How to incorporate personal densities into predictive models: Pairwise Density Distances, Regularized Kernel Estimation and Smoothing Spline ANOVA models.

Grace Wahba

Links to these slides in my website

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

PreAbstract

This talk is some combination of review and speculation, not the usual research talk. It began as appreciation of Manny Parzen, my thesis advisor, who was a key researcher in both density estimation and Reproducing Kernel Hilbert Spaces, of which we will hear more. Its an expansion of the talk that I gave at his memorial session at the 2017 JSM. Next is a picture from 2006.

Manny and party at the Pfizer Colloquium, 2006. l. to r. Nitis Mukhopadhyay, Joe Newton, me, Manny.

Abstract

We are concerned with the use of personal density functions or personal sample densities as subject attributes in prediction and classification models. The situation is particularly interesing when it is desired to combine other attributes with the personal densities in a prediction or classification model.

The procedure is (for each subject) to embed their sample density into a Reproducing Kernel Hilbert Space (RKHS), use this embedding to estimate pairwise distances between densities, use Regularized Kernel Estimation (RKE) with the pairwise distances to embed the subject (training) densities into a Euclidean space, and use the Euclidean coordinates as attributes in a Smoothing Spline ANOVA (SSANOVA) model. Elementary expository introductions to RKHS, RKE and SSANOVA occupy most of this talk.

Outline Part1

- An example of a personal density.
- Introduction to Reproducing Kernel Hilbert Spaces (RKHS)

Outline Part 2 Personal densities as attributes

- Step 1: Embed densities in an RKHS to obtain pairwise distances between densities.
- Step 2: Use Regularized Kernel Estimation (RKE) to map densities into E^r using pairwise distances to get pseudo-attributes.
- Step 3: Use Radial Basis Function kernels to include the pseudo-attributes of densities in SSANOVA Models.

