Reproducing kernels and correspondence matrices

Marianna Bolla
Institute of Mathematics
Technical University of Budapest
(Research supported by the
TÁMOP-4.2.2.C-11/1/KONV-2012-0001 project)
marib@math.bme.hu
András Krámli is 70, Szeged

July 27, 2013

Three dialogs of Hacsek and Sajó in a coffee-house

First dialog: about ML estimation in exponential families

- S: Did you know that in exponential families the ML-equation boils down to

$$\mathbb{E}_\theta(t(\mathbf{X})) = t(sample),$$

  where $\mathbf{X} = (X_1, \ldots, X_n)$ is i.i.d. sample and $t(\mathbf{X})$ is sufficient statistic for the unknown parameter $\theta \in \mathbb{R}^k$?
- H: How can it boil down, and what kind of a family is your exponential?

- S: You are stupid, the likelihood function of the sample $\mathbf{X} = (X_1, \ldots, X_n)$ in exponential family looks like

$$L_\theta(\mathbf{X}) = \frac{1}{a(\theta)} \cdot e^{t(\mathbf{X})\theta^T} \cdot b(\mathbf{X}),$$

  where $\theta = (\theta_1, \ldots, \theta_k)$ is natural parameter, $t(\mathbf{X}) = (t_1(\mathbf{X}), \ldots, t_k(\mathbf{X}))$ is sufficient statistic for it, and ^{T} stands for the transposition.

- H: And what about the $a(\theta)$?
- S: It is the normalizing constant, but can be written as

$$a(\theta) = \int_{\mathcal{X}} e^{t(\mathbf{x})\theta^T} \cdot b(\mathbf{x})\, d\mathbf{x},$$

  where $\mathcal{X} \subset \mathbb{R}^n$ is the sample space. This formula will play a crucial rule in our subsequent calculations.

- H: Haha, let us see those famous calculations!

- S: As you know, the likelihood equation is

$$\nabla_\theta \ln L_\theta(\mathbf{X}) = \mathbf{0},$$

that is

$$-\nabla_\theta \ln a(\theta) + \nabla_\theta(t(\mathbf{X})\theta^T) = \mathbf{0}. \tag{1}$$

Under certain regularity conditions,

$$\nabla_\theta \ln a(\theta) = \int_{\mathcal{X}} t(\mathbf{x}) e^{t(\mathbf{x})\theta^T} \cdot b(\mathbf{x})\, d\mathbf{x} = \mathbb{E}_\theta(t(\mathbf{X})).$$

Therefore, (1) is equivalent to

$$-\mathbb{E}_\theta(t(\mathbf{X})) + t(\mathbf{X}) = \mathbf{0},$$

which finishes the proof.

- H: But this is the idea of **moment estimation**. Is it true that in exponential families the ML-estimator is the same as the moment-estimator?
- S: More or less. When, especially, $t_1(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i$, ..., $t_k(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$, then it is. This is the case when our underlying distribution is Poisson, exponential, or Gaussian.
- H: I will tell it to our colleague, Mogyoro (Imre Toth), who asked whether these two estimators are the same.
- S: Not always, think of the continuous uniform distribution, which does not belong to the exponential family. Anyway, we had a prosperous discussion.
- H: Will you come in tomorrow?

Second dialog: about Correspondence Analysis

- H: My dear Sajó, you told me about correspondence analysis. I have studied it, but believe me, it is a stupid method. How does it come, that the best low-rank approximation of the table has negative entries in most of the cases?
- S: You are right if you consider the best L_2-norm approximation. Nonetheless, I am able to slightly adjust that approximation to obtain a low-rank approximation of positive entries, under very general conditions. My method also reveals the block-structure of the table. I was speaking about these facts at the EMS2013, but I am repeating the most important notions now.
- H: OK, let us see.

