathbf{y}}^{test} = \mathbf{1}\hat{\mu}^{train} + \mathbf{K}^{test \leftrightarrow train}\hat{\boldsymbol{\alpha}}^{train}$$

where $\hat{\mathbf{y}}^{test}$ is the 2499 × 1 vector of predicted responses of bulls in the testing set; $\mathbf{1}$ is a 2499 × 1 vector of ones; $\hat{\mu}^{train}$ is the posterior mean of the intercept estimated from the training set; $\mathbf{K}^{test \leftrightarrow train}$ is a 2499 × 5403 matrix with elements $k(j,i)^{test \leftrightarrow train}$ representing the allelic similarity between bulls in the testing $(j = 1, \cdots, 2499)$ and training $(i = 1, \cdots, 5403)$ sets, with the same bandwidth parameter employed in the training set, and $\hat{\boldsymbol{\alpha}}^{train}$ is the vector of posterior means of 5403 non-parametric regression coefficients obtained from the training set. In the equation above, $\mathbf{K}^{test \leftrightarrow train}$ was either the diffusion or the Gaussian kernel.

In a Bayesian setting, however, one can embed all steps above in a convenient way. Prior to Gibbs sampling, first we construct a full kernel matrix containing both training and testing data sets. We treat the responses of testing set individuals as unobserved, and these values are predicted via a predictive distribution. This is easy to incorporate in the Gibbs sampling scheme. Pearson's correlation between the predicted values (mean of the predictive distribution) and the observed PTA, Cor($\hat{\mathbf{y}}^{test}, \mathbf{y}^{PTA}$), and predictive mean-squared error (MSE) defined as $\sum_{i=1}^{2499}(\hat{y}_i^{test} - y_i^{PTA})^2/n$ were computed to evaluate the predictive ability of the two kernels. Here, $\hat{y}_i^{test}$ is the mean of the predictive distribution of response i in the testing data set, which is the ith element of the $\mathbf{K}^{test \leftrightarrow train}\hat{\boldsymbol{\alpha}}^{train}$.

4 Results

To illustrate the effect of the bandwidth parameter (θ) on the SNP grid kernel, Figure 2 depicts histograms showing how θ controls similarities among individuals based on evaluating the kernel on the SNP data. The larger θ is, the stronger the prior inter-correlation structure. It is important to note that the diagonal elements in our SNP grid kernel matrices are not necessary equal to one, as opposed to what happens in a Gaussian kernel; here, $\mathbf{K}$ is a correlation matrix. Table 2 shows the average of diagonal, $K(x_i, x_i)$, and off-diagonal, $K(x_i, x_j)$, elements for diffusion, Gaussian and two additive genomic relationship kernels at varying bandwidth values. The mean values of the diagonal elements of the 4 diffusion kernels shown in Figure 2 (see Table 2) were 0.369, 0.693, 0.874, and 0.952 for $\theta = 10, 11, 12$, and 13, respectively. This is because in equation (3.13), even when $\mathbf{x} = \mathbf{x}'$, so that $n_1 = n_2 = 0$, m_{11} (the number of 'Aa' genotypes shared by $\mathbf{x}$ and $\mathbf{x}'$) may not be zero. This implies that our diffusion kernel accounts for the degree of heterozygosity in a sample. From the perspective of the kernel as a smoothing function, the diffusion kernel performs smoothing for all elements based on heterozygosity as well as on allelic similarity. As explained below, the larger heterozygosity, the weaker the smoothing, leading to a smaller penalty; this is not so, however, in the Gaussian kernel. In the kernel computation, each factor in (3.13) is < 1, and in particular, the factor corresponding to m_{11} is the largest. Henceforth, if the sample contains few heterozygotes, our $\mathbf{K}$ will be large in value. Consequently, the penalty from the optimizer function f, $||f_{\mathcal{H}}|| = \alpha^T \mathbf{K} \alpha$, will tend to be big. This is interpretable as imposing stronger smoothing for samples with low heterozigosity. As for the "correlation" with itself, an individual with low heterozygosity will have diagonal elements close to one, as in the case of a Gaussian kernel. Therefore, in addition to the 'distance' between genotypes of two individuals, the diffusion kernel takes into account the extant heterozygosity, while the Gaussian kernel incorporates only the former. Also, the two kernels differ in their definition of distance. The diffusion kernel on the SNP grid is based on the Manhattan distance, while the Gaussian kernel is defined on the Euclidean distance. The Manhattan distance is the distance between two points measured by the the sum of the absolute differences of their coordinates.

As shown in Table 2, the average of off-diagonal elements of the diffusion kernel was lower than that of diagonal elements. This is because the first two terms of (3.13) will be different from zero ($n_1, n_2 > 0$) for a pair of individuals. Diffusion kernel evaluations between itself were always larger than kernel evaluated between pairs, that is, diagonal elements had the largest values for each row of **K**. In the Gaussian kernel, diagonal elements are always equal to 1 and a smaller θ value produces a stronger prior correlation. The first type of additive genomic relationship kernel (**G1**) had the average diagonal and off-diagonal elements close to 1 and 0 respectively, as expected. Similarly, **G2** had an average off-diagonal close to 0 but it had smaller average diagonal elements than those of **G1**.

