Robustness and Reproducing Kernel Hilbert Spaces

Grace Wahba

Part 1. Regularized Kernel Estimation RKE. (Robustly)
Part 2. Smoothing Spline ANOVA SS-ANOVA and RKE.
Part 3. Partly missing covariates SS-ANOVA and QPLE(Robustly)

ICORS 2010
Prague, Czech Republic, June 30, 2010

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

Link to the references and all papers/preprints since 1993 at
`http://www.stat.wisc.edu/~wahba/` $->$ TRLIST

References

1. F. Lu, S. Keles, S. Wright, and G. Wahba. A framework for kernel regularization with application to protein clustering. *Proceedings of the National Academy of Sciences*, 102:12332–12337, 2005. RKE.

2. H. Corrada Bravo, G. Wahba, K. E. Lee, B. E. K. Klein, R. Klein, and S. K. Iyengar. Examining the relative influence of familial, genetic and environmental covariate information in flexible risk models. *Proceedings of the National Academy of Sciences*, 106:8128–8133, 2009. RKHS, SS-ANOVA, RKE.

3. 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. TR 1958, Statistics, University of Wisconsin- Madison 2010. Missing covariates, Quadrature Penalized Likelihood Estimates QPLE.

Part 1. Regularized Kernel Estimation RKE

Given scattered noisy non-metric pairwise dissimilarity information d_{ij} between pairs ij of n objects, embed these objects in a Euclidean space that attempts to preserve the dissimilrity information as much as possible. Find an $n \times n$ distance encoding matrix R_{dist} by solving the convex one optimization problem:

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

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 $\hat{d}_{ij}(R) \equiv R(i,i) + R(j,j) - 2R(i,j)$, the natural squared distance induced by R. Robust against dissimilarity data not satisfying the triangle inequality!

Small eigenvalues in the fitted R_{dist} are deleted, leaving r non-zero eigenvalues. $R_{dist}(i,j)$ gives a (unique up to rotation) embedding $z(i)$ in Euclidean r dimensional space of the ith subject by $R_{dist} = \Gamma_{n\times r}\Lambda_r\Gamma'_{r\times n}$, $Z_{n\times r} = \Gamma\Lambda^{1/2}$. The coordinates of the ith object $z(i)$ are given by the ith row of Z, $(z(i), z(j)) = R_{dist,ij}$, $\|z(i) - z(j)\|^2 = \hat{d}_{ij}$.

RKE example: proteins with BLAST scores.

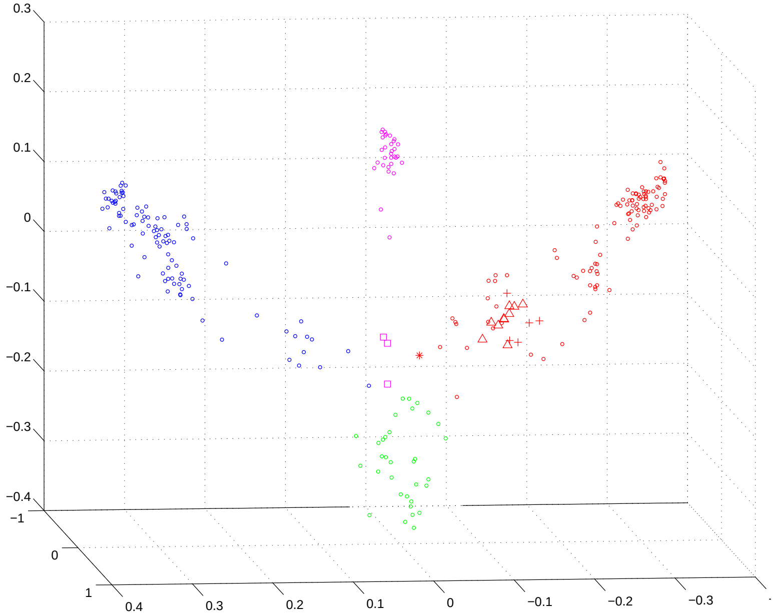


Figure 1: 3D representation of the sequence space for 280 proteins from the globin family. Red: α-globin subfamily, blue: β-globins, purple: myglobin subfamily, and green: a heterogeneous group encompassing proteins from other small subfamilies within the globin family. Note that in this example three, or even two dimensions are enough to separate the subfamilies.

