CHAPTER

Kernel machine regression in neuroimaging genetics

2

T. Ge[1,2], J.W. Smoller[1,2], M.R. Sabuncu[1,3]

Massachusetts General Hospital/Harvard Medical School, Boston, MA, United States[1] Broad Institute of MIT and Harvard, Cambridge, MA, United States[2] Massachusetts Institute of Technology, Cambridge, MA, United States[3]

CHAPTER OUTLINE

Machine Learning and Medical Imaging. http://dx.doi.org/10.1016/B978-0-12-804076-8.00002-5

2.1 INTRODUCTION

The past few years have witnessed a tremendous growth of the amount of biomedical data, including the increasingly accessible medical images and genomic sequences. Techniques that can integrate different resources, extract reliable information from massive data, and reveal true relationships among biological variables have become essential and invaluable in biomedical research.

Kernel methods are a class of machine learning algorithms to study general types of relations in data sets, such as classifications, clusters and correlations, and are particularly powerful in high-dimensional and nonlinear settings (Vapnik, 1998; Cristianini and Shawe-Taylor, 2000; Schölkopf and Smola, 2002). The fundamental idea of kernel methods is built on the observation that, in many situations, relations between data points can be much more easily revealed or modeled if they are transformed from their original representations into a higher dimensional feature space via a user-specified feature map. For example, consider the simple classification problem as shown in Fig. 2.1. The gray and black dots cannot be linearly separated in one-dimensional space (Fig. 2.1, left), but when transformed into a two-dimensional space using the feature map $\varphi : x \mapsto (x, x^2)$, a linear separator can be easily found (eg, the dashed line in Fig. 2.1, right).

However, in practice, the explicit form of this feature map is often unknown, and the evaluation of a high-dimensional mapping can be computationally expensive. Kernel methods resolve this problem by employing a kernel function, which measures the resemblance between pairs of data points in the original space, and implicitly defines a (possibly infinite-dimensional) feature space and a feature map without actually accessing them. This makes kernel methods highly flexible since they can be easily applied to vectors, images, text, graphs, sequences, and any other data types, as long as a valid kernel function that converts any pair of data points into a scalar similarity measure can be defined on the particular data structure. Due to

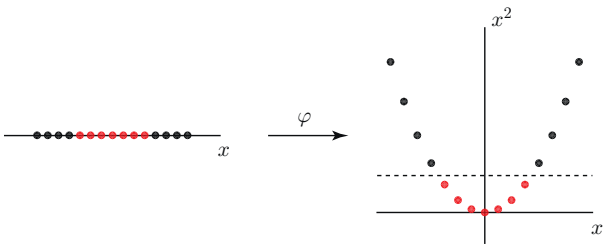

FIG. 2.1

An illustration of the power of feature mapping in a classification problem. Gray and black dots cannot be linearly separated in one-dimensional space (left), but when transformed into a two-dimensional space using the feature map $\varphi : x \mapsto (x, x^2)$, a linear separator can be found (right).

their generality, kernel methods have been applied to a wide range of data problems, such as classification, clustering, dimension reduction and correlation analysis, and have generalized support vector machine (SVM) (Schölkopf and Burges, 1999; Schölkopf and Smola, 2002), principal component analysis (PCA) (Schölkopf et al., 1998, 1997), Gaussian processes (Rasmussen and Williams, 2006), and many other techniques from linear to nonlinear settings.

Kernel machine regression (KMR) is a form of nonparametric regression and an application of the kernel methods to regression analysis. Data that exhibit complex nonlinear relationships in their original representations are implicitly operated in a higher dimensional feature space where a linear regression model is sufficient to describe the transformed data. In practice, KMR essentially regresses the traits (dependent variables) onto the similarity of attributes (independent variables) measured via the kernel function. Previously, KMR was fitted in a penalized regression framework, which can be computationally expensive and produce suboptimal model estimates. Recent theoretical advances have established the connection between KMR and mixed effects models in statistics, leading to elegant model fitting procedures and efficient statistical inferences about model parameters. This has spurred a rapidly expanding literature on applying KMR to biomedical research, especially in the field of genetics to identify cumulative effects of genetic variants on phenotypes, characterize the genetic architecture underlying complex traits, and test gene-by-gene or gene-by-environment interactions. Very recently the idea of KMR has been adapted to imaging genetics, an emerging field that identifies and characterizes genetic influences on brain structure, function and wiring, to dissect the genetic underpinning of features extracted from brain images and to understand the roles genetics plays in brain-related illnesses. In this chapter, we review both the mathematical basis and recent applications of KMR.

The remainder of this chapter is organized as follows. We first explain the intuitions and heuristics behind KMR from the perspective of linear regression analysis and give an illustrative example. We then formally introduce KMR, establish its connection to mixed effects models in statistics, and derive model fitting and statistical inference procedures. We also provide a framework for building and selecting kernel functions in a systematic and objective way. Recent theoretical extensions and applications of KMR are reviewed, with a focus on genetic association studies and imaging genetics. We close the chapter with a discussion of future directions. A rigorous mathematical treatment of the kernel methods and some technical aspects of the material presented in this chapter are included in the appendix.

2.2 MATHEMATICAL FOUNDATIONS

2.2.1 FROM REGRESSION ANALYSIS TO KERNEL METHODS

Kernel machine regression (KMR) is a form of nonparametric regression. We start with a simple model to help understand how it is related to linear regression analysis.

A rigorous mathematical treatment of the theory underlying kernel methods is provided in Appendix A.

Let y_i be a quantitative trait and $\boldsymbol{z}_i$ be a multidimensional attribute for the ith subject. Suppose that y_i is dependent on $\boldsymbol{z}_i$ through an unknown function f:

$$y_i = f(\boldsymbol{z}_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.1}$$

where ϵ_i is assumed to be independently Gaussian distributed with zero mean and homogeneous variance σ^2. The relationship between y_i and $\boldsymbol{z}_i$ can be highly nonlinear and thus the function f can be complex. The idea of kernel methods is to approximate f by a *feature map* φ, which transforms the attributes from their original input space to a higher dimensional *feature space*, $\varphi(\boldsymbol{z}_i) = [\varphi_1(\boldsymbol{z}_i), \varphi_2(\boldsymbol{z}_i), \ldots]^{\mathrm{T}}$, such that the mappings $\{\varphi_k\}$, also known as the *basis functions* of the feature space, can be linearly combined to predict y_i. Expanding a function into the linear combination of a set of basis functions is called the *primal representation* of the function in kernel methods. Frequently used mappings include polynomial functions, spline-based functions (Wahba, 1990; Gu, 2013), and many others. Suppose we employ p different basis functions, model (2.1) then becomes

$$y_i = \sum_{k=1}^{p} \varphi_k(\boldsymbol{z}_i)\omega_k + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.2}$$

where $\omega_k, k = 1, \ldots, p$, are scalar coefficients. If we define vectors $\boldsymbol{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$, $\boldsymbol{\omega} = [\omega_1, \ldots, \omega_p]^{\mathrm{T}}$, $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^{\mathrm{T}}$, and the $n \times p$ *design matrix* $\boldsymbol{Z}_\varphi = [\varphi(\boldsymbol{z}_1), \ldots, \varphi(\boldsymbol{z}_n)]^{\mathrm{T}}$, Eq. (2.2) can be written in the matrix form:

$$\boldsymbol{y} = \boldsymbol{Z}_\varphi \boldsymbol{\omega} + \boldsymbol{\epsilon}, \tag{2.3}$$

which is a linear regression model. If $p \leqslant n$ and the design matrix is full rank, the *ordinary least squares* (OLS) estimator for the regression coefficients $\boldsymbol{\omega}$ is $\widehat{\boldsymbol{\omega}} = (\boldsymbol{Z}_\varphi^{\mathrm{T}}\boldsymbol{Z}_\varphi)^{-1}\boldsymbol{Z}_\varphi^{\mathrm{T}}\boldsymbol{y}$, which minimizes the loss function $\mathcal{L}(\boldsymbol{\omega}) = \|\boldsymbol{y} - \boldsymbol{Z}_\varphi\boldsymbol{\omega}\|^2 = \sum_{i=1}^{n} \left(y_i - \boldsymbol{\omega}^{\mathrm{T}}\varphi(\boldsymbol{z}_i)\right)^2$, and is also the *best linear unbiased estimator* (BLUE) and the *maximum likelihood estimator* (MLE) because we have assumed the residual $\boldsymbol{\epsilon}$ to be normal. When $p > n$, that is, the number of basis functions is greater than the sample size, fitting linear regression model (2.3) becomes ill-posed because the problem is underdetermined, that is, there exists infinitely many $\boldsymbol{\omega}$ that can perfectly fit the model. To avoid the danger of overfitting, an effective and widely used approach is to penalize the norm of the regression coefficients $\boldsymbol{\omega}$ in the loss function. For example, consider the *Ridge regression* (also known as *Tikhonov regularization*), which penalizes the squared error loss by the squared Euclidean norm of $\boldsymbol{\omega}$:

$$\mathcal{J}(\boldsymbol{\omega}) = \frac{1}{2}\sum_{i=1}^{n} \left(y_i - \boldsymbol{\omega}^{\mathrm{T}}\varphi(\boldsymbol{z}_i)\right)^2 + \frac{\lambda}{2}\|\boldsymbol{\omega}\|^2, \tag{2.4}$$

where λ is a tuning parameter, which balances the model fitting and model complexity. Taking the derivative of $\mathcal{J}$ with respect to $\boldsymbol{\omega}$ and equating it to zero gives

$$\widehat{\boldsymbol{\omega}} = (\boldsymbol{Z}_\varphi^{\mathrm{T}}\boldsymbol{Z}_\varphi + \lambda \boldsymbol{I}_{p\times p})^{-1}\boldsymbol{Z}_\varphi^{\mathrm{T}}\boldsymbol{y}, \tag{2.5}$$

where $\boldsymbol{I}_{p\times p}$ is a $p \times p$ identity matrix. A prediction of the trait y given the attribute $\boldsymbol{z}$ from a new subject is

$$\hat{y} = \hat{f}(z) = \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\varphi(z) = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{Z}_\varphi(\boldsymbol{Z}_\varphi^{\mathrm{T}}\boldsymbol{Z}_\varphi + \lambda \boldsymbol{I}_{p\times p})^{-1}\varphi(z). \tag{2.6}$$

It can be seen that for positive λ, the smallest eigenvalue of $\boldsymbol{Z}_\varphi^{\mathrm{T}}\boldsymbol{Z}_\varphi + \lambda \boldsymbol{I}_{p\times p}$ is bounded away from zero, and thus the matrix inverse in Eqs. (2.5) and (2.6) always exists. Ridge regression thus regularizes the linear regression model (2.3) by shrinking the solutions towards zero.

We now rewrite Eqs. (2.5) and (2.6) to provide insights into the connection between kernel-based regression and regularized linear regression. Specifically, we note that

$$(\boldsymbol{Z}_\varphi^{\mathrm{T}}\boldsymbol{Z}_\varphi + \lambda \boldsymbol{I}_{p\times p})\boldsymbol{Z}_\varphi^{\mathrm{T}} = \boldsymbol{Z}_\varphi^{\mathrm{T}}\boldsymbol{Z}_\varphi\boldsymbol{Z}_\varphi^{\mathrm{T}} + \lambda \boldsymbol{Z}_\varphi^{\mathrm{T}} = \boldsymbol{Z}_\varphi^{\mathrm{T}}(\boldsymbol{Z}_\varphi\boldsymbol{Z}_\varphi^{\mathrm{T}} + \lambda \boldsymbol{I}_{n\times n}). \tag{2.7}$$

Since both $\boldsymbol{Z}_\varphi^{\mathrm{T}}\boldsymbol{Z}_\varphi + \lambda \boldsymbol{I}_{p\times p}$ and $\boldsymbol{Z}_\varphi\boldsymbol{Z}_\varphi^{\mathrm{T}} + \lambda \boldsymbol{I}_{n\times n}$ are invertible, we have

$$\widehat{\boldsymbol{\omega}} = (\boldsymbol{Z}_\varphi^{\mathrm{T}}\boldsymbol{Z}_\varphi + \lambda \boldsymbol{I}_{p\times p})^{-1}\boldsymbol{Z}_\varphi^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{Z}_\varphi^{\mathrm{T}}(\boldsymbol{Z}_\varphi\boldsymbol{Z}_\varphi^{\mathrm{T}} + \lambda \boldsymbol{I}_{n\times n})^{-1}\boldsymbol{y} := \boldsymbol{Z}_\varphi^{\mathrm{T}}\widehat{\boldsymbol{\alpha}} = \sum_{i=1}^{n}\hat{\alpha}_i\varphi(z_i), \tag{2.8}$$

where we have defined $\widehat{\boldsymbol{\alpha}} = [\hat{\alpha}_1, \ldots, \hat{\alpha}_n]^{\mathrm{T}} = (\boldsymbol{Z}_\varphi\boldsymbol{Z}_\varphi^{\mathrm{T}} + \lambda \boldsymbol{I}_{n\times n})^{-1}\boldsymbol{y}$. It can be seen that the p-dimensional regression coefficient $\widehat{\boldsymbol{\omega}}$ lies in the span of the n transformed observations $\{\varphi(z_i)\}_{i=1}^n$ even if $p \gg n$. This is expected since the model is linear in the feature space. A more important point can be made by noticing that the $n \times n$ matrix $\boldsymbol{Z}_\varphi\boldsymbol{Z}_\varphi^{\mathrm{T}}$ is *non-negative definite* and its ijth element is the dot product between $\varphi(z_i)$ and $\varphi(z_j)$, which we denote as $\langle\varphi(z_i), \varphi(z_j)\rangle$. Therefore we can define a *kernel function* for any pair of subjects i and j as follows:

$$k(z_i, z_j) = \langle\varphi(z_i), \varphi(z_j)\rangle. \tag{2.9}$$

$\boldsymbol{K} := \boldsymbol{Z}_\varphi\boldsymbol{Z}_\varphi^{\mathrm{T}} = \{k(z_i, z_j)\}_{i,j=1}^n$ is then the *Gram matrix* or *kernel matrix* associated with the kernel function k given the observed attributes $\{z_1, \ldots, z_n\}$. For a new subject with attribute z, the trait can be predicted as

$$\hat{y} = \hat{f}(z) = \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\varphi(z) = \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{Z}_\varphi\varphi(z) := \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{\kappa} = \boldsymbol{y}^{\mathrm{T}}(\boldsymbol{K} + \lambda \boldsymbol{I}_{n\times n})^{-1}\boldsymbol{\kappa}, \tag{2.10}$$

where $\boldsymbol{\kappa} = \boldsymbol{Z}_\varphi\varphi(z) = [k(z_1, z), \ldots, k(z_n, z)]^{\mathrm{T}}$. Here the function f is represented by a linear combination of the kernel function centered at the observed data points, that is, $\hat{f}(z) = \widehat{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{\kappa} = \sum_{i=1}^n \hat{\alpha}_i k(z_i, z)$. This is the *dual representation* of a function in kernel methods. Eq. (2.10) indicates that f can be estimated without accessing the feature map φ, which can be infinite-dimensional, expensive to evaluate, or

difficult to explicitly specify in practice. Instead, we only need to define a kernel function which collapses the (possibly high-dimensional) attributes for each pair of individuals into a scalar similarity measure. Moreover, we can substitute any non-negative definite kernel function $\tilde{k}$ for k defined in Eq. (2.9) to measure the similarity between pairs of individuals in a different way. This technique is called the "*kernel trick*" in the machine learning literature. The theory of kernel methods ensures that any non-negative definite kernel function $\tilde{k}$ implicitly specifies a feature map $\tilde{\varphi}$ (which could be infinite-dimensional) such that $\tilde{k}$ can be expressed as a dot product in the feature space: $\tilde{k}(\boldsymbol{z}_i, \boldsymbol{z}_j) = \langle \tilde{\varphi}(\boldsymbol{z}_i), \tilde{\varphi}(\boldsymbol{z}_j) \rangle$. Therefore kernel methods greatly simplify the specification of a nonparametric model, especially for multidimensional attributes.

