Examining the Relative Influence of Familial, Genetic and Covariate Information In Flexible Risk Models

Grace Wahba

Based on a paper of the same name which has appeared in PNAS May 19, 2009, p 8123-8127 by Hector Corrada Bravo, Grace Wahba, Kristine Lee, Barbara Klein, Ronald Klein and Sudha Iyengar. Open source article, featured in "In This Issue" (http://www.pnas.org/content/106/20/8079.full)

2010 SIAM International Conference on Data Mining

Columbus, Ohio, April 29, 2010

These slides at

`http://www.stat.wisc.edu/~wahba/` → TALKS

Direct link to Corrada Bravo et al and other papers/preprints at

`http://www.stat.wisc.edu/~wahba/` − > TRLIST

Examining the Relative Influence of Familial, Genetic and Covariate Information in Flexible Risk Models.

Smoothing Spline ANOVA (SS-ANOVA) models are a well known approach to penalized likelihood regression given heterogenous attribute variables, with the ability to model their various interactions. In many circumstances, one may observe attributes, along with some relationships between objects in the training set. We describe a new approach to incorporating relationship or (dis)similarity information in an SS-ANOVA model. For the objects under study, we have attributes along with relationship information between (some) pairs of objects in the study. As an example we consider a demographic study with the response a particular disease that is known to run in families. The data includes environmental/clinical observations, genetic data and pedigree information in a study where a large fraction of the population have relatives in the study. One issue is to evaluate the relative influence of the three distinct sources of information.

Outline

1. The Log Likelihood for Bernoulli responses
2. Reproducing Kernel Hilbert Spaces (RKHS)
3. ANOVA decomposition of Functions of Several Variables
4. Smoothing Spline-ANOVA Model and the Beaver Dam Eye Study
5. Modeling Environmental/Clinical, Genetic and Pedigree Data in an extended SS-ANOVA model in the BDES
6. Pedigree (Relationship) Data
7. Relationship Data Encoded by Regularized Kernel Estimation (RKE)
8. Estimating the relative influence of Environmental/Clinical, Genetic and Pedigree Data in the BDES
9. Summary, Conclusions, Further Work, Related Projects

The Log Likelihood for Bernoulli responses

- Given: $y_i, x(i), i = 1, 2, \cdots, n$
  $y \in \{0, 1\}$, $x = (x_1, x_2, \cdots, x_d)$

- Estimate: $p(x) = Prob(y = 1|x)$

- The log odds ratio (logit): $f(x) = \log \frac{p(x)}{1-p(x)}$

- The negative log likelihood:

$$\mathcal{L}(y, f) = \sum_{i=1}^{n} -y_i f(x(i)) + \log(1 + e^{f(x(i))})$$

- Recover $p(x) = e^{f(x)}/(1 + e^f(x))$.

Penalized Log Likelihood Estimate

The penalized log likelihood estimate of f is obtained by finding f in some prescribed function space to minimize

$$I(f) = \mathcal{L}(y, f) + \lambda J(f)$$

where $J(f)$ is a penalty functional on f and λ is a tuning parameter which balances fit to the data and complexity/wiggliness of f. We will fit f in a function space which admits a useful ANOVA decomposition-a Reproducing Kernel Hilbert Space (RKHS), using a Smoothing Spline ANOVA model.

Reproducing Kernel Hilbert Spaces (RKHS)

- f will be in an RKHS (the mother of all kernels!).

- What is an RKHS?

- Let $K(s,t)$ be a positive definite function on $\mathcal{T} \otimes \mathcal{T}$. This means for any $t_1, \cdots, t_k$, $\sum_{r,s=1}^{k} K(t_r, t_s) \geq 0$.

- Moore-Aronszajn Theorem: To every positive definite function $K(\cdot,\cdot)$ there corresponds a unique RKHS and vice versa. $K(\cdot, t*) \in \mathcal{H}_K$ and $\sum_r c_r K(\cdot, t_r) \in \mathcal{H}_K$ and limits $\in \mathcal{H}_K$.

