DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1148
December 27, 2008

Examining the Relative Influence of Familial, Genetic and Environmental Covariate Information in Flexible Risk Models With Application to Ophthalmology Data

Héctor Corrada Bravo[1]
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Grace Wahba[1]
Department of Statistics, Department of Computer Science and
Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI
Kristine E. Lee[2], Barbara E.K. Klein[2], Ronald Klein[2]
Department of Ophthalmology and Visual Science, University of Wisconsin, Madison, WI
Sudha K. Iyengar[3]
Department of Epidemiology and Biostatistics, Department of Genetics and
Department of Ophthalmology, Case Western Reserve University, Cleveland, OH

[1]Research supported in part by NIH Grant EY09946, NSF Grant DMS-0604572 and ONR Grant N0014-06-0095
[2]Support was provided in part by NIH Grant EY06594, and by the Research to Prevent Blindness (R. and B.E.K. Klein) Senior Scientific Investigator Awards, New York, NY.
[3]Research supported in part by NIH Grant 5R01 EY018510

Examining the Relative Influence of Familial, Genetic and Environmental Covariate Information in Flexible Risk Models With Application to Ophthalmology Data

Héctor Corrada Bravo*

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD 21205-2179
e-mail: hcorrada@jhsph.edu

Grace Wahba*

Departments of Statistics and Biostatistics and Medical Informatics
University of Wisconsin-Madison
Madison, WI 53706-1685
e-mail: wahba@stat.wisc.edu

Kristine E. Lee†, Barbara E.K. Klein†, Ronald Klein†

Department of Ophthalmology and Visual Science
University of Wisconsin-Madison
Madison, WI 53706
e-mail: klee; kleinb; kleinr@epi.ophth.wisc.edu

Sudha K. Iyengar‡

Departments of Epidemiology and Biostatistics,
Genetics and Ophthalmology
Case Western Reserve University
Cleveland, OH 44106
e-mail: ski@case.edu

Abstract:

We present a novel method for examining the relative influence of familial, genetic and covariate information in flexible nonparametric risk models. Our goal is investigating the relative importance of these three sources of information as they are associated with a particular outcome. To that end, we developed a method for incorporating arbitrary pedigree information in a smoothing spline ANOVA (SS-ANOVA) model.

By expressing pedigree data as a positive semidefinite kernel matrix, the SS-ANOVA model is able to estimate a log-odds ratio as a multicomponent function of several variables: one or more functional components representing information from environmental covariates and/or genetic marker data and another representing pedigree relationships. We propose two methods for creating positive semidefinite kernels from pedigree information, including the use of Regularized Kernel Estimation (RKE).

We present a case study on models for retinal pigmentary abnormalities in the Beaver Dam Eye Study (BDES). Our model verifies known facts about the epidemiology of this eye lesion — found in eyes with early age-related macular degeneration (AMD) — and shows significantly increased predictive ability in models that include all three of the genetic, environmental and familial data sources. The case study also shows that models that contain only two of these data sources, that is, pedigree-environmental covariates, or pedigree-genetic markers, or environmental covariates-genetic markers, have comparable predictive ability, while less than the model with all three. This result is consistent with the notions that genetic marker data encodes — at least partly — pedigree data, and that familial correlations encode shared environment data as well.

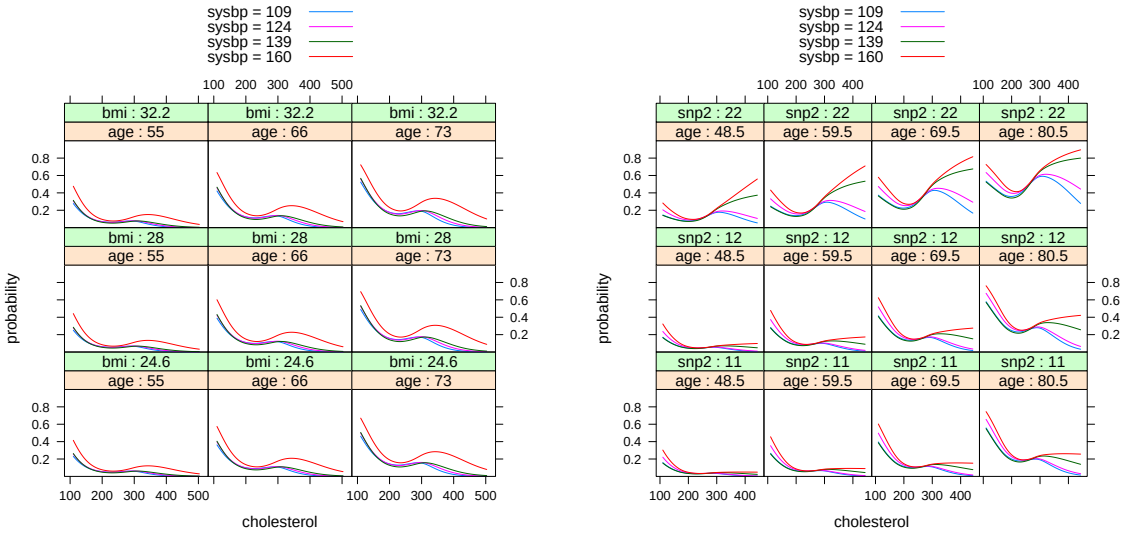
*Research supported in part by NIH Grant EY09946, NSF Grant DMS-0604572 and ONR Grant N0014-06-0095

†Support was provided in part by NIH Grant EY06594, and by the Research to Prevent Blindness (R. and B.E.K. Klein) Senior Scientific Investigator Awards, New York, NY.

‡Research supported in part by NIH Grant 5R01 EY018510

1. Introduction

Smoothing Spline ANOVA (SS-ANOVA) models [1–4] have a successful history modeling ocular traits. In particular, the SS-ANOVA model of retinal pigmentary abnormalities[1] in Lin et al. [2] was able to show an interesting nonlinear protective effect of high total serum cholesterol for a cohort of subjects in the Beaver Dam Eye Study (BDES). We replicate those findings in Figure 1a.[2]



(a) Model with only environmental covariates

(b) Model including SNP rs10490924 from the ARMS2 gene. Label `snp2:22` corresponds to the risk variant of the SNP. Body mass index is fixed at the sample median.

Fig 1: Probability from smoothing spline ANOVA logistic regression model. The x-axis of each plot is total serum cholesterol, each curve is for the indicated fixed value of systolic blood pressure. Each plot fixes body mass index and age to the shown values with $hist = 0$, $horm = 0$ and $smoke = 0$ (see Table 1 for an explanation of model terms).

Multiple studies have reported that risk variants at two loci, near the CFH and ARMS2 genes, show strong association with the development of age-related macular degeneration [5–16], a leading cause of blindness and visual disability [17]. Since retinal pigmentary abnormalities are a precursor to the development of late AMD, we want to make use of genotype data for these two genes to extend the SS-ANOVA model for pigmentary abnormalities risk. For example, by extending the SS-ANOVA model of Lin et al. [2] with SNP rs10490924 in the ARMS2 gene region, we were able to see that the protective effect of total serum cholesterol disappears in older subjects which have the risk variant of this SNP (Figure 1b). Smoothing spline logistic regression models are able to tease out these types of complex nonlinear relationships that would not be detected by more traditional parametric models — linear, or of prespecified form.