The right most columns of Table 2 gives the evaluation of the predictive ability of the kernels measured as correlation between predicted values and observed PTA, and MSE of prediction, for several different bandwidth parameters (**G1** and **G2** do not involve this parameter). The predictive correlation of the diffusion (SNP grid) kernel was best at $\theta = 11$, while with the Gaussian kernel this was achieved at $\theta = 10^{-5}$. Although the averages of diagonal and off-diagonal elements varied substantially with different bandwidth parameters in the diffusion and Gaussian kernels, the influence of this variability on predictive correlations was small. Importantly, no major difference was observed between the diffusion and the Gaussian kernels in terms of predictive performance. Differences among kernels were very minor, probably due to the fact that the response (PTA) is already a smoothed mean based on a large number of daughters of a bull. There was a consistency between the correlation and the MSE, in the sense that the scale of θ with the highest predictive correlation had the smallest MSE. Predictive performance of **G1** was only slightly worse than that of the spatial distance kernels with the best bandwidth parameters.

Values in parentheses in Table 2 were obtained by combining the SNP grid kernel from autosomes and the hypercube kernel from allosomes by applying the same bandwidth parameter. Incorporation of X-chromosome information reduced the average off-diagonal elements slightly, and deteriorated predictive performance to some extent. On the other hand, the average diagonal and off-diagonal elements remained the same in **G1** and **G2**, but a minor reduction of their predictive abilities was observed.

In the wheat data, the superiority of the spatial distance-based kernels over the additive genomic relationship kernels was clear. Table 3 indicates that the diffusion and Gaussian kernels had the best predictive correlations (MSE) at 0.586 (0.685) and 0.582 (0.686), respectively, whereas those of **G1** and **G2** were 0.518 (0.709) and 0.521 (0.708). This is likely due to picking up non-additive genetic variation that this wheat data harbors. With binary markers, the diagonal elements of the diffusion kernel are always 1, since in equation (3.16) the Hamming distance $d(\mathbf{x}, \mathbf{x}')$ is always zero. As with the Holstein data, no apparent difference was observed between the diffusion and the Gaussian kernels.

5 Discussion

Arguably, relationships between phenotypes and genotypes are non-linear and complex [10, 15, 31]. For this reason, ignoring non-additive effects such as dominance and epistasis in a model may lead to an inferior predictive ability of individual phenotypes. A spatial distance-based kernel non-parametric regression is capable of mapping genotypes to phenotypes in a way that accurately reflects underlying, albeit unknown, relationships. These kernel methods incorporate non-linearity of a predictor set $\mathbf{x}$ through a nonlinear transformation of $\mathbf{x}$, subsequently allowing to analyze the response in terms of features $\phi(\mathbf{x})$ in a linear way. This is particularly useful when a response has a linear relationship with respect to the parameters, but is non-linear on covariates, such as in the case of polynomial regression.

The predictive ability of kernel-based genetic models depends on the choice of a kernel and associated bandwidth parameter(s). If the two data points lie in the real line, $x, x^{'} \in \mathbb{R}$, it seems reasonable to compute their distance in terms of Euclidean distance. However, SNP genotypes, coded as dummy variables, take a discrete form. Therefore, it may be worthwhile to consider a kernel designed to capture the discrete structure of the input variables. The best predictive kernel and its performance may vary depending on the underlying genetic architecture, QTL numbers and distribution of effects, data set used and kernel method applied. Here, we investigated the use of ridge regression with a diffusion kernel to assess if this would enhance predictive ability over that of the Gaussian kernel and of two additive genomic relationship counterparts. Kondor and Lafferty [25] obtained promising results when the diffusion kernel was compared with several kernels in classification problems with a set of discrete predictors, and this kernel has been used in a microarray based gene function prediction problem [37]. Ober et al. [18] used the Matérn covariance function, which contains the Gaussian and the exponential kernel as particular cases. Therein, the smoothing parameter controls the actual form of a kernel, and this is directly driven by sample data. Although they obtained a Gaussian form as a choice of the covariance function, the Matérn function is bounded by the Euclidean norm by definition, which may not be suited for discrete genomic data.

A strength of kernels for structured data is their ability of addressing similarities between two data points $x, x' \notin \mathbb{R}$ [38]. The diffusion kernel defines the distance between two data points on graphs, namely vertices, and projects this information into a more interpretable space. As shown in the context of modeling linkage disequilibrium [39, 40], various graph structures can be used to represent sets of discrete random variables, such as genotypes. Coupled with the representer theorem, the diffusion kernel allows casting underlying graph structures into a regression on the real line under a Hilbert space. The main idea behind this kernel is the matrix exponentiation of the graph Laplacian. The p-dimensional grid graph with vertices representing a vector of genotypes was chosen for the graph structure. Each grid conveys information on similarity in terms of the Manhattan distance. Two vertices $\mathbf{x}$ and $\mathbf{x}'$ are connected if $x_i = x_i'$ for all i, except at one coordinate. In the Holstein data, with $n = 7902$ and $p = 42438$, it is unlikely that any of two vertices present in our data are connected. However, what grid graphs embrace is how many "steps" separate a vector of genotypes observed in individual i from an observed vector of genotypes in individual j.