- $f \in \mathcal{H}_K \Rightarrow < f(\cdot), K(\cdot, t*) >= f(t*)$

- $\| \sum c_r K(\cdot, t_r) \|^2_{\mathcal{H}_K} = \sum_{rs} c_r c_s K(t_r, t_s)$

ANOVA Decomposition of Functions of Several Variables

$$x \equiv (x_1, \cdots, x_d) \in \mathcal{X} \equiv \mathcal{X}^{(1)} \otimes \cdots \otimes \mathcal{X}^{(d)}$$

$$f(x) = f(x_1, \cdots, x_d).$$

Let $d\mu_\alpha$ be a probability measure on $\mathcal{X}^{(\alpha)}$ and define the averaging operator $\mathcal{E}_\alpha$ on $\mathcal{X}$ by

$$(\mathcal{E}_\alpha f)(x) = \int_{\mathcal{X}^{(\alpha)}} f(x_1, \cdots, x_d) d\mu_\alpha(x_\alpha).$$

ANOVA Decomposition of Functions of Several Variables (continued)

The averaging operators $\mathcal{E}_\alpha$ give a (unique) ANOVA decomposition of f:

$$f(x_1, \cdots, x_d) = \mu + \sum_\alpha f_\alpha(x_\alpha) + \sum_{\alpha\beta} f_{\alpha\beta}(x_\alpha, x_\beta) + \cdots$$

where

$$\begin{aligned}
\mu &= \prod_\alpha \mathcal{E}_\alpha f = \int \cdots \int f(x_1, \cdots, x_d) d\mu_1(x_1) \cdots d\mu_d(x_d) \\
f_\alpha &= (I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta f \\
f_{\alpha\beta} &= (I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha, \beta} \mathcal{E}_\gamma f \\
\vdots &\quad \vdots \qquad \mathcal{E}_\alpha f_\alpha = 0, \quad \mathcal{E}_\alpha \mathcal{E}_\beta f_{\alpha\beta} = 0, etc.
\end{aligned}$$

ANOVA Decomposition of Functions of Several Variables (continued)

$$f(x) = \mu + \sum_{\alpha=1}^{d} f_\alpha(x_\alpha) + \sum_{\alpha \le \beta} f_{\alpha\beta}(x_\alpha, x_\beta) + \cdots$$

- The series is truncated at some point.
- Terms satisfy ANOVA-like side conditions (identifiable).
- SS-ANOVA representation with weights on kernels :

$$f(\cdot) = \sum_{j=1}^{m} d_j \phi_j(\cdot) + \sum_{j=1}^{n} c_j K_\theta(\cdot, x(j))$$

where the ϕ_j are all unpenalized components + possibly more:

$$K_\theta(\cdot,\cdot) = \sum_{\alpha=1}^{d} \theta_\alpha K_\alpha(\cdot,\cdot) + \sum_{\alpha \le \beta} \theta_{\alpha\beta} K_{\alpha\beta}(\cdot,\cdot) + \cdots$$

- Kernels depend only on components of x in the subscripts.

Since $\|f\|^2_{\mathcal{H}_{\theta K}} = \theta^{-1}\|f\|^2_{\mathcal{H}_K}$ The SS-ANOVA penalty functional has the form:

$$J(f) = \sum_{i,j=1}^{n} c_i c_j \left[\sum_{\alpha=1}^{d} \theta_\alpha^{-1} K_\alpha(x(i), x(j)) + \sum_{\alpha \leq \beta} \theta_{\alpha\beta}^{-1} K_{\alpha\beta}(x(i), x(j)) + \cdots \right]$$

The θs are tuning parameters along with λ and with an identifiability constraint. For each trial set of tuning parameters, the c_i are to be fitted. Calling the fitted result $f_{\lambda\theta}$, the fitted $f_{\lambda\theta}$ are evaluated for the best set of tuning parameters via a tuning criteria. A popular criteria for tuning SS-ANOVA models with RKHS squared norm penalties and Bernoulli data is GACV.

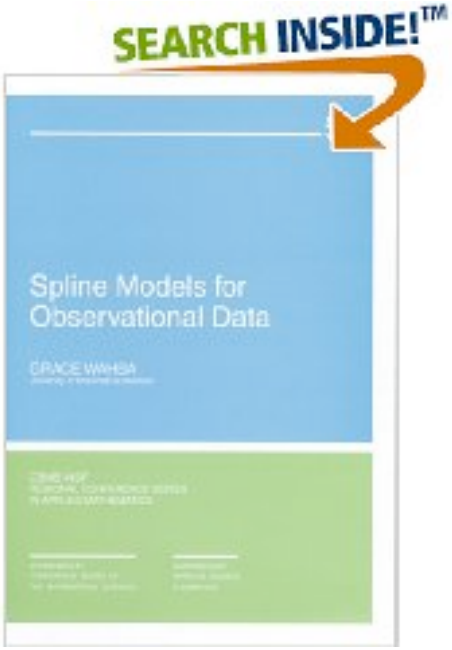


Figure 1: Grace Wahba, Spline Models for Observational Data (1990)

