DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1183
February 27, 2017

Emanuel Parzen and a Tale of Two Kernels

Prepared for the Journal of Time Series Analysis
Special Issue: Emanuel Parzen Memorial,
P. M. Robinson, D. Politis and S. Lahiri, Editors

Grace Wahba[1]
Department of Statistics, Department of Computer Sciences
and Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

[1]Research supported in part by CPCP and NSF Grant DMS-1308847

Emanuel Parzen and a Tale of Two Kernels

Grace Wahba[1]
Department of Statistics, Department of Computer Sciences
and Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison
February 27, 2017

Abstract

I was Manny Parzen's fifth student, according to the Mathematical Genealogy Project, receiving my PhD under his supervision in 1966, and remaining at Stanford with him for a postdoc, leaving for the University of Wisconsin-Madison in 1967, where I remain as I write. To be Manny's PhD student at Stanford in the 60's was nothing short of bliss. Many was full of ideas, the weather was warm and sunny, sometimes classes were held on the grass in front of the old Sequoia Hall with Reproducing Kernel Hilbert Spaces on the improvised blackboard. A lovely, elegant dinner at their Stanford home that Manny and Carol threw for a group of graduate students bring back fond memories. Manny and Carol remained lifelong friends. Manny launched me on a 50 year academic year, buttressed by what I learned about RKHS from him in those days. In this article I first describe many fond memories over the years, and then present some technical results and thoughts relating to RKHS, Parzen density estimates, Statistical Machine learning and related topics.

1 Scholar, teacher, friend

1.1 Manny as PhD advisor

In 1962 I was a single mom working at a D. C. area think tank and also working towards an MS at the University of Maryland-College Park when I read

[1]Research supported in part by NSF Grant DMS-1308847 and a Consortium of NIH Institutes under Award Number U54AI117924

Mod Prob [19] and Stochastic Processes [24] and imagining an impossible dream of relocating to the West Coast, going to Stanford and having Prof. Emanuel Parzen as a thesis advisor. Well, sometimes dreams do come true. Soon I got a new job with a group at IBM in D. C. and shortly thereafter, they moved the whole group to the Bay area. Voila, admission to the IBM work study program and to Stanford and eventually I became Manny's fifth student. Soon we were meeting regularly and he would enthusiastically listen to my initial attempts at research. I have many fond memories of my five years as a student (1962-66 and postdoc 1967). One of my first memories is an elegant dinner that Manny and Carol threw for a group of students - convincing me that academic life was something to be desired. Carol was always part of things, knowing many of Manny's students and being enthusiastic about their academic lives. Once I got through the first year qualifier (at Stanford they had first and second year qualifiers) I didn't really worry about making it through (sort of). Manny was always encouraging, and he was the most positive, outgoing and optimistic person I had ever met. Another fond memory is a class Manny taught and held on nice days on the grass in front of the old Sequoia hall, later demolished and replaced by a more modern new Sequoia hall in 1998. Manny was one of the major figures in time series analysis and this course reflected his contributions to the field at the time, one example being the fundamental paper [21]. For me, I think it was the first time I heard him talk about Reproducing Kernel Hilbert Spaces (RKHS) [1], although he had published several papers utilizing them around that time, for example [20] [25]. In any case, some of his elegant approaches to RKHS remained dormant in my brain for a few years (more on that later) and I went on to write a dissertation on vector valued time series [35, 36] under his supervision. Manny looked after his students. It turned out that E. J. (Ted) Hannan, in Canberra, Australia was one of Manny's colleagues and was working on something similar to my thesis work. Manny sent him what was to become [36]. Recalling that in the 60's it could take three weeks to get a package to Australia, it happened that Hannan sent Manny what was to become [7] and the manuscripts crossed in the mail - a bit different than instant communication around the world today. I think Manny had written Hannan about my work along with sending the manuscript, and although Hannan's paper ultimately was published several years before mine, he generously mentioned my work in his paper. I received the PhD in June of 1966 and, if memory serves, Manny took me and my Dad, who had come from New Jersey for the graduation, to lunch at the faculty club. I went on