Beyond genetic and environmental effects, we want to extend the SS-ANOVA model for pigmentary abnormalities with familial data. Pedigrees (see Section 2) have been ascertained for a large number of subjects of the BDES. The main thrust of this paper is how to incorporate pedigree data into SS-ANOVA models. In fact, we present a general method that is able to incorporate arbitrary relationships encoded as a graph into SS-ANOVA models. This method allows to examine the relative importance of graph relationships compared to other model terms in a predictive model.

The goal of this paper is to estimate models of the log-odds ratio of pigmentary abnormality risk (see Section 3) of the form

$$f(t_i) = \mu + g_1(t_i) + g_2(t_i) + h(z(t_i)),$$

[1]Hereafter we will use the term pigmentary abnormalities when referring to retinal pigmentary abnormalities.

[2]There are minor differences between our plot in Figure 1a and the corresponding plot in Lin et al. [2] since we use a subset of the same cohort. We give details regarding this model in Section 6.

where g_1 is a term that includes only genetic marker data, g_2 is a term containing only environmental covariate data and h is a smooth function over a space encoding relationships, where each subject t_i may be thought of as being represented by a "pseudo-attribute" $z(t_i)$ (see Section 4). In the remainder of the paper we will refer to these model terms as S (for SNP), C (for covariates) and P (for pedigrees); so a model containing all three components will be referred to as S+C+P. In particular, we use models where the g_1 component is an additive linear model, and g_2 is built from cubic splines (see Section 6 for further model details).

An SS-ANOVA model is defined over the tensor sum of multiple reproducing kernel Hilbert spaces (RKHS): one or more components representing information from environmental and/or genetic covariates for each subject and another representing pedigree relationships. It is estimated as the solution of a penalized likelihood problem with an additive penalty including a term for each RKHS in the ANOVA decomposition (Section 3), each weighted by a coefficient. This decomposition allows to measure the relative importance of each model component (S, C or P). Our main tool in extending SS-ANOVA models with pedigree data is the Regularized Kernel Estimation framework [18]. More complex models involving interactions between these three sources of information are possible but beyond the scope of this paper.

The paper is organized as follows: Section 2 defines pedigrees, which encode familial relationships we want to include in the SS-ANOVA model; the SS-ANOVA model itself is discussed in Section 3. The methodology used to extend the SS-ANOVA model with pedigree data is given in Section 4. Results on the extensions of the pigmentary abnormalities risk model in the BDES are given in Section 6. We conclude with a discussion of related and future work in Section 7.

2. Pedigrees

A pedigree is an acyclic graph representing a set of genealogical relationships, where each node corresponds to a member of the family. The graph has an arc from each parent to an offspring, so that each node has two incoming arcs, one for its father and one for its mother, except nodes for founders which have no incoming arcs, and an outgoing arc for each offspring. Figure 2a shows an example of a pedigree.

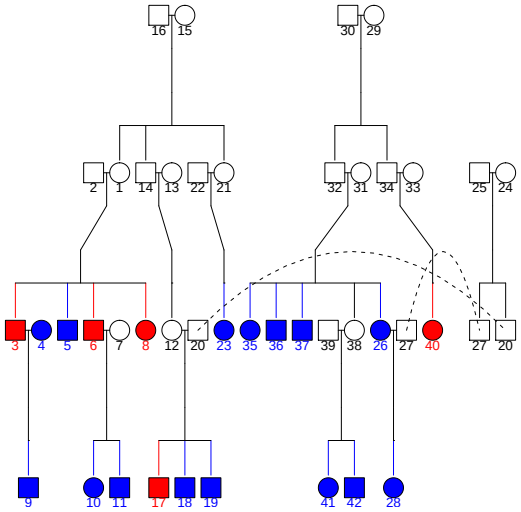
To capture genetic relationships between pedigree members, we use the well-known Malécot kinship coefficient φ [19] to define a pedigree dissimilarity measure. The kinship coefficient between individuals i and j in the pedigree is defined as the probability that a randomly selected pair of alleles, one from each individual, is *identical by descent*, that is, they are derived from a common ancestor. For a parent-offspring pair, $\varphi_{ij} = 1/4$ since there is a 50% chance that the allele inherited from the parent is chosen at random for the offspring, and a 50% chance that the same allele is chosen at random for the parent.

Definition 1 (Pedigree Dissimilarity) *The pedigree dissimilarity between individuals i and j is defined as $d_{ij} = -\log_2(2\varphi_{ij})$, where φ is Malécot's kinship coefficient.*

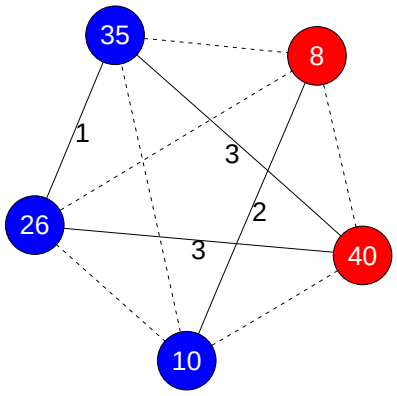
This dissimilarity is also the *degree of relationship* between pedigree members i and j [20]. Another dissimilarity based on the kinship coefficient can be defined as $1 - 2\varphi$. However, since we use Radial Basis Function kernels, defined by an exponential decay with respect to the pedigree dissimilarities, including the exponential decay in φ resulted in overly-diffused kernels (Section 4).

In studies such as the BDES, not all family members are subjects of the study, therefore, the graphs we will use to represent pedigrees in our models only include nodes for subjects rather than the entire pedigree. Furthermore, in our study we want to examine pigmentary abnormality risk in females, so our relationship graphs will only include female subjects. For example, Figure 2b shows the relationship graph for the five female BDES subjects in the pedigree from Figure 2a. Edge labels are the pedigree dissimilarities derived from the kinship coefficient, and dotted lines indicate unrelated pairs.

The main thrust of our methodology is how to incorporate these relationship graphs — derived from pedigrees and weighted by a pedigree dissimilarity that captures genetic relationship — into predictive risk models. In particular, we want to use nonparametric predictive models that incorporate other data, both genetic and environmental. In the next two Sections we will introduce the SS-ANOVA model for Bernoulli data and propose two methods for extending them using relationship graphs.



(a) Example pedigree.



(b) Relationship graph. Edge labels are dissimilarities defined by the kinship coefficient. Dotted edges indicate unrelated pairs.

Fig 2: An example pedigree from the BDES and a relationship graph for five subjects in our cohort. Colored nodes are subjects assessed for retinal pigmentary abnormalities (red encodes a positive result). Circles are females and rectangles are males. Our cohort only includes female subjects assessed for retinal pigmentary abnormalities with full genetic marker and environmental covariate data.

