Analysis of Variance in Reproducing Kernel Hilbert Spaces, Distance Correlation, and Why Mortality Runs in Families

Grace Wahba

Scientific Computing and Imaging Institute
University of Utah
March 27, 2015
Salt Lake City, Utah

Links to these slides in my website
`http://www.stat.wisc.edu/~wahba/` $->$ `TALKS`

Analysis of Variance in Reproducing Keknel Hilbert Spaces, Distance Correlation, and Why Mortality Runs in Families

Abstract

Reproducing Kernel Hilbert Spaces (RKHS) appeared in a theoretical paper (Aronszajn 1950), but their use in applied nonparametric regression, statistical model building, machine learning and classification had to wait for modern computational facilities. We review RKHS and then Analysis of Variance (ANOVA) decompositions of functions of several variables in tensor products of RKHS. We review Distance Correlation, which is a completely nonparametric approach for examining correlation between essentially arbitrary clusters of random variables, based on samples of pairwise distances. We marry these tools to examine how lifestyle and other variables in the Beaver Dam Eye Study correlate with mortality as it runs in families.

Outline

A. 1-5 Positive definite functions and RKHS

B. 1-2 ANOVA decompositions of functions of several variables

C. 1-2 Smoothing Spline ANOVA (SS-ANOVA) decompositions of functions of several variables, and fits in RKHS

D. 1-2 Domains, distance measures, practical issues

E. 1-8 SS-ANOVA joins Distance Correlation (DCOR) to study mortality as it runs in families in the Beaver Dam Eye Study

F. 1 Comments and Conclusions

A1.Positive definite functions

Let $\mathcal{T}$ be some (measurable) domain. $K(s,t), s,t \in \mathcal{T}$ is said to be (strictly) positive definite if, for every n, and every $t_1, \ldots, t_n$ in $\mathcal{T}$, and constants $c_1, \ldots, c_n$,

$$\sum_{j,k} c_j c_k K(t_j, t_k) > 0.$$

The Moore-Aronszajn theorem

The Moore-Aronszajn Theorem: Every positive definite function $K(s,t)$ on $\mathcal{T} \times \mathcal{T}$ corresponds to a unique Reproducing Kernel Hilbert Space $\mathcal{H}_K$ of functions defined on $\mathcal{T}$, and vice versa.

A2. Construction of the unique RKHS

Let $K_s(t) \equiv K(s,t)$ define a function of t on the domain $\mathcal{T}$ with s fixed. Then K_s is in $\mathcal{H}_K$ and the linear manifold spanned by all finite linear combinations of elements of the form

$$f(t) = \sum_\ell c_{s_\ell} K_{s_\ell}(t)$$

is in $\mathcal{H}_K$. Since K is positive definite, it can be shown that

$$< K_s, K_t > \;\; = \;\; K(s,t)$$

defines an inner product and hence a norm on this linear manifold. The Hilbert space posited by the Moore-Aronszajn theorem is the closure of the linear manifold under the induced norm.

A3. The defining property of RKHS

Let $t_* \in \mathcal{T}$ and let $f \in \mathcal{H}_K$. Then $f(t_*)$ can be written as an inner product in $\mathcal{H}_K$ as:

$$f(t_*) \;=\; < f, K_{t_*} >$$

for every $t_* \in \mathcal{T}$ and $f \in \mathcal{H}_K$. K_{t_*} is called the *representer of evaluation* at t_*. RKHS are characterized by the fact that they contain all their representers of evaluation.

A4. The representer theorem (Kimeldorf & Wahba 1971)

Representer Theorem: Given a convex cost function $\mathcal{L}$ and data $y_i, t(i), i = 1, \cdots, n$, the minimizer f_λ of $\sum_{i=1}^n \mathcal{L}[y_i, f(t(i)] + \lambda \|f\|^2_{\mathcal{H}_K}$ is in the span of the representers $K_{t(1)}, K_{t(2)} \cdots K_{t(n)}$. More generally, if $\|f\|^2_{\mathcal{H}_K}$ is replaced by a seminorm $J(f)$ on $\mathcal{H}_K$, (that is, $J(f)$ acts like a square norm but with a null space $\{\phi_\nu\}$), then f_λ can be constructed in span $\{\phi_\nu\} \cup \{K_{t(i)}\}$.

The most familiar case is the cubic smoothing spline on $[0, 1]$, where $\mathcal{H}_K$ is the space of functions on $[0, 1]$ with square integral second derivative, the $\{\phi_\nu\}$ are linear functions and $J(f) = \int_0^1 (f''(t))^2 dt$.