Eigenvalues of R_{dist} from BLAST scores example as λ varies.

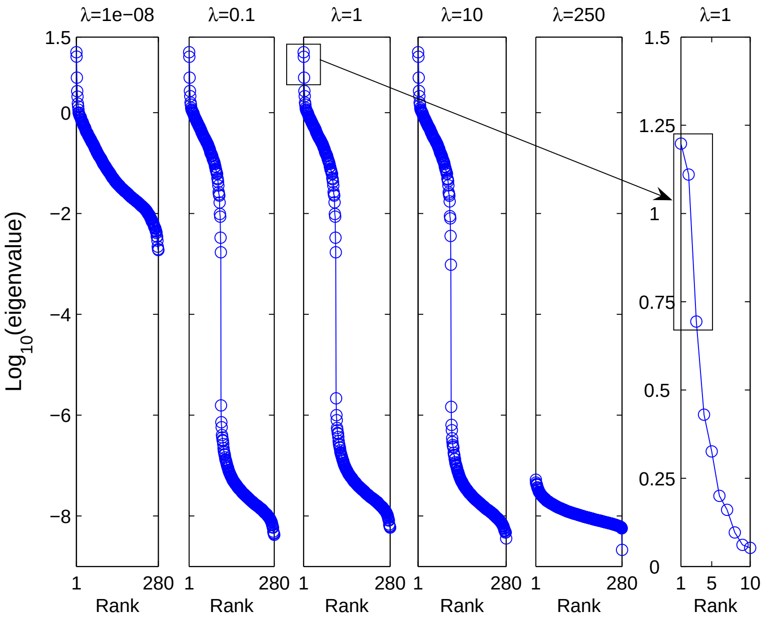


Figure 2: The effect of varying λ on the eigenvalues of R_{dist}. The left five images show log-scale eigensequence plots for five values of λ. As λ increases, smaller eigenvalues begin to shrink. The rightmost image shows the first 10 eigenvalues of the $\lambda = 1$ case displayed on a larger scale. In this example the plots are insensitive to λ over several orders of magnitute.

In this BLAST example the results may be similar to what one might get using multidimensional scaling with a specified dimension of two or three, but perhaps more fault-tolerant. In the next section we will use the RKE in an entirely different context: Wwe have a population from a demographic study of eye diseases where one-third of the population has at least one relative in the study. There is an outcome (pigmentary abnormalities-PA, Yes or No), and attribues-genetic markers, environmental/clinical(E/C) data, and pedigrees. The RKE will be used to embed the subjects in a Euclidean space using the pedigree information, (siblings close, niece-aunt not so close, etc) and the resulting coordinates will be added to the genetic and E/C attributes in a penalized logistic regression Smoothing Spline ANOVA model estimating the risk of PA. From reference 2. Switch gears now to model building with this data.

Part 2. Smoothing Spline ANOVA SS-ANOVA Models.

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 an SS-ANOVA model.

Reproducing Kernel Hilbert Spaces RKHS

- f will be in an RKHS. 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 $\mathcal{H}_K$ and vice versa.
- $K(\cdot, t*) \in \mathcal{H}_K$, all $t* \in \mathcal{T}$. $< K(\cdot, s), K(\cdot, t) >= K(s,t)$.
- All linear combinations of the $K(\cdot, t), t \in \mathcal{T}$ and their limits in the norm induced by the inner product constitute $\mathcal{H}_K$.
- $< f(\cdot), K(\cdot, t*) >= f(t*)$ for all $f \in \mathcal{H}_K$. Important!

To understand Smoothing Spline ANOVA Models:

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 :

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

ϕ_j are unpenalized components (parametric part) and

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

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

The SS-ANOVA penalty functional has the form

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

since $\|f\|^2_{\mathcal{H}_{\theta K}} = \theta^{-1}\|f\|^2_{\mathcal{H}_K}$. The θs are tuning parameters with an identifiability constraint along with λ. We tune this Bernoulli model with RKHS squared norm penalties using the GACV.