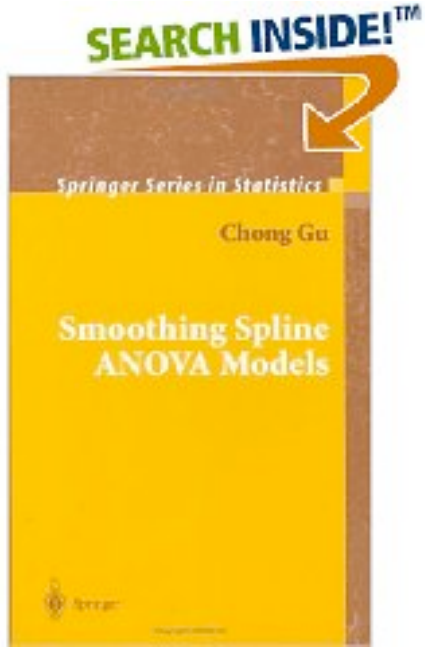


Figure 2: Chong Gu, Smoothing Spline ANOVA Models (2002)

X. Lin et. al. Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist*, 28:1570–1600, 2000.

SS-ANOVA Model in the Beaver Dam Eye Study

- The Beaver Dam Eye Study (BDES) is an ongoing population-based study of age related ocular disorders, begun in 1988.
- An SS-ANOVA model for association of a number of environmental/clinical (E/C) variables based on 2585 women with complete E/C data appears in Lin, Wahba, et al Ann. Statist 28 (2000).
- 684 women have at least one relative also in the study.

- The predictor variables of present interest are:

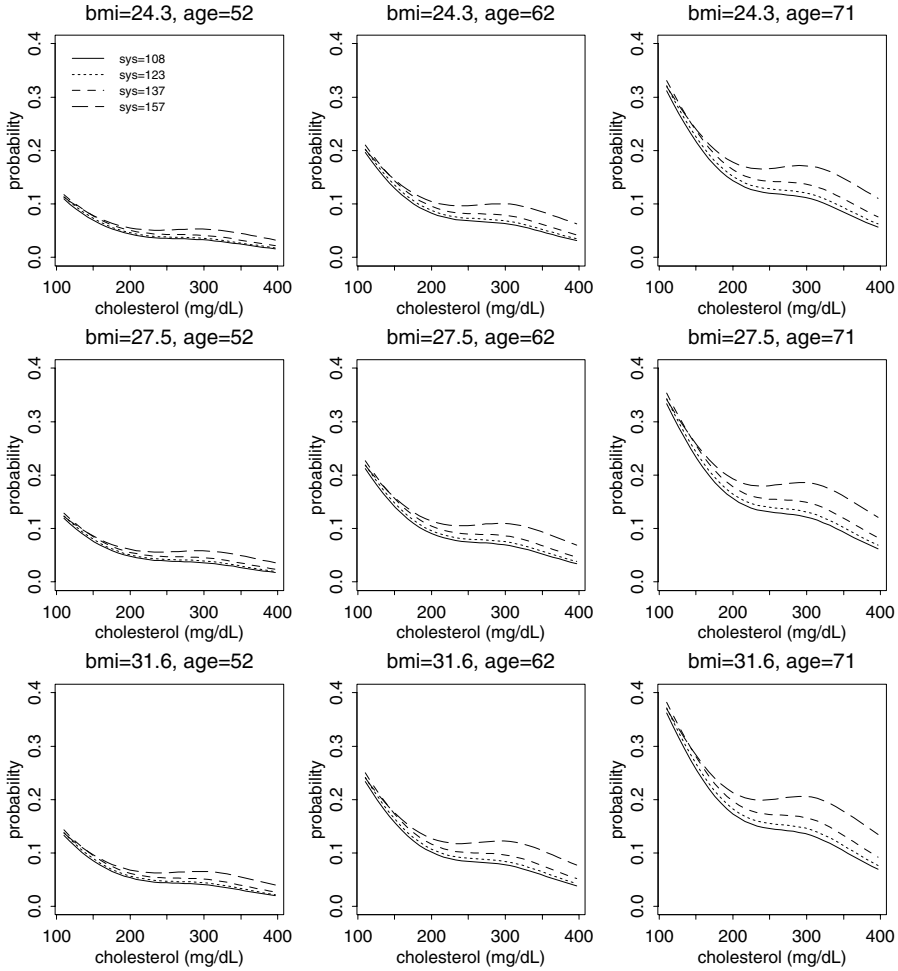
code	units	description
`horm`	yes/no	current usage of hormone replacement therapy
`hist`	yes/no	history of heavy drinking
`bmi`	kg/m^2	body mass index
`age`	years	age at baseline
`sysbp`	$mmmHg$	systolic blood pressure
`chol`	mg/dL	serum cholesterol
`smoke`	yes/no	history of smoking