As an illustration, consider the nonlinear relationship, $y = \sin(z)$, $-\pi \leqslant z \leqslant \pi$, between the trait y and a scalar attribute z. We generated synthetic data for $n = 30$ samples using the model $y_i = \sin(z_i) + \epsilon_i$, $i = 1, 2, \ldots, n$, where each z_i was randomly selected between $-\pi$ and π, and ϵ_i was Gaussian distributed with zero mean and variance $\sigma^2 = 0.01$. The objective is to recover the unknown function $f(\cdot) = \sin(\cdot)$ from the observed trait y_i and attribute z_i for each sample. This can be achieved by evaluating the estimated function $\hat{f}$ on a dense and equally spaced grid between $-\pi$ and π, and plotting the predicted traits.

We first fitted the data using a linear feature map $\varphi_L(z) = [1, z]^{\mathrm{T}}$. This is equivalent to modeling the data using the linear regression $y_i = \omega_0 + z_i\omega_1 + \epsilon_i$, where ω_0 and ω_1 are scalar coefficients. By Eq. (2.9), a linear kernel can be defined as $k_L(z_i, z_j) = \langle \varphi_L(z_i), \varphi_L(z_j) \rangle = z_i z_j + 1$, for any pair of samples i and j. The kernel matrix associated with the observed attributes is $\boldsymbol{K}_L = \{z_i z_j + 1\}_{i,j=1}^n$. The predicted traits can then be computed, using either Eq. (2.6) or Eq. (2.10). The upper left panel of Fig. 2.2 shows $\hat{f}$ estimated at three different values of the tuning parameter λ, along with the ground truth, that is, $f(\cdot) = \sin(\cdot)$, and the observed data points. It can be seen that a linear feature map (and the corresponding linear kernel) can only capture linear relationships and is not sufficient to recover the nonlinear trigonometric function used to generate the data. Moreover, when λ increases, the estimated function tends to be "simpler," which in the linear case means a smaller slope.

Next, we fitted the data by employing a cubic feature map $\varphi_C(z) = [1, \sqrt{3}z, \sqrt{3}z^2, z^3]^{\mathrm{T}}$, which corresponds to the regression model $y_i = \omega_0 + \sqrt{3}z_i\omega_1 + \sqrt{3}z_i^2\omega_2 + z_i^3\omega_3 + \epsilon_i$ in the feature space, where ω_k, $k = 0, 1, 2, 3$, are regression coefficients. This feature map defines the cubic kernel $k_C(z_i, z_j) = \langle \varphi_C(z_i), \varphi_C(z_j) \rangle = (z_i z_j + 1)^3$, for samples i and j. The upper right panel of Fig. 2.2 shows that this nonlinear feature map and the corresponding cubic kernel can represent nonlinear relationships. Also, with the increase of λ, the estimated function tends to be flatter. We note that polynomial kernel functions are monotonic for sufficiently large attributes and thus need to be used with caution when extrapolating values far beyond the range of the observed data.

Lastly, we fitted the data using a Gaussian kernel $k_G(z_i, z_j) = \exp\left\{(z_i - z_j)^2/\rho\right\}$, where ρ is a free parameter that determines the width of the kernel. The Gaussian kernel corresponds to an infinite-dimensional feature map because the exponential

function can be viewed as an infinite sum of polynomials. Thus the feature map cannot be explicitly expressed and evaluated. However, as shown in Eq. (2.10) and discussed above, a major advantage of the kernel methods is that the unknown function can be estimated without assessing the feature map, as long as a kernel function is specified. The infinite-dimensional nature of the Gaussian kernel enables it to model any nonlinear relationship in the data. The lower left panel of Fig. 2.2 shows the estimated function using a Gaussian kernel with $\rho = 0.1$. We observe that when λ is small, the model tends to interpolate the observed data points. Actually, if λ is set to zero, the data will be perfectly fitted without error. However, this is a typical overfitting behavior, where the model describes the noise instead of the true relationship underlying the data, and has poor predictive performance and generalization power. As λ increases, complex models are more heavily penalized, and the estimated function becomes smoother. An alternative approach to alleviate

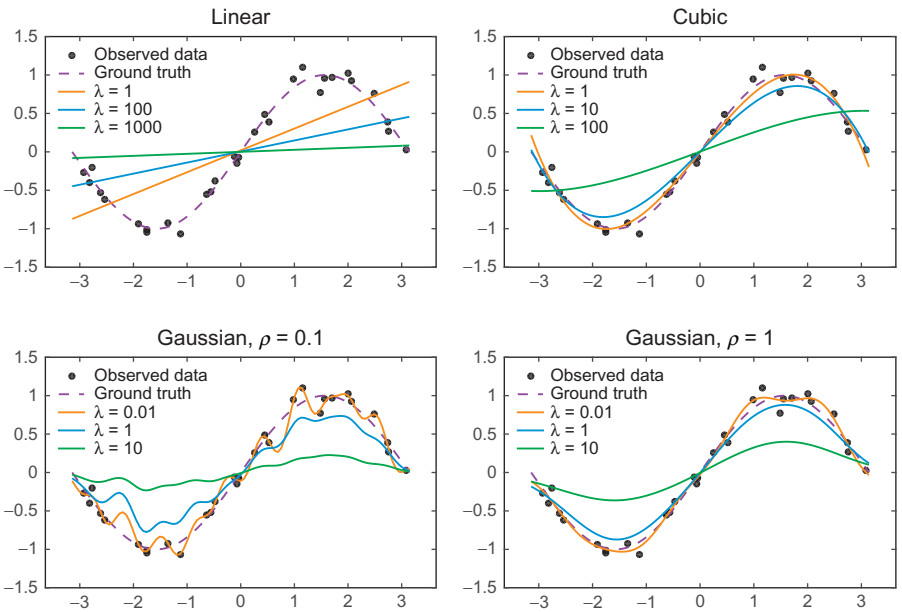

FIG. 2.2

A synthetic example to illustrate the penalized regression and kernel-based regression. Data were generated for 30 samples using the model $y_i = \sin(z_i) + \epsilon_i$, where each z_i was randomly selected between $-\pi$ and π, and ϵ_i was Gaussian distributed with zero mean and variance $\sigma^2 = 0.01$. The function $\sin(\cdot)$ was estimated from the observed traits y_i and attributes z_i using a linear kernel (upper left), a cubic kernel (upper right), a Gaussian kernel with the width parameter $\rho = 0.1$ (lower left), and a Gaussian kernel with $\rho = 1$ (lower right). In each panel, the estimated functions using different tuning parameters λ are shown, along with the ground truth of the function and the observed data points.

the overfitting issue is to increase the width of the Gaussian kernel. This leads to more data points being considered and averaged during model estimation and avoids overfitting the local trend. As shown in the lower right panel of Fig. 2.2, a Gaussian kernel with $\rho = 1$ produced a much better fitting than $\rho = 0.1$. However, a too large width may cause underfitting of the data. In practice, the tuning parameter λ and additional parameters in the kernel function, such as the width of a Gaussian kernel, need to be carefully selected. We will introduce techniques to determine these parameters and perform model selection in the following sections.

2.2.2 KERNEL MACHINE REGRESSION

We now introduce KMR in a more general and rigorous framework by extending model (2.1) as follows:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + f(\boldsymbol{z}_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.11}$$

where $\boldsymbol{x}_i$ is a $q \times 1$ vector of covariates or nuisance variables, and $\boldsymbol{\beta}$ is a $q \times 1$ vector of regression coefficients. Here the covariates $\boldsymbol{x}_i$ are modeled parametrically (linearly), while the attributes $\boldsymbol{z}_i$ are modeled nonparametrically through the unknown function f. As shown in the previous section, we usually expand f into the linear combination of a set of basis functions (primal representation) or the linear combination of a kernel function centered at certain data points (dual representation) with the hope that these representations can be efficiently estimated and provide a good approximation to f. We now solidify this idea by assuming that f lies in a *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$ equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ (Aronszajn, 1950; Saitoh, 1988). An RKHS is a function space defined on the input space $\mathcal{Z}$ where $\boldsymbol{z}_i$ resides, and is uniquely determined by a non-negative kernel function $k(\cdot, \cdot)$ on $\mathcal{Z} \times \mathcal{Z}$. It satisfies that, for any $f \in \mathcal{H}$ and an arbitrary attribute $\boldsymbol{z} \in \mathcal{Z}$, $k(\cdot, \boldsymbol{z})$ as a function belongs to $\mathcal{H}$, and the inner product between f and $k(\cdot, \boldsymbol{z})$ is the evaluation of f at $\boldsymbol{z}$: $\langle f, k(\cdot, \boldsymbol{z}) \rangle_{\mathcal{H}} = f(\boldsymbol{z})$. The latter is known as the *reproducing property* of RKHSs. An RKHS also ensures the existence of the primal and dual representations of the functions belonging to it and implicitly regularizes their smoothness. See Appendix A for a mathematical characterization of RKHSs.

Model (2.11) can be fitted by minimizing the panelized likelihood function:

$$\mathcal{J}(\boldsymbol{\beta}, f) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - f(\boldsymbol{z}_i)\right)^2 + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \tag{2.12}$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm induced by the inner product on the RKHS. The first part of Eq. (2.12) is a loss function quantifying the goodness-of-fit of the model, and the second part of Eq. (2.12) is a regularization term controlling the smoothness of the optimizer. λ is a tuning parameter balancing the model fitting and model complexity. When $\lambda = 0$, the model interpolates the trait, whereas when $\lambda = +\infty$, model (2.11) degenerates to a linear regression model without f. Various other choices of the loss and penalty functions exist and can be used to handle a wide range of problems from

regression to variable selection and to classification (see, eg, Schaid (2010a) for a review). Here we focus on the squared error loss and squared norm penalty because they are widely used for quantitative traits, lead to a closed form solution, and have a strong connection to linear mixed effects models, as shown in the next section.

Minimizing the functional (2.12) is a calculus of variations problem over an infinite-dimensional space of smooth curves, which can be difficult to resolve. However, the *Representer Theorem* (Kimeldorf and Wahba, 1971) shows that the solution of a very general class of optimization problems on RKHSs, which encompasses the problem under investigation here, has a finite-dimensional representation:

$$f(\cdot) = \sum_{j=1}^{n} \alpha_j k(\cdot, z_j), \tag{2.13}$$

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^{\mathrm{T}}$ are unknown parameters. Substituting Eq. (2.13) into Eq. (2.12) gives

$$\begin{aligned} \mathcal{J}(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} - \sum_{j=1}^{n} \alpha_j k(z_i, z_j) \right)^2 + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2 \\ &= \frac{1}{2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}\boldsymbol{\alpha})^{\mathrm{T}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{K}\boldsymbol{\alpha}) + \frac{\lambda}{2} \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{\alpha}, \end{aligned} \tag{2.14}$$

where we have defined the *kernel matrix* $\boldsymbol{K} = \{k(z_i, z_j)\}_{n \times n}$ and used the matrix notation $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^{\mathrm{T}}$, and have made use of the reproducing property in the computation of the penalty term. Specifically,

$$\begin{aligned} \|f\|_{\mathcal{H}}^2 &= \left\| \textstyle\sum_{j=1}^{n} \alpha_j k(\cdot, z_j) \right\|_{\mathcal{H}}^2 = \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \langle k(\cdot, z_i), k(\cdot, z_j) \rangle_{\mathcal{H}}^2 \\ &= \textstyle\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(z_i, z_j) = \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{\alpha}. \end{aligned} \tag{2.15}$$

It can been seen from Eq. (2.14) that the optimization of $\mathcal{J}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is now over finite-dimensional parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Setting the derivatives of $\mathcal{J}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to zero yields the following first-order condition:

$$\begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{K} \\ \boldsymbol{K}^{\mathrm{T}}\boldsymbol{X} & \boldsymbol{K}^{\mathrm{T}}\boldsymbol{K} + \lambda \boldsymbol{K} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\ \boldsymbol{K}^{\mathrm{T}}\boldsymbol{y} \end{bmatrix}. \tag{2.16}$$

It can be verified that the following pair of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is a solution of Eq. (2.16):

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \left[\boldsymbol{X}^{\mathrm{T}} (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X} \right]^{-1} \boldsymbol{X}^{\mathrm{T}} (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y}, \\ \widehat{\boldsymbol{\alpha}} &= (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}). \end{aligned} \tag{2.17}$$

Following Eq. (2.13), the optimizer $\hat{f}$, evaluated at an arbitrary attribute z, can be expressed as

$$\hat{f}(z) = \sum_{j=1}^{n} \hat{\alpha}_j k(z, z_j) = \widehat{\boldsymbol{\alpha}}^{\mathrm{T}} \boldsymbol{\kappa} = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}} (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\kappa}, \tag{2.18}$$

where $\boldsymbol{\kappa} = [k(z_1, z), \ldots, k(z_n, z)]^{\mathrm{T}}$. Specifically, the vector $\widehat{\boldsymbol{f}} = [\hat{f}(z_1), \ldots, \hat{f}(z_n)]^{\mathrm{T}}$, comprising $\hat{f}$ evaluated at the observed attributes $\{z_1, \ldots, z_n\}$, is

$$\widehat{\boldsymbol{f}} = \boldsymbol{K}\widehat{\boldsymbol{\alpha}} = \boldsymbol{K}(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}). \tag{2.19}$$

We emphasize that Eqs. (2.17) and (2.19) only depend on the kernel matrix $\boldsymbol{K}$ and the penalty parameter λ, while the explicit form of the function f does not need to be specified, making this nonparametric modeling approach highly flexible.

We note that the optimal value of λ is usually unknown and, moreover, the kernel function may rely on additional unknown parameters, such as the width in a Gaussian kernel. Few methods exist to jointly estimate these model parameters. In practice, they are usually fixed a priori, resulting in potentially suboptimal solutions. Alternatively, a sequence of tuning and/or kernel parameters can be applied, resulting in a collection of models, and the optimal parameters are determined by selecting the most appropriate model. Information criteria (eg, *Akaike information criterion*, AIC, and *Bayesian information criterion*, BIC) and cross-validation techniques are widely used in model selection. However, a fine search of potential parameter values can be computationally expensive. In a seminal paper, Liu et al. (2007) showed that there is a strong connection between KMR and linear mixed effects models, and thus model estimation and inferences can be conducted within the mixed model framework. In the next section, we review this connection as well as linear mixed effects models.

