Manny Parzen, Cherished Teacher
And a Tale of Two Kernels

Grace Wahba

From "Emanuel Parzen and a Tale of Two Kernels"

Memorial Session for Manny Parzen

Joint Statistical Meetings

Baltimore, MD August 2, 2017

Links to these slides in my website

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

Abstract

According to the Mathematics Genealogy project I was Manny Parzen's fifth student (PhD 1966, Postdoc 1967). Manny was a truly wonderful advisor and mentor and we remained friends for fifty years. Seeing Manny and Carol at JSM and numerous other meetings over the years was always a time of happy reunion. There are many fond memories. I learned about Reproducing Kernel Hilbert Spaces as a student in one of his classes that occasionally met on the lawn in front of the old Sequoia Hall at Stanford in the 60's. In fact I knew Manny by his books Modern Probability and Stochastic Processes before I arrived at Stanford. And his classic paper " Statistical inference on time series by RKHS methods" served as a font of ideas as I embarked on an academic career. Manny was one of the greats and we have lost a beloved friend and colleague.

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

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 his work on kernel density estimation and Reproducing Kernel Hilbert Spaces work in the early 60's, these include time series modeling, spectral density estimation, quantile estimation and others. Manny's work involving these two different kinds of kernels that have played important roles in the development of modern statistical methodology. Thus it might be appropriate to take a short glimpse at some modern ideas related to these two kinds of kernels.

Consider a biostatistical training set where several attributes are observed for each subject, including a personal sample density. We allow the posibility of treating an image which registers intensity as a rescaled ‘density’.

We show a sequence of steps in which densities as attributes could be included in predictive models such as Smoothing Spline ANOVA models, which have main effects, two factor interactions, and so forth.

Outline:

1. Parzen Density Kernels and Reproducing Kernel Hilbert Space Kernels.
2. Step 1: Embed densities in an RKHS to obtain pairwise distances.
3. Step 2: Use Regularized Kernel Estimation to map densities into E^r to get pseudo-attributes.
4. Step 3: Use Radial Basis Function kernels to include the pseudo-attributes of densities/images(?) in SSANOVA Models.

Parzen Density Kernels and RKHS Kernels

Manny was a pioneer in both the theory and practice of density estimation and of RKHS.

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 [Parzen, 1962b] (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, $\sup_{-\infty<y<\infty} K(y) < \infty$ $\int_{-\infty}^{\infty} K(y) = 1$, $\lim_{y\to\infty} |yK(y)| = 0$, and, letting $h = h(n)$, $\lim_{n\to\infty} h(n) = 0$.

This seminal 1962 paper explores in detail the properties of these density estimates. Today we consider multivariate densities/images.

RKHS Kernels

Manny was likely the first statistician to seriously introduce RKHSs to statisticians, certainly highly influential, see [Parzen, 1962a, Parzen, 1963, Parzen, 1970].

- $\mathcal{H}_K$ be an RKHS of functions on a domain $\mathcal{T}$. $\exists$ 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 function, $K(s,t), s,t \in \mathcal{T}$ is defined. $\exists$ a unique RKHS $\mathcal{H}_K$ with K as its reproducing kernel.
- Consider $K_s(t) \equiv K(s,t)$ 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 $\mathcal{H}_K$.

Step 1: Embedding densities in an RKHS

Population case: Let $p(t)$, be a density on some domain $\mathcal{T}$, and let $\mathcal{H}_K$ be an RKHS with kernel $K(\cdot,\cdot)$. Then the embedding of p into $\mathcal{H}_K$ is given by

$$f(\cdot) = \int_{t \in \mathcal{T}} K(\cdot, t) p(t) dt. \tag{2}$$

Here $f \in \mathcal{H}_K$. The sample version of f is given by

$$f_X(\cdot) = \frac{1}{k} \sum_{j=1}^{k} K(X_j, \cdot) \tag{3}$$

where $X_1, \ldots, X_k$ are k iid samples from p. If we were treating p as an image of, say, an x-ray density, then the X_j would be on some regular or otherwise designed grid.

Given a sample from a possibly different distribution q say, we have

$$g_Y(\cdot) = \frac{1}{\ell} \sum_{j=1}^{\ell} K(Y_j, \cdot). \tag{4}$$

Under appropriate conditions on K [Sejdinovic et al., 2012, Sriperumbudur et al., 2011], two different distributions will be mapped into two different elements of $\mathcal{H}_K$. See also p. 727 of [Gretton et al., 2012]. The pairwise distances between these two samples can be taken as

$$\|f_X - g_Y\|_{H_k}^2 = \frac{1}{k^2} \sum_{i,j=1}^{k} K(X_i, X_j) + \frac{1}{\ell^2} \sum_{i,j=1}^{\ell} K(Y_i, Y_j) - \frac{2}{kl} \sum_{i=1,j=1}^{k,\ell} K(X_i, Y_j). \tag{5}$$

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 (!).

Step 2: Using RKE to map densities in E^r. Given the pairwise distances from Step 1 embed the densities in a low dimensional Euclidean space by by using Regularized Kernel Estimation (RKE) [Lu et al., 2005] and then use the results in an SS-ANOVA model.