Table 1: E/C covariates for BDES pigmentary abnormalities SS-ANOVA model

- The fitted E/C model that we are using in the present study is

$$\begin{aligned} f(t) = \mu \quad &+ \quad f_1(\texttt{sys}) + f_2(\texttt{chol}) + f_{12}(\texttt{sys}, \texttt{chol}) \\ &+ \quad d_{\texttt{age}} \cdot \texttt{age} + d_{\texttt{bmi}} \cdot \texttt{bmi} \\ &+ \quad d_{\texttt{horm}} \cdot I_1(\texttt{horm}) + d_{\texttt{drin}} \cdot I_2(\texttt{drin}) + d_{\texttt{smoke}} \cdot I_3(\texttt{smoke}) \end{aligned}$$

- This is the same model that was fitted in *Ann. Statist. 2000* with the exception that `smoke` was not included there.
- f_1, f_2 and f_{12} are splines.



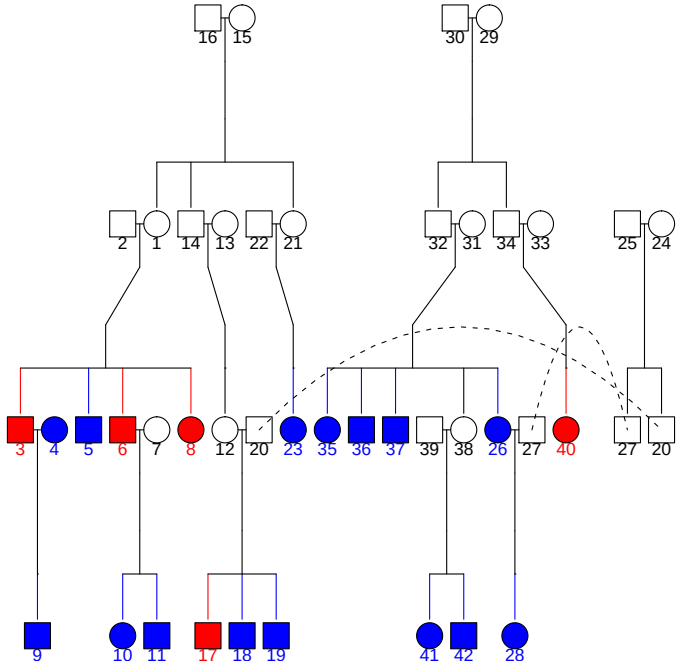
Estimated probability from an SS- ANOVA logistic regression model. Each x-axis is cholesterol, each set of four lines is four values of systolic blood pressure, each plot fixes body mass index and age to the shown values. $hist = 0,\ horm = 0$. From *Ann. Stat. 2000.*

Modeling E/C, genetic and pedigree data in an extended SS-ANOVA model

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

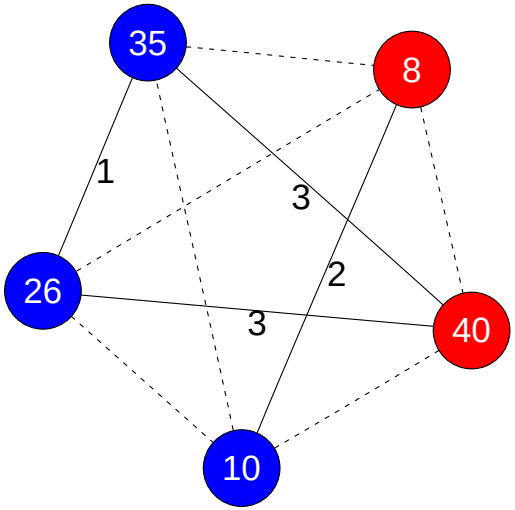
- First two lines: Genetic (SNP) data. Two SNPS each with three levels, (1,1), (1,2), (2,2). (Usual methodology)
- Next three lines E/C variables
- Last line: Pedigree/relationship data goes here. Will explain.