A5. Positive definite functions define distances

Let K be a positive definite function on $\mathcal{T} \times \mathcal{T}$. Then K can be used to define a pairwise distance between any two points s and t in $\mathcal{T}$ by

$$[dist]^2[s,t] = \|K_s - K_t\|^2 = K(s,s) + K(t,t) - 2K(s,t).$$

(sometimes known in the CS literature as the "kernel trick".) Conversely, given noisy, incomplete pairwise distances, one can fit a positive definite (or non-negative definite) kernel (matrix) via Regularized Kernel Estimation that attempts to respect this information while controlling for the trace. (Lu *et al* 2005, Corrada Bravo *et al* 2009.)

B1. ANOVA decompositions of functions of several variables

Let $\mathcal{T}^{(\alpha)}, \alpha = 1, \ldots, d$ be d measurable domains with members $t_\alpha \in \mathcal{T}^{(\alpha)}$. Let

$$t = (t_1, \ldots, t_d) \in \mathcal{T}^{(1)} \times \cdots \times \mathcal{T}^{(d)} = \mathcal{T}.$$

For f satisfying some measurability condition, ANOVA decompositions of f of the form

$$f(t_1, \cdots, t_d) = \mu + \sum_{\alpha} f_\alpha(t_\alpha) + \sum_{\alpha\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots \qquad (1)$$

can always be defined. Our goal is to be able to fit ANOVA models of this form given scattered, noisy observations $\{y_i, t(i), i = 1, \cdots, n\}$ on $\mathcal{T}$.

B2. ANOVA decompositions of functions of several variables (cont.)

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

$$(\mathcal{E}_\alpha f)(t) = \int_{\mathcal{T}^{(\alpha)}} f(t_1, \ldots, t_d) d\mu_\alpha(t_\alpha).$$

Then the identity can be decomposed as

$$I = \prod_\alpha (\mathcal{E}_\alpha + (I - \mathcal{E}_\alpha)) = \prod_\alpha \mathcal{E}_\alpha + \sum_\alpha (I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta +$$

$$\sum_{\alpha < \beta} (I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha, \beta} \mathcal{E}_\gamma + \cdots + \prod_\alpha (I - \mathcal{E}_\alpha),$$

giving

$\mu = (\prod_\alpha \mathcal{E}_\alpha) f, \quad f_\alpha = ((I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta) f$

$f_{\alpha\beta} = ((I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha, \beta} \mathcal{E}_\gamma) f \quad \ldots$ $[\mathtt{Ex : full\ factorial\ designs}]$

C1. Smoothing-Spline ANOVA (SS-ANOVA) decompositions

The idea behind SS-ANOVA is to construct an RKHS $\mathcal{H}$ of functions on $\mathcal{T}$ as the tensor product of RKHSs on each $\mathcal{T}^{(\alpha)}$ that admit an ANOVA decomposition. Let $\mathcal{H}^{(\alpha)}$ be an RKHS of functions on $\mathcal{T}^{(\alpha)}$ with $\int_{\mathcal{T}^{(\alpha)}} f_\alpha(t_\alpha) d\mu_\alpha = 0$ and let $[1^{(\alpha)}]$ be the one dimensional space of constant functions on $\mathcal{T}^{(\alpha)}$. Construct the RKHS $\mathcal{H}$ as

$$\mathcal{H} = \prod_{j=1}^{d} (\{[1^{(\alpha)}]\} \oplus \{\mathcal{H}^{(\alpha)}\})$$

$$= [1] \oplus \sum_{j} \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha<\beta} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \cdots ,$$

where [1] denotes the constant functions on $\mathcal{T}$. Then $f_\alpha \in \mathcal{H}^{(\alpha)}, f_{\alpha\beta} \in [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}]$ and so forth, where the series will usually be truncated at some point. Note that the usual ANOVA side conditions hold here.

C2. Smoothing Spline ANOVA fits

Given data $\{y_i, t(i), i = 1, \ldots, n\}$, we can fit an SS-ANOVA model by minimizing

$$\sum_{i=1}^{n} \mathcal{L}[y_i, f_\lambda(t(i)] + \sum_{\alpha} \lambda_\alpha J_\alpha(f_\alpha) + \sum_{\alpha\beta} \lambda_{\alpha\beta} J_{\alpha\beta}(f_{\alpha\beta}) + \cdots$$

where the Js may be RKHS square norms or seminorms, and the representer theorem used to provide a finite dimensional representation for the minimizer.