SVD of contingency tables and correspondence matrices

$\mathbf{C} = (c_{ij})$: $n \times m$ contingency table, $c_{ij} \geq 0$.
Row set: $Row = \{1, \ldots, n\}$
Column set: $Col = \{1, \ldots, m\}$

$$d_{row,i} = \sum_{j=1}^{m} c_{ij} \quad (i = 1, \ldots, n)$$

$$d_{col,j} = \sum_{i=1}^{n} c_{ij} \quad (j = 1, \ldots, m)$$

$$\mathbf{D}_{row} = \mathtt{diag}\,(d_{row,1}, \ldots, d_{row,n}) \quad \mathbf{D}_{col} = \mathtt{diag}\,(d_{col,1}, \ldots, d_{col,m}).$$

Representation

For a given integer $1 \leq k \leq \min\{n, m\}$, we are looking for k-dimensional representatives $\mathbf{r}_1, \ldots, \mathbf{r}_n$ of the rows and $\mathbf{c}_1, \ldots, \mathbf{c}_m$ of the columns such that they minimize the objective function

$$Q_k = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} \|\mathbf{r}_i - \mathbf{c}_j\|^2 \tag{2}$$

subject to

$$\sum_{i=1}^{n} d_{row,i} \mathbf{r}_i \mathbf{r}_i^T = \mathbf{I}_k, \quad \sum_{j=1}^{m} d_{col,j} \mathbf{c}_j \mathbf{c}_j^T = \mathbf{I}_k. \tag{3}$$

When minimized, the objective function Q_k favors k-dimensional placement of the rows and columns such that representatives of highly associated rows and columns are forced to be close to each other. As we will see, this is equivalent to the problem of correspondence analysis.

Solution

$$\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_k) = (\mathbf{r}_1^T, \ldots, \mathbf{r}_n^T)^T \; n \times k$$
$$\mathbf{Y} := (\mathbf{y}_1, \ldots, \mathbf{y}_k) = (\mathbf{c}_1^T, \ldots, \mathbf{c}_m^T)^T \; m \times k$$

$$Q_k = 2k - \mathtt{tr}\,(\mathbf{D}_{row}^{1/2}\mathbf{X})^T \mathbf{C}_{corr} (\mathbf{D}_{col}^{1/2}\mathbf{Y}) \to \min$$

subject to

$$\mathbf{X}^T \mathbf{D}_{row} \mathbf{X} = \mathbf{I}_k, \quad \mathbf{Y}^T \mathbf{D}_{col} \mathbf{Y} = \mathbf{I}_k,$$

where $\mathbf{C}_{corr} = \mathbf{D}_{row}^{-1/2} \mathbf{C} \mathbf{D}_{col}^{-1/2}$: correspondence matrix (normalized contingency table) belonging to the table $\mathbf{C}$.

Representation theorem

Let $\mathbf{C}_{corr} = \sum_{i=1}^{r} s_i \mathbf{v}_i \mathbf{u}_i^T$ be SVD, where $r \leq \min\{n, m\}$ is the rank of $\mathbf{C}_{corr}$, or equivalently (since there are not identically zero rows or columns), the rank of $\mathbf{C}$ and $1 = s_1 \geq s_2 \geq \cdots \geq s_r > 0$.
$\mathbf{v}_1 = (\sqrt{d_{row,1}}, \ldots, \sqrt{d_{row,n}})^T$ and $\mathbf{u}_1 = (\sqrt{d_{col,1}}, \ldots, \sqrt{d_{col,m}})^T$.
Let $k \leq r$ be a positive integer such that $s_k > s_{k+1}$. Then

$$\min Q_k = 2k - \sum_{i=1}^{k} s_i$$

and it is attained with $\mathbf{X}^* = \mathbf{D}_{row}^{-1/2}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ and $\mathbf{Y}^* = \mathbf{D}_{col}^{-1/2}(\mathbf{u}_1, \ldots, \mathbf{u}_k)$.