3. Smoothing-Spline ANOVA Models

Assume we are given a data set of environmental covariates and/or genetic markers for each of n subjects, represented as numeric feature vectors x_i, along with responses $y_i \in \{0, 1\}$, $i = 1, \ldots, n$. We use the SS-ANOVA model to estimate the log-odds ratio function $f(x_i) = \log \frac{p(x_i)}{1-p(x_i)}$, where $p(x_i) = \Pr(y_i = 1|x_i)$ [1–4]. In particular, we will assume that f is in an RKHS of the form $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ is a finite dimensional space spanned by a set of functions $\{\phi_1, \ldots, \phi_m\}$, and $\mathcal{H}_1$ is an RKHS induced by a given kernel function $k(\cdot, \cdot)$ with the property that $\langle k(x, \cdot), g\rangle_{\mathcal{H}_1} = g(x)$ for $g \in \mathcal{H}_1$, and thus, $\langle k(x_i, \cdot), k(x_j, \cdot)\rangle_{\mathcal{H}_1} = k(x_i, x_j)$. Therefore, f has a semiparametric form given by

$$f(x) = \sum_{j=1}^{m} d_j \phi_j(x) + g(x),$$

for some coefficients d_j, where the functions ϕ_j have a parametric form and $g \in \mathcal{H}_1$. In the SS-ANOVA model, the RKHS $\mathcal{H}_1$ is decomposed in a particular form we discuss below.

The SS-ANOVA estimate of f given data (x_i, y_i), $i = 1, \ldots, n$, is given by the solution of the following penalized likelihood problem:

$$\min_{f \in \mathcal{H}} I_\lambda(f) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)) + J_\lambda(f), \tag{1}$$

where $l(y_i, f(x_i)) = -y_i f(x_i) + \log(1 + e^{f(x_i)})$ is the negative log likelihood of y_i given $f(x_i)$ and $J_\lambda(f)$ is of the form $\lambda \|P_1 f\|^2_{\mathcal{H}_1}$, with $P_1 f$ the projection of f into RKHS $\mathcal{H}_1$ and λ a non-negative regularization parameter. The penalty term $J_\lambda(f)$ penalizes the complexity of the function f using the norm of the RKHS $\mathcal{H}_1$ to avoid over-fitting f to the training data.

By the representer theorem of Kimeldorf and Wahba [21], the minimizer of the problem in Equation (1) has a finite representation of the form

$$f(\cdot) = \sum_{j=1}^{m} d_j \phi_j(\cdot) + \sum_{i=1}^{n} c_i k(x_i, \cdot),$$

in which case $\|P_1 f\|^2_{\mathcal{H}_1} = c^T K c$ for matrix K with $K_{ij} = k(x_i, x_j)$. Thus, for a given value of the regularization parameter λ the minimizer f_λ can be estimated by solving the following convex nonlinear optimization problem

$$\min_{c \in \mathbb{R}^n, d \in \mathbb{R}^m} \sum_{i=1}^{n} -y_i f(x_i) + \log(1 + e^{f(x_i)}) + n\lambda c^T K c, \tag{2}$$

where $f = [f(x_1) \ \ldots \ f(x_n)]^T = Td + Kc$ with $T_{ij} = \phi_j(x_i)$. The fact that the optimization problem is specified completely by the model matrix T and kernel matrix K is essential to the methods we will use below to incorporate pedigree data to this model.

A method is required for choosing the value of the regularization parameter λ that optimizes performance of estimate f_λ on unseen data in general. In this paper, we will use the GACV method, an approximation to the leave-one-out approximation of the comparative Kullback-Leibler distance between the estimate f_λ and the unknown "true" log-odds ratio f [3]. We note that the kernel function may be parametrized by a set of hyper-parameters that can be chosen using the GACV criterion as well. For example, the Gaussian RBF kernel,

$$k(x_i, x_j) = \exp\{-\gamma \|x_i - x_j\|^2\}, \tag{3}$$

has γ as a hyper-parameter, while the Matérn family of RBF kernels described in Section 5 have scale and order hyper-parameters.

In the SS-ANOVA model, the RKHS $\mathcal{H}_1$ is assumed to be the direct sum of multiple RKHSs, so that the function $g \in \mathcal{H}_1$ is defined as

$$g(x) = \sum_{\alpha} g_\alpha(x_\alpha) + \sum_{\alpha<\beta} g_{\alpha\beta}(x_\alpha, x_\beta) + \cdots$$

where $\{g_\alpha\}$ and $\{g_{\alpha\beta}\}$ satisfy side conditions that generalize the standard ANOVA side conditions. Functions g_α encode "main effects", $g_{\alpha\beta}$ encode "second order interactions" and so on. An RKHS $\mathcal{H}_\alpha$ is associated with each component in this sum, along with its corresponding kernel function k_α. We can write the penalty term in (1) as

$$J_{\lambda,\theta}(f) = \lambda \left[\sum_{\alpha} \theta_\alpha^{-1} \|P_\alpha f\|^2_{\mathcal{H}_\alpha} + \sum_{\alpha\beta} \theta_{\alpha\beta}^{-1} \|P_{\alpha\beta} f\|^2_{\mathcal{H}_{\alpha\beta}} + \cdots \right], \tag{4}$$

where the coefficients θ are tunable hyper-parameters that allow weighting the effect of each component's penalty in the total penalty term. For the penalty of Equation (4), the kernel function $k(\cdot,\cdot)$ associated with $\mathcal{H}_1$ can then be itself decomposed as $k(\cdot,\cdot) = \sum_\alpha \theta_\alpha k_\alpha(\cdot,\cdot) + \sum_{\alpha\beta} \theta_{\alpha\beta} k_{\alpha\beta}(\cdot,\cdot) + \cdots$. The hyper-parameters to be chosen, by GACV for example, now include λ and the coefficients θ of the ANOVA decomposition. In models that have genetic, environmental and familial components, the ANOVA decomposition can be used to measure the relative importance of each data component.

For genetic and environmental components, standard kernel functions can be used to define the corresponding RKHS. However, pedigree data is not represented as feature vectors for which standard kernel functions can be used. On the other hand, in order to specify the penalized likelihood problem, only the kernel matrix is required. In the next Section, we give two methods for building kernel matrices that encode familial relationships, and how to include these in the estimation problem.

4. Representing Pedigree Data as Kernels

The requirement for a valid kernel matrix to be used in the penalized likelihood estimation problem of Equation (2) is that the matrix be positive semidefinite: for any vector $\alpha \in \mathbb{R}^n$, $\alpha^T K \alpha \geq 0$, denoted as

$K \succeq 0$. We saw in the previous Section, that there is a close relationship between the inner product of the RKHS $\mathcal{H}_1$ and its associated kernel function k. In fact, the kernel matrix K is the matrix of inner products of the evaluation representers $k(x, \cdot)$ in $\mathcal{H}_1$ of the given data points.

A property of positive semidefinite matrices, is that they may be interpreted as the matrix of inner products of objects in a space equipped with an inner product. Therefore, since $K \succeq 0$ contains the inner products of objects in some space, we can define a distance metric over these objects as $d^2_{ij} = K_{ii} + K_{jj} - 2K_{ij}$. We make use of this connection between distances and inner products in the Regularized Kernel Estimation framework to define a kernel based on the pedigree dissimilarity of Definition 1.

4.1. Regularized Kernel Estimation

The Regularized Kernel Estimation (RKE) framework was introduced by Lu et al. [18] as a robust method for estimating dissimilarity measures between objects from noisy, incomplete, inconsistent and repetitious dissimilarity data. The RKE framework is useful in settings where object classification or clustering is desired but objects do not easily admit description by fixed length feature vectors. Instead, there is access to a source of noisy and incomplete dissimilarity information between objects.

