Statistical Model Building, Machine Learning, and the Ah-Ha Moment.

Grace Wahba

Part 1. Some historical remarks

Part 2. Ah-Ha moments

Part 3. Technical part: Two papers using pairwise dissimilarity information, a topic of growing importance

Women in Statistics

May 15-17, 2014

Cary, NC

Links to these slides in my website

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

Links to the references given and all papers/preprints since 1993 at

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

Abstract

After a few historical remarks, I will describe favorite parts of my career over time which involved serendipitous interactions with colleagues and studets that provided a solution ("the Ah-Ha moment") to some interesting problems. Then I will move to some recent work involving utilization of pairwise dissimilarity/distance information.

Outline

1. A few historical remarks
2. Some "Ah-Ha" moments
3. Two papers on pairwise distances: RKE and DCOR
4. List of references

A few historical remarks

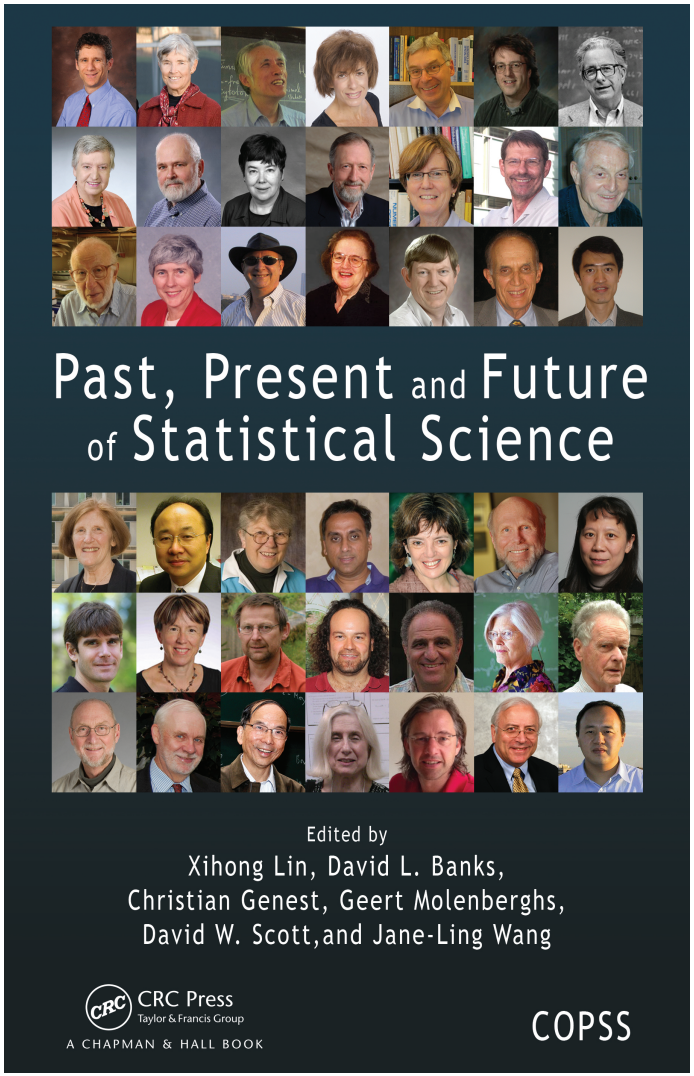
Times have changed:

1. Bell Telephone Labs Summer intern 1955: (my first job) Women earned \$55/week, men \$60
2. Cornell UGrad 1952-56: Two women math majors. Two women in Prof Kac's Advanced Calculus class of about 300. Profs Kac and Rosser were very supportive, however.
3. Operations Research Inc., Silver Spring MD 1959-62. Salary ratio of 2:1 of men/woman(me).
4. IBM 1962-66 Silver Spring MD, San Jose and Palo Alto: Some DC/Silver Spring area restaurants would not serve blacks, including one of my colleagues.
5. Stanford PhD student and Postdoc 1962-67. Few women role models.

6. University of Wisconsin Madison 1967-present. 1967: 12 tenure or tenure track women in a faculty of thousands. 1972: Title IX, Mary Ellen Rudin and Marigold Melli promoted from Lecturer to Full Professor as the nepotism rules went down. No longer did a woman who married an academic have to forget an academic career for herself.

Statistical ModelBuilding, Machine Learning, and the Ah-Ha Moment.

