Examining the Relative Influence of Familial, Genetic and Environmental Covariate Information in Flexible Risk Models (Supplementary Information)

Héctor Corrada Bravo*
Johns Hopkins Bloomberg School of Public Health
Baltimore, MD 21205-2179
hcorrada@cs.wisc.edu

Kristine E. Lee†, Barbara E.K. Klein†, Ronald Klein†
Department of Ophthalmology and Visual Science
University of Wisconsin-Madison
Madison, WI 53706
{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
ski@case.edu

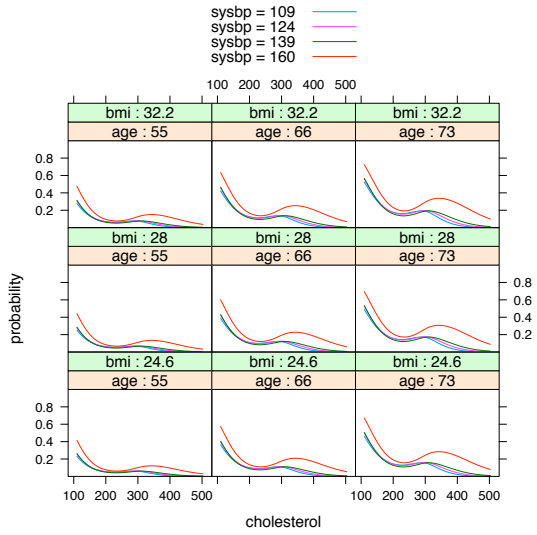
Grace Wahba*
Departments of Statistics and Biostatistics and Medical Informatics
University of Wisconsin-Madison
Madison, WI 53706-1685
wahba@stat.wisc.edu

Including SNP data in an SS-ANOVA model

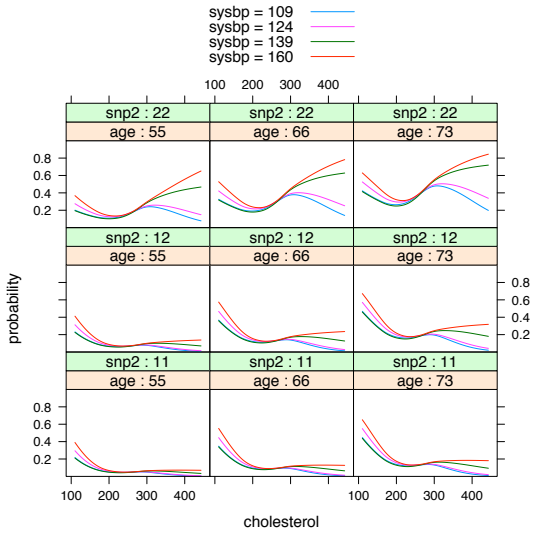
An SS-ANOVA model of retinal pigmentary abnormalities[1] was able to show a nonlinear protective effect of high total serum cholesterol for a cohort of subjects in the Beaver Dam Eye Study [1]. We replicate those findings in Figure S1a.[2] By extending this model with SNP rs10490924 in the ARMS2 gene 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.

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

[2] There are minor differences between our plot in panel (a) and the corresponding plot in Lin et al. [1] since we use a subset of the same cohort. We give details regarding this model in the *Case Study* section of the paper.



(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 (28).

Figure S1: 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 in main paper for an explanation of model terms).