Regular row-column cluster pairs

The Expander Mixing Lemma for edge-weighted graphs naturally extends to this situation (Butler): for all $R \subset Row$ and $C \subset Col$

$$|c(R,C) - \mathtt{Vol}(R)\mathtt{Vol}(C)| \leq s_2\sqrt{\mathtt{Vol}(R)\mathtt{Vol}(C)},$$

where s_2 is the largest but 1 singular value of $\mathbf{C}_{corr}$ and

$$\mathtt{Vol}(R_a) = \sum_{i \in R_a} d_{row,i}, \quad \mathtt{Vol}(C_b) = \sum_{j \in C_b} d_{col,j}.$$

Since the spectral gap of $\mathbf{C}_{corr}$ is $1 - s_2$, in view of the above Expander Mixing Lemma, 'large' spectral gap is an indication of 'small' discrepancy: the weighted cut between any row and column subset of the contingency table is near to what is expected in a random table.

Volume-regular cluster pairs

We extend the notion of discrepancy to volume-regular pairs.

Definition

The row–column cluster pair $R \subset Row$, $C \subset Col$ of the contingency table $\mathbf{C}$ of total volume 1 is γ-volume regular if for all $X \subset R$ and $Y \subset C$ the relation

$$|c(X,Y) - \rho(R,C)\mathtt{Vol}(X)\mathtt{Vol}(Y)| \leq \gamma\sqrt{\mathtt{Vol}(R)\mathtt{Vol}(C)}$$

holds, where $\rho(R,C) = \frac{c(R,C)}{\mathtt{Vol}(R)\mathtt{Vol}(C)}$ is the relative inter-cluster density of the row–column pair R, C.

We will show that for given k, if the clusters are formed via applying the weighted k-means algorithm for the optimal row- and column representatives, respectively, then the so obtained row–column cluster pairs are homogeneous in the sense that they form equally dense parts of the contingency table.

Weighted k-variance

The weighted k-variance of the k-dimensional row representatives is defined by

$$S_k^2(\mathbf{X}) = \min_{(R_1,\ldots,R_k)} \sum_{a=1}^{k} \sum_{j \in R_a} d_{row,j} \|\mathbf{r}_j - \bar{\mathbf{r}}_a\|^2,$$

where $\bar{\mathbf{r}}_a = \frac{1}{\mathtt{Vol}(R_a)} \sum_{j \in R_a} d_{row,j} \mathbf{r}_j$ is the weighted center of cluster R_a $(a = 1, \ldots, k)$. Similarly, the weighted k-variance of the k-dimensional column representatives is

$$S_k^2(\mathbf{Y}) = \min_{(C_1,\ldots,C_k)} \sum_{a=1}^{k} \sum_{j \in C_a} d_{col,j} \|\mathbf{c}_j - \bar{\mathbf{c}}_a\|^2,$$

where $\bar{\mathbf{c}}_a = \frac{1}{\mathtt{Vol}(C_a)} \sum_{j \in C_a} d_{col,j} \mathbf{c}_j$ is the weighted center of cluster C_a $(a = 1, \ldots, k)$. Observe, that the trivial vector components can be omitted, and the k-variance of the so obtained $(k-1)$-dimensional representatives will be the same.

Thm (B, European Journal of Combinatorics 2013)

Theorem

Let $\mathbf{C}$ *be a contingency table of n-element row set Row and m-element column set Col, with row- and column sums* $d_{row,1}, \ldots, d_{row,n}$ *and* $d_{col,1}, \ldots, d_{col,m}$, *respectively. Suppose that* $\sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} = 1$ *and there are no dominant rows and columns:* $d_{row,i} = \Theta(1/n)$, $(i = 1, \ldots, n)$ *and* $d_{col,j} = \Theta(1/m)$, $(j = 1, \ldots, m)$ *as* $n, m \to \infty$. *Let the singular values of* $\mathbf{C}_{corr}$ *be*

$$1 = s_1 > s_2 \geq \cdots \geq s_k > \varepsilon \geq s_i, \quad i \geq k+1.$$

The partition $(R_1, \ldots, R_k)$ *of Row and* $(C_1, \ldots, C_k)$ *of Col are defined so that they minimize the weighted k-variances* $S_k^2(\mathbf{X})$ *and* $S_k^2(\mathbf{Y})$ *of the row and column representatives. Suppose that there are constants* $0 < K_1, K_2 \leq \frac{1}{k}$ *such that* $|R_i| \geq K_1 n$ *and* $|C_i| \geq K_2 m$ $(i = 1, \ldots, k)$, *respectively. Then the* R_i, C_j *pairs are* $\mathcal{O}(\sqrt{2k}(S_k(\mathbf{X})S_k(\mathbf{Y})) + \varepsilon)$*-volume regular* $(i, j = 1, \ldots, k)$.