RKE estimates a symmetric positive semidefinite kernel matrix K which induces a real squared distance admitting of an inner product. K is the solution to an optimization problem with semidefinite constraints that trades-off fit to the observed dissimilarity data and a penalty of the form $\lambda_{rke}\text{trace}(K)$ on the complexity of K, where λ_{rke} is a non-negative regularization parameter.

The solution to the RKE problem is a symmetric positive semidefinite matrix K, which has a spectral decomposition $K = \Gamma\Lambda\Gamma^T$, where Λ is a diagonal matrix with Λ_{ii} equal to the ith leading eigenvalue of K and Γ an orthogonal matrix with eigenvectors as columns in the corresponding order. An embedding $Z \in \mathbb{R}^{n\times r}$ in r-dimensional Euclidean space can be derived from this decomposition by setting $Z = \Gamma_r \Lambda_r^{1/2}$, where the $n \times r$ matrix Γ_r and the $r \times r$ diagonal matrix Λ_r contain the r leading eigenvalues and eigenvectors of K. We refer to the ith row of Z as the vector of "pseudo"-attributes $z(i)$ for subject i. A method for choosing r is required, which we discuss in Section 6.

RKE problem Given a training set of n objects, assume dissimilarity information is given for a subset Ω of the $\binom{n}{2}$ possible pairs of objects. Denote the dissimilarity between objects i and j as $d_{ij} \in \Omega$. We make the requirement that Ω satisfies a connectivity constraint: the undirected graph consisting of objects as nodes and edges between them, such that an edge between nodes i and j is included if $d_{ij} \in \Omega$, is connected. Additionally, optional weights w_{ij} may be associated with each $d_{ij} \in \Omega$.

RKE estimates an n-by-n symmetric positive semidefinite kernel matrix K of size n such that the fitted squared distance between objects induced by K, $\hat{d}^2_{ij} = K_{ii} + K_{jj} - 2K_{ij}$, is as close as possible to the square of the observed dissimilarities $d_{ij} \in \Omega$. Formally, RKE solves the following optimization problem with semidefinite constraints:

$$\min_{K\succeq 0} \sum_{d_{ij}\in\Omega} w_{ij}|d^2_{ij} - \hat{d}^2_{ij}| + \lambda_{rke}\text{trace}(K). \tag{5}$$

The parameter $\lambda_{rke} \geq 0$ is a regularization parameter that trades-off fit of the dissimilarity data, as given by absolute deviation, and a penalty, trace(K), on the complexity of K. The trace may be seen as a proxy for the rank of K, therefore, RKE is regularized by penalizing high dimensionality of the space spanned by K. Note that the trace was used as a penalty function by Lanckriet et al. [22].

As in the SS-ANOVA model, a method for choosing the regularization parameter λ_{rke} is required. However, since our final goal is to build a predictive model that performs well in general, choosing this parameter in terms of prediction performance makes sense. That is, we treat λ_{rke} as a hyper-parameter to the kernel matrix of the SS-ANOVA problem and tune using the GACV criterion.

Figure 3 shows a three-dimensional embedding derived by RKE of the relationship graph in Figure 2b. Notice that the x-axis is order of magnitudes larger than the other two axes and that the unrelated edges in the relationship graph occur along this dimension. That is, the first dimension of this RKE embedding separates the two clusters of relatives in the relationship graph. The remaining dimensions encode the relationship distance.

Not all relationship graphs can be embedded in three-dimensional space, and thus analyzed by inspection as in Figure 3. For example, Figure 5 shows the embedding of a larger relationship graph that requires more

than three-dimensions to embed the pedigree members uniquely. For example, subjects coded 27 and 17 are superposed in this three dimensional embedding, with the fourth dimension separating them.

We may consider the embedding resulting from RKE as providing a set of "pseudo"-attributes $z(i)$ for each subject in this pedigree space. Thus, a smooth predictive function may be estimated in this space. In principle, we should impose a rotational invariance when defining this smooth function since only distance information was used to create the embedding. For this purpose we use the Matérn family of radial basis function kernels described in Section 5.

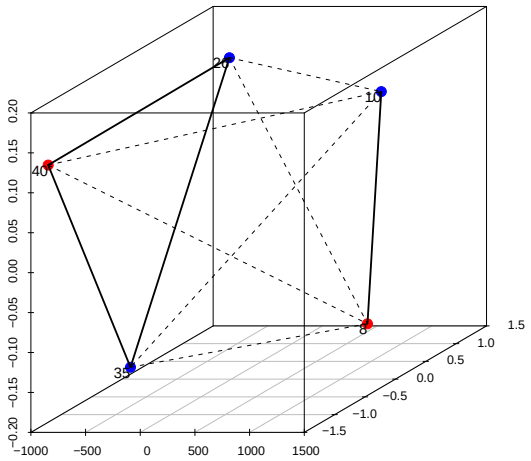


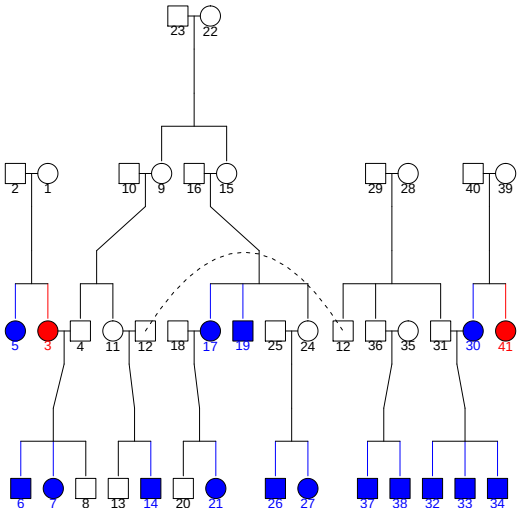
Fig 3: Embedding of pedigree by RKE. The x-axis of this plot is order of magnitudes larger than the other two axes. The unrelated edges in the relationship graph occur along this dimension, while the other two dimensions encode the relationship distance.

The fact that RKE operates on inconsistent dissimilarity data, rather than distances, is significant in this context. The pedigree dissimilarity of Definition 1 is not a distance since it does not satisfy the triangle inequality for general pedigrees. In Figures 4a and 4b we show an example where this is the case, where the dissimilarities between subjects labeled 17, 7 and 5 do not satisfy the triangle inequality. An embedding given by RKE for this graph is shown in Figure 5.

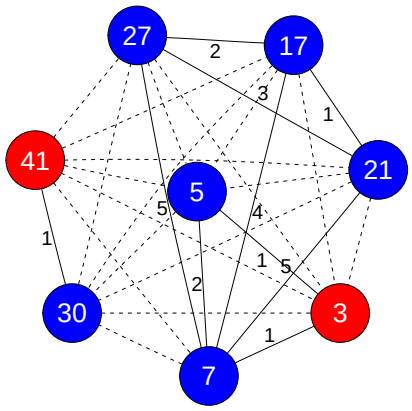
4.2. An alternative to RKE

An alternative to RKE is treating the pedigree dissimilarity as a distance and define a kernel, e.g. a Matérn kernel, directly over it. However, since the pedigree dissimilarity does not satisfy the definitions of a distance, the resulting kernel might not be positive semidefinite. In our implementation, we compute the projection under Frobenius norm of the resulting kernel matrix to the cone of positive semidefinite matrices, easily computed by setting the negative eigenvalues of the matrix to zero.