A Pedigree from BDES



Example pedigree from the Beaver Dam Eye Study. Red nodes-with pigmentary abnormalities, blue nodes-without pigmentary abnormalities. Circles are females, rectangles are males.

A Relationship (Sub)Graph From the Pedigree



Relationship graph for subjects in the pedigree. Edge labels are distances defined by the kinship coefficient. Persons 26 and 35 are siblings [1], persons 8 and 10 are aunt and niece [2] and persons 26 and 40 are cousins [3]. Unrelated pairs have dashed lines.

Relationship Data Encoded with RKE

- To include relationship/pedigree data into an SS-ANOVA model, we encode it with the Regularized Kernel Estimation algorithm (RKE). (Lu et al, *PNAS 2005*)
- Given n objects and pairwise dissimilarity measures d_{ij} between a sufficient number of the $\binom{n}{2}$ pairs, the RKE encodes this information in an $n \times n$ positive definite matrix $R_{dist}(i, j)$ defined on the n objects. The d_{ij} will be obtained from relationship coefficients (will be numbers 1, 2, 3, 4, or 5), by a biologically motivated transformation. ($d_{ij} = -2log_2(2\phi_{ij})$) where ϕ is Malecot's kinship coefficient).

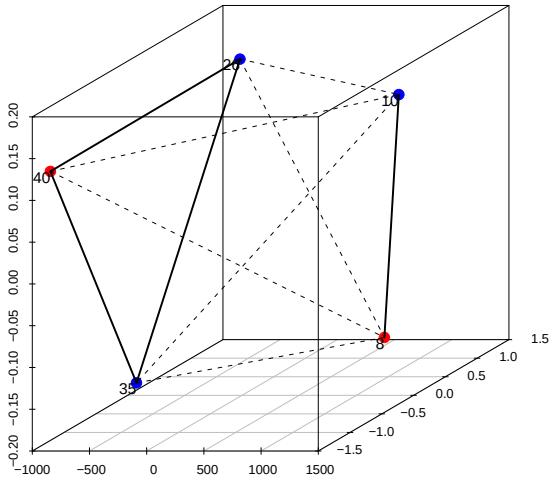
Relationship Data Encoded With RKE (continued)

The distance encoding matrix R_{dist} is obtained by solving the convex cone optimization problem:

$$\min_{R \succeq 0} \sum_{(i,j) \in \Omega} |d_{ij} - B_{ij}(R)| + \lambda_{RKE} trace(R) \tag{1}$$

where $R \succeq 0$ means R is in the convex cone of all real non-negative definite matrices of dimension n, Ω is all or a (sufficiently rich) subset of the $\binom{n}{2}$ pairs of indices, and
$B_{ij}(R) \equiv R(i,i) + R(j,j) - 2R(i,j) \equiv \hat{d_{ij}}$, the natural squared distance induced by R. Small eigenvalues in the fitted R_{dist} are deleted. $R_{dist}(i,j)$ gives a (unique up to rotation) embedding $z(i)$ of the ith subject. This (implicitly) goes into $f_{ped}(z(i))$ in the extended SS-ANOVA model.

Embedding of Pedigree by RKE



$z(i)$ for the five persons in the relationship graph. 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.

Relationship Data Encoded With RKE (continued)

The RKE embedding is unique up to rotation, but only the distances $\hat{d_{ij}}$ are relevant. These distances can be used with any RK that only depends on $\|z(i) - z(j)\|$, that is, $K_{ped}(z(i), z(j)) = k(\|z(i) - z(j)\|$. These kernels are known as Radial Basis Functions (RBFs). A Matern RBF is used in the present work. The Matern family of RBF's is a two-parameter family, and the parameters are to be chosen. Letting t stand for person i, then $f_{ped}(z(i))$ in the extended SS-ANOVA model will be of the form of a liniear combination of $K_{ped}(\cdot, z(j)), j = 1, 2, \cdots, n$. Note that unlike the rest of the SS-ANOVA model, $f_{ped}(z(i))$ is only defined for persons in pedigrees.