See G. Wahba in this: Book available at `http//:www.copss.org`



Statistical Model Building, Machine Learning and the Ah-Ha Moment

The "Ah-Ha" moment: these are moments when the main idea just popped up instantaneously, sparking sequenses of future research activity- for me they crucially involved discussions and interactions with colleagues and student-colleagues. Sometimes the idea came up quickly, sometimes after mulling over the problem from days to months.

1. Aha! (Kimeldorf and Wahba 1971) We proposed and proved the representer theorem. Coffee hour at the Math Research Center. It essentially turns a problem in an infinite dimensional Hilbert space into a finite dimensional problem. Its only a four line proof and at first we thought it was too trivial to publish, but we wrote it up, submitted it to Numerische Mathematik and it was accepted within three weeks. That never happened again. It was important. See Item 6 and Wikipedia.
2. Eureka! Leaving-out-one (Wahba and Wold 1975)
3. GCV (Golub Heath and Wahba 1979, Craven and Wahba 1979)
4. Randomized Trace and the Degrees of Freedom for Signal (Girard 1987, Hutchinson 1989)
5. SSANOVA ANOVA in RKHS (Grahame Wilkinson visits Madison) (Gu 2002, Wang 2011)

6. Penalized Likelihood and the Support Vector Machine (Lin, Wahba, Zhang and Lee 2002)(Vladimir Vapnik and the Hadley meeting).
7. The Multicategory SVM (Lee, Lin and Wahba 2004)
8. RKE and DCOR - see below

Advice: Treasure your colleagues, have great students.

Pairwise Distance/Dissimilarity Information

Many old and new investigations involve pairwise relationships between subjects or objects, and use of this kind of information is something of a 'hot topic'. We will describe two modern ways to utilize pairwise distances, or, more generally, dissimilarities between subjects/objects.

Example 1. Regularized Kernel Estimation RKE (Lu, Keles, Wright and Wahba 2005)

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:

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

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! Generalizes Multidimensional Scaling

This optimization problem can be solved numerically using modern convex cone software.

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.

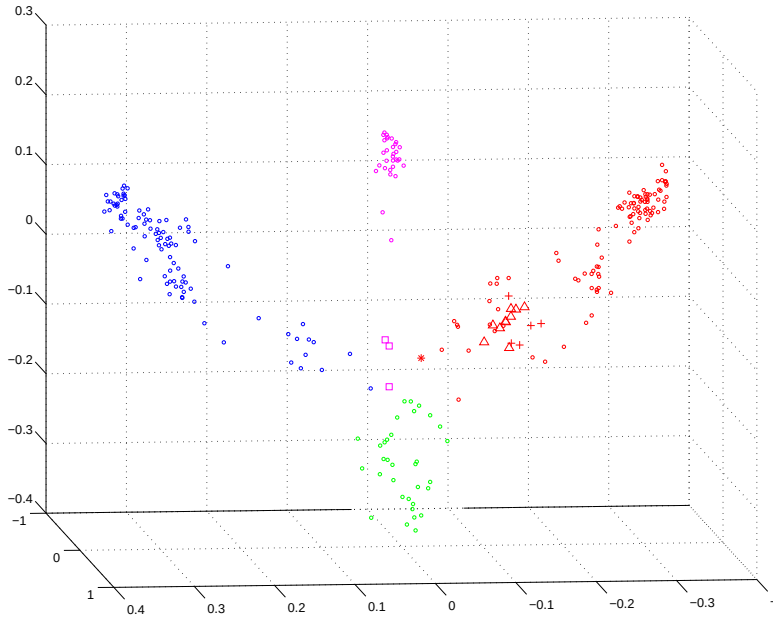


Figure 1: 3D representation of the sequence space for 280 proteins from the globin family. Red: α-globin subfamily, blue: β-globins, purple: myglobin subfamily, and green: a heterogeneous group encompassing proteins from other small subfamilies within the globin family. Note that in this example three,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