to spend a year as a postdoc with Manny. During that year he apparently contacted a number of his friends, including George Box at Madison, resulting in a large number of invitations to give a lecture, and, ultimately eight job offers. The late 60's were a good time to be looking for an academic job, as universities were growing to accommodate the children of the veterans returning from the Second World War. The process was much simpler, too - Manny made a bunch of phone calls, I gave some talks, and eventually I got a letter with a short paragraph saying something like "We would like to offer you a position as an assistant professor with the academic year salary of (say) $10,000. Please let us know by such-and-such a date whether you accept". Today the successful applicant will get a large packet with enough rules and regulations to keep busy reading them for a week. Not to mention the application process whereby the potential hire is usually responding to a job posting, enters a large number of documents into an on line website, while the applicant's references have to enter detailed information into another website. In September of 1967 I left sunny California for the frozen winters of the Midwest, the University of Wisconsin-Madison, where there was a powerful numerical analysis group and a fertile place to nurture the seeds of function estimation in RKHS.

1.2 Manny's influence – density estimation

Manny had written a fundamental paper on density estimation [23] (more on this in the technical section below) and being aware of this work led me to write a bunch of papers on density estimation and spectral density estimation, including [37] [38] [39] [43]. In the early seventies there was a lot of discussion about tuning nonparametric models of various kinds. Manny contributed to this issue in the context of time series, in his CATS tuning criteria [27], and in his influence on those around him, for example, the last three papers above.

1.3 Manny's influence – RKHS

In 1967, when I arrived in Madison, there was a large group of staff and visitors working excitedly in numerical analysis and approximation theory. They were members of the Mathematics Research Center which was located in Stirling Hall, the building that was later blown up in August of 1970 in protest against the Vietnam war. I had a part time appointment there along with

my position in the Statistics Department. Leading researchers in approximation theory and numerical analysis were there, including I. J. Schoenberg, Carl deBoor, Larry Schumaker, Zuhair Nashed and others. There was much interest in splines, the first of which were invented by Schoenberg in the forties. Tea was served mid morning and mid afternoon accompanied by lively discussion.

In the midst of this creativity, that brain space holding memories of RKHS from Manny's class perked up, and George Kimmeldorf, who was a visitor to the MRC at the time, and I together realized that we could derive Schoenberg's polynomial smoothing spline as an optimization problem in an RKHS, and moreover the abstract structure for doing that was highly generalizable. We went on to produce three papers together about RKHS [8] [9] [10], this last paper giving a closed form expression for the solution to the penalized likelihood optimization problem, where the penalty is a square norm or seminorm in an RKHS - the representer theorem. It was accepted within three weeks, something that I never experienced again. RKHS methods seemed to occupy a small niche until around 1996 when it became widely known that the Support Vector Machine (SVM), much appreciated by computer scientists for its classification prowess, could be obtained as the solution to an optimization problem in an RKHS. More on this story can be found in [41] pp486-495. Lin *et al* [16] showed that the SVM was estimating the sign of the log odds ratio, and copious applications of RKHS methods are now part of computer scientists' and statisticians' toolkits.

2 Trips

In 1984, Hirotugu Akaike threw a fun conference in Tokyo focused mostly around a group that was interested in time series and other common interests of Akaike and Manny. There were exciting sightseeing trips and social events almost every evening. I'm sure things are different now, but the gracious ladies of the office staff served us tea and little cakes on conference days. One evening when I surprisingly didn't see anything on the schedule, one of Akaike's younger (and presumably single) colleagues asked me if he could take me out to dinner, and I was quite charmed. I didn't stop to wonder where everyone else was, but some years later I decided that it must have been planned to allow the men to visit some place that didn't expect women e. g. a geisha teahouse, but I'll never know.

Figure 1 is a picture from the Tokyo trip.