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:

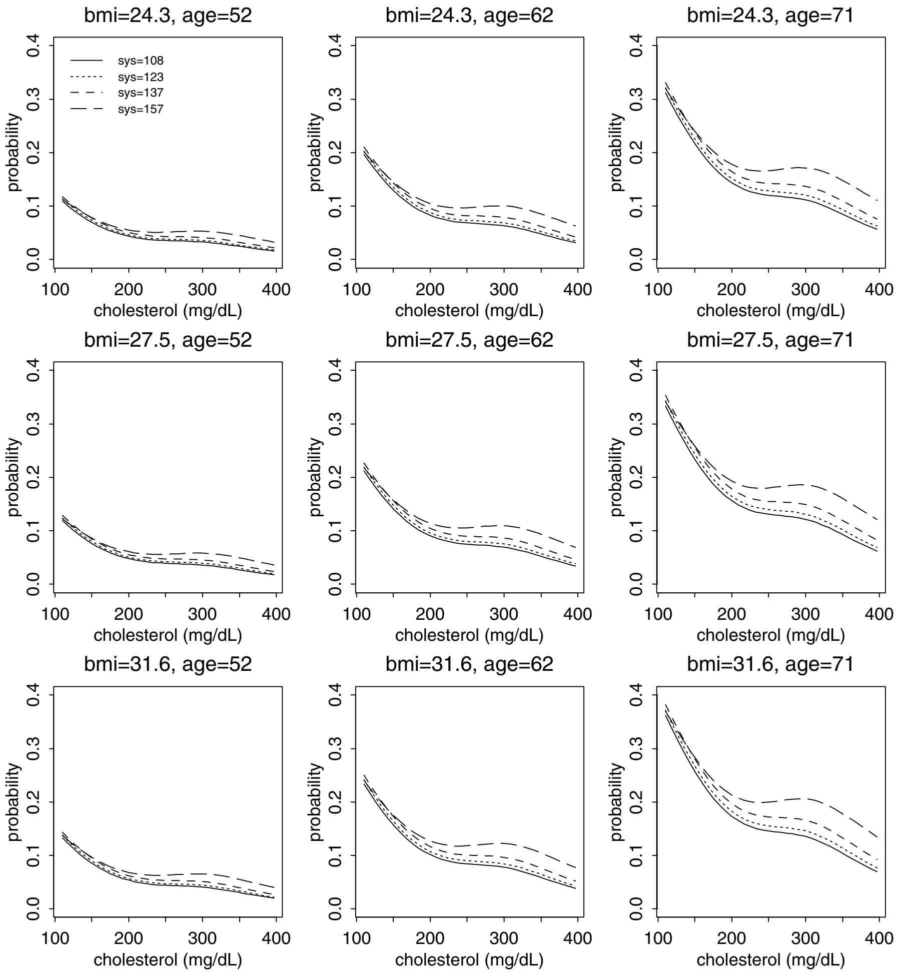
code	units	description
`horm`	yes/no	current usage of hormone replacement therapy
`hist`	yes/no	history of heavy drinking
`bmi`	kg/m^2	body mass index
`age`	years	age at baseline
`sysbp`	$mmmHg$	systolic blood pressure
`chol`	mg/dL	serum cholesterol
`smoke`	yes/no	history of smoking

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.



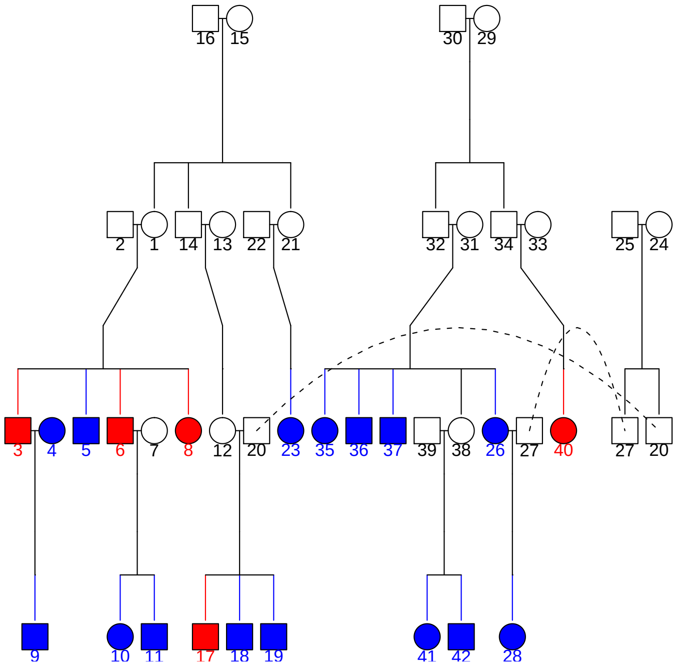
Estimated probability from an SS- ANOVA logistic regression model. Each x-axis is cholesterol, each set of four lines is four values of systolic blood pressure, each plot fixes body mass index and age to the shown values. `hist=0`, `horm =0.` From *Ann. Stat. 2000*.