Our motivation of applying the diffusion kernel stemmed from the assumption that a non-Euclidean distance may be able to more clearly represent genomic similarities. We carried out a matrix exponentiation of two graph Laplacians created from two path graphs (one for SNPs and one for binary markers) for this purpose. This yields a kernel based on the Manhattan distance accounting for the heterozygosity that two individuals share. The two spatial distance kernels resulted in a better predictive performance than the two additive genomic relationship kernels in the wheat data. This agrees with the previous simulation study of Ober et al. [18], in which the Gaussian kernel outperformed **G1** in the presence of non-additive effects. Superiority of the spatial distance kernels was less obvious in the Holstein data. This may be due to the phenotype we chose for this study, since the PTA response variable is a smoothed average using linear mixed models.

As for difference between the diffusion and the Gaussian kernels in terms of predictive ability, the diffusion kernel had the highest predictive correlation and the lowest MSE with $\theta = 11$ in the Holstein data, but the difference with the Gaussian kernel was negligible. The same result was seen in the wheat data. This implies that the Gaussian kernel is robust, even it incorporates genotypes on the real line such as 1.25 or -12.3. Our objective of properly incorporating genotypes into a kernel had a small impact on predictive ability of yet-to-be observed phenotypes. Although certainly the distance between genotypes is not continuous, additional efforts of discretizing the

Euclidean distance may not needed. Another possible reason, might that genotypes do not reside in the Euclidean or in the non-Euclidean spaces explored here, but in a manifold [27].

Incorporation of X chromosome genotypes for building a kernel led to a smaller average diagonal and off-diagonal elements (to some extent) in spatial distance kernels, but no change was observed in the additive genomic relationship kernels. In both types of spatial distance kernels, however, the predictive correlations were worse than when kernels were constructed purely from autosomes. This suggests that applying specific bandwidth parameters for autosomes and allosomes in the spatial distance kernels might be important. A similar decline of predictive performance was observed in the two additive genomic relationship kernels, which do not involve any bandwidth parameter. Further research is need to investigate what produces this drop in predictive performance, although if no markers contribute to PL on chromosome X, this would add extra noise.

To the best of our knowledge, this study involves one of the largest data sets employed for spatial kernel-based genome-enabled selection of agricultural species. The challenge here was the computation of the diffusion kernel, rather than the Gibbs sampler. Approximately, it took 4 days to compute one diffusion kernel on a Linux workstation equipped with the Intel(R) Xeon(R) CPU E5450 3.00GHz and 16GB of RAM. The Gaussian kernel required slightly less time for building, but with several candidates over a grid of values of the bandwidth parameter θ, this was an expensive task for both kernels. One useful approach might be that of multiple kernel learning (MKL) [35, 41], which uses a few kernels with different covariance structure in a single RKHS model. Finally, the SNP grid graph and the hypercube graph used in this study are naive graph structures for modeling discrete inputs. Perhaps developing a graph structure that is more suitable for SNP data might increase predictive correlations.

In conclusion, although the diffusion kernel as a choice of basis function may have potential for use in whole-genome prediction, the results of this study suggest that the simple Gaussian kernel is robust enough, and the scope for enhancing predictive ability via kernel refinement is limited.

References

[1] Zhang Z, Zhang Q, and Ding X: **Advances in genomic selection in domestic animals**. *Chinese Science Bulletin* 2011, **56**(25):2655-2663.

[2] Lorenz AJ, Chao S, Asoro FG, Heffner EL, Hayashi T, Iwata H, Smith KP, Sorrells ME, and Jannink JL: **Chapter two: Genomic Selection in Plant Breeding: Knowledge and Prospects**. *Advances in Agronomy* 2011, **110**:77-123.

[3] de los Campos G, Gianola D, and Allison DB: **Predicting genetic predisposition in humans: the promise of whole-genome markers**. *Nat. Genet. Rev.* 2010, **11**:880-886.

[4] Shao H, et al. (2008) **Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis**. *Proc Natl Acad Sci USA* 2008, **105**:19910-19914.

[5] Mackay TFC, Stone EA, and Ayroles JF: **The genetics of quantitative traits: challenges and prospects**. *Nat. Rev. Genet.* 2009, **10**:565-577.

[6] Xu L, Jiang H, Chen H and Gu Z: **Genetic Architecture of Growth Traits Revealed by Global Epistatic Interactions**. *Genome Biol Evol.* 2011, **3**:909-914.

[7] Loewe L: **A framework for evolutionary systems biology**. *BMC Systems Bio.* 2009, **3**:27.

[8] Meuwissen THE, Hayes BJ, and Goddard ME: **Prediction of total genetic value using genome-wide dense marker maps**. *Genetics* 2001, **157**:1819-1829.

[9] Habier D , Fernando RL, Kizilkaya K and Garrick DJ: **Extension of the Bayesian alphabet for genomic selection**. *BMC Bioinformatics* 2011, **12**:186

[10] Gianola D, Fernando RL, and Stella A: **Genomic-assisted prediction of genetic value with semiparametric procedures**. *Genetics* 2006, **173**(3):1761-1776.

[11] Gianola D, de los Campos G, Hill WG, Manfredi E, and Fernando R. **Additive genetic variability and the Bayesian alphabet**. *Genetics* 2009, **183**(1):347-363.

[12] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, and Visscher PM: **Common SNPs explain a large proportion of the heritability for human height**. *Nat Genet.* 2010, **42**(7):565-569.

[13] Henderson CR: **Applications of linear models in animal breeding**. Guelph, ON: University of Guelph.

[14] VanRaden PM: **Efficient methods to compute genomic predictions**. *J. Dairy Sci.* 2008, **91**:4414-4423.