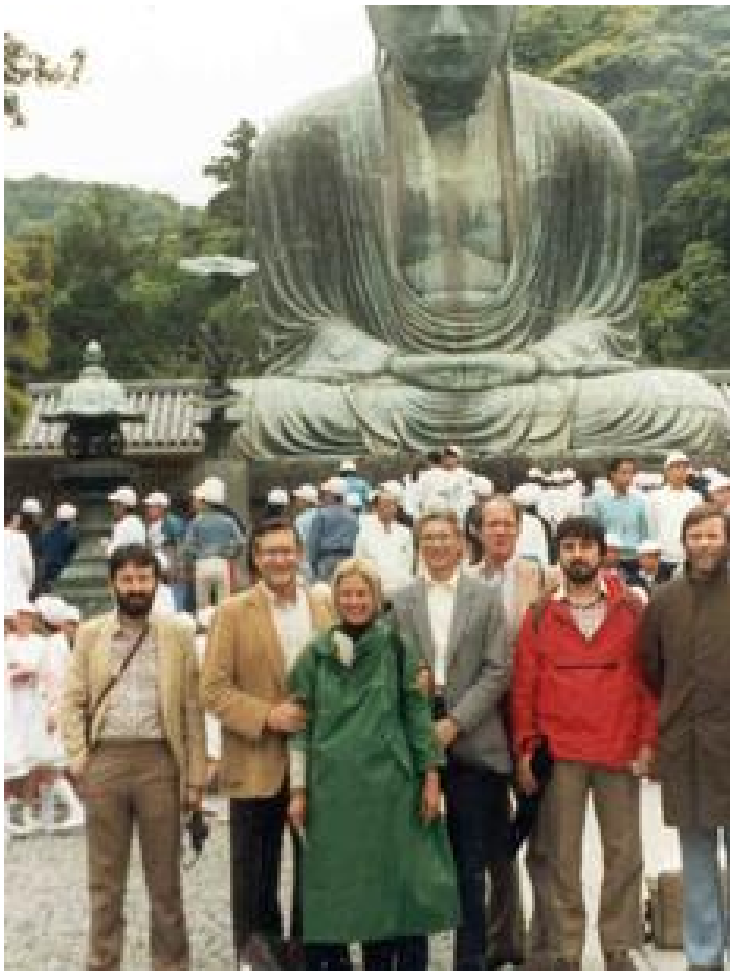

Figure 1: Akaike Time Series Conference, Tokyo 1984. l. to r. Victor Solo, Manny, me, Wayne Fuller, Bill Cleveland, Bob Shumway, David Brillinger

In 1989 there was a swell sixtieth birthday party for Manny, including scientific colleagues, Texas A&M bigwigs and Parzen family. Everyone had a ball, and Figure 2 is a scene from the party - there is Manny outgoing and smiling as ever.

Figure 2: Manny's 60th Birthday, 1989, College Station, TX. l. to r. Don Ylvisaker, me, Joe Newton, Marcello Pagano, Randy Eubank, Manny, Will Alexander, Marvin Zelen, Scott Grimshaw

Marvin Zelen and I attended the JSM2005 Gottfried Noether Scholars Award and are congratulating the winners, Manny and Gerda Claeskins, in Figure 3.

Figure 3: At the Gottfried Noether Senior and Junior Researchers Awards Ceremony, JSM 2005, to Manny Parzen and Gerda Claeskens. l. to r. me, Manny, Gerda, Marvin Zelen

Manny was the featured speaker at the The Pfizer Colloquium 2006 at UConn, with Joe Newton and myself as discussants. Joe and I sat for a "Conversation with Many Parzen", (see Figure 4) which was videotaped, and the main thing I remember about that was the fact that the video was recording off the cuff remarks and I was afraid of making a dumb one. Manny is smiling as usual but I look a bit tense.

Figure 4: Manny, me, Joe Newton, Nitis Mukhopadhyay at the Pfizer Colloquium 2006 in Manny's honor at UConn.

3 Manny, a man of many interests

Manny had a major role in a number of fundamental areas in the development of the Statistical Canon. Aside from density estimation and RKHS, these include time series modeling, spectral density estimation, and in later years, quantile estimation. However, in this chapter we will limit ourselves to Parzen window density estimation and RKHS, the two of Manny's areas I have worked in. Interestingly Manny's work is fundamental to the two different kinds of kernels that have played important roles in the development of modern statistical methodology. Kernels in Parzen window density estimation (to be called density kernels) are typically non-negative symmetric functions integrating to 1 and satisfying some conditions, while kernels in RKHS are positive definite functions, which are not necessarily positive. There are, of course, kernels that are both. We will briefly review both, enough to review some modern results in two emerging fields, density embedding and distance correlation. Density embedding begins with an RKHS and a sample from a density of interest and results in a class of density estimates which include Parzen window estimates. These estimates are elements of the RKHS *so one has a metric for determining pairwise distances between densities, namely the RKHS norm.* This enlarges the class of familiar distance measures between densities (e. g. Hellinger distance, Bhattacharyya distance, Wasserstein distance, etc.) Given pairwise distances between densities, we then describe how these pairwise distances are used to include sample densities *as attributes* in statistical learning models such as Smoothing Spline ANOVA (SS-ANOVA) models, which include penalized likelihood methods and (nonparametric) SVM's. Thus Manny's foundational work in two seemingly diverse areas come together to add another feature to the statistician's tool kit.

