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, by Hector Corrada Bravo, Grace Wahba, Kristine Lee, Barbara Klein, Ronald Klein and Sudha Iyengar, which relies on Lu, Keles, Wright and Wahba, PNAS Aug 30, 2005

XII Latin American Congress of Probability and Mathematical Statistics
March 26, 2012
Viña del Mar, Chile

Links to these slides, and the above papers, in my website http://www.stat.wisc.edu/~wahba/.
Abstract

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 similarity information in an SS-ANOVA model. 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. Sibling, parent-child, uncle-niece, etc. provide relationship data which is incorporated in an SS-ANOVA model penalized likelihood model. One issue is to evaluate the relative influence of the three distinct sources of information.
Outline

1. The Log Likelihood for Bernoulli \((0, 1)\) 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 and Conclusions
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: $p(x) = \text{Prob}(y = 1|x)$

• 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 $\lambda$ 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. What is an RKHS?

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

- Moore-Aronszajn Theorem: To every positive definite function \( K(\cdot, \cdot) \) there corresponds a unique RKHS \( \mathcal{H}_K \) and vice versa.

\[ K(\cdot, t^*) \in \mathcal{H}_K, \text{ all } t^* \in \mathcal{T}. \quad < K(\cdot, s), K(\cdot, t) >= K(s, t). \]

- 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 $E_\alpha$ 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

$$\mu = \prod_\alpha E_\alpha f = \int \cdots \int f(x_1, \cdots, x_d) d\mu_1(x_1) \cdots d\mu_d(x_d)$$

$$f_\alpha = (I - E_\alpha) \prod_{\beta \neq \alpha} E_\beta f$$

$$f_{\alpha\beta} = (I - E_\alpha)(I - E_\beta) \prod_{\gamma \neq \alpha, \beta} E_\gamma f$$

$$\vdots \quad \vdots \quad E_\alpha f_\alpha = 0, \quad E_\alpha E_\beta f_{\alpha\beta} = 0, \text{etc.}$$
ANOVA Decomposition of Functions of Several Variables (continued)

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

- Terms satisfy ANOVA-like side conditions (identifiable).
- SS-ANOVA representation with weights on kernels:

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

\( \phi_j \) are unpenalized components (parametric part) with

\[ K_\theta(\cdot, \cdot) = \sum_{\alpha=1}^{d} \theta_\alpha K_\alpha(\cdot, \cdot), + \sum_{\alpha \leq \beta} \theta_{\alpha\beta} K_{\alpha\beta}(\cdot, \cdot) + \cdots \]

- Kernels depend only on components of \( x \) in the subscripts.
- \( \mathcal{E}_\alpha K_\alpha(x_\alpha, \cdot) = 0 \quad K_{\alpha\beta} = K_\alpha K_\beta \cdots \) etc. (side conditions)
ANOVA Decomposition of Functions of Several Variables (continued)

The SS-ANOVA penalty functional has the form

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

since \[ \|f\|_{\mathcal{H}_{\theta K}}^2 = \theta^{-1} \|f\|_{\mathcal{H}_K}^2 \]. \( \lambda \) and the \( \theta \)s are tuning parameters with an identifiability constraint on the \( \theta \)s.
ANOVA Decomposition of Functions of Several Variables (continued)

To Reprise: For Bernoulli data, find $f$ to min:

$$\mathcal{L}(y, f) + \lambda J(f) \quad (*)$$

where

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

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

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

For fixed $\lambda, \theta$ minimize $(*)$ with respect to $d$ and $c = (c_1, \cdots, c_n)$. Choose $\lambda, \theta$ by minimizing the GACV (Generalized Approximate Cross Validation).
Figure 1: Grace Wahba, Spline Models for Observational Data (1990)

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

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.


- 684 women have at least one relative also in the study.
The predictor variables of present interest are:

<table>
<thead>
<tr>
<th>code</th>
<th>units</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>horm</td>
<td>yes/no</td>
<td>current usage of hormone replacement therapy</td>
</tr>
<tr>
<td>hist</td>
<td>yes/no</td>
<td>history of heavy drinking</td>
</tr>
<tr>
<td>bmi</td>
<td>(kg/m^2)</td>
<td>body mass index</td>
</tr>
<tr>
<td>age</td>
<td>years</td>
<td>age at baseline</td>
</tr>
<tr>
<td>sysbp</td>
<td>(mmHg)</td>
<td>systolic blood pressure</td>
</tr>
<tr>
<td>chol</td>
<td>(mg/dL)</td>
<td>serum cholesterol</td>
</tr>
<tr>
<td>smoke</td>
<td>yes/no</td>
<td>history of smoking</td>
</tr>
</tbody>
</table>

Table 1: E/C covariates for BDES pigmentation abnormalities SS-ANOVA model
• The fitted E/C model that we are using in the present study is

\[ f(t) = \mu + f_1(\text{sys}) + f_2(\text{chol}) + f_{12}(\text{sys, chol}) + d_{\text{age}} \cdot \text{age} + d_{\text{bmi}} \cdot \text{bmi} + d_{\text{horm}} \cdot I_1(\text{horm}) + d_{\text{drin}} \cdot I_2(\text{drin}) + d_{\text{smoke}} \cdot I_3(\text{smoke}) \]

• 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

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

- 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, Keles, Wright and Wahba, 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. ($d_{ij} = -2\log_2(2\phi_{ij})$ where $\phi$ is Malecot’s kinship coefficient).
The distance encoding matrix $R_{dist}$ is obtained by solving the convex cone optimization problem:

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

where $R \succeq 0$ means $R$ is in the convex cone of all real non-negative definite matrices of dimension $n$, $\Omega$ 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 $R$. Small eigenvalues in the fitted $R_{dist}$ are deleted. $R_{dist}(i,j)$ gives a (unique up to rotation) embedding $z(i)$ of the $i$th subject. This $z(i)$ will later appear in $f_{ped}(z(i))$ in the extended SS-ANOVA model.
$z(i)$ 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 DataEncoded 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), $K_{ped}(z(i), z(j)) = k(\|z(i) - z(j)\|)$. We use a Matern RBF
in the present work. Recall that without the pedigree data,

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

The pedigree data enters the model by

$$K_\theta(\cdot, \cdot) \rightarrow K_\theta(\cdot, \cdot) + \theta_{ped} K_{ped}(\cdot, \cdot). \quad (3)$$

The Matern family is a two-parameter family, and the parameters
are to be chosen along with $\lambda$ and the $\theta$s.
Preliminary Qualitative Results

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

- $S = \text{SNPS (genetic data) only}$
- $C = \text{Environmental/Clinical (E/C) data only}$
- $S + C$
- $P = \text{Pedigrees only}$
- $S + P$
- $C + P$
- $S + C + P$

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.
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.
Final Remarks

The RKE and its embedding properties have some relation to spectral clustering as well as multidimensional scaling (not discussed today).

Future promising applications include medical imaging where making comparisons between images could be fruitful. GW’s theorem says that determining appropriate dissimilarity measures is the key in this and other applications.