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

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

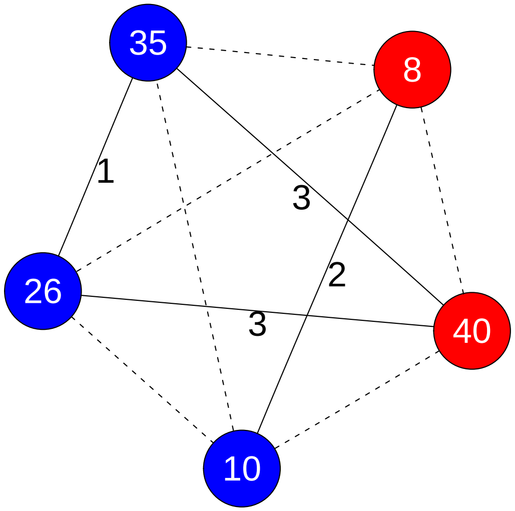
- First two lines: Genetic (SNP) data. Two SNPS each with three levels, (1,1), (1,2), (2,2). (SNP IDs in TR1148)
- Next three lines E/C variables
- Last line: Pedigree/relationship data goes here. Will explain.

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



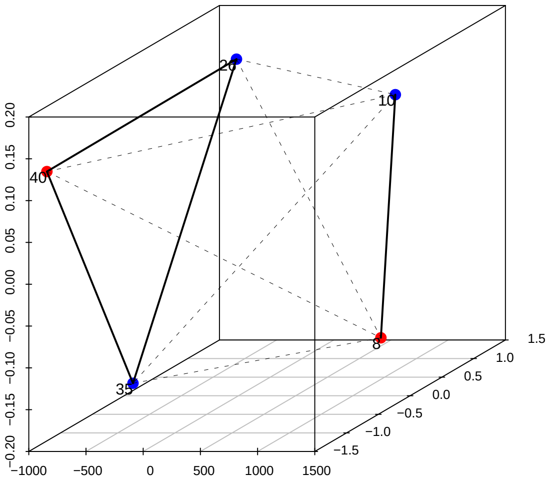
Relationship graph for subjects in the pedigree. Edge labels are distances defined by the kinship coefficient. Persons 26 and 35 are siblings [1], persons 8 and 10 are aunt and niece [2] and persons 26 and 40 are cousins [3]. Unrelated pairs have dashed lines.

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} are obtained based on the relationship coefficients (1, 2, 3, 4, 5, L), where L is "no relation" by a biologically motivated transformation. ($d_{ij} = -2log_2(2\phi_{ij})$) where ϕ is Malecot's kinship coefficient).

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, a radial basis function (RBF), $K_{ped}(z(i), z(j)) = K_{ped}(\|z(i) - z(j)\|)$. We use a Matern RBF in the present work. Recall that without the pedigree data,

$$f(\cdot) = \sum_{j=1}^{m} d_j \phi_j(\cdot) + \sum_{j=1}^{n} c_j K_\theta(\cdot, x(j)). \tag{2}$$

The pedigree data enters the model by

$$K_\theta(\cdot, \cdot) \to K_\theta(\cdot, \cdot) + \theta_{ped} K_{ped}(\cdot, \cdot). \tag{3}$$

The Matern family is a two-parameter family, and the parameters are to be chosen along with λ and the θs.