[15] Gianola D, and van Kaam JB: **Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits**. *Genetics* 2008, **178**(4):2289-2303.

[16] Long N, Gianola D, Rosa GJ, Weigel KA, Kranis A, and Gonzlez-Recio O: **Radial basis function regression methods for predicting quantitative traits using SNP markers**. *Genet Res (Camb).* 2010, **92**(3):209-225.

[17] Long N, Gianola D, Rosa GJ, and Weigel KA: **Application of support vector regression to genome-assisted prediction of quantitative traits**. *Theor Appl Genet.* 2011, **123**(7):1065-1074.

[18] Ober U, Erbe M, Long N, Porcu E, Schlather M, and Simianer H: **Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data**. *Genetics* 2011, **188**(3):695-708.

[19] Saunders C, Gammerman A, and Vovk V: **Ridge regression learning algorithm in dual variables**. *In: Proceedings of the 15th International Conference on Machine Learning* 1998, pp. 515-521.

[20] Hoerl AE and Kennward RW: **Ridge regression: Biased estimation for nonorthogonal problems**. *Technometrics* 1970, **12**:55-67.

[21] Gianola D, and de los Campos G: **Inferring genetic values for quantitative traits non-parametrically**. *Genet Res Camb.* 2008, **90**(6):525-540.

[22] González-Recio O, Gianola D, Long N, Weigel KA, Rosa GJ, and Avendaño S: **Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers**. *Genetics* 2008 **178**(4):2305-2313.

[23] González-Recio O, Gianola D, Rosa GJ, Weigel KA, and Kranis A: **Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens**. *Genet Sel Evol.* 2009 **5**:41:3.

[24] de los Campos G, Gianola D, and Rosa GJ: **Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation**. *J Anim Sci.* 2009 **87**(6):1883-1887.

[25] Kondor IR, and Lafferty J: **Diffusion Kernels on Graphs and and Other Discrete Input Spaces**. *Proceedings of 19th International Conference on Machine Learning (ICML-2002).* 2002.

[26] Smola AJ and Kondor IR: **Kernels and regularization on graphs**. *In B. Schölkopf and M. K. Warmuth, editors, Proc. Annual Conf. Computational Learning Theory, Lecture Notes in Comput. Sci.*, pages 144-158, Heidelberg, Germany, 2003. Springer-Verlag.

[27] Lafferty J and Lebanon G: **Diffusion Kernels on Statistical Manifolds**. *Journal of Machine Learning Research* 2005, **6**:129-163.

[28] Fouss F, Francoisse K, Yen L, Pirotte A, and Saerens M: **An experimental investigation of graph kernels on collaborative recommendation and semi-supervised classification**. http://www.isys.ucl.ac.be/staff/marco/Publications/2008_ExperimentalInvestigation.pdf

[29] Vishwanathan SVN, Schraudolph NN, Kondor IR, and Borgwardt KM: **Graph kernels**. *Journal of Machine Learning Research* 2010, **11**:1201-1242.

[30] Tsuruta S, Misztal I, and Lawlor TJ: **Changing definition of productive life in US Holsteins: Effect on genetic correlations**. *J. Dairy Sci.* 2005, **88**:1156-1165.

[31] Gianola D, Okut H, Weigel KA, and Rosa GJM: **Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat**. *BMC Genetics* 2011, **12**:87.

[32] Kimeldorf G, and Wahba G: **Some results on Tchebycheffian spline functions**. *Journal of Mathematic Analysis and Applications.* 1971, **33**:82-95.

[33] Meuwissen, TH, Solberg, TR, Shepherd, R and Woolliams, JA: **A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value**. *Genet. Sel. Evol.* 2009, **41**, 2.

[34] Strandén I and Christensen OF: **Allele coding in genomic evaluation**. *Genet. Sel. Evo.* 2011, **43**:25.

[35] de los Campos G, Gianola D, Rosa GJ, Weigel KA, and Crossa J: **Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods**. *Genetics Research* 2010, **92**:295-308.

[36] Kimeldorf G, and Wahba G: **A correspondence between Bayesian estimation on stochastic processes and smoothing by splines** . *Ann. Math. Stat.* 1970, **41**:495-502.

[37] Vert J.-P. and Kanehisa M: **Graph driven features extraction from microarray data using diffusion kernels and kernel cca**. *In S. Becker, S. Thrun, and K. Obermayer, editors, Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2003.

[38] Gärtner T: **A survey of kernels for structured data**. *SIGKDD Explorations, Vol 5, No. 1* 2002, pp. S268-S275.

[39] Morota G, Valente BD, Rosa GJM, Weigel KA, and Gianola D: **An assessment of linkage disequilibrium in Holstein cattle using a Bayesian network**. Journal of Animal Breeding and Genetics 2012, DOI: 10.1111/jbg.12002.

[40] Morota G and Gianola D: **Evaluation of linkage disequilibrium in wheat with an L1 regularized sparse Markov network**. 2012, in preparation.

[41] Gönen M. and Alpaydın E: **Multiple Kernel Learning Algorithms**. *Journal of Machine Learning Research* 2011, **12**:2211-2268.

[42] Evans LC: **Partial Differential Equations: Second Edition**. American Mathematical Society 2010.

