Statistical Analysis

in

Reproducing Kernel Hilbert Space

Nov. 25, 2005 at Inst. Statistical Science, AS

Parametric and nonparametric statistical analysis in Euclidean space R^d

– Density estimation

– Regression

– Classification

Statistical analysis in reproducing kernel Hilbert space –the line between parametrics and nonparametrics becomes thin in an RKHS.

– mainly preparatory work and classification in this lecture.

Reproducing kernel

- A real-valued symmetric function $K(x,u) : \mathcal{X} \times \mathcal{X} \to R$ is called a **positive definite kernel** if, for all $n \in N$, $x_1, \ldots, x_n \in \mathcal{X}$, and $\xi_1, \ldots, \xi_n \in R$, we have $\sum_{i,j=1}^{n} K(x_i, x_j)\xi_i\xi_j \geq 0$. K is also called a **reproducing kernel**.

  In matrix notation, $\xi' K(A, A')\xi \geq 0$, $\forall n, \xi, A' = [x_1, \ldots, x_n]$.

- The kernel examples in last lecture are all reproducing kernels.

- Gaussian kernel: $K(x,u) = \exp\{-(x-u)^2/2h^2\}/(\sqrt{2\pi}h)$,

  $K(x,u) = \exp\{-\|x-u\|^2/2h^2\}/(\sqrt{2\pi}h)^d$,

  $K(x,u) = \exp\{-(x-u)'H^{-1}(x-u)/2\}/[(\sqrt{2\pi})^d|H|]$, H: window matrix.

Reproducing kernel Hilbert space -1

- A **reproducing kernel Hilbert space** $\mathcal{H}$ on $\mathcal{X}$ is a Hilbert space of real-valued functions from $\mathcal{X}$ to R where all evaluation functionals* are bounded (or equivalently continuous)†.

  There exists a RK K, for every $x, u \in \mathcal{X}$, $K(x,\cdot), K(\cdot,u) \in \mathcal{H}$ and for every $x, u \in E$ and $f \in \mathcal{H}$, we have the **reproducing property**

$$\langle\ f(\cdot),\ K(x,\cdot)\ \rangle_{\mathcal{H}} = f(x) \quad \text{and} \quad \langle\ f(\cdot),\ K(\cdot,u)\ \rangle_{\mathcal{H}} = f(u).$$

- $K \longleftrightarrow \mathcal{H}$. (**existence and uniqueness**)

*$\ell_x : \mathcal{H} \to R$ such that $\ell_x(f) = f(x)$.

†An RKHS is a Hilbert space of pointwise defined functions, where the $\mathcal{H}$-norm convergence implies pointwise convergence.

Reproducing kernel Hilbert space -2

- K-generated Hilbert space consists of functions of the form $\sum \alpha_i K(x, x_i)$ and completed with limits.

  RKHS: $\mathcal{H} =$ closure$\{\sum \alpha_i K(x, x_i)\}$ wrt the norm below.

- Inner product $\langle K(x, x_i), K(x, x_j)\rangle_{\mathcal{H}} = K(x_i, x_j)$. **easy to compute**

  – Norm: $\|\sum \alpha_i K(x, x_i)\|^2_{\mathcal{H}} = \sum_{i,j=1}^n K(x_i, x_j)\alpha_i\alpha_j = \alpha' K\alpha$.

  – $K(x, u) = \sum_{j=1}^\infty \lambda_j \phi_j(x)\phi_j(u)$, if K induces a compact integral operator on $L_2(\mathcal{X}, d\mu)$, where $\{\phi_j\}$ are orthonormal in $L_2(\mathcal{X}, d\mu)$.

  – In spectral representation: $\langle f, g\rangle_{\mathcal{H}} = \sum_{j=1}^\infty f_j g_j/\lambda_j$, where $f(x) = \sum_{j=1}^\infty f_j\phi_j(x)$ and same for g.

  – $\|f\|^2_{\mathcal{H}} = \sum_{j=1}^\infty f_j^2/\lambda_j < \infty$ for $f \in \mathcal{H}$.

  – **small λ in the denominator causing smoothing effect.**

Let μ be a probability measure on $(\mathcal{X}, \mathcal{B})$. ($\mu$ need not be the uderlying probability distribution of the training inputs.) We assume all the reproducing kernels employed are

- measurable,
- trace type, i.e., $\int_{\mathcal{X}} K(x,x)d\mu < \infty$,
- for $x \neq u$, $K(x,\cdot) \neq K(u,\cdot)$.

Consider a transformation $\gamma : \mathcal{X} \to \mathcal{H}$ given by

$$x \to \gamma(x) := K(x,\cdot). \tag{1}$$

The original input space $\mathcal{X}$ is then embedded into a new input space $\mathcal{H}$ via the transformation γ. Each input point $x \in \mathcal{X}$ is mapped to an element $\gamma(x) = K(x,\cdot) \in \mathcal{H}$.