Since solving the RKE problem is computationally expensive, this is an attractive alternative due to its computational efficiency. On the other hand, the RKE problem gives a sound and principled way of generating a kernel encoding relationships, while this alternative method is ad-hoc. Although this efficient alternative might perform well in some cases, we expect the RKE method to be more robust and work better in the general case. Thus, gains in efficiency must be weighted against possible losses in the general applicability of this alternative method.



(a) Second example pedigree.



(b) Corresponding relationship graph

Fig 4: A different example pedigree and its corresponding relationship graph. The dissimilarities between nodes labeled 17, 7 and 5 in this pedigree show that the pedigree dissimilarity of Definition 1 is not a distance.

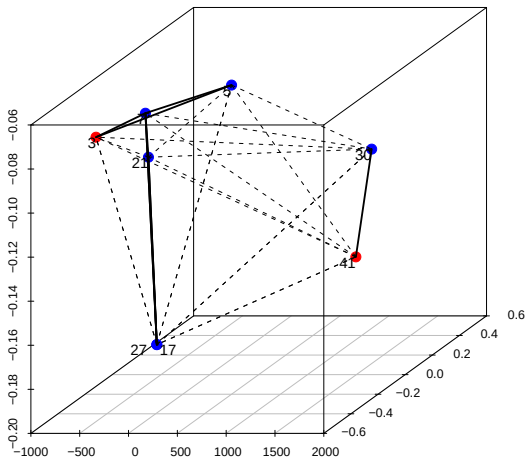


Fig 5: RKE Embedding for second example graph. Subjects 27 and 17 are superimposed in this three dimensional plot, but are separated by the fourth dimension.

5. Matérn Kernel Family

A standard kernel function commonly used is the Gaussian kernel (Equation (3)). This kernel function is a good candidate for this task since it depends only on the distance between objects and is rotationally invariant. However, its exponential decay poses a problem in this setting since the relationship graphs derived from pedigrees are very sparse, and the dissimilarity measure of Definition 1 makes the kernel very diffuse, in that most non-zero entries are relatively small.

The Matérn family of radial basis functions [23, 24] also have the same two appealing features of the Gaussian kernel — dependence only on distance and rotational invariance — while providing a parametrized way of controlling exponential decay. The ν-th order Matérn function is given by

$$k_\nu(i,j) = \exp\{-\alpha d_{ij}\}\pi_\nu(\alpha, d_{ij}), \tag{6}$$

where α is a tunable scale hyper-parameter and π_ν is a polynomial of a certain form. In the results of Sections 6 we use the third order Matérn function:

$$k_3(i,j) = \frac{1}{\alpha^7}\exp\{-\alpha\tau\}[15 + 15\alpha\tau + 6\alpha^2\tau^2 + \alpha^3\tau^3], \tag{7}$$

where $\tau = d_{ij}$. The general recursion relation for the $m+1$-th Matérn function is

$$k_{m+1}(i,j) = \frac{1}{\alpha^{2m+1}}\exp\{-\alpha\tau\}\sum_{i=0}^{m+1} a_{m+1,i}\alpha^i\tau^i, \tag{8}$$

where $a_{m+1,0} = (2m+1)a_{m,0}$, $a_{m+1,i} = (2m+1)a_{m,i} + a_{m,i-1}$, for $i = 1,\ldots,m$ and $a_{m+1,m+1} = 1$. The Matérn family is defined for general positive orders but closed form expressions are available only for integral orders.

6. Case Study: Beaver Dam Eye Study

The Beaver Dam Eye Study (BDES) is an ongoing population-based study of age-related ocular disorders. Subjects at baseline, examined between 1988 and 1990, were a group of 4926 people aged 43-86 years who lived in Beaver Dam, WI. A description of the population and details of the study at baseline may be found in Klein et al. [25]. Although we will only use data from this baseline study for our experiments, five, ten, and fifteen year follow-up data were also obtained [26–28]. Familial relationships of participants were ascertained and pedigrees constructed [29] for the subset of subjects who had at least one relative in the cohort. Genotype data for specific SNPs was subsequently generated for those participants included in the pedigree data.

Our goal was to use this new genetic and pedigree data to extend previous work studying the association between retinal pigmentary abnormalities and a number of environmental covariates using SS-ANOVA models [2]. Retinal pigmentary abnormalities are defined by the presence of retinal depigmentation and increased retinal pigmentation [30, 31]. They are an early sign of age-related macular degeneration (AMD), a leading cause of blindness and visual disability in its late stages [17]. We use genotype data for the Y402H region of the complement factor H (CFH) gene and for SNP rs10490924 in the LOC387715 (ARMS2) gene. Variation in these loci have been shown to significantly alter the risk of AMD [5–16].

Extending the methodology of Lin et al. [2], we estimated SS-ANOVA models of the form

$$\begin{aligned} f(t) = \mu &+ d_{\text{SNP1,1}}\cdot I(X_1 = \texttt{12}) + d_{\text{SNP1,2}}\cdot I(X_1 = \texttt{22}) + \\ &d_{\text{SNP2,1}}\cdot I(X_2 = \texttt{12}) + d_{\text{SNP2,2}}\cdot I(X_2 = \texttt{22}) + \\ &f_1(\text{sysbp}) + f_2(\text{chol}) + f_{12}(\text{sysbp},\text{chol}) + \\ &d_{\text{age}}\cdot \text{age} + d_{\text{bmi}}\cdot \text{bmi} + d_{\text{horm}}\cdot I_1(\text{horm}) + \\ &d_{\text{hist}}\cdot I_2(\text{hist}) + d_{\text{smoke}}\cdot I_3(\text{smoke}) + h(z(t)). \end{aligned} \tag{9}$$

The terms in the first two lines of Equation (9) encode the effect of the two genetic markers (SNPs). A variable for each SNP is coded according to the subject genotype for that SNP as (`11,12,22`). For

identifiability, the `11` level of each SNP is modeled by the intercept μ, while an indicator variable is included for each of the other two levels. This results in each level (other than the `11` level) having its own model coefficient.

The next few terms encode the effect of the environmental covariates listed in Table 1. Functions f_1 and f_2 are cubic splines, while f_{12} uses the tensor product construction [4]. The remaining covariates are modeled as linear terms with I_j as indicator functions. All continuous variables were scaled to lie in the interval $[0, 1]$. A model of pigmentary abnormality risk of this form was shown to report a protective effect of hormone replacement therapy and a suggestion of a nonlinear protective effect of total serum cholesterol [2, and Figure 1a]. The term $h(z(t))$ encodes familial effects and was defined by the kernels presented in Section 4.

TABLE 1
Environmental covariates for BDES pigmentary abnormality risk SS-ANOVA model

code	units	description
`horm`	yes/no	undergoing hormone replacement therapy
`hist`	yes/no	history of heavy drinking
`bmi`	kg/m^2	body mass index
`age`	years	age at baseline
`sysbp`	$mmmHg$	systolic blood pressure
`chol`	mg/dL	total serum cholesterol
`smoke`	yes/no	history of smoking

Models tested include combinations of the following components: 1) P (for pedigree) which defines a function only on an RKHS encoding the pedigree data (term $h(z(t))$ in Equation (9)), 2) S (for SNP) which includes data for the two genetic markers (terms 2 through 5 in Equation (9)), and 3) C (for covariates) which includes the remaining terms in Equation (9) encoding environmental covariates. For example, P-only refers to a model containing only a pedigree component; S+C, to a model containing components for genetic markers and environmental covariates; C-only was the original SS-ANOVA model for pigmentary abnormalities [2]; and P+S+C refers to a model containing components for all three data sources.