3.1 Two kinds of kernels

We now discuss the two kinds of kernels, those used in density estimation, and those that characterize an RKHS. Our goal is to show how sample densities possessed by subjects in RKHS-based prediction models can be treated *as attributes* in these models.

3.2 Parzen density kernels

Let $X_1, X_2, \ldots, X_n$ be a random sample from some (univariate) density $f(x), x \in [-\infty, \infty]$. The kernel density estimates of Manny's seminal 1962 paper [23] (paraphrasing slightly) are of the form

$$f_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right), \tag{1}$$

where $K(y)$ is non-negative,

$$\begin{aligned} \sup_{-\infty<y<\infty} K(y) &< \infty, \\ \int_{-\infty}^{\infty} K(y) &= 1, \\ \lim_{y\to\infty} |yK(y)| &= 0, \end{aligned} \tag{2}$$

and, letting $h = h(n)$,

$$\lim_{n\to\infty} h(n) = 0. \tag{3}$$

This landmark paper explores in detail the properties of these density estimates, and gives a table of a number of K that satisfy the requirements. Looking at the table of the K and their Fourier transforms reveals that several but not all are also positive definite.

3.3 RKHS kernels

Manny was likely the first statistician to seriously introduce RKHSs to statisticians, certainly highly influential, see [22, 25, 26]. As a graduate student and postdoc at Stanford from 1962-1967 I learned about RKHS directly from Manny's lectures. Later the rich and beautiful results in [26] were highly influential in my own life when work on splines at Madison rang a bell that splines were a prototype of a vast class of nonparametric modeling problems that could be solved by RKHS methods, see [10].

Let $\mathcal{H}_K$ be an RKHS of functions on a domain $\mathcal{T}$. Then there exist a unique positive definite function $K(s,t), s,t \in \mathcal{T}$ associated with $\mathcal{H}_K$. Conversely, let $\mathcal{T}$ be a domain on which a positive definite kernel function, $K(s,t), s,t \in \mathcal{T}$ can be defined. Then there exists a unique RKHS $\mathcal{H}_K$ with K as its reproducing kernel. This means the following: Let $K_s(t) \equiv K(s,t)$

be considered as a function of t for each fixed s. Then, letting $< \cdot, \cdot >$ be the inner product in $\mathcal{H}_K$, for $f \in \mathcal{H}_K$ we have $< f, K_s >= f(s)$, and $< K_s, K_t >= K(s,t)$. The square distance between f and g is denoted as $||f-g||^2_{\mathcal{H}_K}$, where $||\cdot||^2_{\mathcal{H}_K}$ is the square norm in $\mathcal{H}_K$. As a special case, if $s, t \in \mathcal{T}$, then the squared distance between s and t can be taken as $||K_s - K_t||^2_{\mathcal{H}_K} = K(s,s) - 2K(s,t) + K(t,t)$. We will be using the fact that K encodes pairwise distances. We note that tensor sums and products of positive definite functions are positive definite functions and have associated RKHS as tensor sums and products of the corresponding component RKHS, see [1] and the references cited below for examples.