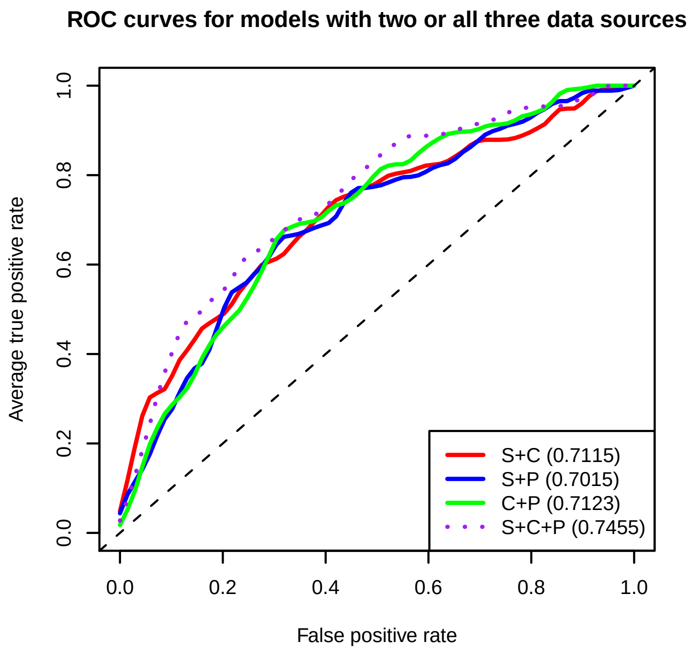
Qualitative Results

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

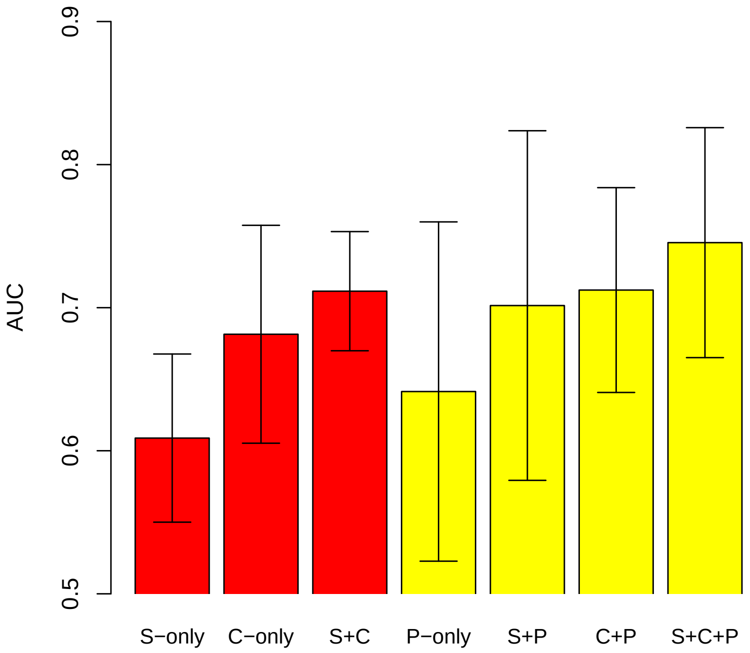
Compare models by evaluating the AUC (Area Under the Curve).

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



ROC curves for the models with two data sources. Plot is constructed by classifying each person in a test set 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.

Part 3. Partly Missing Covariates.

In the third reference, methods for handling partly missing covariates in parametric models have been extended to Smoothing Spline ANOVA models (robustly!).