Pedigrees

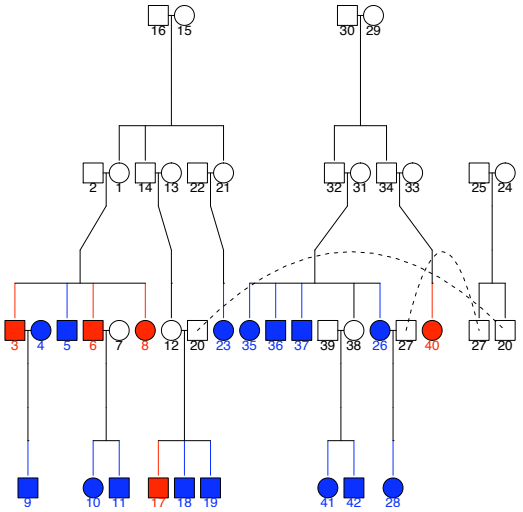
Figure S2a shows an example of a pedigree. The dissimilarity in Definition 1 of the main paper is also the *degree of relationship* between pedigree members i and j [3]. Another dissimilarity based on the kinship coefficient can be defined as $1 - 2\varphi$. However, since we use Radial Basis Function kernels, defined by an exponential decay with respect to the pedigree dissimilarities, including the exponential decay in φ resulted in overly-diffused kernels. For example, Figure S2b shows the relationship graph for five female BDES subjects in the pedigree from Figure S2a. Edge labels are the pedigree dissimilarities derived from the kinship coefficient, and dotted lines indicate unrelated pairs.

Regularized Kernel Estimation

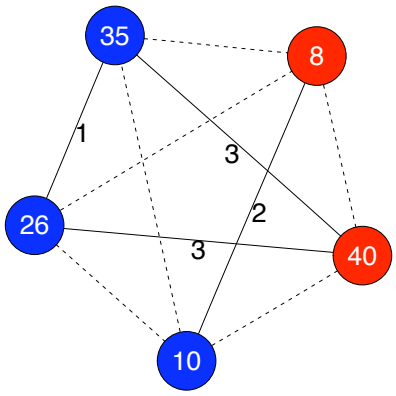
Figure S3 shows a three-dimensional embedding derived by RKE of the relationship graph in Figure S2b. Notice that the x-axis is order of magnitudes larger than the other two axes and that the unrelated edges in the relationship graph occur along this dimension. That is, the first dimension of this RKE 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 S3. For example, Figure S5 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.

The pedigree dissimilarity of Definition 1 is not a distance since it does not satisfy the triangle inequality for general pedigrees. In Figures S4a and S4b 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 S5.



(a) Example pedigree.



(b) Relationship graph. Edge labels are dissimilarities defined by the kinship coefficient (sibling/parental=1, avuncular=2, first cousins=3,...). Dotted edges indicate unrelated pairs.

Figure S2: 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.

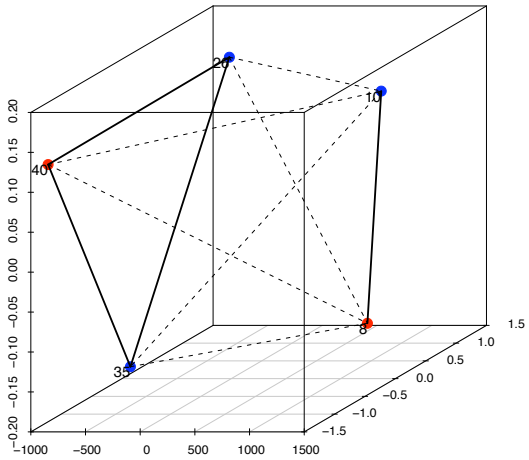
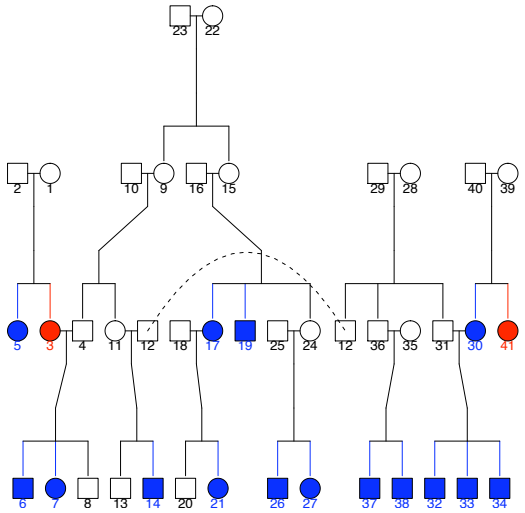
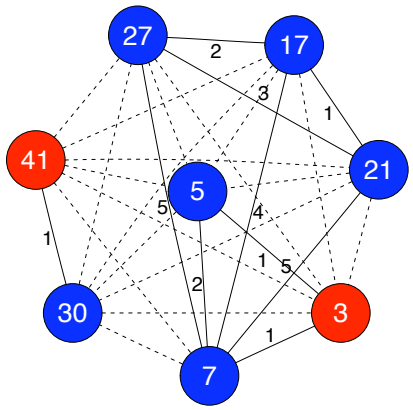


Figure S3: Embedding of relationship graph in Figure S2 by RKE. The x-axis of this plot is order of magnitudes larger than the other two axes. The unrelated edges in the relationship graph occur along this dimension, while the other two dimensions encode the relationship distance.