3.4 Smoothing Spline ANOVA models

Basic references for SS-ANOVA models are [5] and [44], both describe software in the R collection. Numerous applications include [3, 15, 42]

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

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

With each domain we have a positive definite function and associated RKHS. Now let $\mathcal{H}_K$ be the tensor product of the d RKHSs. Its RK is then the tensor product of the component RK's. With some conditions, including that the constant function is in each component space and there is an averaging operator in which the constant function averages to 1, then for $f \in \mathcal{H}_K$ an ANOVA decomposition of f of the form

$$f(t_1, \cdots, t_d) = \mu + \sum_\alpha f_\alpha(t_\alpha) + \sum_{\alpha\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots \tag{4}$$

can always be defined. Then a regularized kernel estimate is the solution to the problem

$$\min_{f \in \mathcal{H}_K} \sum_{i=1}^{n} \mathcal{C}(y, f) + \lambda J(f). \tag{5}$$

where $\mathcal{C}(y, f)$ relates to fit to predict y from f, for example a Gaussian or Bernoulli log likelihood, or a hinge function (Support Vector Machine), and $J(f)$ is a square norm or seminorm in $\mathcal{H}_K$ Given this model for f (generally truncated as warranted) this provides a method for combining heterogenous

domains (attributes) in a regularized prediction model. Note that nothing has been assumed about the domains, other than that a positive definite function can be defined on them. We sketch an outline of facts relating to SS-ANOVA models, partly to set up notation to facilitate our goal of demonstrating how sample densities may be treated as *attributes* in conjunction with SS-ANOVA models.

The choice of kernel class for each variable may be an issue in practice and may be specific to the particular issue and data at hand. Once the kernel form has been chosen, the tuning parameter λ in Equation (4) along with other tuning parameters hidden in $J(f)$ must be chosen and can be important. We are omitting any discussion of these issues here, but applications papers referenced below discuss choice of tuning parameters.

Note that we use the same symbol K for density kernels, positive definite functions and positive definite matrices.

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

$$(\mathcal{E}_\alpha f)(t) = \int_{\mathcal{T}^{(\alpha)}} f(t_1, \ldots, t_d) d\mu_\alpha(t_\alpha). \tag{6}$$

Then the identity operator 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

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

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

Further details in the RKHS context may be found in [6, 40, 42] The idea behind SS-ANOVA is to construct an RKHS $\mathcal{H}$ of functions on $\mathcal{T}$ as the tensor product of RKHS on each $\mathcal{T}^{(\alpha)}$ that admit an ANOVA decomposition. Let $\mathcal{H}^{(\alpha)}$ be an RKHS of functions on $\mathcal{T}^{(\alpha)}$ with $\int_{\mathcal{T}^{(\alpha)}} f_\alpha(t_\alpha) d\mu_\alpha(t_\alpha) = 0$ and let

$[1^{(\alpha)}]$ be the one dimensional space of constant functions on $\mathcal{T}^{(\alpha)}$. Construct the RKHS $\mathcal{H}$ as

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

$$= [1] \oplus \sum_{\alpha} \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha<\beta} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \cdots, \tag{7}$$

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.

3.5 Pairwise distances in data analysis

3.5.1 Regularized Kernel Estimation

Interesting examples of pairwise distances occur in, for example, blast scores [17] which give a pairwise dissimilarity between pairs of protein sequences. The blast score pairwise dissimilarities are not a real distance, but they can be embedded (approximately) in a Euclidean space using Regularized Kernel Estimation (RKE) [17].

For a given $n \times n$ dimensional positive definite matrix K, the pairwise distance that it induces is $\hat{d}_{ij} = K(i,i) + K(j,j) - 2K(i,j) = B_{ij} \cdot K$, where $K(i,j)$ is the (i,j) entry of K and B_{ij} is a symmetric $n \times n$ matrix with all elements 0 except $B_{ij}(i,i) = B_{ij}(j,j) = 1$, $B_{ij}(i,j) = B_{ij}(j,i) = -1$. The RKE problem is as follows: Given observed data d_{ij} find K to

$$\min_{K \succeq 0} \sum_{(i,j)\in\Omega} |d_{ij} - B_{ij} \cdot K| + \lambda \operatorname{trace}(K). \tag{8}$$

Ω may be all pairs, or a connected subset.

The data may be noisy/not Euclidean, but the RKE provides a (non-unique) embedding of the n objects into an r- dimensional Euclidean space as follows: Let the spectral decomposition of K be $\Gamma\Lambda\Gamma^T$. The largest r eigenvalues and eigenvectors of K are retained to give the $n \times r$ matrix $Z = \Gamma_r \Lambda_r^{1/2}$. We let the ith row of Z, an element of R^r, be the pseudo-attribute of the ith subject.