The proof

- H: But how does the positivity of the k-rank approximation follow from this?
- S: I am afraid, you have to understand some basic steps of the proof for this, as follows.

Recall that the largest singular value of $\mathbf{C}_{corr}$ is 1 with corresponding singular vector pair $\mathbf{v}_0 = \mathbf{D}_{row}^{1/2}\mathbf{1}_m$ and $\mathbf{u}_0 = \mathbf{D}_{col}^{1/2}\mathbf{1}_n$, respectively. The optimal k-dimensional representatives of the rows and columns are row vectors of the matrices $\mathbf{X} = (\mathbf{x}_0, \ldots, \mathbf{x}_{k-1})$ and $\mathbf{Y} = (\mathbf{y}_0, \ldots, \mathbf{y}_{k-1})$, where $\mathbf{x}_i = \mathbf{D}_{row}^{-1/2}\mathbf{v}_i$ and $\mathbf{y}_i = \mathbf{D}_{col}^{-1/2}\mathbf{u}_i$, respectively ($i = 0, \ldots, k-1$). (Note that the first columns of equal coordinates can as well be omitted.)

Assume that the minimum k-variance is attained on the k-partition $(R_1, \ldots, R_k)$ of the rows and $(C_1, \ldots, C_k)$ of the columns, respectively. Then

$$S_k^2(\mathbf{X}) = \sum_{i=0}^{k-1} \mathtt{dist}^2(\mathbf{v}_i, F), \quad S_k^2(\mathbf{Y}) = \sum_{i=0}^{k-1} \mathtt{dist}^2(\mathbf{u}_i, G),$$

where $F = \mathtt{Span}\{\mathbf{D}_{row}^{1/2}\mathbf{w}_1, \ldots, \mathbf{D}_{row}^{1/2}\mathbf{w}_k\}$ and $G = \mathtt{Span}\{\mathbf{D}_{col}^{1/2}\mathbf{z}_1, \ldots, \mathbf{D}_{col}^{1/2}\mathbf{z}_k\}$ with the so-called normalized row partition vectors $\mathbf{w}_1, \ldots, \mathbf{w}_k$ of coordinates $w_{ji} = \frac{1}{\sqrt{\mathtt{Vol}(R_i)}}$ if $j \in R_i$ and 0, otherwise; and column partition vectors $\mathbf{z}_1, \ldots, \mathbf{z}_k$ of coordinates $z_{ji} = \frac{1}{\sqrt{\mathtt{Vol}(C_i)}}$ if $j \in C_i$ and 0, otherwise $(i = 1, \ldots, k)$.

Note that the vectors $\mathbf{D}_{row}^{1/2}\mathbf{w}_1, \ldots, \mathbf{D}_{row}^{1/2}\mathbf{w}_k$ and $\mathbf{D}_{col}^{1/2}\mathbf{z}_1, \ldots, \mathbf{D}_{col}^{1/2}\mathbf{z}_k$ form orthonormal systems in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively (but they are, usually, not complete).