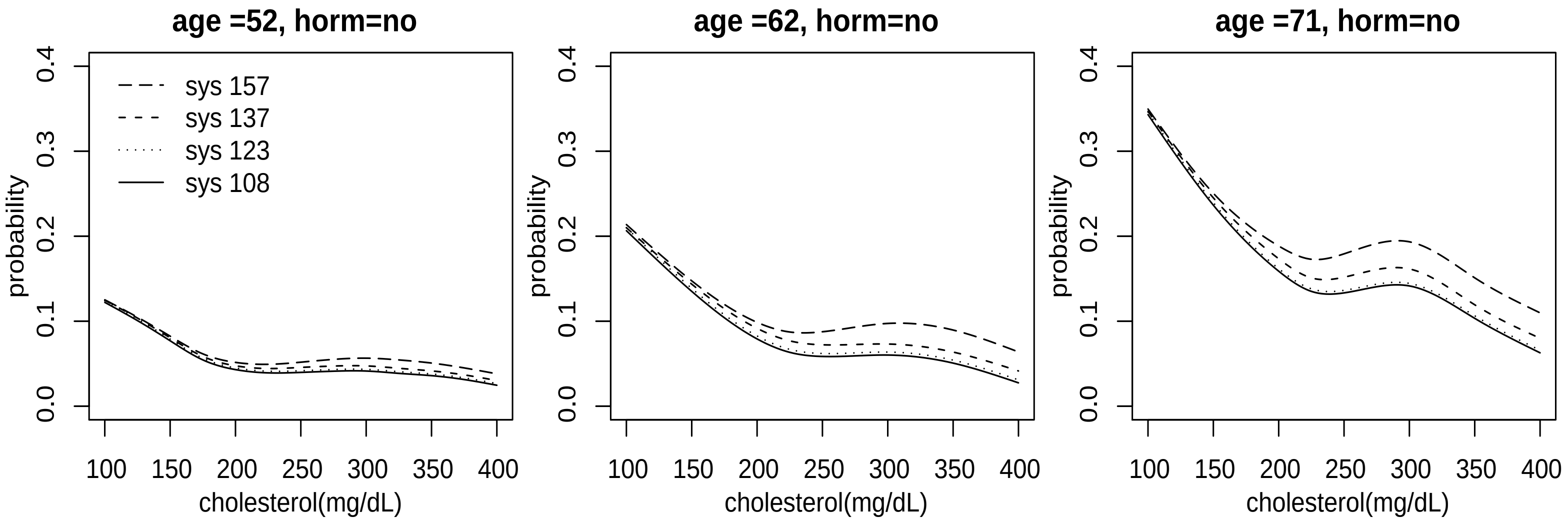


Figure 3: Repeat of the `bmi=27.5` row of the earlier figure from *Ann. Stat. 2000.* This model includes `horm,hist,bmi,age,sysbp,chol`.

To test the proposed method, 517 subjects out of the $n = 2585$ original subjects with `chol` between 250 and 350 have one or more of their covariates `sys,bmi,horm` deleted. 30 subjects missed `sys,bmi,horm`, 109 subjects missed `sys,bmi`, 118 subjects missed `sys,horm` and 260 subjects missed one attribute.