Thus each subject may be identified with an r-dimensional pseudo attribute, where the pairwise distances betwen the pseudo attributes respect

(approximately, depending on r) the original pairwise distances. Even if the original pairwise distances may be Euclidean, the RKE may be used as a dimension reduction procedure where the original pairwise distances have been obtained in a much larger space (e. g. an infinite dimensional RKHS). The rank r may be chosen to retain, say, 95% of the trace, by examining an eigensequence plot for a sharp drop off, or maximizing the predictability in a supervised learning model. Note that if used in a predictive model it is necessary to know how a "newbie" fits in; this is discussed in [17].

In the blast scores example four well separated clusters of known proteins were readily evident in a three dimensional in-depth plot of the pseudo attributes, and it could be seen that the multicategory support vector machine [14] would have classified the clusters nearly perfectly from the these rank 3 pseudo attributes. Note that this embedding is only unique up to a rotation, because rotating the data set does not change the pairwise distance. Therefore in fitting nonparametric models on the embedded data only radial basis function (rbf) kernels may be used, since they depend only on pairwise distances.

Corrada Bravo *et al* [2] built a risk factor model consisting of an SS-ANOVA model with two genetic variables, life style attributes and an additive term involving pairwise distances of subjects in pedigrees. The pedigree pairwise distances were mapped into Euclidean space using RKE, and the Euclidean space of the resulting pseudo attributes used as the domain of an rbf based RKHS. The results were used to examine the relative importance of genetic, lifestyle, and pedigree information. It can be seen that this RKHS is not treated as other terms in the SS-ANOVA model, as there are no constant functions in the rbf based RKHS.

Below we will see how sample densities can be embedded in an RKHS, and pairwise distances and pseudo attributes obtained. Then the sample densities may be used in an SS-ANOVA model in the same way as in Corrada Bravo *et al*.

3.5.2 Pairwise distances reprised

So, pairwise distances, either noisy or exact, may be included in information that can be built into learning models. Applications of RK's in a variety of domains such as texts, images, strings and gene sequences, dynamical systems, graphs and structured objects of various kinds have been defined. Recent examples include [11] [29]. We now proceed to examine pairwise

distances for sample densities.

3.6 Pairwise distances and kernel embedding for densities

Many definitions of pairwise distance between densities have appeared in the literature, in the context of testing for equality, including Wasserstein distance, Bhattacharyya distance, Hellinger distance, Mahalanobis distance, among others.

Smola *et al* [30] proposed to embed distributions into an RKHS, and, once this is done, pairwise distances between a pair of distributions can be taken as the RKHS norm of the difference between the two embedded distributions.

Let $\mathcal{H}_K$ be an RKHS of functions on $\mathcal{T}$ with RK $K(s,t), s,t \in \mathcal{T}$. Let $X_1, X_2, \cdots, X_k$ be an iid sample from some density p_X. A map from this sample to H_K is given by

$$f_X(\cdot) = \frac{1}{k}\sum_{j=1}^{k} K(X_j, \cdot). \tag{9}$$

Given a sample from a possibly different distribution, we have

$$g_Y(\cdot) = \frac{1}{\ell}\sum_{j=1}^{\ell} K(Y_j, \cdot) \tag{10}$$

It is required that K be universal, among other things, [28, 31], which guarantees that two different distributions will be mapped into two different elements of $\mathcal{H}_K$. See also p. 727 of [4].

The pairwise distances between these two samples can be taken as $\|f_X - g_Y\|$, see [31], where

$$\|f_X - g_Y\|_{H_k} = \frac{1}{k^2}\sum_{i,j=1}^{k} K(X_i, X_j) + \frac{1}{\ell^2}\sum_{i,j=1}^{\ell} K(Y_i, Y_j) - \frac{2}{kl}\sum_{i=1,j=1}^{k,\ell} K(X_i, Y_j) \tag{11}$$

thus providing a distance measure for each universal kernel to the other pairwise distances already noted. Note that if K is a nonnegative, bounded radial basis function, then (up to scaling) we have mapped f_X and g_Y into Parzen type density estimates (!). The univariate version of a Gaussian rbf appears in Table 1 of Parzen [23].