However, we can find orthonormal systems $\tilde{\mathbf{v}}_0, \ldots, \tilde{\mathbf{v}}_{k-1} \in F$ and $\tilde{\mathbf{u}}_0, \ldots, \tilde{\mathbf{u}}_{k-1} \in G$ such that

$$S_k^2(\mathbf{X}) \leq \sum_{i=0}^{k-1} \|\mathbf{v}_i - \tilde{\mathbf{v}}_i\|^2 \leq 2S_k^2(\mathbf{X}), \quad S_k^2(\mathbf{Y}) \leq \sum_{i=0}^{k-1} \|\mathbf{u}_i - \tilde{\mathbf{u}}_i\|^2 \leq 2S_k^2(\mathbf{Y}).$$

Let $\mathbf{C}_{corr} = \sum_{i=0}^{r-1} s_i \mathbf{v}_i \mathbf{u}_i^T$ be SVD, where $r = \mathtt{rank}\,(\mathbf{C}) = \mathtt{rank}\,(\mathbf{C}_{corr})$. We approximate $\mathbf{C}_{corr}$ by the rank k matrix $\sum_{i=0}^{k-1} s_i \tilde{\mathbf{v}}_i \tilde{\mathbf{u}}_i^T$.

The vectors $\hat{\mathbf{v}}_i := \mathbf{D}_{row}^{-1/2}\tilde{\mathbf{v}}_i$ are stepwise constants on the partition $(R_1, \ldots, R_k)$ of the rows, whereas the vectors $\hat{\mathbf{u}}_i := \mathbf{D}_{col}^{-1/2}\tilde{\mathbf{u}}_i$ are stepwise constants on the partition $(C_1, \ldots, C_k)$ of the columns, $i = 0, \ldots, k-1$. The matrix

$$\sum_{i=0}^{k-1} s_i \hat{\mathbf{v}}_i \hat{\mathbf{u}}_i^T$$

is therefore an $n \times m$ block-matrix on $k \times k$ blocks corresponding to the above partition of the rows and columns. Let $\hat{c}_{ab}$ denote its entries in the ab block $(a, b = 1, \ldots, k)$.
This is the rank k approximation of the matrix $\mathbf{C}$ with a block-matrix.

The point is: The entries $\hat{c}_{ij}$'s of the block-matrix will already be positive provided the weighted k-variances $S_k^2(\mathbf{X})$ and $S_k^2(\mathbf{Y})$ are 'small' enough. Let us discuss this issue more precisely.
In accord with the notation used in the proof, denote by ab in the lower index the matrix restricted to the $R_a \times C_b$ block (otherwise it has zero entries). Then for the squared Frobenius norm of the rank k approximation of $\mathbf{D}_{row}^{-1}\mathbf{C}\mathbf{D}_{col}^{-1}$, restricted to the ab block, we have that

$$\left\| \mathbf{D}_{row,a}^{-1}\mathbf{C}_{ab}\mathbf{D}_{col,b}^{-1} - \left(\sum_{i=0}^{k-1} s_i \hat{\mathbf{v}}_i \hat{\mathbf{u}}_i^T\right)_{ab} \right\|_2^2 = \sum_{i\in R_a}\sum_{j\in C_b}\left(\frac{c_{ij}}{d_{row,i}d_{col,j}} - \hat{c}_{ab}\right)^2$$

$$= \sum_{i\in R_a}\sum_{j\in C_b}\left(\frac{c_{ij}}{d_{row,i}d_{col,j}} - \bar{c}_{ab}\right)^2 + |R_a||C_b|(\bar{c}_{ab} - \hat{c}_{ab})^2.$$

Here we used the Steiner equality with the average $\bar{c}_{ab}$ of the entries of $\mathbf{D}_{row}^{-1}\mathbf{C}\mathbf{D}_{col}^{-1}$ in the ab block. We can estimate the above Frobenius norm by a constant multiple of the spectral norm. In this way,

$$(\bar{c}_{ab}-\hat{c}_{ab})^2 \leq \frac{1}{\max\{|R_a|,|C_b|\}} \cdot \max_{i\in R_a} \frac{1}{d_{row,i}} \cdot \max_{j\in C_b} \frac{1}{d_{col,j}} \cdot [\sqrt{2k}(S_k(\mathbf{X})+S_k(\mathbf{Y}))$$

But using the conditions on the block sizes and the row- and column-sums of the Theorem, provided

$$\sqrt{2k}(S_k(\mathbf{X}) + S_k(\mathbf{Y})) + \varepsilon) = \mathcal{O}\left(\frac{1}{(\min\{m,n\})^{\frac{1}{2}+\tau}}\right)$$

holds with some 'small' $\tau > 0$, the relation $\bar{c}_{ab} - \hat{c}_{ab} \to 0$ also holds as $n, m \to \infty$. Therefore, both $\hat{c}_{ab}$ and $\hat{c}_{ab} d_{row,i} d_{col,j}$ are positive over such blocks that are not constantly zero in the original table if m and n are large enough.