Outline Part 3 Summary and Comments

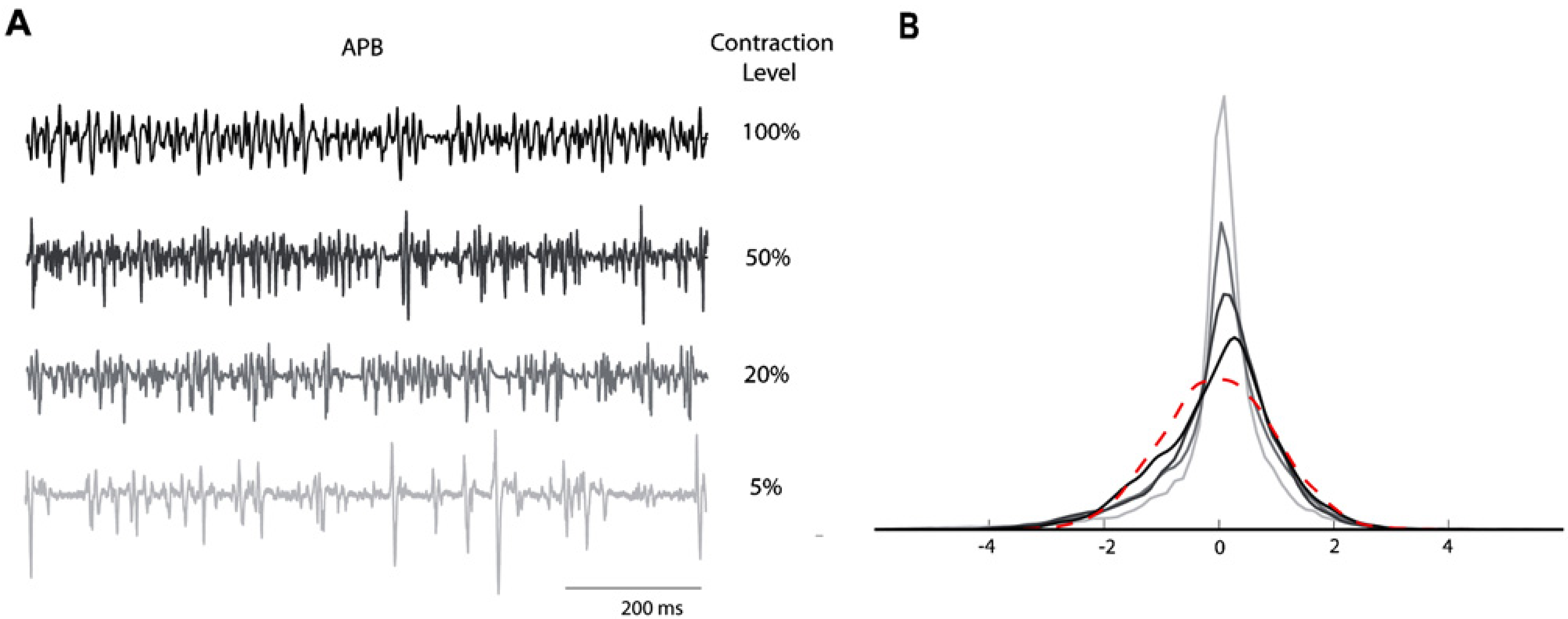
An example of a personal density

"A note on the probability distribution function of the surface electromyogram signal ". [Nazapour et al., 2013]

A surface electromyogram signal is the electrical manifestation of neuromuscular activity, recorded at the surface of the skin. The left figure is the trace at the Abductor Pollicis Brevis, the muscle whose job is to move the thumb away from the palm. The hand was restrained, and the signal was measured under four coditions of activity, amplified, filtered and sampled at 10kHz. Density estimates were obtained from the four sets of samples using Parzen kernel density estimates. [Parzen, 1962b]

An example (cont.)

K. Nazarpour et al. / Brain Research Bulletin 90 (2013) 88–91



Other biological time series where useful density information can be captured by high frequency sampling suggest themselves.

Introduction to RKHS, a trivial example

Ordinary ridge regression in d dimensions is a trivial example of an RKHS. We use a simple form of ridge regression to explain this.

Let $y = (y_1, y_2, \cdots, y_d)$ and $f = (f_1, f_2, \cdots, f_d)$ be d dimensional vectors and let Σ be a $d \times d$ (strictly) positive definite matrix. We can define a square norm on vectors in E^d by $\|f\|_\Sigma^2 = f\Sigma^{-1}f'$.

Let the eigenvectors and eigenvalues of Σ be ϕ_ν, λ_ν, we have $\Sigma_{ij} = \sum_{\nu=1}^{d} \lambda_\nu \phi_\nu(i)\phi_\nu(j)$. Then

$$\|f\|_\Sigma^2 = \sum_{\nu=1}^{d} \frac{f_\nu^2}{\lambda_\nu} \text{ where } f_\nu = (f, \phi_\nu). \tag{1}$$

Supposing $y = f + e$, where e is white Gaussian noise, then the ridge regression estimate of f is the minimizer in E^d of

$$\sum_{j=1}^{d} (y_j - f_j)^2 + \lambda\|f\|_\Sigma^2. \tag{2}$$

Some geometry which will generalize from matrices to kernels

Let $< f, g >\equiv f'\Sigma^{-1}g$, let σ_j be the jth column of Σ, and let σ_{jk} be the jk entry.

$$\Sigma^{-1}\Sigma = I \Rightarrow \Sigma^{-1}\sigma_j = (0, .., 0, 1, 0, .., 0)' \qquad (3)$$

with 1 in the jth position, and

$$f'\Sigma^{-1}\sigma_j \equiv< f, \sigma_j >= f_j. \qquad (4)$$

σ_j is the "representer" of the value of the jth component of f.
Furthermore,

$$< \sigma_j, \sigma_k >= \sigma_{jk}. \qquad (5)$$

This is the "reproducing" property (!). Our vector space E^d is the span of the representers. Next: From matrices to kernels.

Introduction to RKHS, continued

Let $\mathcal{T}$ be some domain of interest, examples are $[0,1]$, the d dimensional unit cube, the sphere, more complex domains to be discussed. $K(s,t)$ is a (strictly) positive definite kernel on $\mathcal{T}$ if

$$\sum_{i,j=1}^{n} a_i a_j K(t_i, t_j) > 0. \tag{6}$$

for all $\{a_i, a_j\}, t_i, t_j \in \mathcal{T}, n = 1, 2, ...$

Note that nothing is being assumed about the domain, other than the existence of a positive definite function on it.

Introduction to RKHS, continued

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

Moore-Aronszajn Theorem:

Let $\mathcal{T}$ be a domain on which a positive definite kernel, $K(s,t), s,t \in \mathcal{T}$ is defined. Then there exists a unique RKHS $\mathcal{H}_K$ associated with K, and vice versa, for every RKHS there exists a unique positive definite K. [Aronszajn, 1950] We just did the case $\mathcal{T}$ is $(1, 2, \cdots, d)$.