We also compared the two methods presented for incorporating pedigree data described in Section 4. We refer to the method using a kernel defined over an embedding resulting from RKE (Section 4.1) as RKE/MATERN, and to the kernel defined over the pedigree dissimilarities directly (Section 4.2), as MATERN. Therefore, the abbreviation P+S+C (MATERN) refers to a model containing all three data sources, where pedigree data were incorporated using the graph kernel method with Matérn third order kernel. We also tested the Gaussian kernel, but it consistently showed similar or worse performance as the Matérn kernels and thus is not reported.

The penalized likelihood problem of Equation (2) was solved by the quasi-Newton method implemented in the `gss R` package [32] using the function `gssanova0` with slight modifications to address some numerical instabilities. The RKE semidefinite problem of Equation (5) was solved using the CSDP library [33] with input dissimilarities given by Definition 1. A number of additional edges between unrelated individuals encoding the "infinite" dissimilarity were added randomly to the graph. The dissimilarity encoded by these edges was arbitrarily chosen to be the sum of all dissimilarities in the entire cohort, while the number of additional edges was chosen such that each subject had an edge to at least twenty-five other subjects in the cohort (including all members of the same pedigree). The kernel matrix obtained from RKE was then truncated to those leading eigenvalues that account for 90% of the matrix trace to create a "pseudo"-attribute embedding. A third order Matérn kernel was then built over the points in the resulting embedding. Pedigree dissimilarities were derived from kinship coefficients calculated using the `kinship R` package [34].

The cohort used were female subjects of the BDES baseline study for which we had full data for genetic markers, environmental covariates and pedigrees. The cohort was further restricted to those from pedigrees containing two or more subjects within the cohort ($n = 684$). This resulted in 175 pedigrees in the data set, with sizes ranging from 2 to 103 subjects. More than a third of the subjects were in pedigrees with 8 or more observations. We chose to only include female subjects in this study to make our model a direct extension to that of [2] which used only female subjects in their cohort and included exposure to hormone replacement therapy as a covariate.

We used area under the ROC curve [35, referred to as AUC], to compare predictive performance of model/method combinations estimated using ten-fold cross-validation. The cross-validation folds were created such that for every subject in each test fold, at least one other member of their pedigree was included in

the labeled training set. Pedigree kernels were built on all members of the study cohort with hyper-parameters chosen independently for each fold using GACV on the labeled training set.

TABLE 2

Ten-fold cross-validation mean for area under ROC curve. Columns correspond to models indexed by components: P (pedigrees), S (genetic markers), C (environmental covariates). Rows correspond to methods tested (NO/PED are SS-ANOVA models without pedigree data). Numbers in parentheses are standard deviations.

		S-only	C-only	S+C
NO/PED		0.6089 (0.05876)	0.6814 (0.07614)	0.7115 (0.04165)
	P-only	S+P	C+P	S+C+P
MATERN	0.6414 (0.11856)	0.7015 (0.12218)	0.7123 (0.07158)	0.7455 (0.08037)
RKE/MATERN	0.6267 (0.10851)	0.6676 (0.08086)	0.6947 (0.10354)	0.7332 (0.06469)

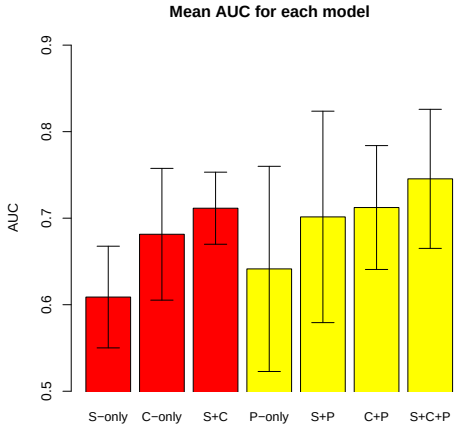


Fig 6: AUC comparison of models. S-only is a model with only genetic markers, C-only is a model with only environmental covariates and S+C is a model containing both data sources. P-only is a model with only pedigree data, P+S is a model with both pedigree data and genetic marker data, P+C is a model with both pedigree data and environmental covariates, P+S+C is a model with all three data sources. Error bars are one standard deviation from the mean. Yellow bars indicate models containing pedigree data. For models containing pedigrees, the best AUC score for each model is plotted. All AUC scores are given in Table 2.

Table 2 shows the resulting mean and standard deviations of the AUC over the ten cross-validation folds of each individual model/method combination. Figure 6 summarizes the same result by plotting the mean AUC of the best method for each model type. We can observe large variation in the AUC reported for most model/method combinations over the cross-validation folds, however some features are apparent: for example, the model with highest overall mean AUC is the S+C+P (MATERN) model. We carried out pairwise t-tests on a few model comparisons and report p-values from estimates where variance is calculated from the differences in AUC between the pair of models being compared over the ten cross-validation folds. Although there was high variability in AUC over the ten cross-validation folds for most individual models, in general, there was much less variation in the difference in AUC between the pairs of models we compare below across the ten cross-validation folds. The next few paragraphs summarize and discuss the results from these tests. [3]

For pedigree-less models, the S+C model containing both markers and covariates had better AUC than either the S-only or C-only models (p-values: 0.00250 and 0.065 respectively). This means that combining genetic markers and environmental covariates yields a better model than either data source by itself, a result

[3]Pedigree results refer to the best scoring method for each model type.

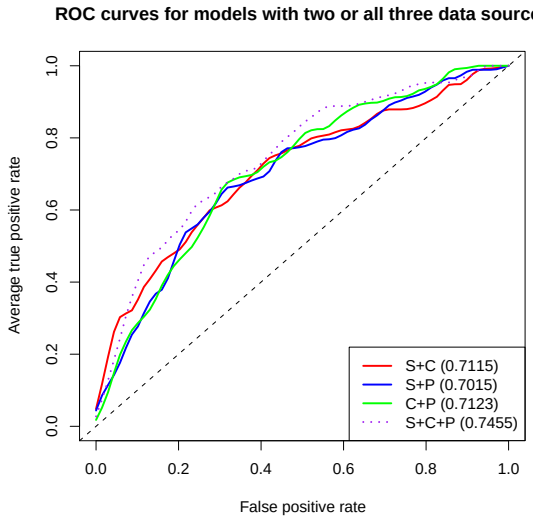


Fig 7: ROC curves for models with two or all three data sources. Although all three models with two data sources have comparative AUC (shown in parenthesis in the legend), the relationship between their curves varies across ROC space. The S+C model dominates the low false positive rate portion of space, while models including pedigree data dominate in the high true positive rate portion. The model with all three data sources (dotted curve) dominates through the entire ROC space.

consistent with the known epidemiology of pigmentary abnormalities, where risk is associated with both the genetic markers and environmental covariates included in this model.