2.2.3 LINEAR MIXED EFFECTS MODELS

Linear mixed effects models (LMMs) (also known as variance component models) are widely used in statistics to model dependent data structures such as clustered data (McCulloch and Neuhaus, 2001; Snijders, 2011) and longitudinal data (Diggle et al., 2002; Verbeke and Molenberghs, 2009). Here we consider the following LMM:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f} + \boldsymbol{\epsilon}, \tag{2.20}$$

where $\boldsymbol{y}$ is a vector of traits from n individuals, $\boldsymbol{X}$ is an $n \times q$ covariate matrix, $\boldsymbol{\beta}$ is a $q \times 1$ vector of regression coefficients, $\boldsymbol{f}$ is an $n \times 1$ random vector following the normal distribution $\mathrm{N}(\boldsymbol{0}, \tau^2 \boldsymbol{K})$, in which $\boldsymbol{K}$ is an $n \times n$ kernel matrix, $\tau^2 = \lambda^{-1}\sigma^2$, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of residuals following $\mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. $\boldsymbol{f}$ is assumed to be independent of $\boldsymbol{\epsilon}$. Here, τ^2 is expressed as a function of the tuning parameter λ and the variance of the residual σ^2 in the previous section in order to build the connection between KMR and LMMs, which will soon be made clear. Also, note that although $\boldsymbol{K}$ can be an arbitrary non-negative definite matrix, as we will see below, it is connected to the

kernel matrix in the previous section. Hence we use the somewhat confusing notation $\boldsymbol{K}$ to represent the covariance of the random effects in the LMM. In the context of mixed models, the regression coefficients $\boldsymbol{\beta}$ are called *fixed effects* since they model population-average effects, whereas $\boldsymbol{f}$ is a vector of *random effects* since it models subject-specific effects, which are assumed to be randomly sampled from a general population. The conditional distribution of $\boldsymbol{y}$ given the random effects $\boldsymbol{f}$ is normal:

$$\boldsymbol{y} \mid \boldsymbol{f} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f}, \sigma^2\boldsymbol{I}). \tag{2.21}$$

Marginally (averaged across individuals),

$$\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, \tau^2\boldsymbol{K} + \sigma^2\boldsymbol{I}). \tag{2.22}$$

We denote the marginal covariance of $\boldsymbol{y}$ as $\boldsymbol{V} = \tau^2\boldsymbol{K} + \sigma^2\boldsymbol{I}$. It is clear from Eq. (2.22) that $\boldsymbol{\beta}$ can be estimated by *generalized least squares*: $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{y}$, which coincides with the solution in Eq. (2.17) since $\boldsymbol{V} = \tau^2(\boldsymbol{K} + \lambda\boldsymbol{I})$. $\boldsymbol{f}$ can be estimated by noticing the fact that $\boldsymbol{y}$ and $\boldsymbol{f}$ are jointly normal and their covariance is $\tau^2\boldsymbol{K}$. Thus by making use of the conditional distribution of multivariate normal, the expectation of $\boldsymbol{f}$ given the observation $\boldsymbol{y}$ can be estimated as $\tau^2\boldsymbol{K}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$, which is the same as Eq. (2.19). Alternatively, $\boldsymbol{\beta}$ and $\boldsymbol{f}$ can be estimated by jointly maximizing the log likelihood of $[\boldsymbol{y}^{\mathrm{T}}, \boldsymbol{f}^{\mathrm{T}}]^{\mathrm{T}}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{f}$. This gives the Henderson *mixed model equation* (MME):

$$\begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{R}^{-1} \\ \boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{R}^{-1} + \tau^{-2}\boldsymbol{K}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{f} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}, \tag{2.23}$$

where $\boldsymbol{R} = \sigma^2\boldsymbol{I}$. The solutions to the MME, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{f}}$, are the *best linear unbiased estimator* (BLUE) and the *best linear unbiased predictor* (BLUP) for $\boldsymbol{\beta}$ and $\boldsymbol{f}$, respectively. It is easy to verify that Eq. (2.23) and Eq. (2.16) are identical by using the identity $\boldsymbol{f} = \boldsymbol{K}\boldsymbol{\alpha}$ (see Eq. 2.13). Therefore the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{f}$ obtained by minimizing the penalized likelihood function in Eq. (2.12) are equivalent to the BLUE and BLUP of the LMM defined in Eq. (2.20). This connection bridges machine learning and statistics, specifically KMR and LMMs, and allows for a unified framework of model fitting and statistical inferences. In the mixed model framework, the tuning parameter λ can be interpreted as a ratio between the variance component parameters: $\lambda = \sigma^2/\tau^2$. When the kernel matrix explains a large portion of the trait variation, λ tends to be small and thus, in the context of KMR, the nonparametric function f is less penalized. When the kernel matrix captures little variation of the trait, λ tends to be large, and the KMR approaches a parametric linear regression.

The variance component parameters τ^2 and σ^2, and any unknown parameter in the kernel function, can now be estimated by maximizing the likelihood of the LMM.

Specifically, denoting $\boldsymbol{\theta}$ as a vector comprising all the unknown parameters in the marginal covariance structure $\boldsymbol{V}$, the log likelihood function of the LMM is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\theta})| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{2.24}$$

and the profile log likelihood is

$$\ell_{\mathrm{P}}(\boldsymbol{\theta}) = -\frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\theta})| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}), \tag{2.25}$$

where we have replaced $\boldsymbol{\beta}$ in Eq. (2.24) by its generalized least squares estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{y}$. Maximizing ℓ_{P} with respect to $\boldsymbol{\theta}$ gives the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$. However, $\widehat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ is biased since it does not account for the loss in degrees of freedom resulting from estimating the unknown fixed effects $\boldsymbol{\beta}$. In contrast, restricted maximum likelihood (ReML) estimation (Patterson and Thompson, 1971; Harville, 1977; Lindstrom and Bates, 1988) produces unbiased estimates of variance component parameters by applying a transformation to the LMM to remove the effect of covariates and calculating the log likelihood function based on the transformed data. The restricted log likelihood can be written as a correction to the profile log likelihood:

$$\ell_{\mathrm{R}}(\boldsymbol{\theta}) = \ell_{\mathrm{P}}(\boldsymbol{\theta}) - \frac{1}{2}\log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}|. \tag{2.26}$$

We review the Newton-Raphson method commonly employed to maximize ℓ_{R} with respect to $\boldsymbol{\theta}$ in Appendix B. Once the ReML estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{ReML}}$ has been obtained, it can be plugged into Eqs. (2.17) and (2.19), producing *empirical* BLUE and BLUP of the fixed and random effects, respectively. The covariance matrices of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{f}}$ can be calculated as

$$\begin{aligned} \mathbf{cov}(\widehat{\boldsymbol{\beta}}) &= \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}\right)^{-1}, \\ \mathbf{cov}(\widehat{\boldsymbol{f}}) = \mathbf{cov}(\widehat{\boldsymbol{f}} - \boldsymbol{f}) &= \tau^2\boldsymbol{K} - (\tau^2\boldsymbol{K})\boldsymbol{P}(\tau^2\boldsymbol{K}), \end{aligned} \tag{2.27}$$

where $\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}$. Analogously, inserting $\widehat{\boldsymbol{\theta}}_{\mathrm{ReML}}$ into Eq. (2.27) gives empirical estimates of the covariance matrices. However, we note that these empirical covariance estimates are derived under the assumption that the covariance structure $\boldsymbol{V}$ is known. Therefore they are expected to underestimate the variation of the fixed and random effects as the variation in $\widehat{\boldsymbol{\theta}}_{\mathrm{ReML}}$ is not taken into account. A full Bayesian analysis (Gelman et al., 2013) based on sampling methods such as the Markov chain Monte Carlo (MCMC) can produce more accurate approximations. However, in this chapter, we will focus on making inferences about the kernel function and the corresponding variance component parameters. As we will see in the next section, efficient statistical tests have been devised, which can avoid expensive computation for fitting the full LMM.

2.2.4 STATISTICAL INFERENCE

Hypothesis testing on elements of the fixed effects $\widehat{\boldsymbol{\beta}}$ can be performed using the standard *likelihood ratio test* (LRT) or the *Wald test*, under the assumption that $\widehat{\boldsymbol{\beta}}$ is asymptotically normal with mean $\boldsymbol{\beta}$ and covariance calculated in Eq. (2.27). For example, to test the null hypothesis $\mathcal{H}_0 : \boldsymbol{C\beta} = \boldsymbol{d}$ versus the alternative $\mathcal{H}_1 : \boldsymbol{C\beta} \neq \boldsymbol{d}$, where $\boldsymbol{C}$ is a *contrast matrix* with rank r, the LRT statistic is

$$\mathcal{D} = -2\left[\ell(\widehat{\boldsymbol{\beta}}_{\mathrm{R}}, \widehat{\boldsymbol{\theta}}_{\mathrm{R}}) - \ell(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})\right] \sim \chi^2_r, \tag{2.28}$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ are estimates in the full, unrestricted model, and $\widehat{\boldsymbol{\beta}}_{\mathrm{R}}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{R}}$ are estimates in the restricted model, that is, when $\boldsymbol{C\beta} = \boldsymbol{d}$ is satisfied. Under the null hypothesis, $\mathcal{D}$ approximately follows a chi-squared distribution with r degrees of freedom. Alternatively, the Wald test statistic is

$$\mathcal{W} = \left(\boldsymbol{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{d}\right)^{\mathrm{T}} \left(\boldsymbol{C}\boldsymbol{A}(\widehat{\boldsymbol{\theta}})\boldsymbol{C}^{\mathrm{T}}\right)^{-1} \left(\boldsymbol{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{d}\right) \sim \chi^2_r, \tag{2.29}$$

where $\boldsymbol{A}(\widehat{\boldsymbol{\theta}}) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}(\widehat{\boldsymbol{\theta}})^{-1}\boldsymbol{X}\right)^{-1}$. Under the null hypothesis, $\mathcal{W}$ also approximately follows a chi-squared distribution with r degrees of freedom.

However, in many applications, covariates or nuisance variables are not of primary interest, and people are more interested in testing whether the nonparametric function f significantly contributes to the trait in Eq. (2.11), that is, testing the null hypothesis $\mathcal{H}_0 : f(\cdot) = 0$. Using the LMM representation (2.20), it can be seen that testing this null hypothesis is equivalent to testing $\mathcal{H}_0 : \tau^2 = 0$ against the alternative $\mathcal{H}_1 : \tau^2 > 0$. Note that under the null hypothesis, τ^2 lies on the boundary of the parameter space (since it cannot be negative), and the kernel matrix $\boldsymbol{K}$ is not block diagonal, making standard LRT inapplicable (Self and Liang, 1987). Liu et al. (2007) proposed a *score test* to address this issue. The score test (also known as the *Lagrange multiplier test*) assesses whether a parameter of interest θ is equal to a particular value θ_0 under the null using a test statistic that in general takes the form:

$$\mathcal{S} = \left.\frac{\mathcal{U}(\theta)}{\mathcal{I}_{\mathrm{E}}(\theta)^{1/2}}\right|_{\theta=\theta_0}, \tag{2.30}$$

where $\mathcal{U}(\theta) = \partial\ell(\theta)/\partial\theta$ is the derivative of the log likelihood with respect to θ, known as the *score*, and $\mathcal{I}_{\mathrm{E}}(\theta)$ is the *Fisher information* (or expected information) of θ. Here, using the results derived in Eqs. (B.3) and (B.6) in Appendix B, an ReML version of the score and Fisher information can be calculated as

$$\begin{aligned}
\mathcal{U}(\tau^2)\Big|_{\tau^2=0} &= \left.\frac{\partial\ell_{\mathrm{R}}}{\partial\tau^2}\right|_{\tau^2=0} = -\frac{1}{2\sigma_0^2}\mathbf{tr}\{\boldsymbol{P}_0\boldsymbol{K}\} + \frac{1}{2\sigma_0^4}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{K}\boldsymbol{P}_0\boldsymbol{y}, \\
\mathcal{I}_{\mathrm{E}}(\tau^2)\Big|_{\tau^2=0} &= \mathbf{E}\left[-\frac{\partial^2\ell_{\mathrm{R}}}{(\partial\tau^2)^2}\right]_{\tau^2=0} = \frac{1}{2\sigma_0^4}\mathbf{tr}\{\boldsymbol{P}_0\boldsymbol{K}\boldsymbol{P}_0\boldsymbol{K}\},
\end{aligned} \tag{2.31}$$

where $\mathbf{E}[\cdot]$ denotes expectation, σ_0^2 is the variance of the residual $\boldsymbol{\epsilon}_0$ under the null regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}_0$, and $\boldsymbol{P}_0 = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ is the projection matrix under the null. Since $\mathbf{tr}\{\boldsymbol{P}_0\boldsymbol{K}\}$ and $\mathbf{tr}\{\boldsymbol{P}_0\boldsymbol{K}\boldsymbol{P}_0\boldsymbol{K}\}$ are constants independent of $\boldsymbol{y}$, the null hypothesis $\mathcal{H}_0 : \tau^2 = 0$ can be tested using the following score test statistic:

$$\mathcal{S}(\sigma_0^2) = \frac{1}{2\sigma_0^2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{K}\boldsymbol{P}_0\boldsymbol{y} = \frac{1}{2\sigma_0^2}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_0\right)^{\mathrm{T}}\boldsymbol{K}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_0\right), \tag{2.32}$$

which is a measure of the association between the residuals estimated from the null model and the kernel matrix. $\mathcal{S}(\sigma_0^2)$ is a quadratic function of $\boldsymbol{y}$ and thus follows a mixture of chi-squares under the null. Specifically,

$$\mathcal{S}(\sigma_0^2) = \sum_{i=1}^{n}\lambda_i\chi_{1,i}^2, \tag{2.33}$$

where $\{\lambda_i\}_{i=1}^n$ are eigenvalues of the matrix $\frac{1}{2}\boldsymbol{P}_0^{1/2}\boldsymbol{K}\boldsymbol{P}_0^{1/2}$, and $\chi_{1,i}^2$ are independent and identically distributed (i.i.d.) random variables following chi-squared distributions with 1 degree of freedom. The p-value of an observed score test statistic can be analytically computed by the Davies (1980) method or Kuonen's saddle point method Kuonen (1999) using Eq. (2.33). Liu et al. (2007) proposed to use the *Satterthwaite method* to approximate the distribution of $\mathcal{S}(\sigma_0^2)$ by a scaled chi-squared distribution $\delta\chi_\nu^2$, where δ is the scale parameter and ν denotes the degrees of freedom. The two parameters are estimated by matching the first two moments, mean and variance, of $\mathcal{S}(\sigma_0^2)$ with those of $\delta\chi_\nu^2$:

$$\begin{cases} \zeta := \mathbf{E}[\mathcal{S}(\sigma_0^2)] = \frac{1}{2}\mathbf{tr}\{\boldsymbol{P}_0\boldsymbol{K}\} = \mathbf{E}[\delta\chi_\nu^2] = \delta\nu, \\ \xi := \mathbf{var}[\mathcal{S}(\sigma_0^2)] = \frac{1}{2}\mathbf{tr}\{\boldsymbol{P}_0\boldsymbol{K}\boldsymbol{P}_0\boldsymbol{K}\} = \mathbf{var}[\delta\chi_\nu^2] = 2\delta^2\nu. \end{cases} \tag{2.34}$$

Solving the two equations yields $\delta = \xi/2\zeta$ and $\nu = 2\zeta^2/\xi$. In practice, σ_0^2 is often unknown and is replaced by its maximum likelihood estimate $\hat{\sigma}_0^2$ under the null model. To account for this substitution, ξ needs to be replaced by $\hat{\xi}$ based on efficient information (Zhang and Lin, 2003): $\hat{\xi} = \widehat{\mathcal{I}}_{\tau\tau} = \mathcal{I}_{\tau\tau} - \mathcal{I}_{\sigma\sigma}^{-1}\mathcal{I}_{\tau\sigma}^2$, where $\mathcal{I}_{\tau\tau} = \mathbf{tr}\{\boldsymbol{P}_0\boldsymbol{K}\boldsymbol{P}_0\boldsymbol{K}\}/2$, $\mathcal{I}_{\tau\sigma} = \mathbf{tr}\{\boldsymbol{P}_0\boldsymbol{K}\boldsymbol{P}_0\}/2$ and $\mathcal{I}_{\sigma\sigma} = \mathbf{tr}\{\boldsymbol{P}_0\boldsymbol{P}_0\}/2$. Here $\mathcal{I}_{\tau\tau}$, $\mathcal{I}_{\tau\sigma}$ and $\mathcal{I}_{\sigma\sigma}$ are proportional to elements in the Fisher information matrix of τ^2 and σ^2 (see Appendix B). With the adjusted parameters $\hat{\delta} = \hat{\xi}/2\zeta$ and $\hat{\nu} = 2\zeta^2/\hat{\xi}$, the p-value of an observed score statistic $\mathcal{S}(\hat{\sigma}_0^2)$ is then computed using the scaled chi-squared distribution $\hat{\delta}\chi_{\hat{\nu}}^2$. One advantage of this score test is that it only requires fitting of a linear fixed effects model under the null hypothesis, and thus can be highly computationally efficient and suitable for analyzing a large number of traits. Pan (2011) showed that when no covariate needs to be adjusted other than an intercept,

the score test statistic is equivalent to the genomic distance-based regression (GDBR) (Wessel and Schork, 2006), which is based on the Gower distance (Gower, 1966) and the pseudo-F statistic (McArdle and Anderson, 2001). The F-statistic is also closely related to the sum of squared score (SSU) test (Pan, 2011) and the Goeman's test (Goeman et al., 2006; Pan, 2009).

2.2.5 CONSTRUCTING AND SELECTING KERNELS

Kernel methods are appealing for their flexibility and generality; any non-negative definite kernel function can be used to measure the similarity between attributes from pairs of individuals and explain the trait variation. However, this flexibility can sometimes make the selection and comparison of kernels challenging. For example, it can be difficult to select a kernel that best captures the characteristics of the data or most powerfully detects a specific mechanism from a collection of valid kernels. In most studies, a kernel is a priori selected from commonly used candidates, such as the linear kernel, polynomial kernel and Gaussian kernel, or specifically designed in order to address a unique scientific question. See Schaid (2010b) and Hofmann et al. (2008) for reviews of kernel functions proposed in genomic studies and machine learning, respectively. However, there are existing methods that can build, compare and select kernels in a more systematic and objective way.