- S: This is the end of the long story.
- H: Will you come in tomorrow?

Third dialog: about Reproducing Kernel Hilbert Spaces

- H: My dear Sajó, there are fairy tales about some fictitious spaces, where everything is 'smooth' and 'linear'.
- S: Such spaces really exist, the hard part is that we should adopt them to our data. Good news is that it is not necessary to actually map our data into them.
- H: Then how can we use them?
- S: It suffices to treat only a kernel function, but the bad news is that the kernel must be appropriately selected so that the underlying nonlinearity could be detected.
- H: They must be the Reproducing Kernel Hilbert Spaces. Tell me more about them!

History

- S: Reproducing Kernel Hilbert Spaces were introduced in the middle of the 20th century by Aronszajn,Parzen, and others, but the theory itself is an elegant application of already known theorems of functional analysis, first of all the Riesz–Fréchet representation theorem and the theory of integral operators (see Fréchet, Riesz, Szőkefalvi-Nagy) tracing back to the beginning of the 20th century. Later on, in the last decades of the 20th century and even in our days, Reproducing Kernel Hilbert Spaces are several times reinvented and applied in modern statistical methods and data mining (for example, Bach,Baker).
- H: But what is the mystery of reproducing kernels and what is the diabolic kernel trick?

Definition of an RKHS

A stronger condition imposed on a Hilbert space $\mathcal{H}$ of functions $\mathcal{X} \to \mathbb{R}$ (where $\mathcal{X}$ is an arbitrary set, for the time being) is that the following so-called evaluation mapping be a continuous, or equivalently, a bounded linear functional. The evaluation mapping $L_x : \mathcal{H} \to \mathbb{R}$ works on an $f \in \mathcal{H}$ such that $L_x(f) = f(x)$.

Definition

A Hilbert space $\mathcal{H}$ of (real) functions on the set $\mathcal{X}$ is an RKHS if the point evaluation functional L_x exists and is continuous for all $x \in \mathcal{X}$.

- H: And where does the name RKHS come from?
- S: From the Riesz–Fréchet representation theorem. This theorem states that a Hilbert space (in our case $\mathcal{H}$) and its dual (in our case the set of $\mathcal{H} \to \mathbb{R}$ continuous linear functionals, e.g. L_x) are isometrically isomorphic. Therefore, to any L_x there uniquely corresponds a $K_x \in \mathcal{H}$ such that

$$L_x(f) = \langle f, K_x \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}. \tag{4}$$

Since K_x is itself an $\mathcal{X} \to \mathbb{R}$ function, it can be evaluated at any point $y \in \mathcal{X}$. We define the bivariate function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as

$$K(x, y) := K_x(y) \tag{5}$$

and call it the reproducing kernel for the Hilbert space $\mathcal{H}$.

- H: But why is this kernel positive semidefinite?
- S: Because by using formulas (4) and (5), we get that on the one hand,

$$K(x,y) = K_x(y) = L_y(K_x) = \langle K_x, K_y \rangle_{\mathcal{H}},$$

and on the other hand,

$$K(y,x) = K_y(x) = L_x(K_y) = \langle K_y, K_x \rangle_{\mathcal{H}}.$$

By the symmetry of the (real) inner product it follows that the reproducing kernel is symmetric and it is also reproduced as the inner product of special functions in the RKHS:

$$K(x,y) = \langle K_x, K_y \rangle_{\mathcal{H}} = \langle K(x,.), K(.,y) \rangle_{\mathcal{H}},$$

hence, K is positive semidefinite.