The model with the highest overall mean AUC was the S+C+P model (MATERN), with statistically significant differences at the 90% level for all except the S+P (MATERN) (p-value 0.108) model. This is the main result of our paper: a full model containing genetic marker data along with environmental covariates and pedigree data is the best performing model for predictive purposes over models that only contain a subset of these data sources.

Models with only two data sources, i.e., S+C, S+P and C+P, performed statistically similarly. That is, there were no statistically significant differences in the predictive performance of these models, although the C+P model performed slightly better. This finding is consistent with the notions that SNP data — at least partly — encode pedigree data and that familial correlations encode shared environment data as well. We can see the former since SNP and pedigree data each add, relatively speaking, about the same amount of information when combined with covariate data; neither is strongly more informative than the other in the present context. In contrast, we can see the latter since combining SNP and pedigree data is as informative as combining SNP and covariate data. In summary, models that contain only two of these data sources, that is, pedigree-environmental covariates, or pedigree-genetic markers, or environmental covariates-genetic markers, have comparable predictive ability, while less than the model with all three.

The ROC curves for models using only two data sources (Figure 7) show an interesting trend. We can see that in the high sensitivity portion of the curve (false positive rate between 0 and 0.2), the S+C model, which does not contain any pedigree data, dominated the other two models. On the other hand, we see that the pedigree models dominated the S+C model on the other extreme portion of the curve (true positive rate higher than 0.8). The ROC curve for the S+C+P model dominates these three curves throughout ROC space.

Another observation we can make is that the P-only (MATERN) model had greater AUC than the S-only model but without a statistically significant difference (p-value 0.207). On the other hand, combining the two terms improves predictive performance significantly, indicating that the genetic influence on pigmentary abnormality risk is not properly modeled by either data source alone.

We conclude this case study by looking at diagnostics of the resulting models to illustrate the effect

of including pedigree data in the pigmentary abnormalities risk model. Cosine diagnostics [4, 36] are an illustrative way of displaying the relative importance of model terms in the SS-ANOVA decomposition defined on the vector of fitted "pseudo"-gaussian responses $\hat{f}$ for the entire cohort. The π diagnostic decomposes the norm of $\hat{f}$ according to the additive terms of the SS-ANOVA decomposition assigning a relative weight to each term in the model. We compare the π diagnostic for the S+C+P and S+C models in Table 3.

TABLE 3
π diagnostic comparison of the S+C+P and S+C model fits. Diagnostic π gives relative weight to each of the model terms indexed in the columns. Row and column labels correspond to those defined in Table 2 and Figure 6.

	S	C	P
S+C	0.34	0.66	
S+C+P	0.17	0.26	0.53

In the pedigree-less S+C model, the environmental covariates have two thirds of the decomposition weights. In the full S+C+P model, more than half of the decomposition weight is given to the pedigree term, while the relative weight of the other two terms are essentially unchanged: for example, the SNP terms (S) have $0.17/(0.17 + 0.26) = .39$ of the weights of the S and C terms. The fact that the C term in the full data S+C+P model has more weight than the S term may explain why the C+P model slightly outperforms the S+P model.

Considering that the full S+C+P model had the best prediction performance and that the pedigree term had a large relative weight in the model, we may conclude that incorporating familial relationship data in an SS-ANOVA model as described by our methodology not only improves the predictive performance of existing models of pigmentary abnormality risk, but also partly describes how these three sources of data relate in a predictive model. Refining this statistical methodology to further understand the interaction of these data sources would be of both technical and scientific interest.

7. Discussion

Throughout our experiments and simulations we have used genetic marker data in a very simple manner by including single markers for each gene in an additive model. A more realistic model should include multiple markers per gene and would include interaction terms between these markers. While we have data on two additional markers for each of the two genes included in our case study (CFH and ARMS2) for a total of six markers (three per gene), we chose to use the additive model on only two markers since, for this cohort, this model showed the same predictive ability as models including all six markers with interaction terms (analysis not shown). Furthermore, due to some missing entries in the genetic marker data, including multiple markers reduced the sample size.

Along the same lines, we currently use a very simple inheritance model to define pedigree dissimilarity. Including, for example, dissimilarities between unrelated subjects should prove advantageous. A simple example would be including a spousal relationship when defining dissimilarity since this would be capturing some shared environmental factors. Extensions to this methodology that include more complex marker models and multiple or more complex dissimilarity measures are fertile grounds for future work.

Other methods for including graph-based data in predictive models have been proposed recently, especially in the Machine Learning community. They range from semi-supervised methods that regularize a predictive model by applying smoothness penalties over the graph [37–39], to discriminative graphical models [40–43], and methods closer to ours which define kernels from graph relationships [44, 45].

There are issues in the risk modelling setting with general pedigrees, where relationship graphs encode relationships between a subset of a study cohort, that are usually not explicitly addressed in the general graph-based setting. Most important is the assumption that, while graph structure has some influence in the risk model, it is not necessarily an overwhelming influence. Thus, a model that produces relative weights between components of the model, one being graph relationships, is required. That is the motivation for using the SS-ANOVA framework in this paper. While graph regularization methods have a parameter that controls the influence of the graph structure in the predictive model, it is not directly comparable to the influence of other model components, e.g. genetic data or environmental covariates. On the other hand, graphical model techniques define a probabilistic model over the graph to define the predictive model. This gives the

graph relationships too much influence over the predictive model in the sense that it imposes conditional independence properties over subjects determined by the relationship graph that might not be valid for the other data sources, e.g. environmental covariates.

The relationship graphs in this setting lead to kernels that are highly diffuse in the sense that, due to the nature of the pedigree dissimilarity, there is rapid decay as the radial basis functions extend away from each subject. The use of the third order Matérn kernel function significantly improved the predictive ability of our methods over the Gaussian kernel (results not shown), probably due to the Matérn kernel softening this diffusion effect. Tuning the order of the Matérn kernel could further improve our models. Further understanding of the type of situations in which the Matérn kernel would perform better than the Gaussian is another direction for further research.

References

[1] G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist*, 23: 1865–1895, 1995.

[2] X. Lin, G. Wahba, D. Xiang, F. Gao, R. Klein, and B. Klein. Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist*, 28:1570–1600, 2000.

[3] D. Xiang and G. Wahba. A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statistica Sinica*, 6(3):675–692, 1996.

[4] C. Gu. *Smoothing Spline Anova Models*. Springer, 2002.

[5] P.N. Baird, F.M.A. Islam, A.J. Richardson, M. Cain, N. Hunt, and R. Guymer. Analysis of the Y402H variant of the Complement Factor H gene in age-related macular degeneration. *Investigative Ophthalmology & Visual Science*, 47(10):4194–4198, 2006.

[6] A.O. Edwards, R. Ritter, K.J. Abel, A. Manning, C. Panhuysen, and L.A. Farrer. Complement Factor H polymorphism and age-related macular degeneration. *Science*, 308(5720):421–424, 2005.

[7] S.A. Fisher, G.R. Abecasis, B.M. Yashar, S. Zareparsi, A. Swaroop, S.K. Iyengar, B.E.K. Klein, R. Klein, K.E. Lee, J. Majewski, et al. Meta-analysis of genome scans of age-related macular degeneration. *Human Molecular Genetics*, 14(15):2257–2264, 2005.