First, new kernels can be created by using existing non-negative definite kernel functions as building blocks. Assuming that $\{k_t\}_{t=1}^{+\infty}$ is a sequence of kernels defined on $\mathcal{Z} \times \mathcal{Z}$ with the associated kernel matrices $\{\boldsymbol{K}_t\}_{t=1}^{+\infty}$ evaluated at a set of attributes $\{z_1, \ldots, z_n\}$, and $z, z' \in \mathcal{Z}$ are arbitrary attributes, then:

- For any $\gamma_1, \gamma_2 \geqslant 0$, the linear combination $(\gamma_1 k_1 + \gamma_2 k_2)(z, z') := \gamma_1 k_1(z, z') + \gamma_2 k_2(z, z')$ is a new kernel function with the associated kernel matrix $\gamma_1 \boldsymbol{K}_1 + \gamma_2 \boldsymbol{K}_2$. This property can be useful when jointly modeling data from different sources and/or at different scales using multiple kernels.
- The point-wise product $(k_1 k_2)(z, z') := k_1(z, z') \cdot k_2(z, z')$ is a new kernel function with the associated kernel matrix $\boldsymbol{K}_1 \circ \boldsymbol{K}_2$, where $\circ$ is the Hadamard product (element-wise product) of two matrices. This can be useful when modeling the interaction of two kernels defined on the same space.
- $k(z, z') := \lim_{t \to +\infty} k_t(z, z')$ is a new kernel function with the associated kernel matrix $\boldsymbol{K} = \lim_{t \to +\infty} \boldsymbol{K}_t$, if the limit exists for arbitrary z and z'.

Many more kernels can be created using a combination of these mathematical operations. For example, any polynomial of a kernel, $\tilde{k}(z, z') := \sum_l \gamma_l k^l(z, z')$ with $\gamma_l \geqslant 0$, gives a new kernel; the exponentiation of a kernel, $\tilde{k}(z, z') := \exp\{k(z, z')\}$, is also a kernel since it can be expanded into a convergent sequence of polynomials. Moreover, given two kernels k_1 and k_2 defined on $\mathcal{Z}_1 \times \mathcal{Z}_1$ and $\mathcal{Z}_2 \times \mathcal{Z}_2$, with the associated kernel matrices $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ evaluated at $\{z_{1,1}, \ldots, z_{1,n}\}$ and $\{z_{2,1}, \ldots, z_{2,n}\}$, respectively, and arbitrary attributes $z_1, z_1' \in \mathcal{Z}_1, z_2, z_2' \in \mathcal{Z}_2$, then:

- The tensor product $(k_1 \otimes k_2)\left((z_1, z_2), (z_1', z_2')\right) := k_1(z_1, z_1') \cdot k_2(z_2, z_2')$ is a new kernel on the product domain $(\mathcal{Z}_1 \times \mathcal{Z}_2) \times (\mathcal{Z}_1 \times \mathcal{Z}_2)$, with the associated kernel matrix $\boldsymbol{K}_1 \circ \boldsymbol{K}_2$. This can be useful to construct kernels on the tensor product of two RKHSs.

In practice, it can be difficult to directly specify a kernel to capture complex relationships between pairs of attributes. However, the properties presented above suggest that if the effect of interest can be decomposed into components that can be well characterized by primitive kernels, an advanced kernel can be constructed using a bottom-up approach.

Second, specifying a kernel function requires a similarity measure between pairs of attributes, but sometimes it is more natural to measure dissimilarity or distance. In this case, we note that any distance measure, $d = d(z, z')$, can be converted into a similarity measure $-\frac{1}{2}d^2$, which is known as the Gower distance (Gower, 1966). For arbitrary attributes $\{z_1, \ldots, z_n\}$, this gives a similarity matrix $-\frac{1}{2}\boldsymbol{D} \circ \boldsymbol{D}$, where $\boldsymbol{D} = [d_{ij}]_{n \times n}$ with $d_{ij} = d(z_i, z_j)$. We now center the Gower distance matrix and define $\boldsymbol{K} = \boldsymbol{H}\left[-\frac{1}{2}\boldsymbol{D} \circ \boldsymbol{D}\right]\boldsymbol{H}$, where $\boldsymbol{H}$ is a centering matrix with the ijth entry $\boldsymbol{H}_{ij} = \delta_{ij} - 1/n$, δ_{ij} being the Kronecker delta. Note that $\boldsymbol{K}$ is not guaranteed to be non-negative definite. However, when the Euclidean distance, $d_{ij} = \|z_i - z_j\| = \sqrt{\sum_l (z_{il} - z_{jl})^2}$, is used, we have

$$\boldsymbol{K} = \boldsymbol{H}\left[-\frac{1}{2}\boldsymbol{D} \circ \boldsymbol{D}\right]\boldsymbol{H} = \boldsymbol{H}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{H}, \tag{2.35}$$

where $\boldsymbol{Z} = (z_1, \ldots, z_n)^{\mathrm{T}}$. Clearly, $\boldsymbol{K}$ is now non-negative definite, and more specifically, $\boldsymbol{K}$ is a centered linear kernel of the attributes. This offers a way to induce a valid kernel from Euclidean distance, perhaps the most widely used distance measure in practice.

Finally, Liu et al. (2007) pointed out that both model selection and variable selection are special cases of kernel selection within the KMR framework. They proposed AIC and BIC in the context of KMR, which can evaluate candidate kernels in a systematic and objective way. The calculation of AIC/BIC normally requires the number of estimated parameters in the model. However, KMR models are nonparametric and thus the number of model parameters is not explicitly defined. To address this issue, Liu et al. (2007) noted that, using Eqs. (2.17) and (2.19), the estimated trait, $\widehat{\boldsymbol{y}}$, can be expressed as

$$\begin{aligned}\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{f}} &= \left\{\hat{\tau}^2\boldsymbol{K}\widehat{\boldsymbol{V}}^{-1} + \hat{\sigma}^2\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\right\}\boldsymbol{y} \\ &= \left(\boldsymbol{I} - \hat{\sigma}^2\widehat{\boldsymbol{P}}\right)\boldsymbol{y} := \boldsymbol{S}\boldsymbol{y},\end{aligned} \tag{2.36}$$

where $\widehat{\boldsymbol{V}} = \boldsymbol{V}(\widehat{\boldsymbol{\theta}}_{\text{ReML}})$ and $\widehat{\boldsymbol{P}} = \boldsymbol{P}(\widehat{\boldsymbol{\theta}}_{\text{ReML}}) = \widehat{\boldsymbol{V}}^{-1} - \widehat{\boldsymbol{V}}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{V}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\text{T}}\widehat{\boldsymbol{V}}^{-1}$. The matrix $\boldsymbol{S}$ smoothes the observed data $\boldsymbol{y}$ to produce the predicted trait $\widehat{\boldsymbol{y}}$, and its trace, $\mathbf{tr}\{\boldsymbol{S}\}$, can be interpreted as a measure of model complexity. To see this, we notice that Eq. (B.3) gives

$$\frac{\partial \ell_{\text{R}}}{\partial \sigma^2} = -\frac{1}{2}\mathbf{tr}\{\boldsymbol{P}\} + \frac{1}{2}\boldsymbol{y}^{\text{T}}\boldsymbol{P}\boldsymbol{P}\boldsymbol{y}. \tag{2.37}$$

Thus at the ReML estimate $\hat{\sigma}^2$ that maximizes the restricted likelihood ℓ_{R}, the derivative is zero and we have $\mathbf{tr}\{\widehat{\boldsymbol{P}}\} = \boldsymbol{y}^{\text{T}}\widehat{\boldsymbol{P}}\widehat{\boldsymbol{P}}\boldsymbol{y}$. Also, by the definition of $\boldsymbol{S}$, we have $\mathbf{tr}\{\boldsymbol{S}\} = \mathbf{tr}\left\{\boldsymbol{I} - \hat{\sigma}^2\widehat{\boldsymbol{P}}\right\} = n - \hat{\sigma}^2\mathbf{tr}\{\widehat{\boldsymbol{P}}\}$. The residual sum of squares (RSS) can then be computed as

$$\text{RSS} = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^{\text{T}}(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \hat{\sigma}^4\boldsymbol{y}^{\text{T}}\widehat{\boldsymbol{P}}\widehat{\boldsymbol{P}}\boldsymbol{y} = \hat{\sigma}^4\mathbf{tr}\{\widehat{\boldsymbol{P}}\} = \hat{\sigma}^2(n - \mathbf{tr}\{\boldsymbol{S}\}), \tag{2.38}$$

and thus $\hat{\sigma}^2 = \text{RSS}/(n - \mathbf{tr}\{\boldsymbol{S}\})$, indicating that $\mathbf{tr}\{\boldsymbol{S}\}$ is the loss in degrees of freedom resulting from estimating the fixed effects $\boldsymbol{\beta}$ and the random effects $\boldsymbol{f}$ when estimating σ^2. Therefore Liu et al. (2007) proposed the KMR-based AIC and BIC as

$$\begin{aligned} \text{AIC} &= n\log(\text{RSS}) + 2\mathbf{tr}\{\boldsymbol{S}\}, \\ \text{BIC} &= n\log(\text{RSS}) + \mathbf{tr}\{\boldsymbol{S}\}\cdot\log(n). \end{aligned} \tag{2.39}$$

Both measures reward goodness-of-fit (the first term) and penalize complex models (the second term) in order to avoid overfitting. BIC has a larger penalty term than AIC for large n, and thus favors simpler models. Models with smaller AIC/BIC values are selected and believed to be better descriptions of the data.

2.2.6 THEORETICAL EXTENSIONS

In previous sections, we have introduced the basics of KMR, focusing on univariate (scalar) and quantitative traits collected from unrelated individuals whose attributes are modeled by a single nonparametric function. In this section, we review recent theoretical developments that generalized the classical KMR model to handle more complex data structures.

2.2.6.1 Generalized kernel machine regression

KMR can be extended to handle a much wider class of data types, such as binary and count data, whose distribution lies in the *exponential family* (McCullagh and Nelder, 1989; Liu et al., 2008; Wu et al., 2010). Specifically, suppose that the density function of the trait for the ith subject, y_i, takes the *canonical form* (or *natural form*):

$$p(y_i|\eta_i, \phi) = h(y_i, \phi)\exp\left\{\frac{1}{\phi}\left[\eta_i y_i - a(\eta_i)\right]\right\}, \tag{2.40}$$

where η_i is the *natural parameter*, ϕ is a scale or dispersion parameter, and $h(\cdot,\cdot)$ and $a(\cdot)$ are known functions. $a(\cdot)$ is a normalization factor (known as the *log-partition*

function) that ensures the distribution sums or integrates to one, and is connected to the moments of y_i:

$$\mu_i = \mathbf{E}[y_i|\eta_i] = a'(\eta_i), \quad \mathbf{var}[y_i|\eta_i,\phi] = \phi a''(\eta_i) := \phi v(\mu_i), \tag{2.41}$$

where we have defined the variance function $v(\cdot)$ that associates the first and second moments of y_i. The generalized kernel machine regression (GKMR) then connects the mean of the distribution function, μ_i, with the covariates $\boldsymbol{x}_i$ and attributes $\boldsymbol{z}_i$, using a monotonic *link function* g:

$$g(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + f(\boldsymbol{z}_i), \quad i = 1,2,\ldots,n, \tag{2.42}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients, and f is an unknown function that lies in an RKHS $\mathcal{H}$ defined by a kernel function k. When $g^{-1}(\cdot) = a'(\cdot)$, g is the *canonical link* and we have $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + f(\boldsymbol{z}_i)$. Table 2.1 lists the natural parameter, scale parameter, mean of the trait, variance function and canonical link for linear, logistic and Poisson regressions, which can handle normal, binary and count data, respectively. In the derivation below, we always assume that the canonical link is used.

Model (2.42) can be fitted by minimizing the panelized likelihood function:

$$\mathcal{J}(\boldsymbol{\beta},f) = -\sum_{i=1}^{n}[\eta_i y_i - a(\eta_i)] + \frac{\lambda}{2}\|f\|_{\mathcal{H}}^2, \tag{2.43}$$

where the first part is proportional to the minus log likelihood of the model, ignoring a constant independent of $\boldsymbol{\beta}$ and f, while the second part is a regularization term penalizing rough functions. By the Representer theorem, the minimizer takes the form $f(\cdot) = \sum_{j=1}^{n}\alpha_j k(\cdot,\boldsymbol{z}_j)$, and $\mathcal{J}$ has a finite-dimensional representation:

$$\mathcal{J}(\boldsymbol{\beta},\boldsymbol{\alpha}) = -\sum_{i=1}^{n}[\eta_i y_i - a(\eta_i)] + \frac{\lambda}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}, \quad \eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\kappa}_i^{\mathrm{T}}\boldsymbol{\alpha}, \tag{2.44}$$

Table 2.1 The Natural Parameter, Scale Parameter, Mean of the Trait, Variance Function and Canonical Link for Linear, Logistic and Poisson Regressions

Regression	η	ϕ	μ	$v(\mu)$	$g(\cdot)$
Linear	μ	σ^2	μ	1	id$(\cdot)$
Logistic	$\log\left(\frac{\pi}{1-\pi}\right)$	1	π	$\mu(1-\mu)$	logit$(\cdot)$
Poisson	$\log(\lambda)$	1	λ	μ	log$(\cdot)$

id$(\cdot)$ is the identity function. π is the probability of an outcome of interest or event. logit$(\cdot)$ is the logit function. λ is the rate of occurrence of an event.

where $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]^{\mathrm{T}}$ and $\boldsymbol{\kappa}_i = [k(\boldsymbol{z}_1, \boldsymbol{z}_i), \ldots, k(\boldsymbol{z}_n, \boldsymbol{z}_i)]^{\mathrm{T}}$. Setting the derivatives of $\mathcal{J}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to zero yields the following equations:

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)\boldsymbol{x}_i}{v(\mu_i)g'(\mu_i)} = \mathbf{0}, \qquad \sum_{i=1}^{n} \frac{(y_i - \mu_i)\boldsymbol{\kappa}_i}{v(\mu_i)g'(\mu_i)} = \lambda \boldsymbol{K}\boldsymbol{\alpha}. \tag{2.45}$$

Using the canonical link implies that $g'(\mu_i) = 1/v(\mu_i)$, and the denominators in Eq. (2.45) thus vanish. We note that μ_i depends on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ through the link g, which is in general nonlinear. Thus the solution of Eq. (2.45) does not have a closed form and need to be estimated numerically. Breslow and Clayton (1993) showed that the solution to Eq. (2.45) via Fisher scoring is the iterative solution to the following linear system:

$$\begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{K} \\ \boldsymbol{W}\boldsymbol{X} & \lambda\boldsymbol{I} + \boldsymbol{W}\boldsymbol{K} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}\tilde{\boldsymbol{y}} \\ \boldsymbol{W}\tilde{\boldsymbol{y}} \end{bmatrix}, \tag{2.46}$$

where $\boldsymbol{W}$ is an $n \times n$ diagonal matrix with the ith diagonal element $v(\mu_i)$, and $\tilde{\boldsymbol{y}}$ is an $n \times 1$ working vector with the ith element $\tilde{y}_i = \eta_i + (y_i - \mu_i)g'(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\kappa}_i^{\mathrm{T}}\boldsymbol{\alpha} + (y_i - \mu_i)/v(\mu_i)$. Note that both $\boldsymbol{W}$ and $\tilde{\boldsymbol{y}}$ depend on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and thus Eq. (2.46) needs to be solved iteratively. More specifically, we first fix $\boldsymbol{W}$ and $\tilde{\boldsymbol{y}}$ to solve Eq. (2.46), and then update $\boldsymbol{W}$ and $\tilde{\boldsymbol{y}}$ using the new estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This is repeated until convergence.