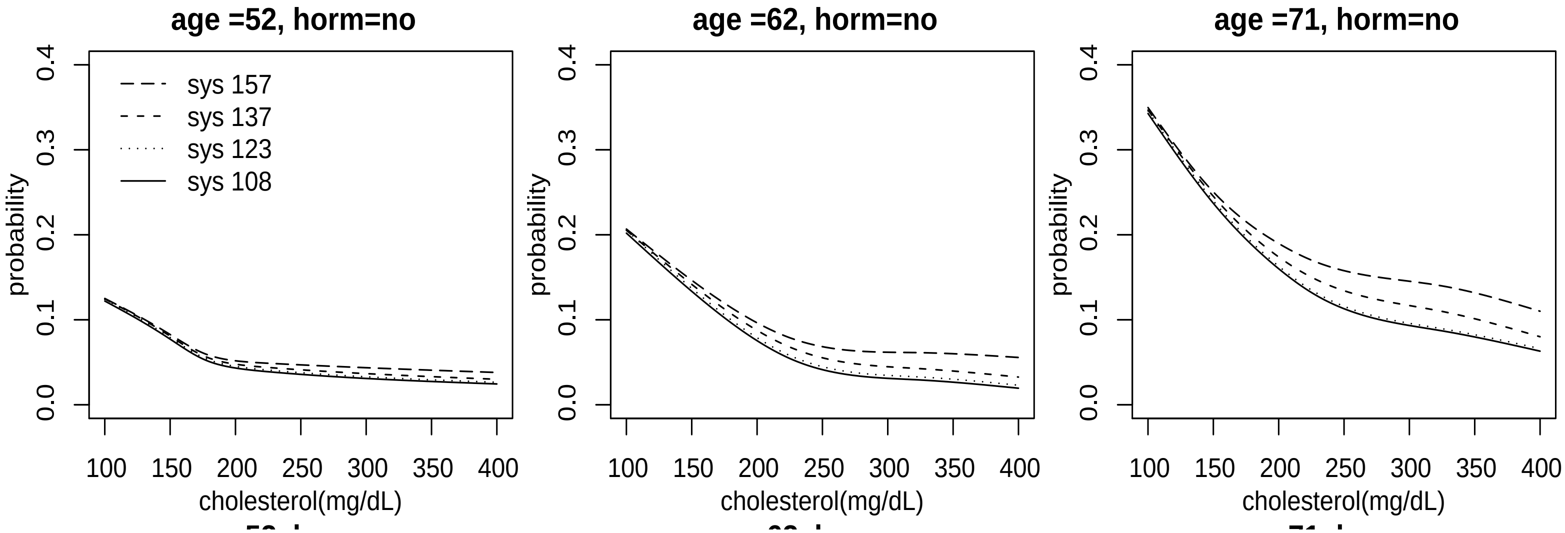


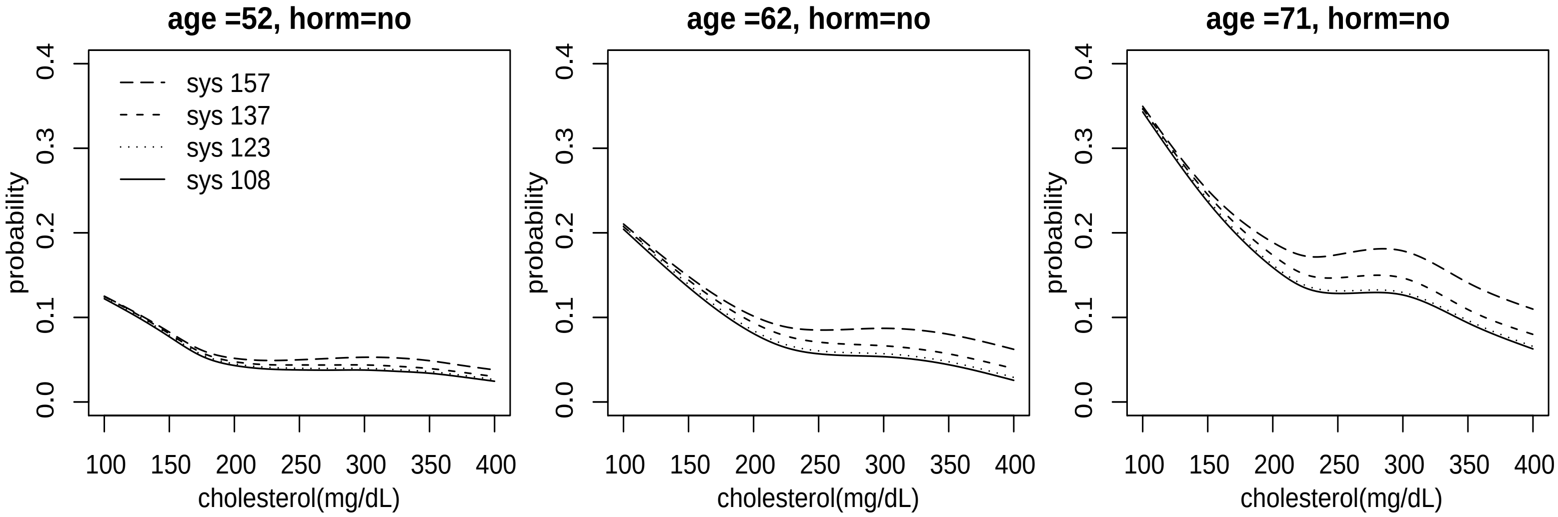
Figure 4: Model using only the 2068 subjects with no missing data in the method test data set. Note the missing ‘bump’.