Zhou *et al* [45] used pairwise embedding to consider samples from two different data sources. They only observed transformed versions $h(X_j), j = 1, 2, \ldots, k$ and $g(Y_j), j = 1, 2 \ldots, \ell$ for some known function class containing $h(\cdot)$ and $g(\cdot)$. The goal was to perform a statistical test whether the two sources are the same while removing the distortions induced by the transformations.

We already noted how Corrada Bravo *et al* [2] used pairwise distances between pedigrees to include pedigree information as an additive term in an SS-ANOVA model. Now, suppose we have a study where subjects have various attributes, including a sample density for each. One such example can be seen in [18]. Now that we now have pairwise distances between pairs of the sample densities, the densities can be included in an SS-ANOVA model as an additive term, using the same approach as in [2].

3.7 Is density correlated with other variables?

Distance Correlation (DCOR) [32] is key to an important area of recent research that uses pairwise distances only, to estimate a correlation-like quantity which behaves much like the Pearson correlation in the case of Gaussian variables, but provides a fully nonparametric test of independence of two random variables. See [32, 34]. Recent contributions in the area include [33].

For a random sample $(X, Y) = \{(X_k, Y_k) : k = 1, ..., n\}$ of n iid 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

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

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

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

Distribution of the sample distance correlation under the null hypothesis of independence is easily found by scrambling the data.

Kong *et al* [12] used DCOR and SS-ANOVA to assess associations of familial relations and lifestyle factors, diseases and mortality, by examining the strength of the rejection of the null hypothesis of independence. Later, [13] used distance covariance as a greedy variable selector for learning a model with an extremely large number of candidate genetic variables.

3.8 Including densities as attributes in an SS-ANOVA model

Suppose you have a population, each member having a (personal) sample density and several other attributes, and you find using DCOR that the individual sample densities are correlated with another variable in the model. The way to think about this is, when densities are close, so is the other variable, and vice versa. Interacting terms in the SS-ANOVA model which include an rbf for the density RKHS can be included: As in [2], the densities are to be embedded in some (generally infinite dimensional) rbf based RKHS, and pairwise distances in this RKHS are determined. RKE is then used to obtain pseudo attributes, which are r dimensional vectors, and a second rbf based RKHS is chosen to model functions of the pseudo attributes. The dimension r of the pseudo attributes can be controlled by the tuning parameter in the RKE. As noted earlier, the rbfs over R^r do not in general contain a constant function, so they are treated a little differently than the function spaces in the SS-ANOVA model that do. However, tensor product spaces consisting of the density RKHS $\mathcal{H}^{(dens)}$ and other RKHS in the SS-ANOVA model after they have been stripped of their constant functions may clearly

be added to the model — for example, suppose the density variable is correlated with the α variable, then $[\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(dens)}]$ can be added to the model in Equation (7), and similarly for higher order interactions.

3.9 We have come full circle

So, we have come full circle. Manny proposed and investigated the properties of of Parzen kernel density estimates. Then Manny initiated an investigation into the various properties and importance of RKHS in new statistical methodology, and inspired me and many others to study these wonderful objects. So now we are able to include kernel density estimates as attributes in SS-ANOVA models based on RKHS, a modeling approach whose foundation lies in two of Manny's major contributions to Statistical Science: density estimates, and Reproducing Kernel Hilbert Spaces !

4 Summary

In summary, I have been blessed to be one of Manny's students and lifelong friends, and inspired by his path breaking work. He is terribly missed.

References

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

[2] H. C. Bravo, K. Lee, B. E. K. Klein, R. Klein, S. 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(20):8128–8133, 2009.

[3] 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.

[4] A. Gretton, K. Borgwardt, M. Rasch, B. Scholkopf, and A. Smola. A kernel two-sample test. *J. Machine Learning Research*, 13:723–773, 2012.

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

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

[7] E. Hannan. The estimation of a lagged regression relation. *Biometrika*, 54:409–418, 1967.

[8] G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41:495–502, 1970.

[9] G. Kimeldorf and G. Wahba. Spline functions and stochastic processes. *Sankya Ser. A*, 32, Part 2:173–180, 1970b.