It can be seen that Eq. (2.46) also depends on the tuning parameter λ, which is usually unknown. This can be resolved by establishing the connection between GKMR and the following generalized linear mixed effects model (GLMM):

$$g(\mu_i) = g\left(\mathbf{E}[y_i|\eta_i]\right) = \eta_i, \quad \eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + f_i, \quad i = 1, 2, \ldots, n, \tag{2.47}$$

where $\boldsymbol{\beta}$ is a vector of fixed effects and $\boldsymbol{f} = [f_1, \ldots, f_n]^{\mathrm{T}}$ is a vector of random effects following $\mathrm{N}(\mathbf{0}, \tau^2\boldsymbol{K})$ with $\tau^2 = \lambda^{-1}\phi$. The full likelihood-based estimation in GLMM is difficult due to the integral over the random effects. For LMMs, this integral can be analytically computed but in general the problem is intractable and numerical approximations have to be employed. A number of different approaches exist, which can be broadly categorized into integrand approximation (eg, the Laplace approximation), integral approximation (eg, the Gaussian quadrature approximation), and data approximation (eg, the penalized quasi-likelihood methods). Here we show that the penalized quasi-likelihood (PQL) approach is tightly connected to GKMR.

PQL does not specify a full probability distribution but approximates the data by the mean and variance: $y_i \approx \mu_i + \epsilon_i$, where ϵ_i is an error term with $\mathbf{var}[\epsilon_i] = \mathbf{var}[y_i|\eta_i, \phi] = \phi v(\mu_i)$. Then using a Taylor expansion, we have

$$\begin{aligned} y_i &\approx \mu_i + \epsilon_i = g^{-1}\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + f_i\right) + \epsilon_i \\ &\approx g^{-1}\left(\boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \hat{f}_i\right) + a''\left(\boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \hat{f}_i\right)\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + a''\left(\boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \hat{f}_i\right)(f_i - \hat{f}_i) + \epsilon_i \\ &\approx \hat{\mu}_i + v(\hat{\mu}_i)\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + v(\hat{\mu}_i)(f_i - \hat{f}_i) + \epsilon_i, \end{aligned} \tag{2.48}$$

where we have used the identity $[g^{-1}(\eta_i)]' = a''(\eta_i) = v(\mu_i)$ for canonical links. We note that this expansion is exact for LMMs, in which g is an identity function. Reorganizing Eq. (2.48) gives

$$\tilde{y}_i := \hat{\eta}_i + (y_i - \hat{\mu}_i)/v(\hat{\mu}_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + f_i + \epsilon_i/v(\hat{\mu}_i) := \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + f_i + \tilde{\epsilon}_i, \tag{2.49}$$

where $\mathbf{var}[\tilde{\epsilon}_i] = \phi/v(\hat{\mu}_i)$ if we evaluate $\mathbf{var}[\epsilon_i]$ at $\hat{\mu}_i$. Therefore we have approximated the GLMM (2.47) by a working LMM for the pseudo-data $\tilde{y}_i$:

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f} + \tilde{\boldsymbol{\epsilon}}, \quad \boldsymbol{f} \sim \mathrm{N}(\boldsymbol{0}, \tau^2\boldsymbol{K}), \quad \tilde{\boldsymbol{\epsilon}} \sim \mathrm{N}(\boldsymbol{0}, \phi\widehat{\boldsymbol{W}}^{-1}), \tag{2.50}$$

in which $\tilde{\boldsymbol{y}} = [\tilde{y}_1, \ldots, \tilde{y}_n]^{\mathrm{T}}$, $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^{\mathrm{T}}$, $\tilde{\boldsymbol{\epsilon}} = [\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n]^{\mathrm{T}}$, and $\widehat{\boldsymbol{W}}$ is an $n \times n$ diagonal matrix with the ith diagonal element $v(\hat{\mu}_i)$. It is easy to verify that Henderson's MME of the LMM (2.50) is exactly the same as the linear system (2.46), by using the fact that $\boldsymbol{f} = \boldsymbol{K}\boldsymbol{\alpha}$. We have thus bridged GKMR and GLMMs by showing that the minimizer of the panelized likelihood function (2.43) is equivalent to the PQL estimate of the GLMM (2.47). The vector $\boldsymbol{\theta}$, containing ϕ, τ^2, and any unknown parameter in the kernel function, can be estimated by maximizing the restricted likelihood of the working LMM:

$$\ell_{\mathrm{R}}(\boldsymbol{\theta}) = -\frac{1}{2}\log|\boldsymbol{V}(\boldsymbol{\theta})| - \frac{1}{2}\log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}| - \frac{1}{2}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\theta})^{-1}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}), \tag{2.51}$$

where $\boldsymbol{V}(\boldsymbol{\theta}) = \tau^2\boldsymbol{K} + \phi\widehat{\boldsymbol{W}}^{-1}$ and $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}\boldsymbol{V}(\boldsymbol{\theta})^{-1}\tilde{\boldsymbol{y}}$. Note that the pseudo-data $\tilde{\boldsymbol{y}}$ and the matrix $\widehat{\boldsymbol{W}}$ depend on the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{f}$, and thus the GKMR model needs to be fitted by iteratively solving the working linear system (2.46) and maximizing the working restricted likelihood (2.51) until convergence.

Statistical inferences of the GKMR can also be conducted within the framework of GLMMs. Specifically, once the GLMM has been fitted, standard likelihood ratio test (LRT) and Wald test can be applied to the fixed effects $\widehat{\boldsymbol{\beta}}$. To test the null hypothesis $\mathcal{H}_0 : f(\cdot) = 0$, or equivalently $\mathcal{H}_0 : \tau^2 = 0$ against the alternative $\mathcal{H}_1 : \tau^2 > 0$ in the GLMM, a score statistic can be derived by using the linear approximation (2.50) and the null model $g(\mu_{0,i}) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_0$:

$$\mathcal{S}(\phi_0) = \frac{1}{2\phi_0^2}\tilde{\boldsymbol{y}}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{K}\boldsymbol{P}_0\tilde{\boldsymbol{y}} = \frac{1}{2\phi_0^2}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_0)^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}_0), \tag{2.52}$$

where ϕ_0 is the scale parameter of the null GLMM, $\boldsymbol{P}_0 = \widehat{\boldsymbol{W}}_0 - \widehat{\boldsymbol{W}}_0\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{W}}_0\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{W}}_0$ is the null projection matrix, $\widehat{\boldsymbol{W}}_0$ is an $n \times n$ diagonal matrix with the ith element $v(\hat{\mu}_{0,i})$, $\hat{\mu}_{0,i} = g^{-1}\left(\boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_0\right)$ is the estimated mean under the null, $\widehat{\boldsymbol{\mu}}_0 = [\hat{\mu}_{0,1}, \ldots, \hat{\mu}_{0,n}]^{\mathrm{T}}$. $\mathcal{S}(\phi_0)$ follows a mixture of chi-squares under the null and can be approximated by a scaled chi-squared distribution using the Satterthwaite method. The estimated scale parameter and degrees of freedom of the chi-squared distribution literally take the same form as in the linear case if we assume that $\tilde{\boldsymbol{y}}$ is approximately normal. For very skewed data, taking into account the influences of

high-order moments such as the *kurtosis* can provide more accurate approximations to the distribution of the score test statistic (Lin, 1997).

2.2.6.2 Multiple kernel functions

The KMR can also be extended to include multiple kernel functions (Gianola and van Kaam, 2008). These kernel functions can either be defined on the same input space, capturing different aspects of the same attribute or jointly modeling multiple attributes, or be defined on different input spaces, integrating data from different domains. In general, assuming quantitative traits, a KMR model with multiple kernel functions can be written as

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \sum_{l=1}^{\varsigma} f_l(z_{il}) + \epsilon_i, \tag{2.53}$$

where ς is the number of kernel functions in the model, each attribute z_{il} belongs to an input space $\mathcal{Z}_l$, and f_l is an unknown function lying in an RKHS $\mathcal{H}_l$ defined by the kernel function k_l. A parallel proof of the classical KMR theory shows that fitting model (2.53) by minimizing its penalized likelihood function is equivalent to fitting the LMM:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{l=1}^{\varsigma} \boldsymbol{f}_l + \boldsymbol{\epsilon}, \tag{2.54}$$

where $\boldsymbol{\epsilon} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, $\boldsymbol{f}_l \sim \mathrm{N}(\boldsymbol{0}, \tau_l^2\boldsymbol{K}_l)$, and $\lambda_l = \sigma^2/\tau_l^2$ is the tuning parameter controlling the penalization on the lth nonparametric function f_l. Variance component parameters σ^2, τ_l^2, $l = 1, 2, \ldots, \varsigma$, and other unknown parameters in the kernel functions can be estimated by maximizing the restricted likelihood (B.1), in which the marginal covariance structure is $\boldsymbol{V} = \sum_{l=1}^{\varsigma} \tau_l^2\boldsymbol{K}_l + \sigma^2\boldsymbol{I}$.

A score test statistic can be computed to assess whether at least one function in a subset of the nonparametric functions $\{f_l\}_{l=1}^{\varsigma}$ is significantly different from zero. More specifically, let ϖ be any subset of the indexes: $\varpi \subset \{1, 2, \ldots, \varsigma\}$. Suppose we test the null hypothesis that all functions whose indexes in ϖ are zero, that is, $\mathcal{H}_0 : f_l(\cdot) = 0$, for all $l \in \varpi$, against the alternative that at least one of these functions is significantly different from zero, the null model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \sum_{l \notin \varpi} \boldsymbol{f}_l + \boldsymbol{\epsilon}_0, \quad \boldsymbol{f}_l \sim \mathrm{N}(\boldsymbol{0}, \tau_{0,l}^2\boldsymbol{K}_l), \quad \boldsymbol{\epsilon}_0 \sim \mathrm{N}(\boldsymbol{0}, \sigma_0^2\boldsymbol{I}), \tag{2.55}$$

and the score test statistic can be constructed as

$$\mathcal{S}(\sigma_0^2; \tau_{0,l}^2, l \notin \varpi) = \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{K}\boldsymbol{P}_0\boldsymbol{y}, \tag{2.56}$$

where $\boldsymbol{P}_0 = \boldsymbol{V}_0^{-1} - \boldsymbol{V}_0^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_0^{-1}$ is the projection matrix under the null, and $\boldsymbol{V}_0 = \sum_{l \notin \varpi} \tau_{0,l}^2\boldsymbol{K}_l + \sigma_0^2\boldsymbol{I}$ is the marginal covariance matrix under the null. The Satterthwaite method can be used to approximate the distribution of $\mathcal{S}(\sigma_0^2; \tau_{0,l}^2, l \notin$

ϖ) by a scaled chi-squared distribution. The unknown model parameters σ_0^2 and $\tau_{0,l}^2, l \notin \varpi$, can be replaced by their ReML estimates under the null model in practice, with the scale parameter and the degrees of freedom of the chi-squared distribution being adjusted using efficient information to account for this substitution.

2.2.6.3 Correlated phenotypes

Previous sections have focused on phenotypes collected from individuals in the general population, and thus can be reasonably assumed to be independent. However, correlated phenotypes are frequently seen in practice. For example, family or pedigree-based designs have been widely used in many fields, and are particularly valuable in health-related research such as the study of rare diseases. Phenotypes of closely related individuals are often highly correlated, due to shared genetics and common environmental factors. Direct application of classical KMR to family/pedigree data by ignoring the familial structure leads to misspecified models and may result in inflated type I error in statistical inferences. The KMR model (2.11) can be easily extended, by incorporating a random effect, to appropriately account for familial correlation (Schifano et al., 2012; Chen et al., 2013):

$$y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \pi_i + f(\boldsymbol{z}_i) + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2.57}$$

where $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_n]^{\mathrm{T}} \sim \mathrm{N}(\mathbf{0}, \sigma_g^2\boldsymbol{\Pi})$ is a vector of random effects modeling the familial correlation, σ_g^2 is the total additive genetic variance, and $\boldsymbol{\Pi} = 2\boldsymbol{\Phi}$ is twice the *kinship matrix* and indicates expected genetic covariance among individuals. The ijth entry of the kinship matrix, ϕ_{ij}, known as the kinship coefficient, defines genetic relatedness for subjects i and j, and can in general be derived from pedigree information. For example, for identical (monozygotic) twins $\phi_{ij} = 1/2$, for full siblings and parent-offspring $\phi_{ij} = 1/4$, for half siblings and grandparent-grandchild $\phi_{ij} = 1/8$. Model (2.57) can then be converted into an LMM, which has been used for decades in the field of quantitative genetics for human pedigree analysis:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\pi} + \boldsymbol{f} + \boldsymbol{\epsilon}, \tag{2.58}$$

where $\boldsymbol{\pi} \sim \mathrm{N}(\mathbf{0}, \sigma_g^2\boldsymbol{\Pi})$, $\boldsymbol{f} \sim \mathrm{N}(\mathbf{0}, \tau^2\boldsymbol{K})$, and $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$. Testing the null hypothesis $\mathcal{H}_0 : f(\cdot) = 0$, or equivalently $\mathcal{H}_0 : \tau^2 = 0$ using a score test falls in the framework set in the above section by noticing that the null covariance matrix is $\boldsymbol{V}_0 = \sigma_{0,g}^2\boldsymbol{\Pi} + \sigma_0^2\boldsymbol{I}$, where $\sigma_{0,g}^2$ and σ_0^2 are parameters in the null model and can be estimated using the ReML approach.

2.2.6.4 Multidimensional traits

With the rapid advances in phenotyping technology, it is now not uncommon to collect multiple secondary phenotypes that characterize a health-related outcome from different angles. These phenotypes can be highly related and describe a common mechanism underlying biological processes, and thus a joint analysis that

accounts for their correlation structure may improve statistical power relative to conducting analysis on individual phenotypes. The classical KMR can be extended to model multiple traits (Maity et al., 2012). In particular, consider a total of ϱ traits collected on each individual, each trait modeled by the following KMR:

$$y_{il} = \boldsymbol{x}_{il}^{\mathrm{T}}\boldsymbol{\beta}_l + f_l(\boldsymbol{z}_i) + \epsilon_{il}, \quad l = 1, 2, \ldots, \varrho, \tag{2.59}$$

where f_l is an unknown function lying in an RKHS $\mathcal{H}_l$ defined by the kernel function k_l. The residuals are independent across individuals but are correlated across trait dimensions: $[\epsilon_{i1}, \ldots, \epsilon_{i\varrho}]^{\mathrm{T}} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, reflecting the covariance among phenotypes. It is easy to see, using the classical KMR theory, that each trait has a corresponding LMM:

$$\boldsymbol{y}_l = \boldsymbol{X}_l\boldsymbol{\beta}_l + \boldsymbol{f}_l + \boldsymbol{\epsilon}_l, \quad l = 1, 2, \ldots, \varrho, \tag{2.60}$$

where $\boldsymbol{y}_l = [y_{1l}, \ldots, y_{nl}]^{\mathrm{T}}$, $\boldsymbol{X}_l = [\boldsymbol{x}_{1l}, \ldots, \boldsymbol{x}_{nl}]^{\mathrm{T}}$, $\boldsymbol{f}_l \sim \mathrm{N}(\boldsymbol{0}, \tau_l^2\boldsymbol{K}_l)$, and $\boldsymbol{\epsilon}_l = [\epsilon_{1l}, \ldots, \epsilon_{nl}]^{\mathrm{T}}$. By stacking these individual LMMs, we have the following joint LMM:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{f} + \boldsymbol{\epsilon}, \tag{2.61}$$

where $\boldsymbol{y} = [\boldsymbol{y}_1^{\mathrm{T}}, \ldots, \boldsymbol{y}_\varrho^{\mathrm{T}}]^{\mathrm{T}}$, $\boldsymbol{X} = \mathbf{diag}\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_\varrho\}$ is a block diagonal matrix, $\boldsymbol{\beta} = [\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_\varrho^{\mathrm{T}}]^{\mathrm{T}}$, $\boldsymbol{f} = [\boldsymbol{f}_1^{\mathrm{T}}, \ldots, \boldsymbol{f}_\varrho^{\mathrm{T}}]^{\mathrm{T}} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Lambda}\boldsymbol{K})$, $\boldsymbol{\Lambda} = \mathbf{diag}\{\tau_1^2\boldsymbol{I}, \ldots, \tau_\varrho^2\boldsymbol{I}\}$, $\boldsymbol{K} = \mathbf{diag}\{\boldsymbol{K}_1, \ldots, \boldsymbol{K}_\varrho\}$, and $\boldsymbol{\epsilon} = [\boldsymbol{\epsilon}_1^{\mathrm{T}}, \ldots, \boldsymbol{\epsilon}_\varrho^{\mathrm{T}}]^{\mathrm{T}} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{I})$, $\otimes$ being the Kronecker product between matrices.