Table 1: Example of diffusion on a graph. $x = (0, 1, 2)$ are genotype codes; $\alpha = (0.1, 0.2)$ is the diffusion rate; $k_{\tilde{x}}(t, x)$ is the time t diffusion of the influence of genotype $\tilde{x}$ on genotype x.

α= 0.1				$\alpha = 0.2$				$\alpha = 0.2$			
$x =$	0	1	2	x=	0	1	2	x=	0	1	2
$k_1(0, x)$	0	1	0	$k_1(0, x)$	0	1	0	$k_2(0, x)$	0	0	1
$k_1(1, x)$	0.1	0.8	0.1	$k_1(1, x)$	0.2	0.6	0.2	$k_2(1, x)$	0	0.2	0.8
$k_1(2, x)$	0.17	0.66	0.17	$k_1(2, x)$	0.28	0.44	0.28	$k_2(2, x)$	0.04	0.28	0.68
$k_1(3, x)$	0.219	0.562	0.219	$k_1(3, x)$	0.312	0.376	0.312	$k_2(3, x)$	0.171	0.330	0.498
$k_1(15, x)$	0.331	0.336	0.331	$k_1(15, x)$	0.333	0.333	0.333	$k_2(15, x)$	0.324	0.333	0.342

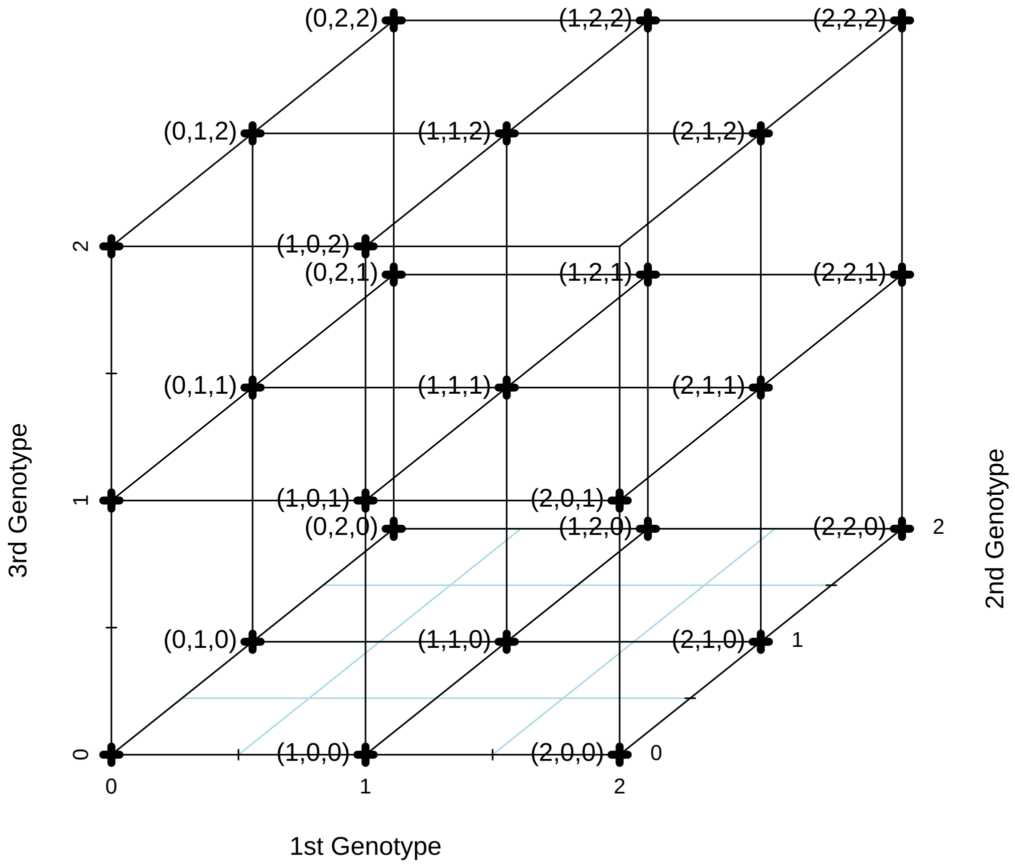

Figure 1: A SNP grid graph with 3 genotypes ($p = 3$). It has $3^3 = 27$ vertices.