RKHS belonging to a kernel

- S: Vice versa, if we are given a positive definite kernel function on $\mathcal{X} \times \mathcal{X}$ at the beginning, then there exists an RKHS such that with appropriate elements of it, the inner product relation holds.

Definition

A symmetric two-variate function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite kernel (equivalently, admissible, valid, or Mercer kernel) if for any $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathcal{X}$, the symmetric matrix of entries $K(x_i, x_j) = K(x_j, x_i)$ $(i, j = 1, \ldots n)$ is positive semidefinite.

We remark that a symmetric real matrix is positive semidefinite if and only if it is a Gram matrix, and hence, its entries become inner products, but usually not of the entries in its arguments. However, the simplest kernel function, the so-called linear kernel, does this job: $K_{\text{lin}}(x, y) = \langle x, y \rangle_{\mathcal{X}}$, where $\mathcal{X}$ is subset of a Euclidean space.

- H: Show me other positive definite kernels!
- S: You can get a lot of them with the following operations:
  1. If $K_1, K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are positive definite kernels, then the kernel K defined by $K(x, y) = K_1(x, y) + K_2(x, y)$ is also positive definite.
  2. If $K_1, K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are positive definite kernels, then the kernel K defined by $K(x, y) = K_1(x, y)K_2(x, y)$ is also positive definite. Especially, if K is a positive definite kernel, then so does cK with any $c > 0$.

  Consequently, if h is a polynomial with positive coefficients and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite kernel, then the kernel $K_h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$K_h(x, y) = h(K(x, y))$$

  is also positive definite. Since the exponential function can be approximated by polynomials with positive coefficients and the positive definiteness is closed under pointwise convergence, the same is true if h is the exponential function: $h(x) = e^x$, perhaps some transformation of it.

- H: Then putting these facts together and using the formula

$$\|x - y\|^2 = \langle x, x\rangle + \langle y, y\rangle - 2\langle x, y\rangle,$$

we can easily verify that the Gaussian kernel is positive definite:

$$K_{\text{Gauss}}(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}},$$

where $\sigma > 0$ is a parameter.

- S: You are getting more and more clever, Hacsek. Now we are able to formulate the converse statement.

Theorem

For any positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ there exists a unique, possibly infinite-dimensional Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$, for which K is a reproducing kernel.

If we want to emphasize that the RKHS corresponds to the kernel K, we will denote it by $\mathcal{H}_K$.

- H: Cannot we realize the elements of $\mathcal{H}_K$ in a more straightforward Hilbert space?
- S: Oh, yes, this is the feature space $\mathcal{F}$. Assume that there is a (usually not linear) map $\phi : \mathcal{X} \to \mathcal{F}$ such that when $x \in \mathcal{X}$ is mapped into $\phi(x) \in \mathcal{F}$, then

$$K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$$

is the desired positive definite kernel.
Let T be a linear operator from $\mathcal{F}$ to the space of functions $\mathcal{X} \to \mathbb{R}$ defined by

$$(Tf)(y) = \langle f, \phi(y) \rangle_{\mathcal{F}}, \quad y \in \mathcal{X}, \, f \in \mathcal{F}.$$

Then

$$T\phi(x) = K_x, \quad \forall x \in \mathcal{X}$$

and hence, $\mathcal{H}_K$ becomes the range of T.

- H: Could you tell me an example of an RKHS? What an animal is it?
- S: Yes, a couple. I'll give the theoretical construction for $\mathcal{H}_k$ and $\mathcal{F}$, together with the functions K_x and the features $\phi(x)$.