To test the null hypothesis $\mathcal{H}_0 : f_l(\cdot) = 0$, $l = 1, 2, \ldots, \varrho$, or equivalently $\mathcal{H}_0 : \tau_l^2 = 0$, $l = 1, 2, \ldots, \varrho$, the score test statistic is

$$\mathcal{S}(\boldsymbol{\Sigma}_0) = \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{K}\boldsymbol{P}_0\boldsymbol{y}, \tag{2.62}$$

where $\boldsymbol{P}_0 = \boldsymbol{V}_0^{-1} - \boldsymbol{V}_0^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}_0^{-1}$ is the projection matrix under the null, and $\boldsymbol{V}_0 = \boldsymbol{\Sigma}_0 \otimes \boldsymbol{I}$ is the marginal covariance matrix under the null. When all the traits adjust for a common set of covariates and all the unknown functions f_l lie in an RKHS defined by the same kernel function, that is, $\boldsymbol{X}_1 = \boldsymbol{X}_2 = \cdots = \boldsymbol{X}_\varrho := \boldsymbol{X}_{\mathrm{c}}$ and $\boldsymbol{K}_1 = \boldsymbol{K}_2 = \cdots = \boldsymbol{K}_\varrho := \boldsymbol{K}_{\mathrm{c}}$, we have $\boldsymbol{X} = \boldsymbol{I} \otimes \boldsymbol{X}_{\mathrm{c}}$, $\boldsymbol{K} = \boldsymbol{I} \otimes \boldsymbol{K}_{\mathrm{c}}$, $\boldsymbol{P}_0 = \boldsymbol{\Sigma}_0^{-1} \otimes \boldsymbol{P}_{0,\mathrm{c}}$, where $\boldsymbol{P}_{0,\mathrm{c}} = \boldsymbol{I} - \boldsymbol{X}_{\mathrm{c}}(\boldsymbol{X}_{\mathrm{c}}^{\mathrm{T}}\boldsymbol{X}_{\mathrm{c}})^{-1}\boldsymbol{X}_{\mathrm{c}}^{\mathrm{T}}$, and the score test statistic can be simplified as

$$\begin{aligned}
\mathcal{S}(\boldsymbol{\Sigma}_0) &= \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\left[\boldsymbol{\Sigma}_0^{-2} \otimes (\boldsymbol{P}_{0,\mathrm{c}}\boldsymbol{K}_{\mathrm{c}}\boldsymbol{P}_{0,\mathrm{c}})\right]\boldsymbol{y} \\
&= \frac{1}{2}\mathbf{vec}(\boldsymbol{Y})^{\mathrm{T}}\left[\boldsymbol{\Sigma}_0^{-2} \otimes (\boldsymbol{P}_{0,\mathrm{c}}\boldsymbol{K}_{\mathrm{c}}\boldsymbol{P}_{0,\mathrm{c}})\right]\mathbf{vec}(\boldsymbol{Y}) \\
&= \frac{1}{2}\mathbf{tr}\left\{\boldsymbol{P}_{0,\mathrm{c}}\boldsymbol{K}_{\mathrm{c}}\boldsymbol{P}_{0,\mathrm{c}}\boldsymbol{Y}\boldsymbol{\Sigma}_0^{-2}\boldsymbol{Y}^{\mathrm{T}}\right\},
\end{aligned} \tag{2.63}$$

where $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_\varrho]$. Here $\mathcal{S}(\boldsymbol{\Sigma}_0)$ essentially quantifies the association between the attribute similarity matrix $\boldsymbol{K}_{\mathrm{c}}$ and the phenotypic similarity measured by the

matrix $\boldsymbol{Y}\boldsymbol{\Sigma}_0^{-2}\boldsymbol{Y}^{\mathrm{T}}$. When the trait is a scalar, $\mathcal{S}(\boldsymbol{\Sigma}_0)$ degenerates to the score test statistic for the classical KMR. The distribution of $\mathcal{S}(\boldsymbol{\Sigma}_0)$ can again be approximated using the Satterthwaite method. In practice, we fit individual linear regression models under the null to obtain estimates of the residuals, $\widehat{\boldsymbol{\epsilon}}_{0,l} = \boldsymbol{y}_l - \boldsymbol{X}_l\widehat{\boldsymbol{\beta}}_{0,l}$, and then estimate the residual covariance matrix by $\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\widehat{\mathcal{E}}^{\mathrm{T}}\widehat{\mathcal{E}}$, where $\widehat{\mathcal{E}} = [\widehat{\boldsymbol{\epsilon}}_{0,1}, \ldots, \widehat{\boldsymbol{\epsilon}}_{0,Q}]$.

2.3 APPLICATIONS

In this section, we review recent applications of KMR with a focus on biomedical research. Most of the work has a genetic component since kernel methods are particularly useful to flexibly model the joint effect of a collection of genetic variants on a phenotype of interest.

2.3.1 GENETIC ASSOCIATION STUDIES

The past decade has witnessed tremendous scientific and biological discoveries made through genome-wide association studies (GWASs), an experimental design that detects the association between millions of individual single nucleotide polymorphisms (SNPs, a DNA sequence variation occurring at a single nucleotide in the genome) and a wide range of clinical conditions (Visscher et al., 2012; Gratten et al., 2014; Psychiatric Genomics Consortium, 2014). However, GWASs require large sample size (hundreds of thousands) to achieve stringent statistical significance and identify robust and replicable associations, and can only be applied to common genetic variants (occurring more than 1% in a population). To date, a majority of the SNPs identified by GWASs are not in protein-coding regions, and thus do not have direct implications for disease diagnosis or treatment.

SNP set-based analyses offer a complementary method to GWASs by modeling a collection of genetic variants, which open opportunities to test the cumulative effect of rare genetic variants and dissect complex interactions in the genome. By grouping SNPs based on a priori biological knowledge such as genes, pathways, functional annotations and previous GWAS findings, set-based association studies can alleviate the burden of multiple testing correction, improve robustness, reproducibility and statistical power relative to univariate methods, and provide more interpretable results and biological insights.

Conventional methods model a collection of SNPs in a genomic region within the multiple regression framework, and use regularization techniques, such as the ridge regression, LASSO and elastic net (Kohannim et al., 2011, 2012b,a), or PCA (Hibar et al., 2011), to handle collinearity of SNP regressors due to linkage disequilibrium (LD, statistical associations between co-segregated SNPs). Recently, a large body of the literature has been devoted to the development of *burden tests*, which collapse a set of genetic variants in a genomic region into a single burden variable. The burden variable can be dichotomous, indicating the presence of any rare variant within a

region (eg, cohort allelic sum test, CAST) (Morgenthaler and Thilly, 2007), the count of rare variants in the SNP set (Morris and Zeggini, 2010), a weighted average of the number of minor alleles (eg, weighted sum test, WST) (Madsen and Browning, 2009), or a hybrid of these methods (Li and Leal, 2008). Pan et al. (2014) showed that both the sum test and the sum of squared score (SSU) test are special cases of a family of tests, termed sum of powered score (SPU) tests, which also has an adaptive version (aSPU) (Pan et al., 2014, 2015). A major criticism of burden tests is that they rely on the strong assumptions that all variants being modeled are causal and the effects are in the same direction, and may suffer from dramatic power loss when these assumptions are violated.

KMR offers an alternative way to conduct SNP set-based association studies by employing a *sequence kernel* that defines the similarity between a pair of strings. Kernel-base association tests belong to the class of *nonburden tests* that is robust to the direction of SNPs and the proportion of causal SNPs in the set. A variety of kernels can be used to characterize different genetic contributions to the trait. For example, a linear kernel models the additive effects of SNPs (Yang et al., 2010, 2011), while a nonparametric identity-by-state (IBS) kernel provides a biologically informed way to capture the epistasis (interactions) in the SNP set. As a concrete illustration, consider two genetic variants `A` and `B`, with the corresponding minor and major alleles (the less and more common alleles in the population) represented with lower and upper case, respectively. Now, suppose that two subjects have the following genotypes at the two loci: [`AA`; `bb`] and [`aa`; `Bb`]. For the linear kernel, assuming that the minor allele is considered as the reference allele, the above genotypes can be coded as [0; 2] and [2; 1]. The linear similarity between the two subjects is then computed as an inner product between these two vectors, normalized by the number of loci, giving $(0 * 2 + 2 * 1)/2 = 1$. Note that, with this definition, the choice of the reference allele impacts the similarity measure and thus there is an implicit directionality. However, in practice, the genotypes are often standardized (subtracted by the mean and divided by the standard deviation across subjects) before the linear kernel is computed, and the linear similarity for the standardized genotype is independent of the choice of the reference allele. For the IBS kernel, the similarity between a pair of subjects is calculated as the number of identical alleles, normalized by the total number of alleles. Therefore the IBS similarity between the above two subjects is $(0 + 1)/4 = 1/4$. Note that here the reference allele does not need to be specified and no directionality is assumed. Different weighting strategies can also be used when building kernel functions to up-weight or down-weight SNPs based on their allele frequencies or a priori biological knowledge. When testing the aggregated genetic effect on the trait, the degrees of freedom of the fitted chi-squared distribution is adaptive to the correlation structure of the SNPs, and thus the KMR approach allows for modeling and testing highly correlated variants.

Liu et al. (2007) laid the theoretical foundation for the kernel-based association tests and applied the technique to testing the pathway effect of multiple gene expressions on prostate cancer. They termed the method sequence kernel association test (SKAT), which has become the basis of many follow-up theoretical extensions

and biomedical applications. Specifically, SKAT has been generalized to handle binary data (case-control studies) (Liu et al., 2008), multidimensional traits (Maity et al., 2012), family data (Schifano et al., 2012; Chen et al., 2013; Ionita-Laza et al., 2013a; Jiang et al., 2014), and survival outcomes (Cai et al., 2011; Lin et al., 2011a; Chen et al., 2014). Kwee et al. (2008) and Wu et al. (2010) introduced SKAT to the genetic community and inspired a series of papers on testing the cumulative effects of rare variants (Wu et al., 2011, 2015; Lee et al., 2013; Ionita-Laza et al., 2013b; Lee et al., 2014), and modeling and detecting gene-by-gene and gene-by-environment (G×E) interactions (Maity and Lin, 2011; Li and Cui, 2012; Lin et al., 2013, 2015; Broadaway et al., 2015; Marceau et al., 2015). To maximize the statistical power of kernel-based association tests, Cai et al. (2012) developed an adaptive score test that up-weights or down-weights the contributions from individual SNPs based on their marginal effects. Lee et al. (2012b,a) proposed an optimal association test that combines SKAT with conventional burden tests, known as SKAT-O.

2.3.2 IMAGING GENETICS

Imaging genetics is an emerging field that identifies and characterizes the genetic basis of brain structure, function and wiring, which play important roles in fundamental cognitive, emotional and behavioral processes, and may be altered in brain-related illnesses (Meyer-Lindenberg and Weinberger, 2006; Thompson et al., 2013). Revealing the true relationship between neuroimaging and genetic variables is challenging because both data types are extremely high-dimensional (millions of genetic variants spanning the genome and hundreds of thousands of voxels/vertices across the image) and have complex covariance structures (genetic variants that are physically close in the genome are often correlated due to LD, and imaging data that are spatially close are also correlated), and sample sizes are typically limited (hundreds or thousands of subjects). Kernel-based methods, due to their modeling flexibility and computational efficiency, have shown promising potential to dissect the genetic underpinnings of the human brain.

Ge et al. (2012) were the first to introduce KMR to the context of imaging genetics. They modeled the aggregated effects and potential interactions of SNPs located in each gene using an IBS kernel, and conducted a voxel-wise, whole-genome, gene-based association study. The kernel method was also combined with a suite of other approaches including a fast implementation of the random field theory that takes use of the spatial information in images, and an efficient permutation procedure. The authors demonstrated that this multivariate analysis framework has boosted statistical power relative to previous massive univariate approaches (Stein et al., 2010). For the first time, some genes were identified to be significantly associated with local volumetric changes in the brain.

Recently, Ge et al. (2015a) proposed a flexible KMR-based method for detecting the interactive effects between multidimensional variable sets, which is particularly useful to identify G×E interactions. Specifically, they introduced three kernels in the KMR framework: one for modeling the joint and epistatic effect of a set of SNPs,

one for accommodating multiple factors that potentially moderate genetic influences, and a third one, which is the Hadamard product of the first two kernels, for capturing the overall interactions between two sets of variables. An initial application of this method to imaging genetics has identified interactive effects between candidate late-onset Alzheimer's disease (AD) risk genes and a collection of cardiovascular disease (CVD) risk factors on hippocampal volume derived from structural brain magnetic resonance imaging (MRI) scans, an imaging biomarker associated with AD risk and future AD progression.

Lastly, Ge et al. (2015b) noticed that using a linear kernel function to combine all the SNPs spanning the genome assesses the total additive genetic effects on the trait and essentially gives a narrow-sense heritability estimate. Leveraging the efficient score test of the KMR, Ge et al. (2015b) proposed a statistical method, termed massively expedited genome-wide heritability analysis (MEGHA), for high-dimensional heritability analysis using genome-wide SNP data from unrelated individuals. This method is thousands of times faster than existing tools and makes heritability-based prioritization of millions of phenotypes tractable for the first time. The authors also developed a permutation-based nonparametric sampling technique within the KMR framework that enables flexible and accurate inferences for arbitrary statistics of interest. As a demonstration of application, Ge et al. (2015b) investigated the genetic basis of morphometric measurements derived from structural MRI, and have created and distributed high-resolution surface maps (containing approximately 300,000 vertices across the two hemispheres) for the heritability estimates and their significance of cortical thickness, sulcal depth, curvature and surface area (https://surfer.nmr.mgh.harvard.edu/fswiki/HeritabilityAnalysis_Ge2015). These maps can be useful to define regions of interest (ROIs) that are under substantial genetic influences. As an example, Fig. 2.3 shows the vertex-wise surface map for the heritability estimates of cortical thickness measurements constructed by MEGHA. The method can also be applied to qualify the heritability of other imaging modalities, or any other types of big data in a variety of settings.

2.4 CONCLUSION AND FUTURE DIRECTIONS

Kernel machine regression (KMR) is a powerful machine learning method, which allows for flexible modeling of multidimensional and heterogeneous data by implicitly specifying the complex relationship between traits and attributes via a knowledge-based similarity measure that characterizes the resemblance between pairs of attributes. Recent technical advances have bridged KMR with mixed effects models in statistics, enabling unified model fitting procedures, and accurate and efficient statistical inferences about model parameters. In this chapter, we have introduced the theoretical basis of KMR and highlighted some of its key extensions. Although we have focused the review on genetic research, which constitutes a large body of the expanding literature on the application of KMR, the method

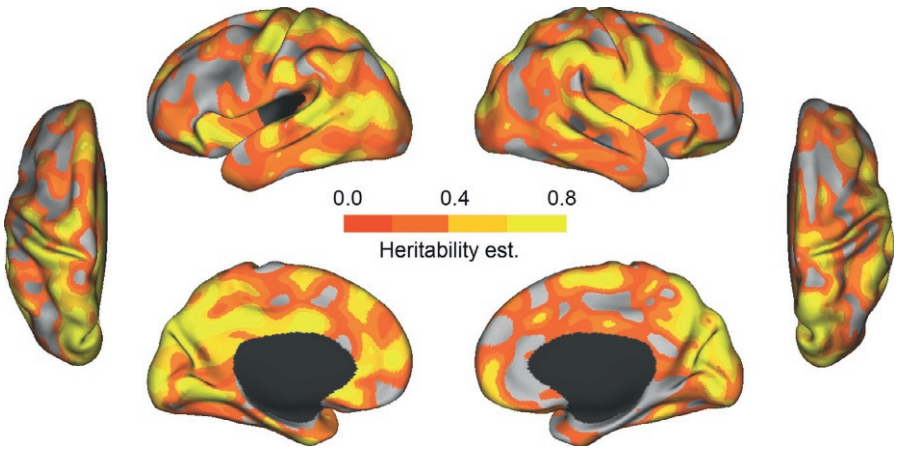

FIG. 2.3

Vertex-wise surface map for the heritability estimates of cortical thickness measurements constructed by MEGHA.

is general enough to explore the relationship between other data types, such as the association between neuroimaging measurements and cognitive, behavioral or diagnostic variables. In fact, the exponential progress in biological technologies is generating massive amounts of data spanning multiple levels of a biological system, from genomic sequences, to intermediate phenotypes such as medical images, and to high-level symptomatic variables. The KMR framework can potentially be used to integrate and jointly analyze different data sources, or be extended to respect the hierarchical structure of these data (Lin et al., 2011b; Huang et al., 2014). With the increasing availability of longitudinal imaging scans (Bernal-Rusiel et al., 2013a,b), KMR seems promising to exploit the high-dimensional imaging space and identify biomarkers that are related to the progression of a brain-related illness and the timing of a clinical event of interest. Last but not least, the research and application of kernel-based methods are not restricted to association detection, but can also encompass prediction, classification, clustering, learning, dimension reduction and variable selection problems, opening vast opportunities for both theoretical advancement and biological discoveries.