For a given $n \times n$ dimensional positive definite matrix Σ, the pairwise distance that it induces is $\hat{d}_{ij} = \Sigma(i,i) + \Sigma(j,j) - 2\Sigma(i,j)$

The RKE problem is as follows: Given observed data d_{ij} find Σ to

$$\min_{\Sigma \succeq 0} \sum_{(i,j) \in \Omega} |d_{ij} - \hat{d_{ij}}| + \lambda \, \text{trace}(\Sigma) \tag{6}$$

where $\hat{d}_{ij} = \Sigma(i,i) + \Sigma(j,j) - 2\Sigma(i,j)$.

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 (determined by λ) as follows: Let the spectral decomposition of Σ be $\Gamma\Lambda\Gamma^T$. The largest r eigenvalues and eigenvectors of Σ 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 E^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). Note that if used in a predictive model it is necessary to know how a "newbie" fits in; this is discussed in [Lu et al., 2005].

Step 3: SSANOVA models with densities as attributes, using Radial Basis Function Kernels. Briefly, Smoothing Spline ANOVA models of functions of d variables are of the form

$$f(t_1, \ldots, t_d) = \mu + \sum_{\alpha} f_\alpha(t_\alpha) + \sum_{\alpha\beta} f_{\alpha\beta}(t_\alpha, t_\beta) + \cdots \tag{7}$$

and the terms satisfy ANOVA-like side conditions.

f is assumed to be in a tensor product space

$$\mathcal{H} = \Pi_{\alpha=1}^{d} \otimes \mathcal{H}_\alpha.$$

Each $\mathcal{H}_\alpha$ is an RKHS of functions on $\mathcal{T}_\alpha$ that admits a decomposition of the form

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

with an averaging operator $\mathcal{E}_\alpha$ such that $\mathcal{E}_\alpha 1^{(\alpha)} = 1$ and $\mathcal{E}_\alpha f_\alpha = 0$ for $f_\alpha \in \mathcal{H}^{(\alpha)}$.

Expanding $\mathcal{H}$ gives

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

$$= [1] \oplus \sum_{\alpha} \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha<\beta} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \cdots , \qquad (8)$$

where [1] denotes the constant functions on $\mathcal{T} = \Pi_{\alpha=1}^{d} \otimes \mathcal{T}_{\alpha}$. Then $f_{\alpha} \in \mathcal{H}^{(\alpha)}, f_{\alpha\beta} \in [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}]$ and so forth. Extensive literature and software exists for fitting these models, examples include [Gu, 2002, Wang, 2011, Wahba et al., 1995].

To use the pseudo-attributes in E^r found via RKE in an RKHS we must confine ourselves to radial basis function kernels (RBF's), which depend only on pairwise distances between the arguments: thus $K(s,t) = k(\|s-t\|)$. Let $\mathcal{H}^{(\alpha)}$ be the RKHS associated with $k(\cdot)$ and let k be (for example) the multivariate Gaussian with argument $\|s-t\|$. The constant function over E^r is not in this space with the Gaussian RBF kernel. Adjoin $[1^{(\alpha)}]$ to this space and define the averaging operator $\mathcal{E}_\alpha$ needed for the ANOVA decomposition as

$$\mathcal{E}_\alpha f_\alpha = \lim_{A\to\infty} \frac{1}{A^r} \int_A^A \dots \int_A^A f_\alpha(s) ds.$$

See that $\mathcal{E}_\alpha 1^{(\alpha)} = 1$ and $\mathcal{E}_\alpha f_\alpha = 0$ for f_α in $\mathcal{H}^{(\alpha)}$. Thus, we have the decomposition

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

and this term can be combined into the SSANOVA model.

Thus training sets with observed or coded pairwise distances as pseudo-attributes ma be treated like other, direct, observations in SSANOVA models.

Note that the r-variate Gaussian an be used as a density , or, as a positive definite function, and any other multivariate density which is an RBF when condsidered as a function of two arguments would work.

The Bottom Line The bottom line is that training sets with variables (attributes) where you only have pairwise distances between samples may be included in a Smoothing Spline ANOVA Model, either addtively or with interactions, and, in particular, when the attribute is a density, then pairwise distances between densities may be obtained by embedding the densities in an RKHS to get pairwise distances, and then mapping the pairwise distances into a low(er) dimensional Euclidean space to get pseudo-attributes, and thence into an SSANOVA model.

So Manny's work on both density kernels and RKHS kernels can be brought together to include densities/(images?) as attributes in an SSANOVA model.

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

References

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

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

[Lu et al., 2005] Lu, F., Keles, S., Wright, S., and Wahba, G. (2005). A framework for kernel regularization with application to protein clustering. *Proceedings of the National Academy of Sciences*, 102:12332–12337. Open Source at www.pnas.org/content/102/35/12332, PMCID: PMC118947.

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

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

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

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

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

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

[Wahba et al., 1995] Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.*, 23:1865–1895. Neyman Lecture.

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