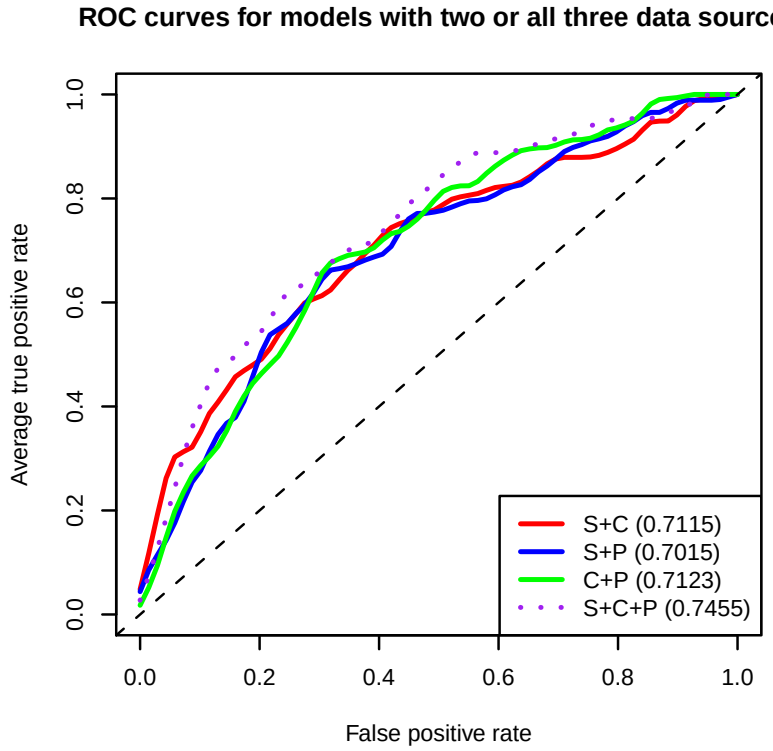
Qualitative Results

An important goal of the study is to explore the relative contribution of each source of data. Since there three sources of information: (S=SNPS, P=Pedigrees,C= Environmental/Clinical) there are seven models we can consider:

- S = SNPS (genetic data) only
- C = Environmental/Clinical (E/C) data only
- S + C
- P = Pedigrees only
- S + P
- C + P
- S + C + P

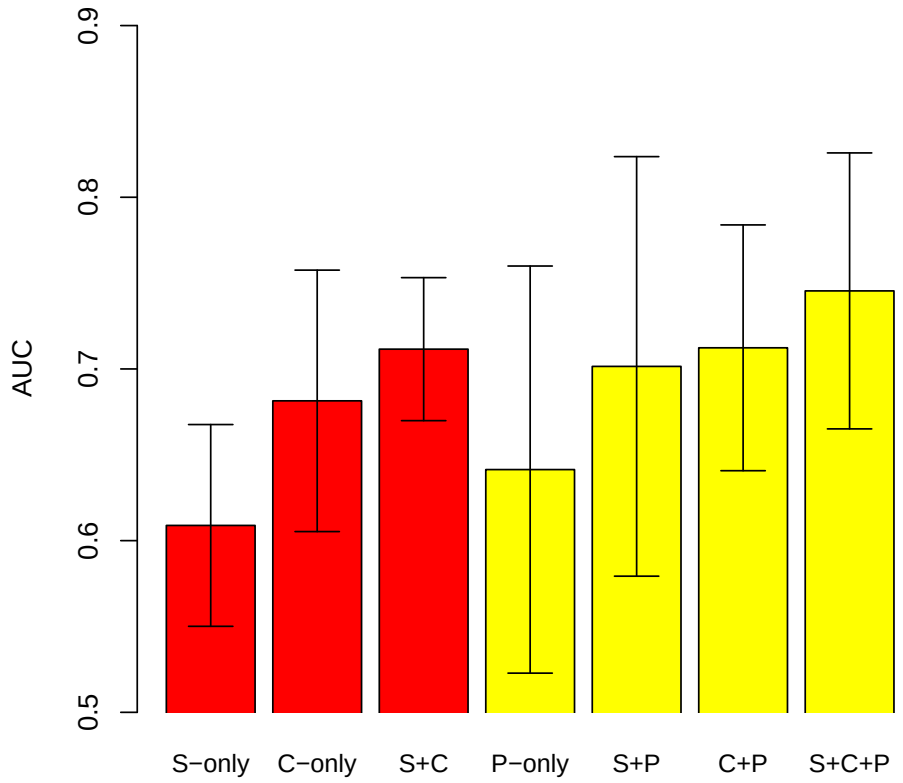
Compare models by evaluating the AUC (Area Under the Curve).