More:

Let $K_s(t) \equiv K(s,t)$ 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

$$< f, K_s > = \ f(s), \tag{7}$$

K_s is the "representer" of the value of f at s and

$$< K_s, K_t > = \ K(s,t). \tag{8}$$

This is the "reproducing" property of K.

The RKHS $\mathcal{H}_K$ is the closure of the span of the representers.

The Mercer theorem gives a class of kernels which are analogues of Σ that appeared in the ridge regession case.

Mercer Theorem: Let $\mathcal{T}$ be a compact domain in E^d, and K positive definite on $\mathcal{T}$. Suppose $\int_{\mathcal{T}} \int_{\mathcal{T}} K^2(s,t) ds dt = C < \infty$, then there exists an eigenfunction-eigenvalue decomposition

$$K(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu \phi_\nu(s) \phi_\nu(t). \tag{9}$$

[Riesz and Nagy, 1955] p243. Here, the λ_ν are eigenvalues and the ϕ_ν (orthonormal) eigenfunctions with $\sum_{\nu=1}^{\infty} \lambda_\nu^2 = C < \infty$.

The squared norm of f in this case is

$$\|f\|_K^2 = \sum_{\nu=1}^{\infty} \frac{f_\nu^2}{\lambda_\nu}, \text{ where } f_\nu = \int_{\mathcal{T}} f(s) \phi_\nu(s) ds. \tag{10}$$

Other reproducing kernels can be quite different, for example so called radial basis functions (RBF's), which depend only on the (Euclidean) distance between pairs of points- *e.g.* the Gaussian RBF:

$$K(s,t) = e^{-\frac{1}{\sigma^2}\|s-t\|^2}. \tag{11}$$

Functions in this RKHS are infinitely differentiable. The Matern class of RBF's is a useful class of RBF's, see [Bravo et al., 2009]. The squared norms can be expressed in terms of Fourier transforms.

Irrespective of the nature of the positive definite functions, let K_1 be a positive definite function on the domain $\mathcal{T}_1$ and K_2 be a positive definite function on $\mathcal{T}_2$ then $K = K_1 \otimes K_2$ is a positive definite function on the domain $\mathcal{T} = [\mathcal{T}_1 \otimes \mathcal{T}_2]$.

With $s_1, t_1 \in \mathcal{T}_1, s_2, t_2 \in \mathcal{T}_2$, $K(s_1, s_2; t_1, t_2) = K(s_1, t_1)K(s_2, t_2)$.

Let

$$y_i = f(t_i) + e_i, i = 1, 2, \cdots, n; \; t_i \in \mathcal{T} \tag{12}$$

where e is white Gaussian noise. The penalized likelihood estimate f_λ of $f \in \mathcal{H}_K$ is the solution to:

$$\min_{f \in \mathcal{H}_K} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \|f\|_K^2 \tag{13}$$

There may be other parameters hidden inside of K. For classification, the sum of squares is replaced by a sum of hinge functions (sometimes called the "kernel trick"). Or, it can be replaced by a log likelihood from the exponential family.

In all these cases, the representer theorem [Kimeldorf and Wahba, 1971] says that the minimizer will be in the span of the $K_{t_i}(t), i = 1, 2, \cdots, n$.

How to use personal densities as attributes

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

$$f(\cdot) = \int_{t\in\mathcal{T}} K(\cdot,t)p(t)dt. \tag{14}$$

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

$$f_X(\cdot) = \frac{1}{k}\sum_{j=1}^{k} K(X_j,\cdot) \tag{15}$$

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

$$g_Y(\cdot) = \frac{1}{\ell} \sum_{j=1}^{\ell} K(Y_j, \cdot). \tag{16}$$

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 is taken as $\|f_X - g_Y\|_K$, which is:

$$\frac{1}{k^2} \sum_{i,j=1}^{k} K(X_i, X_j) + \frac{1}{\ell^2} \sum_{i,j=1}^{\ell} K(Y_i, Y_j) - \frac{2}{kl} \sum_{i=1,j=1}^{k,\ell} K(X_i, Y_j). \tag{17}$$