(a) Second example pedigree.



(b) Corresponding relationship graph

Figure S4: 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.

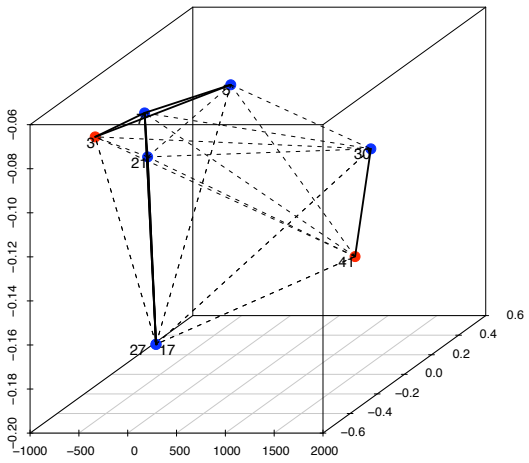


Figure S5: RKE Embedding for second example graph. Subjects 27 and 17 are superimposed in this three dimensional plot, but are separated by the fourth dimension.

Matérn Kernel Family

A standard kernel function commonly used is the Gaussian kernel:

$$k(x_i, x_j) = \exp\{-\gamma \|x_i - x_j\|^2\}. \tag{1}$$

This kernel function is a good candidate for this setting 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 [10, 11] 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

$$k_\nu(i, j) = \exp\{-\alpha d_{ij}\}\pi_\nu(\alpha, d_{ij}), \tag{2}$$

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

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

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

$$k_{m+1}(i, j) = \frac{1}{\alpha^{2m+1}} \exp\{-\alpha\tau\} \sum_{i=0}^{m+1} a_{m+1,i}\alpha^i\tau^i, \tag{4}$$

where $a_{m+1,0} = (2m + 1)a_{m,0}$, $a_{m+1,i} = (2m + 1 - i)a_{m,i} + a_{m,i-1}$, 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.

The Pedigree Term

In this section we specify completely the pedigree term $h(z(t))$ in Equation (5) of the paper. For subject t, $z(t)$ is the pseudo-attribute obtained as the corresponding row of embedding matrix Z where $K = ZZ^T$ and K is the solution to the RKE problem in Equation (4) as described in the Section *Representing Pedigree Data as Kernels*. The smooth function h is then defined by a kernel function k, e.g. from the Matérn family described in the previous Section. Thus for subject t, $h(z(t)) = \theta_h \sum_{i=1}^n c_i k(z(t), z(i))$, where c_i is obtained as the solution of the penalized likelihood problem of Equation (3) and θ_h is the coefficient of the pedigree term in the SS-ANOVA decomposition.

Case Study

We also compared the two methods presented for incorporating pedigree data described in the paper. We refer to the method using a kernel defined over an embedding resulting from RKE as RKE/MATERN, and to the alternative method, that defines a kernel over the pedigree dissimilarities directly 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 a Matérn third-order kernel.

Table S1 shows the resulting mean and standard deviations of the AUC over the ten cross-validation folds of each individual model/method combination summarized in Figure 1 of paper. The values of the π diagnostic discussed in the text are given in Table S2. In Figure 1 of the paper, the AUC for models that include pedigrees is shown for the best scoring method in Table S1.

We also tested the Gaussian kernel, but it consistently showed similar or worse performance as the Matérn kernels and thus is not reported. 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

Table S1: 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)

Table S2: 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 S1.

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

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] 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.

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

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

[4] 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.

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

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

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

[8] 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.

[9] 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.

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

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