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

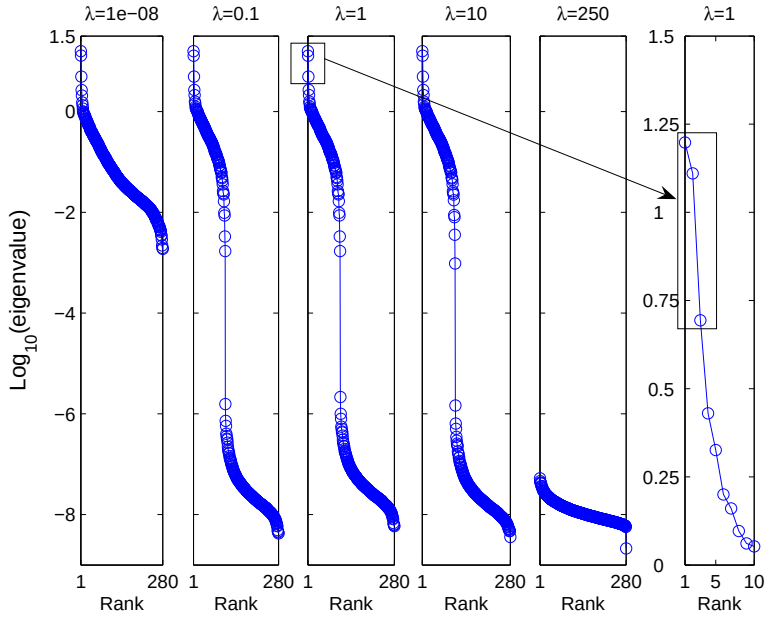


Figure 2: The effect of varying λ on the eigenvalues of R_{dist}. The left five images show log-scale eigensequence plots for five values of λ. As λ increases, smaller eigenvalues begin to shrink. The rightmost image shows the first 10 eigenvalues of the $\lambda = 1$ case displayed on a larger scale.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

RKE Embedding Newbies.

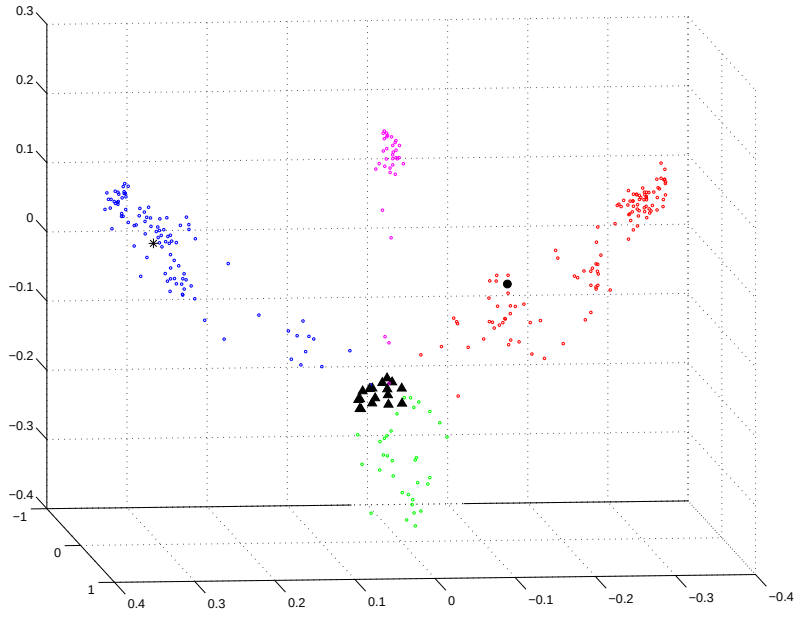


Figure 3: Positioning test globin sequences in the coordinate system of 280 training sequences. The newbie algorithm is used to locate one Hemoglobin zeta chain (black circle), one Hemoglobin theta chain (black star), and seventeen Leghemoglobins (black triagles) into the coordinate system of the training global sequence data.

Example 2. Distance Correlation DCOR (Kong, Klein, Klein, Lee and Wahba 2012)

Does Life Span Run in Families, and If So, Why? The Beaver Dam Eye study (BDES), starting with about 5000 subjects in 1988 from ages 43-84 years, and about 2400 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 compaired pairwise death ages between relatives and between unrelated subjects and it is clear that mortality runs in families. Distance Correlation is used to quantify this.