Comparing Models by Their Area Under the (ROC) Curve (AUC)



Each person in a test set is classified by thresholding their value of $p(x)$. As the threshold goes from 0 to 1, plot "True positive rate" against "False positive rate". Dashed line-random classification.

Results



The mean AUC for each of the seven models is given in the plot above, in order: Red: S-only, C-only and S+C. Pedigrees are added in yellow: P-only, S+P, C+P and S+C+P.

Summary and Conclusions

We have described the log likelihood for Bernoulli responses, Reproducing Kernel Hilbert Spaces, and Smoothing Spline ANOVA models. We discussed how Smoothing Spline ANOVA models were originally applied to data from the Beaver Dam Eye Study - to examine association of clinical/environmental variables with pigmentary abnormalities. Pigmentary abnormalities are a precursor to Age Related Maculopathy, which is known to run in families. We described some of the the pedigree data from the Eye Study, and we developed a new method for incorporating this information into a Smoothing Spline ANOVA model, using Regularized Kernel Estimation. We can see the relative importance of clinical/environmental variables, certain genetic information, and pedigree information in modeling risk of pigmentary abnormalities. The approach has promise for many other applications where relationship or (dis)similarity information is available.

Related and Further Work:

Enhancements of the existing model

1. Include interactions between Environmental/Clinical variables and genetic markers (and pedigrees?).
2. "Tune" the dissimilarity information, that is, instead of 1,2,3,4,5 as dissimilarity levels in the pedigree data, use a "tuned" monotone function of them. For example for subjective data, one might ask an evaluator to assign labels of one of "very close", "close" "distant", "very distant" and it is desired to assign numerical values, e. g. $1, 1+\delta_1, 1+\delta_1+\delta_2, 1+\delta_1+\delta_2+\delta_3$ where the δs are three positive values to be chosen according to some prediction criteria

Combine variable selection with dissimilarity data

1. Methods for dealing with a very large number of candidate SNP or other genetic marker patterns. See W. Shi, G. Wahba, S. Wright, K. Lee, B. Klein, and R. Klein. LASSO Patternsearch algorithm with applications to ophthalmology and genomic data. (LPS method). *Statistics and Its Interface*, 1:137–153, 2008. $f(\cdot) = \sum_{r=1}^{p} e_r B_r(\cdot)$, where $B_r(\cdot)$ is the ith pattern (groups) of SNPs, which typically have three values. $J(f) = \sum_{r=1}^{p} |e_r| = J_{LPS}(f)$

2. Efficient software for p extremely large, from Steve Wright: `http://www.cs.wisc.edu/~swright/LPS`

3. Combine LASSO Patternsearch with SS-SANOVA - continued

Combine variable selection with dissimilarity data continued

4. Now can set

$$f(\cdot) = \sum_{j=1}^{m} d_j \phi_j(\cdot) + \sum_{j=1}^{n} c_j K_\theta(\cdot, x(j)) + \sum_{r=1}^{p} e_r B_r(\cdot)$$

   and with penalty $\lambda_{LPS} J_{LPS} + \lambda_{SS-ANOVA} J_{SS-ANOVA}$.

5. Best prediction is not the same as best variable selection (Leng, Lin and Wahba, A note on the LASSO and related procedures in model selection, Statistica Sinica, 16 (4) 1273-1284 (2006), perhaps because it costs more to leave out an important variable then to include an unimportant one. What is the right tradeoff between prediction (SS-ANOVA) and sparsity(LASSO-Patternsearch) λs?.

Combine variable selection with dissimilarity data continued

In multiple correlated Bernoulli responses-

6. Extend F. Gao, G. Wahba, R. Klein and B. Klein, Smoothing spline ANOVA for multivariate Bernoulli observations, with applications to ophthalmology data, with discussion. *J. Amer. Statist. Assoc.*, 96:127–160, 2001.

What is the appropriate choice of distance/dissimilarity in any project-very important.

Noisy covariates, errors in variables, missing data

7. Missing Data: X. Ma, B. Dai, R. Klein, B. Klein K. Lee and G. Wahba, Penalized Likelihood Regression in Reproducing Kernel Hilbert Spaces with Randomized Covariate Data, UWisconsin Statistics Department TR1158, April, 2010.