[8] L.G. Fritsche, T. Loenhardt, A. Janssen, S.A. Fisher, A. Rivera, C.N. Keilhauer, and B.H.F. Weber. Age-related macular degeneration is associated with an unstable ARMS2 (LOC387715) mRNA. *Nature Genetics*, 40(7):892–896, 2008.

[9] G.S. Hageman, D.H. Anderson, L.V. Johnson, L.S. Hancox, A.J. Taiber, L.I. Hardisty, J.L. Hageman, H.A. Stockman, J.D. Borchardt, K.M. Gehrs, et al. A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proceedings of the National Academy of Sciences*, 102(20):7227, 2005.

[10] J.L. Haines, M.A. Hauser, S. Schmidt, W.K. Scott, L.M. Olson, P. Gallins, K.L. Spencer, S.Y. Kwan, M. Noureddine, J.R. Gilbert, et al. Complement Factor H variant increases the risk of age-related macular degeneration. *Science*, 308(5720):419–421, 2005.

[11] A. Kanda, W. Chen, M. Othman, K.E.H. Branham, M. Brooks, R. Khanna, S. He, R. Lyons, G.R. Abecasis, and A. Swaroop. A variant of mitochondrial protein LOC387715/ARMS2, not HTRA1, is strongly associated with age-related macular degeneration. *Proceedings of the National Academy of Sciences*, 104(41):16227, 2007.

[12] R.J. Klein, C. Zeiss, E.Y. Chew, J.Y. Tsai, R.S. Sackler, C. Haynes, A.K. Henning, J.P. SanGiovanni, S.M. Mane, S.T. Mayne, et al. Complement Factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

[13] M. Li, P. Atmaca-Sonmez, M. Othman, K.E.H. Branham, R. Khanna, M.S. Wade, Y. Li, L. Liang, S. Zareparsi, A. Swaroop, et al. CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nature Genetics*, 38(9):1049, 2006.

[14] K.P. Magnusson, S. Duan, H. Sigurdsson, H. Petursson, Z. Yang, Y. Zhao, P.S. Bernstein, J. Ge, F. Jonasson, E. Stefansson, et al. CFH Y402H confers similar risk of soft drusen and both forms of advanced AMD. *PLoS Med*, 3(1):e5, 2006.

[15] C.L. Thompson, G. Jun, B.E.K. Klein, R. Klein, J. Capriotti, K.E. Lee, and S.K. Iyengar. Genetics of

pigment changes and geographic atrophy. *Investigative Ophthalmology & Visual Science*, 48(7):3005–3013, 2007.

[16] C.L. Thompson, B.E.K. Klein, R. Klein, Z. Xu, J. Capriotti, T. Joshi, D. Leontiev, K.E. Lee, R.C. Elston, and S.K. Iyengar. Complement Factor H and Hemicentin-1 in age-related macular degeneration and renal phenotypes. *Human Molecular Genetics*, 16(17):2135, 2007.

[17] R. Klein, T. Peto, A. Bird, and M.R. Vannewkirk. The epidemiology of age-related macular degeneration. *American Journal of Ophthalmology*, 137(3):486–495, 2004.

[18] F. Lu, S. Keles, S.J. Wright, and G. Wahba. Framework for kernel regularization with application to protein clustering. *Proceedings of the National Academy of Sciences*, 102(35):12332–12337, 2005.

[19] G. Malécot. *Les mathématiques de l'hérédité*. Masson, 1948.

[20] D.C. Thomas. *Statistical Methods in Genetic Epidemiology*. Oxford Univ Press, 2004.

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

[22] G.R.G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 5:27–72, 2004.

[23] B. Matérn. *Spatial variation, number 36 in lectures notes in statistics*. Springer, 1986.

[24] M.L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.

[25] R. Klein, BE Klein, KL Linton, and DL De Mets. The Beaver Dam Eye Study: visual acuity. *Ophthalmology*, 98(8):1310–5, 1991.

[26] R. Klein, BE Klein, SC Jensen, and SM Meuer. The five-year incidence and progression of age-related maculopathy: the Beaver Dam Eye Study. *Ophthalmology*, 104(1):7–21, 1997.

[27] R. Klein, B.E.K. Klein, S.C. Tomany, S.M. Meuer, and G.H. Huang. Ten-year incidence and progression of age-related maculopathy: The Beaver Dam eye study. *Ophthalmology*, 109(10):1767–1779, 2002.

[28] R. Klein, B.E.K. Klein, M.D. Knudtson, S.M. Meuer, M. Swift, and R.E. Gangnon. Fifteen-year cumulative incidence of age-related macular degeneration: The Beaver Dam Eye Study. *Ophthalmology*, 114 (2):253–262, 2007.

[29] K.E. Lee, B.E.K. Klein, R. Klein, and M.D. Knudtson. Familial aggregation of retinal vessel caliber in the Beaver Dam Eye Study. *Investigative Ophthalmology & Visual Science*, 45(11):3929, 2004.

[30] R Klein, B E Klein, and K L Linton. Prevalence of age-related maculopathy. The Beaver Dam Eye Study. *Ophthalmology*, 99(6):933–43, Jun 1992.

[31] B E Klein, R Klein, S C Jensen, and L L Ritter. Are sex hormones associated with age-related maculopathy in women? The Beaver Dam Eye Study. *Transactions of the American Ophthalmological Society*, 92:289–95; discussion 295–7, Jan 1994.

[32] Chong Gu. *gss: general smoothing splines*, 2007. R package version 1.0-0.

[33] B. Borchers. CSDP, A C library for semidefinite programming. *Optimization Methods and Software*, 11 (1):613–623, 1999.

[34] B. Atkinson and T. Therneau. *kinship: mixed-effects Cox models, sparse matrices, and modeling data from large pedigrees*, 2007. R package version 1.1.0-18.

[35] T. Fawcett. ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, 31, 2004.

[36] C. Gu. Diagnostics for nonparametric regression models with additive terms. *Journal of the American Statistical Association*, 87(420):1051–1058, 1992.

[37] V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. *ACM International Conference Proceeding Series*, 119:824–831, 2005.

[38] A.B. Goldberg, X. Zhu, and S. Wright. Dissimilarity in graph-based semi-supervised classification. In *Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2007.

[39] X. Zhu. Semi-supervised learning literature survey. Technical Report TR 1530, Computer Science, University of Wisconsin-Madison, 2005.

[40] W. Chu, V. Sindhwani, Z. Ghahramani, and S.S. Keerthi. Relational learning with gaussian processes. *Advances in Neural Information Processing Systems: Proceedings of the 2006 Conference*, 2007.

[41] L. Getoor. Link-based classification. *Advanced Methods for Knowledge Discovery from Complex Data*, 2005.

[42] B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. *Advances in Neural Information Processing Systems*, 16:51, 2004.

[43] J. Lafferty, X. Zhu, and Y. Liu. Kernel conditional random fields: representation and clique selection. *ACM International Conference Proceeding Series*, 2004.

[44] A. Smola and R. Kondor. Kernels and regularization on graphs. *Conference on Learning Theory, COLT/KW*, 2003.

[45] X. Zhu, J. Kandola, J. Lafferty, and Z. Ghahramani. *Semi-Supervised Learning*, chapter Graph Kernels by Spectral Transforms. MIT Press, 2006.