First example

Let K be the continuous kernel of a positive definite Hilbert–Schmidt operator which is an integral operator working on the $L^2(\mathcal{X})$ space, where $\mathcal{X}$ is a compact set in $\mathbb{R}$ for simplicity. Due to the positive definiteness of K, and the Mercer theorem, K can be expanded into the following uniformly convergent series:

$$K(x,y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(y), \quad \forall x, y \in \mathcal{X}$$

by the eigenfunctions and the eigenvalues of the integral operator. The RKHS defined by K is the following:

$$\mathcal{H}_K = \{f : \mathcal{X} \to \mathbb{R} \ : \ f(x) = \sum_{i=1}^{\infty} c_i \psi_i(x)\}$$

such that $\sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i} < \infty$.

$$K_x = K(x, .) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i$$

Therefore,

$$\langle K_x, K_y \rangle_{\mathcal{H}_K} = \sum_{i=1}^{\infty} \frac{\lambda_i \psi_i(x) \lambda_i \psi_i(y)}{\lambda_i} = K(x, y).$$

Here the feature space $\mathcal{F}$ is the following:

$$\phi(x) = (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \dots), \quad x \in \mathcal{X}$$

and the inner product is naturally defined by

$$\langle \phi(x), \phi(y) \rangle_{\mathcal{F}} = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(y) = \langle K_x, K_y \rangle_{\mathcal{H}_K}.$$

The functions $\sqrt{\lambda_1}\psi_1, \sqrt{\lambda_2}\psi_2, \dots$, which form an orthonormal basis in $\mathcal{H}_k$, and because of this transformation, a function $f \in L^2(\mathcal{X})$ is in $\mathcal{H}_k$ if $\|f\|^2_{\mathcal{H}_k} = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i} < \infty$.

Second example

Now $\mathcal{X}$ is a Hilbert space of finite dimension, say R^p, and its elements will be denoted by boldface $\mathbf{x}$, stressing that they are vectors.
If we used K_{lin} on $\mathcal{X} \times \mathcal{X}$, then $K_{\mathbf{x}} = \langle \mathbf{x}, . \rangle_{\mathcal{X}}$, and by the Riesz–Fréchet representation theorem, $\phi(\mathbf{x}) = \mathbf{x}$ would reproduce the kernel, as $K_{\text{lin}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{X}}$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Now the RKHS induced by K_{lin} is identified with the feature space, which is $\mathcal{X} = \mathbb{R}^p$ itself.
In case of more sophisticated kernels, $\mathcal{H}_K$ contains non-linear functions, and therefore, the features $\phi(\mathbf{x})$ can be realized usually in much higher dimension than that of $\mathcal{X}$.

For $\mathbf{x} = (x_1, x_2) \in \mathcal{X} = \mathbb{R}^2$ let

$$\phi(\mathbf{x}) := (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

$\mathcal{F} \subset \mathbb{R}^3$. We want to separate data points allocated along two concentric circles, and therefore $\mathbb{R}^2 \to \mathbb{R}$ quadratic functions are applied.

$$\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{F}} = x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2 = (x_1y_1 + x_2y_2)^2 = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{X}}^2,$$

hence, the new kernel is the square of the linear one, which is also positive definite (polynomial kernel).
The RKHS $\mathcal{H}_K$ corresponding to the feature space $\mathcal{F}$ now consists of homogeneous degree quadratic functions $\mathbb{R}^2 \to \mathbb{R}$, with the functions $f_1 : (x_1, x_2) \to x_1^2$, $f_2 : (x_1, x_2) \to x_2^2$, and $f_3 : (x_1, x_2) \to \sqrt{2}x_1x_2$ forming an orthonormal basis in $\mathcal{H}_k$ such that $K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^3 f_i(\mathbf{x})f_i(\mathbf{y})$.

- S: Observe that in both examples $\phi(x)$ is a vector with coordinates which are the basis vectors of the RKHS evaluated at x. In the first exercise $\phi(x)$ is an infinite, whereas in the second one, a finite dimensional vector. Note that $\mathcal{H}_K$ is an affine and sparsified version of the Hilbert space of $\mathcal{X} \to \mathbb{R}$ functions, between which the inner product is adopted to the requirement that it would reproduce the kernel.
- H: For me it means that non-linear features can be represented by a more complicated Hilbert space, and not the original one. Am I right, my dear Sajó?
- S: Not exactly, but probably this is the point of the whole RKHS story.

The end