Advantages

- Computational advantages: inner products calculated as kernel values, optimization tool, etc.

- View from $\mathcal{H}$: linear algorithm, a single global linear model.

  View from $\mathcal{X}$: nonlinear algorithm, mixture of many local models.

  – Space $\mathcal{H}$ has richer algebraic and topological structure than $\mathcal{X} \subset R^d$ to allow, e.g., linear separation of clusters.

  – Nonparametric modelling, while fitting data via a certain parametric notion.

- Linear in $\{x_i\}_{i=1}^n$: $\sum_i \alpha_i x_i \in R^d$;

  Linear in $\{K(x_i, \cdot)\}_{i=1}^n$: $\sum_i \alpha_i K(x_i, \cdot) \in \mathcal{H}$.

  Linear in x: $v'x$;

  Linear in $\{K(x, \cdot)\}_{i=1}^n$: $\langle h(\cdot), K(x, \cdot)\rangle_{\mathcal{H}}$, kernel mixture.

Isometrical isomorphism

Let $\mathcal{J}$ be a map from one feature space $\Phi(\mathcal{X})$ to another $\gamma(\mathcal{X}) \subset \mathcal{H}$ defined by $\mathcal{J}(\Phi(x)) = \gamma(x) \in \mathcal{H}$. Note that $\mathcal{J}$ is a one-to-one linear transformation satisfying

$$\|\Phi(x)\|_{\mathcal{Z}}^2 = K(x,x) = \|\gamma(x)\|_{\mathcal{H}}^2.$$

Thus, $\Phi(\mathcal{X})$ and $\gamma(\mathcal{X})$ are isometrically isomorphic, and the two feature representations

- $x \to \gamma(x) := K(x,\cdot)$: explicitly defined,

- $x \to \Phi(x)$: implicitly defined,

are equivalent in the sense of isometrical isomorphism.

Gaussian measure on a Hilbert space

- Let $\mathcal{H}$ be an arbitrary real separable* Hilbert space. A probability measure $P_{\mathcal{H}}$ defined on $\mathcal{H}$ is said to be Gaussian, if the distribution of $\langle f, h\rangle_{\mathcal{H}}$ is a one-dimensional normal for any $f \in \mathcal{H}$, where h denotes the random element having the probability measure $P_{\mathcal{H}}$.

- It can be shown that for any m and any $\{f_1, \ldots, f_m \in \mathcal{H}\}$, the joint distribution of $\langle f_1, h\rangle_{\mathcal{H}}, \ldots, \langle f_m, h\rangle_{\mathcal{H}}$ is normal.

- In binary classification, the SVM-type algorithms (linear in $\mathcal{H}$) have **effective working subspace of dimensionality one**. For a k-group classification, they have effective working subspace of dimensionality at most $k - 1$.

- **Low dimensional normal approximation will be enough.**

*i.e., with a countable dense subset. In a separable Hilbert space countable orthonormal systems are used to expand any element as an infinite sum.

Covariance operator

- For a probability measure $P_{\mathcal{H}}$ on $\mathcal{H}$ satisfying $E\langle h, h\rangle_{\mathcal{H}} < \infty$, there exists $m \in \mathcal{H}$, the mean, and a covariance operator Λ such that
  - $\langle m, f\rangle_{\mathcal{H}} = E\langle h, f\rangle_{\mathcal{H}},\ \forall f \in \mathcal{H}$ and
  - $\langle \Lambda f, g\rangle_{\mathcal{H}} = E\langle h - m, f\rangle_{\mathcal{H}}\langle h - m, g\rangle_{\mathcal{H}},\ \forall f, g \in \mathcal{H}$.
  - Λ is of trace type and trace$(\Lambda) = E\langle h, h\rangle_{\mathcal{H}}$.

- It plays a similar role as a covariance matrix in Euclidean space.

Linear classifier. Consider a binary classification in a Hilbert space $\mathcal{H}$. We say that a classifier is linear if and only if its decision boundary is given by

$$\ell(h) + b = 0,$$

where $\ell(\cdot)$ is a bounded linear functional, b is a real scalar and h is an element in $\mathcal{H}$.

- By Riesz Representation Theorem, there exists a unique $g \in \mathcal{H}$ such that the decision boundary is given by

$$\langle g, h \rangle_{\mathcal{H}} + b = 0.$$

- Recall the transformation $\gamma : \mathcal{X} \to \mathcal{H}$, which is equipped with a richer algebraic and topological structure. The idea is to look for a *functional normal direction* g, which is **optimal*** in a certain sense in separating the two groups.

*e.g., maximum margin for SVM, maximum likelihood ratio for KFDA, etc.