Quadrature Penalized Likelihood Estimation QPLE

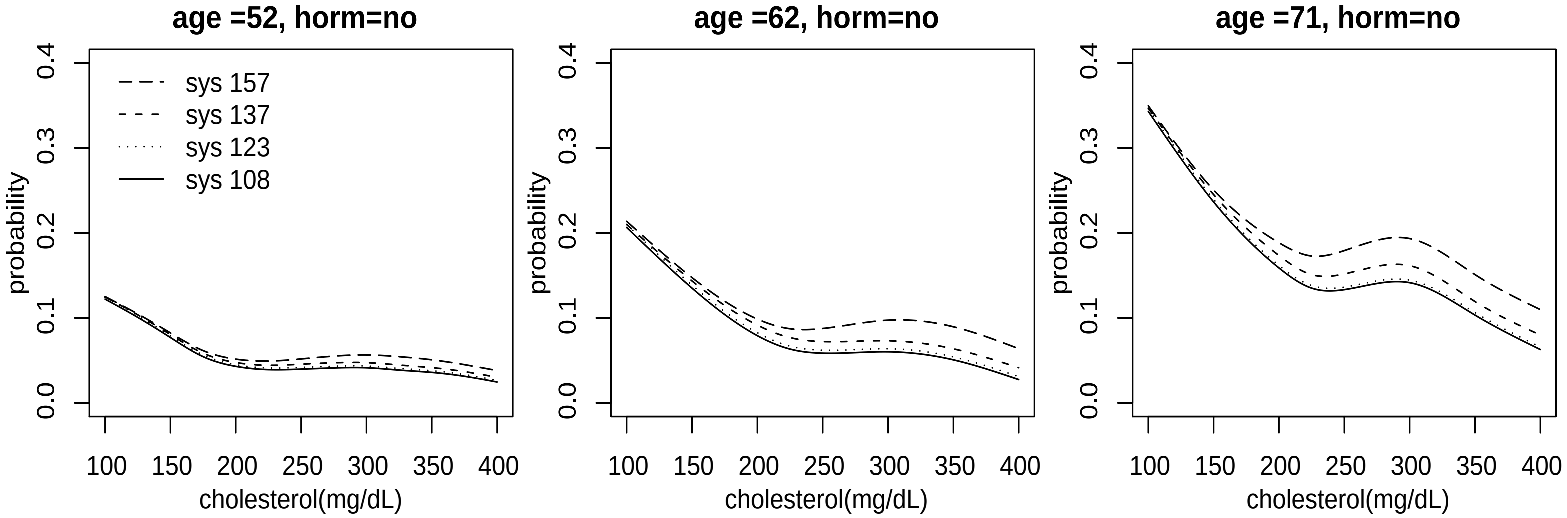
The QPLE method for missing data in SS-ANOVA models generalizes results on missing data for parametric models (many references). It begins by assuming conditional probability distributions for each cluster of missing observations, conditional on available observations, with parameters to be estimated. The probability distributions for continuous data appear in integrals for the penalized log likelihood. The integrals are replaced by quadrature formulae. Thus, the continuous prior probability distributions are replaced by distributions (weights) on mass points, and (updated) parameters of the distributions are used to update the weights via the EM algorithm. The process is carried out for each trial set of tuning parametera and at convergence a tuning score (GACV) can be obtained in a manner similar to that for complete data.

In this example the conditional distribution for `sys,bmi` given `age,chol` was chosed as bivariate normal with means linear in `age,chol`, with unknown coefficients and unknown bivariate covariance matrix, and `horm` was modeled conditional on `age, chol,sys,bmi`, as the conditional logit linear in these variables with unknown coefficients.

Here is the resulting model



which can be compared with the model with complete data:.



The result is rather amazing, but depends on the attributes being well correlated, very common in medical data.

Summary and Conclusions

In Part 1 we have described the RKE estimate which embeds noisy, non-metric, incomplete dissimilarity information into a Euclidean space. In Part 2 we first reviewed penalized log likelihood estimates for Bernoulli responses. Then we showed how pedigree data could be embedded into a Euclidean space via an RKE estimate, and then combined with genetic and environmentl/clinical variables to build a penalized logistic regression Smoothing Spline ANOVA model. Part 3 notes recent work dealing with partially missing covariate data in Smoothing Spline ANOVA models using a Quadrature Penalized Estimate QPLE in a robust manner.

Robustness is everywhere and we thank the organizers for prompting us to think along these lines!