ACKNOWLEDGMENTS

This research was carried out at the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital (MGH), using resources provided by the Center for Functional Neuroimaging Technologies, P41EB015896, a P41 Biotechnology Resource Grant

supported by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), National Institutes of Health (NIH).

This research was also funded in part by an MGH Executive Committee on Research (ECOR) Tosteson Postdoctoral Fellowship Award (to TG); NIH grants R01 NS083534, R01 NS070963, and NIBIB 1K25EB013649-01 (to MRS); K24 MH094614 and R01 MH101486 (to JWS); and a BrightFocus Foundation grant AHAF-A2012333 (to MRS). JWS is a Tepper Family MGH Research Scholar.

Appendix A REPRODUCING KERNEL HILBERT SPACES

In this appendix, we briefly review the mathematical foundations of the kernel methods. We restrict our discussion to real vector spaces and kernel functions. For a more detailed introduction to kernel methods, see, for example, Cristianini and Shawe-Taylor (2000) and Schölkopf and Smola (2002).

Appendix A.1 INNER PRODUCT AND HILBERT SPACE

A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is an *inner product* on the *vector space* (or *linear space*) $\mathcal{H}$ if the following conditions are satisfied:

- Symmetry: $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$, for any $f, g \in \mathcal{H}$.
- Bilinearity: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$, for any $f_1, f_2, g \in \mathcal{H}$ and $\alpha_1, \alpha_2 \in \mathbb{R}$.
- Positive definiteness: $\langle f, f \rangle_{\mathcal{H}} \geqslant 0$, for any $f \in \mathcal{H}$, with equality if and only if $f = 0$.

A vector space equipped with an inner product is called an *inner product space* or *pre-Hilbert space*. An inner product induces a *metric* or a *norm* by $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}, f \in \mathcal{H}$, and a *distance* between f and g in $\mathcal{H}$ by $d(f, g) = \|f - g\|_{\mathcal{H}} = \sqrt{\langle f - g, f - g \rangle_{\mathcal{H}}}$.

In a vector space $\mathcal{H}$ with a metric $\| \cdot \|_{\mathcal{H}}$, a sequence $\{f_i\}_{i=1}^{+\infty}$ in $\mathcal{H}$ is said to be a *Cauchy sequence* if for any $\epsilon > 0$, there exists a positive integer $N(\epsilon)$ such that for all positive integers $m, n > N$, $\|f_m - f_n\|_{\mathcal{H}} < \epsilon$. A metric space $\mathcal{H}$ in which every Cauchy sequence converges to an element in $\mathcal{H}$ is *complete*. Intuitively a complete metric space suggests that when the terms of the sequence are getting closer and closer, a limit always exists and it never escapes from the space, that is, the space has no "holes." A *Hilbert space* is a complete inner product space with respect to the norm induced by the inner product.

Appendix A.2 KERNEL FUNCTION AND KERNEL MATRIX

Let $\mathcal{Z}$ be a nonempty set. A function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is called a *kernel function* if there exists a Hilbert space $\mathcal{H}$ with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a map $\varphi : \mathcal{Z} \to \mathcal{H}$ such that for any z and z' in the space $\mathcal{Z}$,

$$k(z, z') = \langle \varphi(z), \varphi(z') \rangle_{\mathcal{H}}. \tag{A.1}$$

Here φ is called a *feature map*, which transforms the data from the input space $\mathcal{Z}$ to a *feature space* $\mathcal{H}$, and can be highly complex and even infinite-dimensional. Kernel methods capture nonlinear patterns in the data by mapping the input to higher dimensions where linear models can be applied.

A function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is *non-negative definite* (or *positive semidefinite*) if for any finite subset $\{z_1, \ldots, z_n\}$ chosen from $\mathcal{Z}$, the *Gram matrix* (or *kernel matrix*) $\boldsymbol{K} = \{k(z_i, z_j)\}_{i,j=1}^n$ is symmetric and non-negative definite, ie, for any real numbers $a_1, \ldots, a_n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(z_i, z_j) \geqslant 0. \tag{A.2}$$

Any kernel function k is clearly symmetric and we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(z_i, z_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \varphi(z_i), a_j \varphi(z_j) \rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{n} a_i \varphi(z_i) \right\|_{\mathcal{H}}^2 \geqslant 0. \tag{A.3}$$

Therefore all kernel functions are non-negative definite. The reverse direction of the statement is also true, that is, for any non-negative definite function k, there exists a Hilbert space $\mathcal{H}$ and a feature map φ, such that Eq. (A.1) is satisfied. This is remarkable because the feature map is often expensive to compute or difficult to explicitly specify, while the kernel function may be easily evaluated and arbitrarily selected as long as it is non-negative definite. We note that a kernel function may rely on additional parameters, such as the width in a Gaussian kernel, as long as it is non-negative definite once the parameters are fixed. A widely used technique in the machine learning community is to substitute a kernel function $k(z, z')$ for a dot product between $\varphi(z)$ and $\varphi(z')$, which implicitly defines a feature map. This is known as the *kernel trick*.

Appendix A.3 REPRODUCING KERNEL HILBERT SPACE

Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on a nonempty set $\mathcal{Z}$. A function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$, and $\mathcal{H}$ is a *reproducing kernel Hilbert space* (RKHS) on $\mathcal{Z}$ (Aronszajn, 1950; Saitoh, 1988), if the followings are satisfied:

- For any $z \in \mathcal{Z}$, $k_z(\cdot) = k(\cdot, z)$ as a function on $\mathcal{Z}$ belongs to $\mathcal{H}$.
- The *reproducing property*: For any $z \in \mathcal{Z}$ and any $f \in \mathcal{H}$, $\langle f(\cdot), k(\cdot, z) \rangle_{\mathcal{H}} = f(z)$.

The reproducing property states that the evaluation of f at z can be expressed as an inner product in the feature space. By applying this property, we have, for any $z, z' \in \mathcal{Z}$,

$$k(z, z') = \langle k(\cdot, z), k(\cdot, z') \rangle_{\mathcal{H}}. \tag{A.4}$$

Since $\varphi(z) = k(\cdot, z)$ is a valid feature map of k, it can be seen from Eq. (A.4) that every reproducing kernel is indeed a kernel as defined in Eq. (A.1), and is thus non-negative definite.

It can be shown that if a Hilbert space $\mathcal{H}$ of functions on $\mathcal{Z}$ admits a reproducing kernel, then the reproducing kernel is uniquely determined by $\mathcal{H}$. Conversely, given any non-negative definite kernel $k(\cdot, \cdot)$ on $\mathcal{Z}$, there exists a uniquely determined Hilbert space $\mathcal{H}$ of functions on $\mathcal{Z}$, which admits the reproducing kernel k. In fact, the Hilbert space $\mathcal{H}$ can be constructed by completing the function space $\mathcal{H}_0$ spanned by $\{k(\cdot, z) \mid z \in \mathcal{Z}\}$, that is,

$$\mathcal{H}_0 = \left\{ f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, z_i) \;\middle|\; n \in \mathbb{N}, \alpha_i \in \mathbb{R}, z_i \in \mathcal{Z} \right\}, \tag{A.5}$$

with the inner product of the functions f and $g(\cdot) = \sum_{j=1}^{m} \beta_j k(\cdot, y_j)$ from $\mathcal{H}_0$ defined as

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \langle k(\cdot, z_i),\ k(\cdot, y_j) \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(z_i, y_j), \tag{A.6}$$

where $m \in \mathbb{N}$, $\beta_j \in \mathbb{R}$, and $y_j \in \mathcal{Z}$. Therefore there is a one-to-one correspondence between non-negative definite kernels and RKHSs. Since $\mathcal{H}_0$ is dense in $\mathcal{H}$, any function in $\mathcal{H}$ can be represented as $f(z) = \sum_{i=1}^{+\infty} \alpha_i k(z, z_i)$, with $\alpha_i \in \mathbb{R}$ and $z_i \in \mathcal{Z}$. The infinite summation is due to the fact that a function in $\mathcal{H}$ may be a limit point of a sequence of functions in $\mathcal{H}_0$. This is the *dual representation* of functions in RKHSs.

Appendix A.4 MERCER'S THEOREM

Mercer's theorem (Mercer, 1909; Schölkopf and Smola, 2002; Cristianini and Shawe-Taylor, 2000) is essentially an analog of the singular value decomposition (SVD) of a matrix in an infinite-dimensional space. It states that under certain regulatory conditions, a kernel function k can be expanded in terms of eigenvalues and orthonormal eigenfunctions of an operator induced by k. More formally, suppose k is a continuous non-negative definite kernel function on a compact set $\mathcal{Z}$. Let $L^2(\mathcal{Z})$ be the space of square-integrable real-valued functions on $\mathcal{Z}$. Define the integral operator $T_k : L^2(\mathcal{Z}) \to L^2(\mathcal{Z})$ by

$$(T_k f)(\cdot) = \int_{\mathcal{Z}} k(\cdot, z) f(z) \mathrm{d}z. \tag{A.7}$$

Then $k(z, z')$ can be expanded into a set of orthonormal basis $\{\psi_i\}$ of $L^2(\mathcal{Z})$ consisting of the eigenfunctions of T_k, and the corresponding sequence of non-negative eigenvalues $\{\lambda_i\}$:

$$k(z, z') = \sum_{i=1}^{\infty} \lambda_i \psi_i(z) \psi_i(z'), \tag{A.8}$$

where the convergence is absolute and uniform. Any squared-integrable function on $\mathcal{Z}$ can thus be represented as $f(z) = \sum_{i=1}^{\infty} \omega_i \sqrt{\lambda_i}\psi_i(z)$, with $\omega_i \in \mathbb{R}$ and $\sum_{i=1}^{\infty} \omega_i^2 < +\infty$. This is the *primal representation* of functions in RKHSs. The inner product of the functions f and $g(\cdot) = \sum_{j=1}^{\infty} \upsilon_j \sqrt{\lambda_j}\psi_j(\cdot)$ from $\mathcal{H} = L^2(\mathcal{Z})$ is defined as $\langle f, g\rangle = \sum_{i=1}^{\infty} \omega_i \upsilon_i$. We thus have

$$\langle f(\cdot), k(\cdot, z)\rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} \frac{\omega_i \lambda_i \sqrt{\lambda_i}\psi_i(z)}{\lambda_i} = \sum_{i=1}^{\infty} \omega_i \sqrt{\lambda_i}\psi_i(z) = f(z), \tag{A.9}$$

that is, the reproducing property of the kernel function. Moreover, a feature map can be explicitly written as $\varphi(z) = [\cdots, \lambda_\ell \psi_\ell(z), \cdots]^{\mathrm{T}}$.

Appendix A.5 REPRESENTER THEOREM

The representer theorem (Kimeldorf and Wahba, 1971) shows that solutions of a large class of optimization problems can be expressed as linear combinations of kernel functions centered on the observed data. Specifically, let $\mathcal{L} : \{\mathcal{Z} \times \mathbb{R}^2\}^n \to \mathbb{R}$ be an arbitrary loss function, and $\Omega : [0, \infty) \to \mathbb{R}$ be a strictly monotonic increasing function. Then each solution of the optimization problem

$$\operatorname{argmin}_{f \in \mathcal{H}} \mathcal{L}\left((z_1, f(z_1), y_1), \ldots, (z_n, f(z_n), y_n)\right) + \Omega(\|f\|_{\mathcal{H}}^2) \tag{A.10}$$

can be represented in the form $f = \sum_{i=1}^{n} \alpha_i k(\cdot, z_i)$. Moreover, if the loss function $\mathcal{L}$ is convex, a global minimum is ensured. Here $\Omega(\|f\|_{\mathcal{H}}^2)$ is a regularization term and controls smoothness of the minimizer. To see this, suppose that f can be expanded as $f(z) = \sum_{i=1}^{\infty} \omega_i \sqrt{\lambda_i}\psi_i(z)$ using its primal representation. We note that $\|f\|_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}} = \sum_{i=1}^{+\infty} \omega_i^2 < +\infty$, which indicates that ω_i decays with increasing i and thus de-emphasizes nonsmooth functions. The representer theorem is significant because it states that the solution of an optimization problem in an infinite-dimensional space $\mathcal{H}$, containing linear combinations of kernels centered on arbitrary points of $\mathcal{Z}$ (see Eq. A.5), has a finite-dimensional representation, which is the span of n kernel functions centered on the observed data points.

Appendix B RESTRICTED MAXIMUM LIKELIHOOD ESTIMATION

In this appendix, we review the Newton-Raphson method (Lindstrom and Bates, 1988; Kenward and Roger, 1997) that is often used to maximize the log restricted likelihood, ℓ_{R}, of linear mixed effects models (LMMs), producing unbiased estimates of the variance component parameters. Recall that ℓ_{R} can be written as a correction to the profile log likelihood:

$$\begin{aligned}\ell_{\mathrm{R}}(\boldsymbol{\theta}) &= \ell_{\mathrm{P}}(\boldsymbol{\theta}) - \frac{1}{2}\log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}| \\ &= -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{y}, \end{aligned} \tag{B.1}$$

where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is the marginal covariance of $\boldsymbol{y}$, which is dependent on an s-dimensional vector of unknown parameters $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{y}$, and we have defined $\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}$. The last equality in Eq. (B.1) is based on the identities $\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{P}\boldsymbol{y}$ and $\boldsymbol{P}\boldsymbol{V}\boldsymbol{P} = \boldsymbol{P}$. By making use of the following results on matrix derivatives:

$$\frac{\partial \log|\boldsymbol{V}|}{\partial\theta_i} = \mathbf{tr}\left\{\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial\theta_i}\right\}, \quad \frac{\partial \boldsymbol{V}^{-1}}{\partial\theta_i} = -\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial\theta_i}\boldsymbol{V}^{-1}, \quad \frac{\partial \boldsymbol{P}}{\partial\theta_i} = -\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_i}\boldsymbol{P}, \tag{B.2}$$

where $\mathbf{tr}\{\cdot\}$ is the trace of a matrix, we have the gradient or the score:

$$\frac{\partial \ell_{\mathrm{R}}}{\partial\theta_i} = -\frac{1}{2}\mathbf{tr}\left\{\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_i}\right\} + \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_i}\boldsymbol{P}\boldsymbol{y}, \quad i = 1, 2, \ldots, s, \tag{B.3}$$

and the ijth element of the $s \times s$ observed information matrix $\mathcal{I}_{\mathrm{O}}$:

$$\begin{aligned}[\mathcal{I}_{\mathrm{O}}]_{ij} = -\frac{\partial^2 \ell_{\mathrm{R}}}{\partial\theta_i\partial\theta_j} &= -\frac{1}{2}\mathbf{tr}\left\{\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_i}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_j}\right\} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_i}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_j}\boldsymbol{P}\boldsymbol{y} \\ &\quad + \frac{1}{2}\mathbf{tr}\left\{\boldsymbol{P}\frac{\partial^2 \boldsymbol{V}}{\partial\theta_i\partial\theta_j}\right\} - \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\frac{\partial^2 \boldsymbol{V}}{\partial\theta_i\partial\theta_j}\boldsymbol{P}\boldsymbol{y}, \quad i, j = 1, 2, \ldots, s. \end{aligned} \tag{B.4}$$