- What is DCOR?
- Variable Descriptions, the 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

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

where

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

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

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 $f_{X,Y}, f_X$ and f_Y be the characteristic functions of $(X:Y), X$ and Y. (The characteristic function of a distribution F_U is $f_U(t) = \int e^{itu} dF_U$). Let

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

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

Death Age Scoring

Death age as a function of fixed, lifestyle and disease variables will be modeled 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:

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

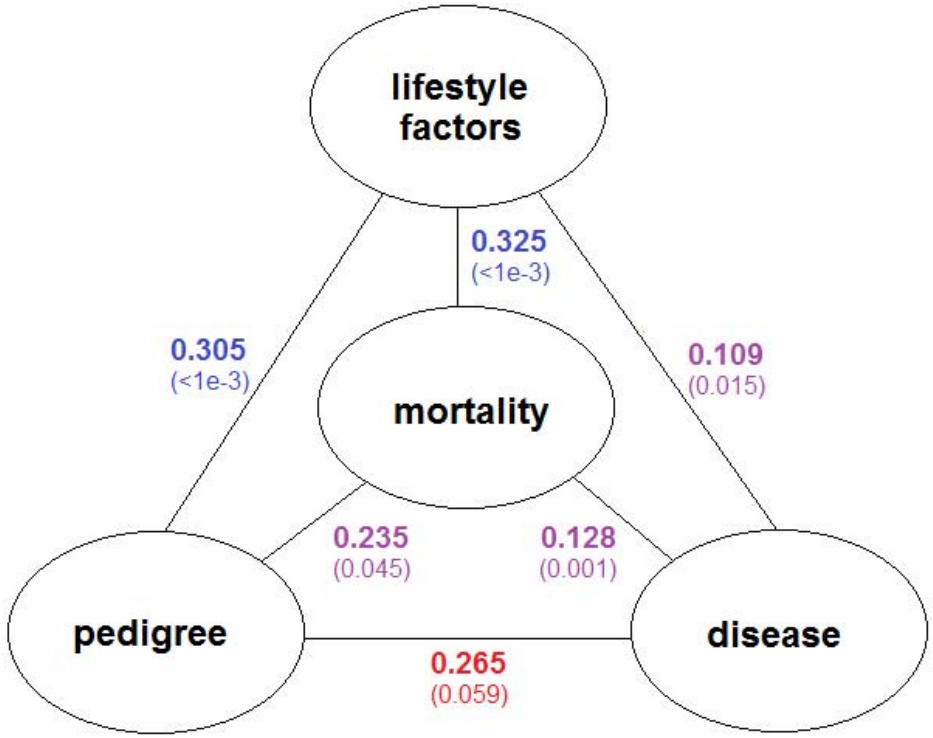
Determining Distance Correlation (DCOR)

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 which shows promise to help in answering these questions in future studies.

Example 3. More DCOR incorporated in SSANOVA
Corrada Bravo, Lee, Klein, Klein, Iyengar and Wahba 2009

Examining the relative influence of familial, genetic and environmental covariate information in flexible risk models. (Not discussed here)

Summary and Conclusions

We have discussed

- Historical Remarks: How things have changed since 1955!
- A few examples of the Ah-Ha moment, when answers popped out suddenly as a result of conversations with faculty colleagues and student colleagues.
- Two papers of the use of pairwise dissimilarity information, are described, a topic of growing importance.

In today's world, women are enjoying the joys and headaches of scientific research of all kinds, including in Statistics, often sharing with partners the solution of the two body problem and the responsibilities of parenting

References

[1] G. Wahba. Statistical model building, machine learning and the ah-ha moment. In X. Lin *et al*, editor, *Past, Present and Future of Statistical Science*, pages xx–toapppear. Committee of Presidents of Statistical Societies, 2013. Available at http://www.stat.wisc.edu/~wahba/ftp1/tr1173.pdf.

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

[3] G. Wahba and S. Wold. A completely automatic French curve. *Commun. Stat.*, 4:1–17, 1975.

[4] G. Golub, M. Heath, and G. Wahba. Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–224, 1979.

[5] P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31:377–403, 1979.

[6] D. Girard. A fast 'Monte Carlo cross validation' procedure for large least squares problems with noisy data. Technical Report RR 687-M, IMAG, Grenoble, France, 1987.

[7] M. Hutchinson. A stochastic estimator for the trace of the influence matrix for Laplacian smoothing splines. *Commun. Statist.-Simula.*, 18:1059–1076, 1989.

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

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

[10] Y. Lin, G. Wahba, H. Zhang, and Y. Lee. Statistical properties and adaptive tuning of support vector machines. *Machine Learning*, 48:115–136, 2002.

[11] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *J. Amer. Statist. Assoc.*, 99:67–81, 2004.

[12] 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.

[13] 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.

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

[15] 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.