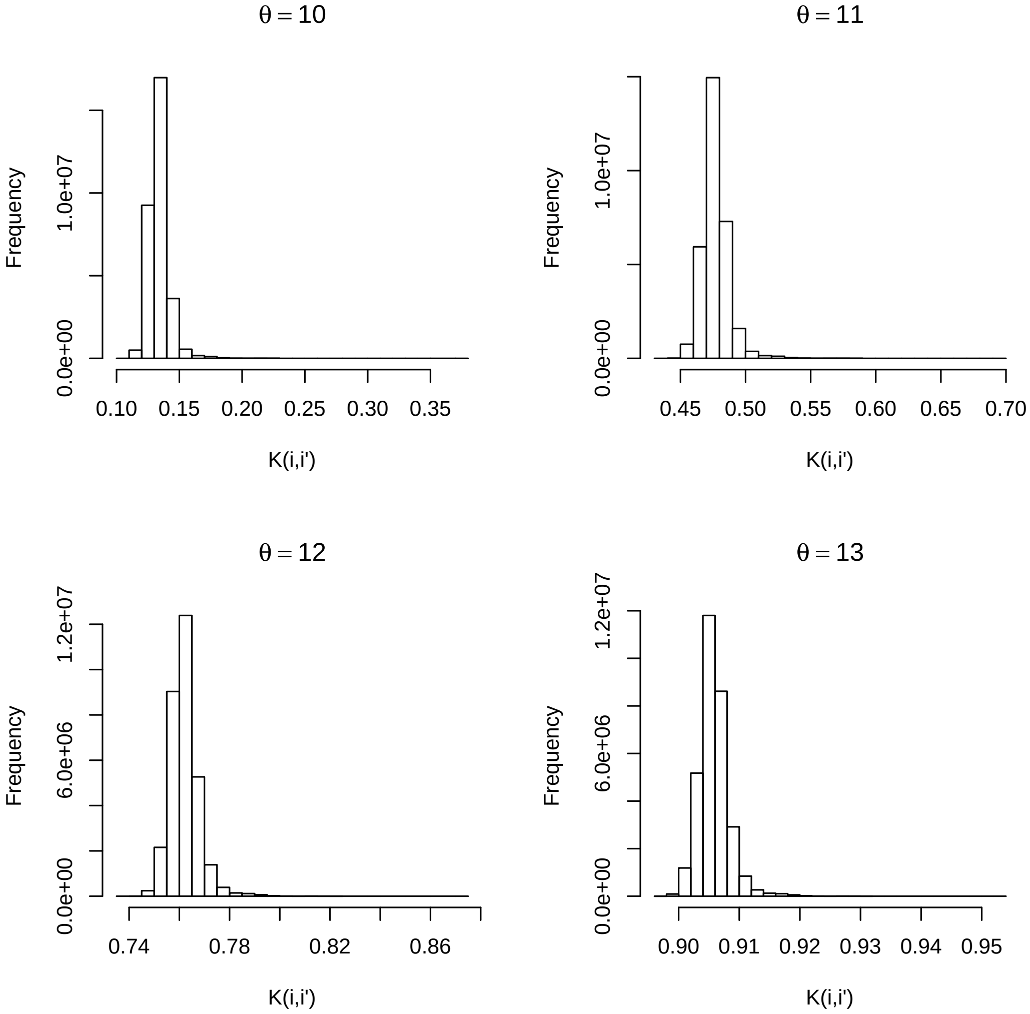

Figure 2: Lower triangular elements of four diffusion kernels based on four different bandwidth parameters (θ).

Table 2: Averages of diagonal $K(x_i, x_i)$ and off-diagonal $K(x_i, x_j)$ kernel elements, predictive correlation, and mean-squared error of prediction (MSE) for the diffusion, Gaussian, and two additive genomic relationship kernels (**G1** and **G2**) with different values of the bandwidth parameter θ for the Holstein data. Values in parentheses were obtained by combining the SNP grid and the hypercube kernels by applying a same bandwidth parameter. **G1** and **G2** do not involve bandwidth parameters. The best prediction within a same kernel is underlined.

Kernel	θ	$k(x_i, x_i)$	$k(x_i, x_j)$	Cor($\hat{\mathbf{y}}^{test}, \mathbf{y}^{PTA}$)	MSE
Diffusion	10	0.369 (0.369)	0.138 (0.134)	0.727 (0.726)	215.93 (216.61)
	11	0.693 (0.693)	0.483 (0.477)	$\underline{0.745}$ (0.741)	$\underline{204.36}$ (208.68)
	11.5	0.801 (0.801)	0.644 (0.639)	0.739 (0.732)	207.93 (212.97)
	12	0.874 (0.874)	0.765 (0.762)	0.739 (0.728)	210.54 (215.08)
	13	0.952 (0.952)	0.907 (0.906)	0.734 (0.725)	211.50 (217.61)
	14	0.982 (0.982)	0.966 (0.965)	0.729 (0.723)	214.29 (218.70)
Gaussian	5×10^{-5}	1 (1)	0.237 (0.225)	0.721 (0.702)	220.675 (233.21)
	2×10^{-5}	1 (1)	0.551 (0.542)	0.736 (0.733)	213.41 (213.95)
	1×10^{-5}	1 (1)	0.749 (0.742)	$\underline{0.742}$ (0.736)	$\underline{210.14}$ (211.24)
	5×10^{-6}	1 (1)	0.866 (0.861)	0.736 (0.729)	210.24 (214.47)
	3×10^{-6}	1 (1)	0.917 (0.914)	0.734 (0.726)	211.51 (216.42)
	1×10^{-6}	1 (1)	0.971 (0.971)	0.729 (0.724)	214.37 (217.93)
G1	NA	0.992 (1.009)	-0.000126 (-0.000128)	0.729 (0.722)	214.36 (219.27)
G2	NA	0.894 (0.909)	-0.000113 (-0.00012)	0.730 (0.723)	213.64 (218.31)

Table 3: Average of diagonal $K(x_i, x_i)$ and off-diagonal $K(x_i, x_j)$ kernel elements, predictive correlation, and mean-squared error of prediction (MSE) for the diffusion, Gaussian, and two additive genomic relationship kernels at different values of the bandwidth parameter θ for the wheat data. The predictive correlation and the MSE were obtained from a 10-fold cross-validation. Additive genomic relationship kernels (**G1** and **G2**) do not involve bandwidth parameters. The best prediction within a same kernel is underlined.