We notice that

$$\begin{aligned}\mathbf{E}\left\{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_i}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_j}\boldsymbol{P}\boldsymbol{y}\right\} &= \mathbf{tr}\left\{\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_i}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_j}\right\}, \\ \mathbf{E}\left\{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\frac{\partial^2 \boldsymbol{V}}{\partial\theta_i\partial\theta_j}\boldsymbol{P}\boldsymbol{y}\right\} &= \mathbf{tr}\left\{\boldsymbol{P}\frac{\partial^2 \boldsymbol{V}}{\partial\theta_i\partial\theta_j}\right\}, \end{aligned} \tag{B.5}$$

and thus the ijth element of the $s \times s$ Fisher information (expected information) matrix $\mathcal{I}_{\mathrm{E}}$ is

$$[\mathcal{I}_{\mathrm{E}}]_{ij} = \mathbf{E}\left[-\frac{\partial^2 \ell_{\mathrm{R}}}{\partial\theta_i\partial\theta_j}\right] = \frac{1}{2}\mathbf{tr}\left\{\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_i}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_j}\right\}. \tag{B.6}$$

When the covariance structure $\boldsymbol{V}(\boldsymbol{\theta})$ is a linear function of the unknown parameters $\boldsymbol{\theta}$, the second-order derivatives in Eq. (B.4) vanish, and the ijth element of the average information matrix, $\mathcal{I}_{\mathrm{A}} = (\mathcal{I}_{\mathrm{O}} + \mathcal{I}_{\mathrm{E}})/2$, has a simple form:

$$[\mathcal{I}_{\mathrm{A}}]_{ij} = \frac{1}{2}[\mathcal{I}_{\mathrm{O}}]_{ij} + \frac{1}{2}[\mathcal{I}_{\mathrm{E}}]_{ij} = \frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_i}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial\theta_j}\boldsymbol{P}\boldsymbol{y}. \tag{B.7}$$

Then given estimates of the unknown variance component parameters at the kth iteration $\boldsymbol{\theta}^{(k)}$, the parameters are iteratively updated by

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \left[\mathcal{I}_{\bullet}^{(k)}\right]^{-1} \left.\frac{\partial \ell_{\mathrm{R}}}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}^{(k)}}, \quad k = 1, 2, \ldots, \tag{B.8}$$

where $\mathcal{I}_{\bullet}$ is either the expected information matrix $\mathcal{I}_{\mathrm{E}}$, leading to the Fisher scoring ReML, or the average information matrix $\mathcal{I}_{\mathrm{A}}$, leading to the average information ReML (Gilmour et al., 1995). At the beginning of the iteration process, all the variance component parameters need to be initialized to reasonable values $\theta_i^{(0)}$. An initial step of the expectation maximization (EM) algorithm (Laird et al., 1987), which is robust to poor starting values, may be used to determine the direction of the updates:

$$\theta_i^{(1)} = \frac{1}{n}\left[\left[\theta_i^{(0)}\right]^2 \boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \theta_i}\boldsymbol{P}\boldsymbol{y} + \mathbf{tr}\left\{\theta_i^{(0)}\boldsymbol{I} - \left[\theta_i^{(0)}\right]^2 \boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \theta_i}\right\}\right], \quad i = 1, 2, \ldots, s. \tag{B.9}$$

The Newton-Raphson algorithm is terminated until the difference between successive log restricted likelihoods, or the gradient of the log restricted likelihood, is smaller than a predefined tolerance, such as 10^{-4}. In the iteration process, parameter estimates may escape from the parameter space (eg, negative estimates of variance parameters), in which case they should be reset to a feasible value close to the boundary of the parameter space.

REFERENCES

Aronszajn, N., 1950. Theory of reproducing kernels. Trans. Am. Math. Soc. 68 (3), 337–404.

Bernal-Rusiel, J.L., Greve, D.N., Reuter, M., Fischl, B., Sabuncu, M.R., et al., 2013a. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. NeuroImage 66, 249–260.

Bernal-Rusiel, J.L., Reuter, M., Greve, D., Fischl, B., Sabuncu, M.R., et al., 2013b. Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. NeuroImage 81, 358–370.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. 88 (421), 9–25.

Broadaway, K.A., Duncan, R., Conneely, K.N., Almli, L.M., Bradley, B., et al., 2015. Kernel approach for modeling interaction effects in genetic association studies of complex quantitative traits. Genet. Epidemiol. 39 (5), 366–375.

Cai, T., Tonini, G., Lin, X., 2011. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics 67 (3), 975–986.

Cai, T., Lin, X., Carroll, R.J., 2012. Identifying genetic marker sets associated with phenotypes via an efficient adaptive score test. Biostatistics 13 (4), 776–790.

Chen, H., Meigs, J.B., Dupuis, J., 2013. Sequence kernel association test for quantitative traits in family samples. Genet. Epidemiol. 37 (2), 196–204.

Chen, H., Lumley, T., Brody, J., Heard-Costa, N.L., Fox, C.S., et al., 2014. Sequence kernel association test for survival traits. Genet. Epidemiol. 38 (3), 191–197.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, Cambridge, MA.

Davies, R.B., 1980. The distribution of a linear combination of χ^2 random variables. J. R. Stat. Soc. C 29, 323–333.

Diggle, P., Heagerty, P., Liang, K.Y., Zeger, S., 2002. Analysis of Longitudinal Data. Oxford University Press, Oxford, UK.

Ge, T., Feng, J., Hibar, D.P., Thompson, P.M., Nichols, T.E., 2012. Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. NeuroImage 63 (2), 858–873.

Ge, T., Nichols, T.E., Ghosh, D., Mormino, E.C., Smoller, J.W., et al., 2015a. A kernel machine method for detecting effects of interaction between multidimensional variable sets: an imaging genetics application. NeuroImage 109, 505–514.

Ge, T., Nichols, T.E., Lee, P.H., Holmes, A.J., Roffman, J.L., et al., 2015b. Massively expedited genome-wide heritability analysis (MEGHA). Proc. Natl. Acad. Sci. USA 112 (8), 2479–2484.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2013. Bayesian Data Analysis. Chapman & Hall, New York.

Gianola, D., van Kaam, J.B.C.H.M., 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178 (4), 2289–2303.

Gilmour, A.R., Thompson, R., Cullis, B.R., 1995. Average information ReML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51 (4), 1440–1450.

Goeman, J.J., Van De Geer, S.A., Van H.C., 2006. Testing against a high dimensional alternative. J. R. Stat. Soc. B 68 (3), 477–493.

Gower, J.C., 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53 (3–4), 325–338.

Gratten, J., Wray, N.R., Keller, M.C., Visscher, P.M., 2014. Large-scale genomics unveils the genetic architecture of psychiatric disorders. Nat. Neurosci. 17 (6), 782–790.

Gu, C., 2013. Smoothing Spline ANOVA Models. Springer Science & Business Media, New York.

Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Stat. Assoc. 72 (358), 320–338.

Hibar, D.P., Stein, J.L., Kohannim, O., Jahanshad, N., Saykin, A.J., et al., 2011. Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. NeuroImage 56 (4), 1875–1891.

Hofmann, T., Schölkopf, B., Smola, A.J., 2008. Kernel methods in machine learning. Ann. Stat. 36 (3), 1171–1220.

Huang, Y.T., VanderWeele, T.J., Lin, X., 2014. Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. Ann. Appl. Stat. 8 (1), 352.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D., Lin, X., 2013a. Family-based association tests for sequence data, and comparisons with population-based association tests. Eur. J. Hum. Genet. 21 (10), 1158–1162.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D., Lin, X., 2013b. Sequence kernel association tests for the combined effect of rare and common variants. Am. J. Hum. Genet. 92 (6), 841–853.

Jiang, Y., Conneely, K.N., Epstein, M.P., 2014. Flexible and robust methods for rare-variant testing of quantitative traits in trios and nuclear families. Genet. Epidemiol. 38 (6), 542–551.

Kenward, M.G., Roger, J.H., 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53 (3), 983–997.

Kimeldorf, G., Wahba, G., 1971. Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. 33 (1), 82–95.

Kohannim, O., Hibar, D.P., Stein, J.L., Jahanshad, N., Jack Jr, C.R., et al., 2011. Boosting power to detect genetic associations in imaging using multi-locus, genome-wide scans and ridge regression. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE, Piscataway, NJ, pp. 1855–1859.

Kohannim, O., Hibar, D.P., Jahanshad, N., Stein, J.L., Hua, X., et al., 2012a. Predicting temporal lobe volume on MRI from genotypes using l1-l2 regularized regression. In: 2012 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. IEEE, Piscataway, NJ, pp. 1160–1163.

Kohannim, O., Hibar, D.P., Stein, J.L., Jahanshad, N., Hua, X., et al., 2012b. Discovery and replication of gene influences on brain structure using LASSO regression. Front. Neurosci. 6, Article 115.

Kuonen, D., 1999. Saddlepoint approximations for distributions of quadratic forms in normal variables. Biometrika 86 (4), 929–935.

Kwee, L.C., Liu, D., Lin, X., Ghosh, D., Epstein, M.P., 2008. A powerful and flexible multilocus association test for quantitative traits. Am. J. Hum. Genet. 82 (2), 386–397.

Laird, N., Lange, N., Stram, D., 1987. Maximum likelihood computations with repeated measures: application of the EM algorithm. J. Am. Stat. Assoc. 82 (397), 97–105.

Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., et al., 2012a. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. 91 (2), 224–237.

Lee, S., Wu, M.C., Lin, X., 2012b. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13 (4), 762–775.

Lee, S., Teslovich, T.M., Boehnke, M., Lin, X., 2013. General framework for meta-analysis of rare variants in sequencing association studies. Am. J. Hum. Genet. 93 (1), 42–53.

Lee, S., Abecasis, G.R., Boehnke, M., Lin, X., 2014. Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. 95 (1), 5–23.

Li, B., Leal, S.M., 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. 83 (3), 311–321.

Li, S., Cui, Y., 2012. Gene-centric gene-gene interaction: a model-based kernel machine method. Ann. Appl. Stat. 6 (3), 1134–1161.

Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73 (1), 13–22.

Lin, X., 1997. Variance component testing in generalized linear models with random effects. Biometrika 84 (2), 309–326.

Lin, X., Cai, T., Wu, M.C., Zhou, Q., Liu, G., et al., 2011a. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. Genet. Epidemiol. 35 (7), 620–631.

Lin, Y.Y., Liu, T.L., Fuh, C.S., 2011b. Multiple kernel learning for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. 33 (6), 1147–1160.

Lin, X., Lee, S., Christiani, D.C., Lin, X., 2013. Test for interactions between a genetic marker set and environment in generalized linear models. Biostatistics 14 (4), 667–681.

Lin, X., Lee, S., Wu, M.C., Wang, C., Chen, H., et al., 2015. Test for rare variants by environment interactions in sequencing association studies. Biometrics 72 (1), 156–164. doi: 10.1111/biom. 12368.

Lindstrom, M.J., Bates, D.M., 1988. Newton-Raphson and EM algorithms for linear mixed–effects models for repeated-measures data. J. Am. Stat. Assoc. 83 (404), 1014–1022.

Liu, D., Lin, X., Ghosh, D., 2007. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 63 (4), 1079–1088.

Liu, D., Ghosh, D., Lin, X., 2008. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Bioinformatics 9 (1), 292.

Madsen, B.E., Browning, S.R., 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 5 (2), e1000384.

Maity, A., Lin, X., 2011. Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines. Biometrics 67 (4), 1271–1284.

Maity, A., Sullivan, P.E., Tzeng, J., 2012. Multivariate phenotype association analysis by marker-set kernel machine regression. Genet. Epidemiol. 36 (7), 686–695.

Marceau, R., Lu, W., Holloway, S., Sale, M.M., Worrall, B.B., et al., 2015. A fast multiple-kernel method with applications to detect gene-environment interaction. Genet. Epidemiol. 39 (6), 456–468.

McArdle, B.H., Anderson, M.J., 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82 (1), 290–297.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. CRC Press, Boca Raton, FL.

McCulloch, C.E., Neuhaus, J.M., 2001. Generalized Linear Mixed Models. Wiley Online Library.

Mercer, J., 1909. Functions of positive and negative type, and their connection with the theory of integral equations. Philos. Trans. R. Soc. Lond. A 209, 415–446.

Meyer-Lindenberg, A., Weinberger, D.R., 2006. Intermediate phenotypes and genetic mechanisms of psychiatric disorders. Nat. Rev. Neurosci. 7 (10), 818–827.

Morgenthaler, S., Thilly, W.G., 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat. Res./Fund. Mole. Mech. Mutagen. 615 (1), 28–56.

Morris, A.P., Zeggini, E., 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. 34 (2), 188.

Pan, W., 2009. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet. Epidemiol. 33 (6), 497.

Pan, W., 2011. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. Genet. Epidemiol. 35 (4), 211–216.

Pan, W., Kim, J., Zhang, Y., Shen, X., Wei, P., 2014. A powerful and adaptive association test for rare variants. Genetics 197 (4), 1081–1095.

Pan, W., Kwak, I.Y., Wei, P., 2015. A powerful pathway-based adaptive test for genetic association with common or rare variants. Am. J. Hum. Genet. 97 (1), 86–98.

Patterson, H.D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58 (3), 545–554.

Psychiatric Genomics Consortium, 2014. Biological insights from 108 schizophrenia-associated genetic loci. Nature 511 (7510), 421–427.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA.
Saitoh, S., 1988. Theory of Reproducing Kernels and its Applications. Longman, Harlow.
Schaid, D.J., 2010a. Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum. Hered. 70 (2), 109–131.
Schaid, D.J., 2010b. Genomic similarity and kernel methods II: methods for genomic information. Hum. Hered. 70 (2), 132–140.
Schifano, E.D., Epstein, M.P., Bielak, L.F., Jhun, M.A., Kardia, S.L.R., et al., 2012. SNP set association analysis for familial data. Genet. Epidemiol. 36 (8), 797–810.
Schölkopf, B., Burges, C.J.C., 1999. Advances in Kernel Methods: Support Vector Learning. MIT Press, Cambridge, MA.
Schölkopf, B., Smola, A.J., 2002. Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA.
Schölkopf, B., Smola, A., Müller, K.R., 1997. Kernel principal component analysis. In: Artificial Neural Networks—ICANN'97. Springer, New York, pp. 583–588.
Schölkopf, B., Smola, A., Müller, K.R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 10 (5), 1299–1319.
Self, S.G., Liang, K.Y., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. 82 (398), 605–610.
Snijders, T.A.B., 2011. Multilevel Analysis. Springer, Berlin.
Stein, J.L., Hua, X., Lee, S., Ho, A.J., Leow, A.D., et al., 2010. Voxel-wise genome-wide association study (vGWAS). NeuroImage 53 (3), 1160–1174.
Thompson, P.M., Ge, T., Glahn, D.C., Jahanshad, N., Nichols, T.E., 2013. Genetics of the connectome. NeuroImage 80, 475–488.
Vapnik, V., 1998. Statistical Learning Theory. Wiley, New York.
Verbeke, G., Molenberghs, G., 2009. Linear Mixed Models for Longitudinal Data. Springer Science & Business Media, New York.
Visscher, P.M., Brown, M.A., McCarthy, M.I., Yang, J., 2012. Five years of GWAS discovery. Am. J. Hum. Genet. 90 (1), 7–24.
Wahba, G., 1990. Spline Models for Observational Data. SIAM Press, Philadelphia, PA.
Wessel, J., Schork, N.J., 2006. Generalized genomic distance-based regression methodology for multilocus association analysis. Am. J. Hum. Genet. 79 (5), 792–806.
Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., et al., 2010. Powerful SNP-set analysis for case-control genome-wide association studies. Am. J. Hum. Genet. 86 (6), 929–942.
Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., et al., 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. 89 (1), 82–93.
Wu, B., Pankow, J.S., Guan, W., 2015. Sequence kernel association analysis of rare variant set based on the marginal regression model for binary traits. Genet. Epidemiol. 39 (6), 399–405.
Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., et al., 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42 (7), 565–569.
Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., 2011. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88 (1), 76–82.
Zhang, D., Lin, X., 2003. Hypothesis testing in semiparametric additive mixed models. Biostatistics 4 (1), 57–74.