D1. Domains, distance measures, practical issues etc.

- Essentially nothing is assumed about individual domains $\mathcal{T}^{(\alpha)}$ other than that averaging operators can be defined on them.

- Typical historical domains include the unit interval, Euclidean d space, the sphere and other Riemannian manifolds, typically with J_α depending on the Laplacian.

- More recently RK's (positive definite functions) have been defined on peptides (Shen *et al* 2013), anatomical (airway) trees (Feragen *et al* 2013), and many other objects, including images, paragraphs, networks, graphs.

- The choice of RK is equivalent to a choice of distance measure. If only pairwise (Euclidean) distances are known, then radial basis functions may be used as RK's. If scattered noisy pairwise dissimilarities are observed, Regularized Kernel Estimation may be used to embed a training set into a Euclidean space.

D2. Domains, distance measures, practical issues (cont.)

- The kernel trick may be used to transform one distance measure to another.
- The SS-ANOVA class of methods provides a principled way of combining information from qualitatively different variables in a predictive model, including interactions.
- When it comes to applications, many practical issues remain. The key to a successful application may well depend on a meaningful choice of distances (as well as pruning, tuning, approximations for extremely large data sets, not discussed).

R codes: `gss` and `assist`. Books: Wahba 1990, Gu 2002, Berlinet and Thomas-Agnan 2003, Wang 2011. Gu and Wang provide many examples. More references at end.

E. SS-ANOVA meets DCOR
(Kong, Klein, Klein, Lee and Wahba 2012)

Does Life Span Run in Families, and If So, Why? The Beaver Dam Eye study (BDES) started with about 5000 subjects in 1988 between the ages of 43-84 years and about 2400 of these had relatives in the study. The study has a large amount of covariate information, and pedigree (relationship) information, along with mortality information through 2011. We compared pairwise death ages between relatives and between unrelated subjects and it is clear that mortality runs in families. SS-ANOVA combined with Distance Correlation is used to quantify this.

- What is DCOR?
- Variable Descriptions, the SS-ANOVA Deathage Scoring Model
- Determining DCOR from the Deathage Scoring Model
- DCOR results

Distance Correlation (DCOR) (Szekely and Rizzo 2009)

For a random sample $(X, Y) = \{(X_k, Y_k) : k = 1, ..., n\}$ of n i.i.d. random vectors (X, Y) from the joint distribution of random vectors X in R^p and Y in R^q, the Euclidean distance matrices $(a_{ij}) = (|X_i - X_j|_p)$ and $(b_{ij}) = (|Y_i - Y_j|_q)$ are computed. Define the double centering distance matrices

$$A_{ij} = a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}, \quad i, j = 1, ..., n,$$

where

$$\bar{a}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \quad \bar{a}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}, \quad \bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,j=1}^{n} a_{ij},$$

similarly for $B_{ij} = b_{ij} - \bar{b}_{i\cdot} - \bar{b}_{\cdot j} + \bar{b}_{\cdot\cdot}, \quad i, j = 1, ..., n.$

The sample distance covariance $\mathcal{V}_n(X,Y)$ is defined by

$$\mathcal{V}_n^2(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}B_{ij}.$$

The sample distance correlation $\mathcal{R}_n(X,Y)$ (DCOR) is defined by

$$\mathcal{R}_n^2(X,Y) = \begin{cases} \dfrac{\mathcal{V}_n^2(X,Y)}{\sqrt{\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y)}}, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) > 0; \\ 0, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) = 0, \end{cases}$$

where the sample distance variance is defined by

$$\mathcal{V}_n^2(X) = \mathcal{V}_n^2(X,X) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}^2.$$

What is the Sample Distance Covariance $\mathcal{V}_n^2(X,Y)$ estimating?

Let f_{XY} be the characteristic function of the joint distribution of X and Y, and let f_X and f_Y be the characteristic functions of X and of Y. Let

$$\mathcal{V}^2(X,Y) = \int_{R^{p+q}} |f_{XY}(s,t) - f_X(t)f_Y(s)|^2 \omega_{pq}(t,s)dtds$$

where

$$\omega_{pq} = [c_p c_q |t|_p^{1+p} |s|_q^{1+q}]^{-1}.$$

Amazing Theorem: (Szekely and Rizzo).