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

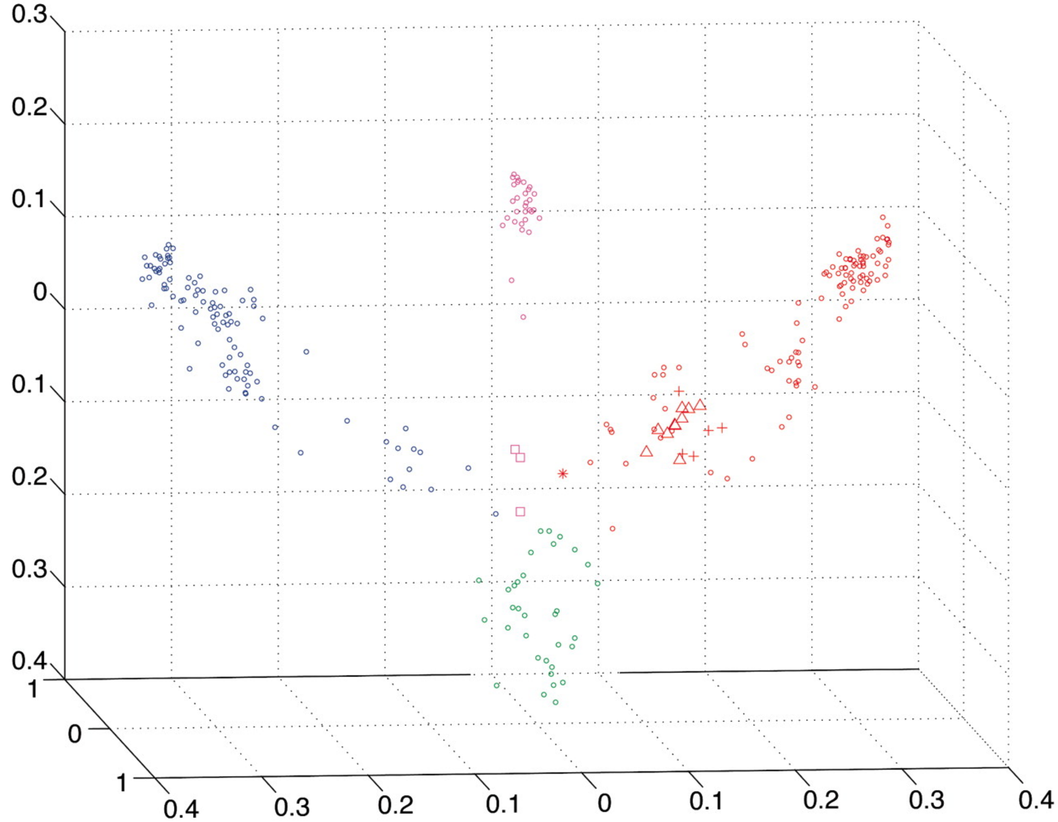
$$\hat{d}_{ij} = \Sigma(i,i) + \Sigma(j,j) - 2\Sigma(i,j). \tag{18}$$