Kernel	θ	$k(x_i, x_i)$	$k(x_i, x_j)$	$\mathrm{Cor}(\hat{\mathbf{y}}^{test}, \mathbf{y}^{train})$	MSE
Diffusion	3	1	0.136	$\underline{0.586}$	0.685
	3.25	1	0.289	0.580	0.673
	3.5	1	0.466	0.577	$\underline{0.681}$
	4	1	0.752	0.547	0.704
	5	1	0.962	0.522	0.721
Gaussian	0.005	1	0.134	$\underline{0.582}$	$\underline{0.686}$
	0.003	1	0.290	0.579	0.697
	0.002	1	0.434	0.562	0.697
	0.001	1	0.655	0.558	0.703
	0.0005	1	0.809	0.556	0.673
G1	NA	2	-0.003	0.518	0.709
G2	NA	2	-0.003	0.521	0.708

Appendix A

Connection between a diffusion and a Gaussian kernel

Intuitively, consider again (3.4) with a one locus case. In order to make space continuous, an infinite number of 'fake' genotypes between and outside of 0 and 2 is needed. That is, instead of the discrete graph $0-1-2$, the interval between 0 and 2, and also outside of it, will be viewed as a 'continuous' graph containing genotypes such as 1.23 or -10.5, for example. While the fundamental structure of the graph remains the same, each genotype is connected just to its immediate neighbors, i.e., each genotype x is connected to only two genotypes, $x+dx$ and $x-dx$ for some infinitesimal dx. Then, $\mathbf{H}$ in (3.5) becomes an infinite-dimensional matrix, and $H(x,x')$ is -2 for $x'=x$ and 1 for $x+dx$, $x-dx$, because each genotype is connected to its neighboring genotypes at both sides. With the vector of genotypes being now infinite-dimensional, $\mathbf{x} = (-\infty, \cdots, x-dx, x, x+dx, \cdots, \infty)$, define a function f that returns an "influence" of genotypes, $\mathbf{f} = (f(-\infty), \cdots, f(x-dx), f(x), f(x+dx), \cdots, f(\infty))$. Approximating dx by h, it can be seen that

$$\begin{aligned}
\frac{1}{h^2}[\mathbf{H}(x,\cdot)\cdot\mathbf{f}] &= \frac{f(x+h)-2f(x)+f(x-h)}{h^2} \\
&= \frac{\frac{f(x+h)-f(x)}{h} - \frac{f(x)-f(x-h)}{h}}{h} \\
&\cong f^{''}(x),
\end{aligned}$$

where $f^{''}(x=x_0)$ is the second derivative of f evaluated at x_0. Thus, with space continuity, $\mathbf{H}$ acts like a second derivative [25]. Using this analogy back in (3.11), we get

$$\frac{d}{d\theta}\mathcal{K}_\theta(x) = \frac{d^2}{dx^2}\mathcal{K}_\theta(x)$$

This equation is called the continuous diffusion equation: the first derivative in "time" is equal to the second derivate in "space". The solution to this partial differential equation (PDE) with a Dirac delta [42] initial condition of concentration on $x=0$, $k_0(x) = 1_{x=0}$, is given by

$$G_\theta(x) = \frac{1}{\sqrt{4\pi\theta}}\exp\left(-\frac{x^2}{4\theta}\right)$$

This is a Gaussian density in a one-dimensional space where $\sigma_e^2 = 2\theta$ is the variance of the distribution. With the initial condition $K_0(x) = f(x)$, the solution to this PDE is

$$K_\theta(x) = \int_{\mathbb{R}} f(x')G_\theta(x - x')dx'$$

where $g_\theta(x, x') = G(x-x')$ is called a Gaussian kernel with bandwidth θ. Thus, the Gaussian kernel is the 'space' continuous analogue of the diffusion kernel as described on the graph. This analogy works exactly the same in higher dimensions.

Appendix B

Proof of (3.13)

Proof. Consider a graph with one locus, Γ_0; this graph has form $0 - 1 - 2$. We compute $\exp(\theta\mathbf{H})$ where exponentiation is defined as the Taylor expansion (3.10), differing from componentwise exponentiation. For Γ_0, $\mathbf{H}$ is given by

$$\mathbf{H} = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix}.$$

We make use of the eigendecomposition of matrix $\mathbf{H} = \mathbf{PDP^{-1}}$ and take note of the fact that $\mathbf{H}^n = \mathbf{PD}^n\mathbf{P}^{-1}$. Plugging this $\mathbf{H}^n$ in (3.10), we obtain $\exp(\theta\mathbf{H}) = \mathbf{P}\exp(\theta\mathbf{D})\mathbf{P}^{-1}$. Here $\exp(\theta\mathbf{D})$ becomes simple componentwise exponentiation because $\mathbf{D}$ is a diagonal matrix of eigenvalues. For this specific matrix,

$$\mathbf{P} = \begin{bmatrix} 1 & 1 & 1 \\ -2 & 0 & 1 \\ 1 & -1 & 1 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} -3 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Thus, the kernel for a one-dimensional grid graph is

$$\begin{aligned}
\mathcal{K}_\theta &= \exp(\theta \mathbf{H}) \\
&= \mathbf{P}\exp(\theta \mathbf{D})\mathbf{P}^{-1} \\
&= \begin{bmatrix} 1 & 1 & 1 \\ -2 & 0 & 1 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} e^{-3\beta} & 0 & 0 \\ 0 & e^{-\beta} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ -2 & 0 & 1 \\ 1 & -1 & 1 \end{bmatrix}^{-1} \cdot \\
&= \frac{1}{6}\begin{bmatrix} e^{-3\theta}+3e^{-\theta}+2 & -2e^{-3\theta}+2 & e^{-3\theta}-3e^{-\theta}+2 \\ -2e^{-3\theta}+2 & 4e^{-3\theta}+2 & -2e^{-3\theta}+2 \\ e^{-3\theta}-3e^{-\theta}+2 & -2e^{-3\theta}+2 & e^{-3\theta}+3e^{-\theta}+2 \end{bmatrix}
\end{aligned} \tag{5.1}$$