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

[11] R. Kondor and H. Pan. The multiscale Laplacian graph kernel. In *NIPS Proceedings 2016*. Neural Information Processing Society, 2016.

[12] 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.

[13] J. Kong, S. Wang, and G. Wahba. Using distance covariance for improved variable selection with application to learning genetic risk models. *Statistics in Medicine*, 34:1708–1720, 2015. PMCID: PMC 4441212.

[14] 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.

[15] X. Lin, G. Wahba, D. Xiang, F. Gao, R. Klein, and B. Klein. Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist.*, 28:1570–1600, 2000.

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

[17] 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.

[18] K. Nazapour, A. Al-Timemy, G. Bugmann, and A. Jackson. A note on the probability distribution function of the surface electromyogram signal. *Brain Res. Bull.*, 90:88–91, 2013.

[19] E. Parzen. *Modern Probability Theory and its Applications*. Wiley, 1960.

[20] E. Parzen. Regression analysis of continuous parameter time series. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pages 469–489, Berkeley, California, 1960. University of California Press.

[21] E. Parzen. An approach to time series analysis. *Ann. Math. Statist.*, 32:951–989, 1961.

[22] E. Parzen. Extraction and detection problems and Reproducing Kernel Hilbert Spaces. *J. SIAM Series A Control*, 1:35–62, 1962.

[23] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33:1065–1076, 1962.

[24] E. Parzen. *Stochastic Processes*. Holden-Day, San Francisco, 1962.

[25] E. Parzen. Probability density functionals and reproducing kernel Hilbert spaces. In M. Rosenblatt, editor, *Proceedings of the Symposium on Time Series Analysis*, pages 155–169. Wiley, 1963.

[26] E. Parzen. Statistical inference on time series by RKHS methods. In R. Pyke, editor, *Proceedings 12th Biennial Seminar*, pages 1–37, Montreal, 1970. Canadian Mathematical Congress.

[27] E. Parzen. Some recent advances in time series modeling. *IEEE Trans. Automatic Control*, AC-19:723–730, 1974.

[28] D. Sejdinovic, A. Gretton, B. Sriperumbudur, and K. Fukumizu. Hypothesis testing using pairwise distances and associated kernels. arXiv:1205.0411v2, 2012.

[29] 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 951–984, 2014.

[30] A. Smola, A. Gretton, L. Song, and B. Scholkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory, 18th International Conference*, pages 13–31. Springer Lecture Notes in Artificial Intelligence, 2007.

[31] B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *J. Machine Learning Research*, 12:2389–2410, 2011.

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

[33] G. Szekely and M. Rizzo. Partial distance correlation with methods for dissimilarities. *Ann. Statist.*, 42:2382–2412, 2014.

[34] G. Szekely, M. Rizzo, and N Bakirov. Measuring and testing independence by correlation of distances. *Ann. Statist.*, 35:2769–2794, 2007.

[35] G. Wahba. On the distribution of some statistics useful in the analysis of jointly stationary time series. *Ann. Math. Statist.*, 39:1849–1862, 1968.

[36] G. Wahba. Estimation of the coefficients in a multi-dimensional distributed lag model. *Econometrica*, 37:398–407, 1969.

[37] G. Wahba. Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation. *Ann. Statist.*, 3:15–29, 1975.

[38] G. Wahba. Optimal smoothing of density estimates. In J. VanRyzin, editor, *Classification and Clustering*, pages 423–458. Academic Press, 1977b.

[39] G. Wahba. Automatic smoothing of the log periodogram. *J. Amer. Statist. Assoc.*, 75:122–132, 1980c.

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

[41] 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 481–495. Committee of Presidents of Statistical Societies, 2013. Available at http://www.stat.wisc.edu/˜wahba/ftp1/tr1173.pdf.

[42] 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.

[43] G. Wahba and S. Wold. Periodic splines for spectral density estimation: The use of cross-validation for determining the degree of smoothing. *Commun. Statist.*, 2:125–141, 1975.

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

[45] H. Zhou, S. Ravi, V. Ithapu, S. Johnson, G. Wahba, and V. Singh. Hypothesis testing in unsupervised domain adaptation with applications in Alzheimer's disease. In *NIPS Proceedings 2016*. Neural Information Processing Society, 2016.