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

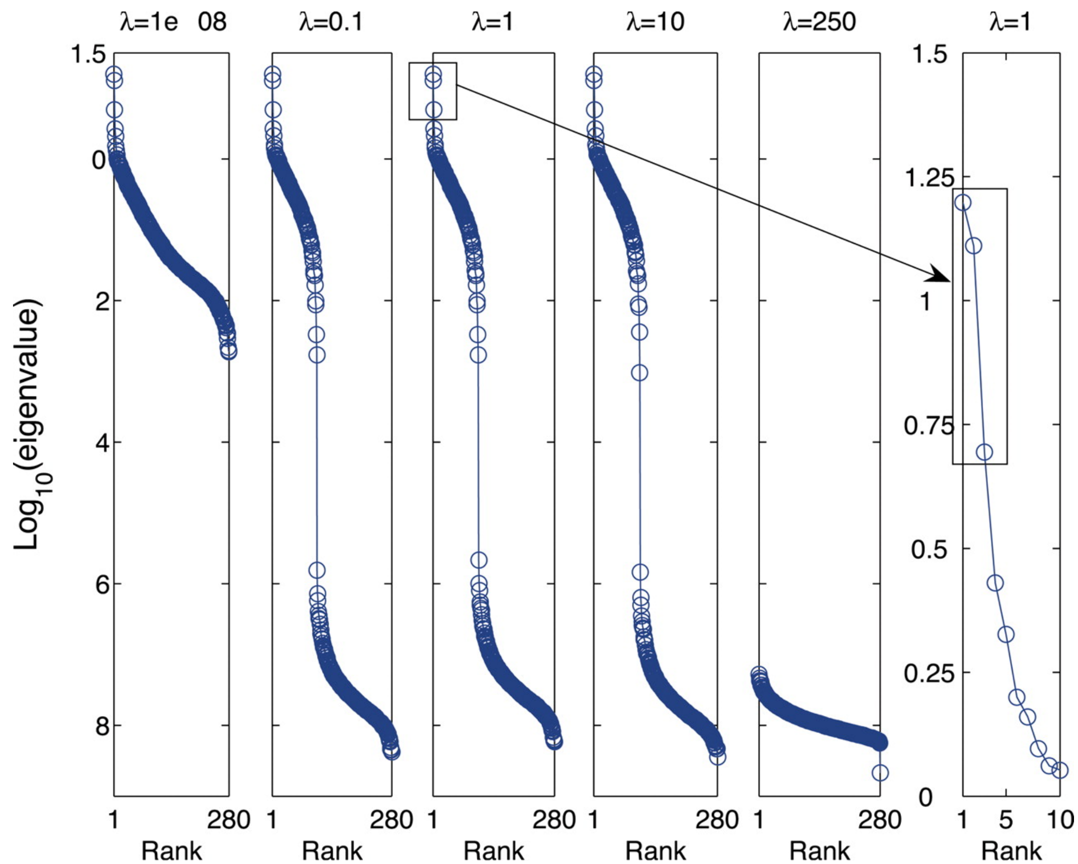
$$\min_{\Sigma \succeq 0} \sum_{(i,j)\in\Omega} |d_{ij} - \hat{d_{ij}}| + \lambda \, \mathrm{trace}(\Sigma). \tag{19}$$

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].



From [Lu et al., 2005] 3D representation of pairwise dissimilarity scores between 280 protein sequences obtained from pairwise alignment scores. RKE was used to get the Euclidean embedding and λ was chosed to capture 95% of the trace of the fitted matrix.



The effect of varying λ on the eigenvalues of the regularized estimate of Σ, log scale.

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

$$f(t_1, \ldots, t_d) = \mu + \sum_{\alpha} f_{\alpha}(t_{\alpha}) + \sum_{\alpha\beta} f_{\alpha\beta}(t_{\alpha}, t_{\beta}) + \cdots \tag{20}$$

and the terms satisfy ANOVA-like side conditions.

f is assumed to be in a tensor product space

$$\mathcal{H} = \otimes_{\alpha=1}^{d} \mathcal{H}_{\alpha}. \tag{21}$$

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

$$\mathcal{H}_{\alpha} = [1^{(\alpha)}] \oplus \mathcal{H}^{(\alpha)} \tag{22}$$

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

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

$$= [1] \oplus \sum_{\alpha} \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha<\beta} [\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \cdots , \qquad (23)$$

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

$$\mathcal{E}_\alpha f_\alpha = \lim_{A\to\infty} \frac{1}{A^r} \int_{-A/2}^{A/2} \dots \int_{-A/2}^{A/2} f_\alpha(s)ds. \tag{24}$$

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

$$\mathcal{H}_\alpha = [1^{(\alpha)}] \oplus \mathcal{H}^{(\alpha)} \tag{25}$$

and this term can be combined into the SSANOVA model.

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

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

Summary and Comments

We have given an elementary introduction to RKHS and showed how it can be used to estimate pairwise distances between densities. We did not discuss how to choose kernels or how to choose the tuning parameter(s) and other parameters inside K. We did not discuss seminorms. We demonstrated how a large set of the pairwise distances can be mapped into Euclidean space by using RKE to get pseudo attributes, and how the pseudo attributes can be used in a Smoothing Spline ANOVA model to incorporate them along with other attributes in a penalized likelihood estimate for prediction (or a support vector machine for classification.) It remains to apply this way of looking at densities as attributes in an analysis of an observational data set where personal densities can interact with other variables in complex ways.

References

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

[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.

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

[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.

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

[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). *Stochastic Processes*. Holden-Day, San Francisco.

[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.

[Riesz and Nagy, 1955] Riesz, F. and Nagy, B. S. (1955). *Functional Analysis*. Ungar, New York.

[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.