$$\mathcal{V}_n^2(X,Y) \text{ is the sample version of } \mathcal{V}^2(X,Y)$$

Table 1. Variable Descriptions: Fixed:Lifestyle:Diseases (from BDES)

variable	units	description
deathage	years	death age
baseage	years	age at baseline
gender	F/M	gender
....		
edu	years	highest year school/college completed
bmi	kg/m^2	body mass index
smoke	yes/no	history of smoking
inc	yes/no	household personal income > 20T
....		
diabetes	yes/no	history of diabetes
cancer	yes/no	history of cancer
heart	yes/no	history of cardiovascular disease
kidney	yes/no	history of chronic kidney disease

SS-ANOVA Death Age Scoring Model

Death age as a function of fixed, lifestyle and disease variables will be modeled using SS-ANOVA as

$$\begin{aligned} death\ age_i = & g_0(baseline\ age_i, gender_i) + \\ & g_1(lifestyle\ factors_i) + g_2(diseases_i), \end{aligned}$$

where g_0 is a term involves fixed characteristics, baseline age and gender for individual i, g_1 is a term that includes only lifestyle factors, namely edu, bmi, smoke, inc, and g_2 is a term containing only disease variables, namely diabetes, cancer, cardiovascular disease and chronic kidney disease. In the paper, the fitted values of g_1 and g_2 are treated as scores for the individuals and to be used to assess the association with familial relationships. Do g_1 and g_2 scores, both high and low, run in families, thus partially explaining why mortality runs in families?

The SSANOVA Death Age Scoring Model

The SSANOVA death age scoring model is:

$$
\begin{aligned}
deathage =& \mu + f_1(baseage) + \beta_{gender} I_{\{gender=F\}} && \big\}\textit{fixed} \\
& + f_2(edu) + f_{12}(baseage : edu) + f_3(bmi) && \multirow{2}{*}{$\Bigg\}\textit{lifestyle } (g_1)$} \\
& + \beta_{smoke} I_{\{smoke=no\}} + \beta_{inc} I_{\{inc>20T\}} && \\
+ \beta_{diabetes} I_{\{diabetes=no\}} &+ \beta_{cancer} I_{\{cancer=no\}} && \multirow{2}{*}{$\Bigg\}\textit{disease } (g_2)$} \\
+ +\beta_{heart} I_{\{heart=no\}} &+ \beta_{kidney} I_{\{kidney=no\}} &&
\end{aligned}
$$

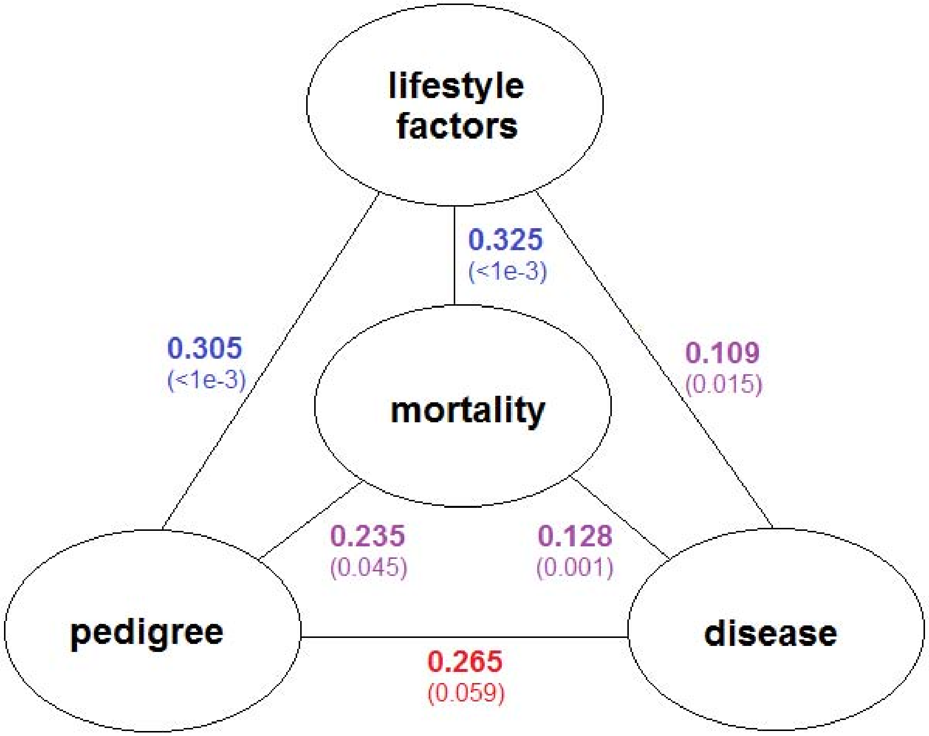
Determining Distance Correlation (DCOR)

All six DCOR values between mortality, pedigree, lifestyle factors and diseases will be computed.