Taking the exponential of eigenvalues always yields a positive real value, so if $\mathbf{H}$ is symmetric, $\exp(\theta\mathbf{H})$ is positive definite, suggesting that the diffusion kernel is a valid kernel. Expression (5.1) is symmetric and in particular,

$$\mathcal{K}_\theta(\mathbf{x}, \mathbf{x}^{'}) = \begin{cases} -2e^{-3\theta}+2 & \text{if} \quad |x_i - x_i^{'}| = 1 \\ e^{-3\theta}-3e^{-\theta}+2 & \text{if} \quad |x_i - x_i^{'}| = 2 \\ e^{-3\theta}+3e^{-\theta}+2 & \text{if} \quad x_i = x_i^{'}, x^{'} \neq 1 \\ 4e^{-3\theta}+2 & \text{if} \quad x_i = x_i^{'} = 1 \end{cases} \tag{5.2}$$

Computing every entry of $\mathcal{K}$ is computationally unfeasible and unnecessary. We only need to compute entries corresponding to the pair of genotypes appearing in the sample. In particular, if $k_i(x_i, y_i)$ is the contribution of the ith locus, then

$$K_\theta(\mathbf{x}, \mathbf{x}^{'}) = \prod_{i=1}^{p} k_i(x_i, x_i^{'}),$$

where $k_i(x_i, x_i^{'})$ is determined by the relation between x_i and $x_i^{'}$, and can take only one of the four values specified above. Thus we can write $k_i(x_i, x_i^{'})$ as

$$(e^{-3\beta}-3e^{-\beta}+2)\delta_{|x_i-y_i|=2} + (-2e^{-3\beta}+2)\delta_{|x_i-y_i|=1} + (e^{-3\beta}+3e^{-\beta}+2)\delta_{x_i=y_i\neq 1} + (4e^{-3\beta}+2)\delta_{x_i=y_i=1}$$

where δ is the indicator function. Therefore,

$$
\begin{aligned}
K_\theta^{\otimes p}(\mathbf{x},\mathbf{x}') \propto \prod_{i=1}^{p} \Big\{ & (e^{-3\theta} - 3e^{-\theta} + 2)\delta_{|x_i - x_i'|=2} + (-2e^{-3\theta} + 2)\delta_{|x_i - x_i'|=1} \\
& + (e^{-3\theta} + 3e^{-\theta} + 2)\delta_{x_i = x_i' \neq 1} + (4e^{-3\theta} + 2)\delta_{x_i = x_i' = 1} \Big\}
\end{aligned}
$$

This can be simplified by using the fact that

$$n_1 + n_0 + n_2 = p,$$

so that

$$
\begin{aligned}
K_\theta^{\otimes p}(\mathbf{x},\mathbf{x}') &= (-2e^{-3\theta} + 2)^{n_1}(e^{-3\theta} - 3e^{-\theta} + 2)^{n_2}(e^{-3\theta} + 3e^{-\theta} + 2)^{n_0 - m_{11}}(4e^{-3\theta} + 2)^{m_{11}} \\
& (-2e^{-3\beta} + 2)^{n_1}(e^{-3\beta} - 3e^{-\beta} + 2)^{n_2}(e^{-3\beta} + 3e^{-\beta} + 2)^{n_0 - m_{11}}(4e^{-3\beta} + 2)^{m_{11}} \cdot \frac{(e^{-3\beta} + 3e^{-\beta} + 2)^p}{(e^{-3\beta} + 3e^{-\beta} + 2)^p} \\
& \propto \frac{(-2e^{-3\beta} + 2)^{n_1}(e^{-3\beta} - 3e^{-\beta} + 2)^{n_2}(e^{-3\beta} + 3e^{-\beta} + 2)^{n_0 - m_{11}}(4e^{-3\beta} + 2)^{m_{11}}}{(e^{-3\beta} + 3e^{-\beta} + 2)^p} \\
& = \frac{(-2e^{-3\beta} + 2)^{n_1}(e^{-3\beta} - 3e^{-\beta} + 2)^{n_2}(e^{-3\beta} + 3e^{-\beta} + 2)^{n_0 - m_{11}}(4e^{-3\beta} + 2)^{m_{11}}}{(e^{-3\beta} + 3e^{-\beta} + 2)^{n_0 + n_1 + n_2}} \\
& = \frac{(-2e^{-3\theta} + 2)^{n_1}(e^{-3\theta} - 3e^{-\theta} + 2)^{n_2}(4e^{-3\theta} + 2)^{m_{11}}}{(e^{-3\theta} + 3e^{-\theta} + 2)^{n_1 + n_2 + m_{11}}}
\end{aligned}
\tag{5.3}
$$

Note that one does not need to count n_0. □