The lifestyle factor score g_1 for an individual is based on the four-vector of the fitted effects for smoke, bmi, edu and inc. Similarly the disease score g_2 is based on the four-vector of fitted effects for the four disease variables.

It is well known that the pedigree distance $(1 - 2\phi)$ based on the kinship coefficient is Euclidean, so that pairwise pedigree distances can be used directly in DCOR.

DCOR Results, Entire Pedigrees



very signif-signif

lifestyle:pedigree

lifestyle:mortality

disease:mortality

mortality:pedigree

disease:lifestyle

disease:pedigree

DCOR results using pedigree distance. Numbers in parens are significance levels to test independence, based on a permutation test with 1000 replicates.

More questions than answers

- We have shown that pairwise differences in lifestyle factors that run in families correlate well with pairwise differences in death age that also run in families, partially accounting for the familial death age effect. This leads to new questions to be asked about the complex relations between genetics, family structure, lifestyle factors, and other variables. We provide here an overall methodological approach joining SS-ANOVA with DCOR which shows promise to help in answering these questions in future studies.

Comments and conclusions

- We have reviewed Smoothing Spline ANOVA models, obtained by the same geometry that decomposes full factorial designs (Fisherian ANOVA), by constructing ANOVA decompositions of tensor product RKHS. The approach allows for the joint modeling of heterogenous variables including interactions. The flexibility of the models allows for highly diverse applications. In applications, many practical issues remain, including model pruning and tuning, numerical approximations for extremely large data sets, subject matter informed choice of distance measures, and graphical displays of complex results.
- A marriage of an SS-ANOVA model with Distance Correlation is used to examines pairwise correlations between variables of interest in a completely distribution free way. Again, applications with big data sets show great potential as well as presenting many challenging issues.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Am. Math. Soc.*, 68:337–404, 1950.

[2] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–95, 1971.

[3] 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. Open Source at www.pnas.org/content/102/35/12332, PMCID: PMC118947.

[4] H. Corrada Bravo, K. E. Lee, B. E. K. Klein, R. Klein, S. K. Iyengar, and G. Wahba. 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. Open Source at `www.pnas.org/content/106/20/8128.full.pdf+html`, PMCID: 2677979.

[5] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.

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

[7] A.Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2003.

[8] Y. Wang. *Smoothing Splines: Methods and Applications*. Chapoman & Hall/CRC Monographs on Statistics & Applied Probability, 2011.

[9] J. Kong, B. Klein, R. Klein, K. Lee, and G. Wahba. Using distance correlation and Smoothing Spline ANOVA to assess associations of familial relationships, lifestyle factors, diseases and mortality. *PNAS*, pages 20353–20357, 2012. PMCID: 3528609.

[10] G. Szekely and M. Rizzo. Brownian distance covariance. *Ann. Appl. Statist.*, 3:1236–1265, 2009.

[11] A. Feragen, J. Petersen, D. Grimm, A. Dirksen, J. Pederse, K. Borgwardt, and M. deBruijne. Geometric tree kernels: classification of COPD from airway tree geometry. In *IMPMI'13 Proceedings of the 23rd international conference on Information Processing in Medical Imaging*, pages 171–183, Heidelberg, 2013. Springer.

[12] H-J. Shen, H-S. Wong, Q-W. Xiao, X. Guo, and S. Smale. Introduction to the peptide binding problem of computational immunology: New results. *Foundations of Computational Mathematics*, pages 1–34, 2013.

[13] C. Gu and G. Wahba. Smoothing spline ANOVA with component-wise Bayesian "confidence intervals". *J. Computational and Graphical Statistics*, 2:97–117, 1993.

[14] 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. Neyman Lecture.

[15] B. Brumback and J. Rice. Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Amer. Statist. Assoc.*, 93:961–991, 1998.

[16] W. Guo. Inference in Smoothing Spline Analysis of Variance. *J. Roy. Stat. Soc. B*, 64:887–889, 2002.

[17] H. Zhang and Y. Lin. Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, 34:2272–2297, 2006.

[18] X. Sun, P. Ma, and R. Mumm. Nonparametric method for genomics-based prediction of performance of quantitative traits involving epistasis in plant breeding. *PLOS One*, November, 2012.

[19] S. Touzani and D. Busby. Smoothing spline analysis of variance approach for global sensitivity analysis of computer code. *Reliability Engineering and System Safety*, 112:67–81, 2013.