Theorem. (Grenander, 1950.) Assume that $P_{1,\mathcal{H}}$ and $P_{2,\mathcal{H}}$ are two equivalent Gaussian measures on $\mathcal{H}$ with means m_1 and m_2 and a common nonsingular covariance operator Λ. Let $L_{2,1} = \log(dP_{2,\mathcal{H}}/dP_{1,\mathcal{H}})$ and h be an element in $\mathcal{H}$. Let $m_a = (m_1 + m_2)/2$ and $m_d = m_2 - m_1$. A necessary and sufficient condition for the log-likelihood ratio $L_{2,1}$ being linear is that $m_d \in R(\Lambda^{1/2})$, where $R(\Lambda^{1/2})$ is the range of $\Lambda^{1/2}$. The log-likelihood ratio is then given by

$$L_{2,1}(h) = \langle h - m_a, \Lambda^{-1} m_d \rangle_{\mathcal{H}}. \tag{2}$$

To separate two Gaussian populations in $\mathcal{H}$, the log-likelihood ratio leads to an ideal optimal linear decision boundary.

Fisher linear discriminant : $(x - (\mu_1 + \mu_2)/2)' \Sigma^{-1} (\mu_2 - \mu_1)$

Remark 1 (Bayesian interpretation) *If prior probabilities q_1 and q_2 are considered, there is an adjustment $\rho = \log(q_2/q_1)$ should be added to the log-likelihood ratio. This prior adjusted log-likelihood ratio provides a Bayesian interpretation.*

Maximum likelihood estimates. Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$. Assume that $\{h_j\}_{j=1}^n$ are iid random elements from a Gaussian measure on $\mathcal{H}$ with mean m and nonsingular covariance operator Λ. Then, for any $g, f \in \mathcal{H}$, the maximum likelihood estimate for $\langle g, m\rangle_{\mathcal{H}}$ is given by $\langle g, \widehat{m}\rangle_{\mathcal{H}_\kappa}$ with

$$\widehat{m} = \frac{1}{n}\sum_{j=1}^{n} h_j, \tag{3}$$

and the maximum likelihood estimate for $\langle g, \Lambda f\rangle_{\mathcal{H}}$ is given by $\langle g, \widehat{\Lambda} f\rangle_{\mathcal{H}}$ with

$$\widehat{\Lambda} = \frac{1}{n}\sum_{j=1}^{n} (h_j - \widehat{m}) \otimes (h_j - \widehat{m}), \tag{4}$$

where $\otimes$ denotes the tensor product.

Classical multivariate statistical analysis v.s. kernel methods

classical	**kernel methods**
Gaussianity on raw data	Gaussianity on low-dim'l projections of kernel data
classical procedures on raw data	classical procedures on kernel data
FDA, CCA, PCA, d.r., etc.	KFDA, KCCA, KPCA, kernel d.r.
parametric in Euclidean space	nonparametric in Euclidean space parametric in $\mathcal{H}$
statistical optimalities on $(\mathcal{X}, P)$	statistical optimalities on $(\mathcal{H}, P_{\mathcal{H}})$

Three kernel methods for multivariate statistical analysis

- Fisher discriminant analysis $\overset{kernel}{--\longrightarrow}$ KFDA
- Canonical correlation analysis $\overset{kernel}{--\longrightarrow}$ KCCA
- Slice inverse regression $\overset{kernel}{--\longrightarrow}$ KSIR

Softwares: Matlab (canoncorr for CCA, classify for FDA)
Splus, R, SAS

Prepare your data in "kernel form"*. Next, standard statistical softwares areready for use.

*may involve discretization, bases selection and dimension reduction in $\mathcal{H}$

Matlab codes for preparing kernel data

function K = KGaussian(A, B, ν)
% Input
% A: Data A; B: Data B; $\nu = 1/2\sigma^2$
% Output
% K: Gaussian Kernel
% Author: Y.J. Lee

[rowA, colA] = size(A); [rowB, colB] = size(B);
K = zeros(rowA, rowB);

for i = 1:rowA; for j = 1:rowB
    dis=A(i,:)-B(j,:);
    K(i,j) = exp(-ν * dis *dis');
end; end;

Low rank approximation, or dimension reduction

- Optimization: linear or quadratic programs,
- Various **eigen problems**, matrix (or operator) decompostion, sigular value decomposition (matrix or operator).
- Feed $\tilde{K}$, as if it is the data design, into standard statistical packages.
- Nonparametric modelling in $\mathcal{X}$, but parametric notion (in $\mathcal{H}$) for fitting data.

  Extra efforts: prepare kernel data $\tilde{K}$.

When classical procedures work

The FDA, PCA, SIR, dimension reduction, or CCA is good for data

- which are approximately Gaussian (normal), or
- whose distribution is approximately elliptically symmetric.

Why kernel methods work

Kernel map (referring to its low-dimensional projection) can bring the data closer to

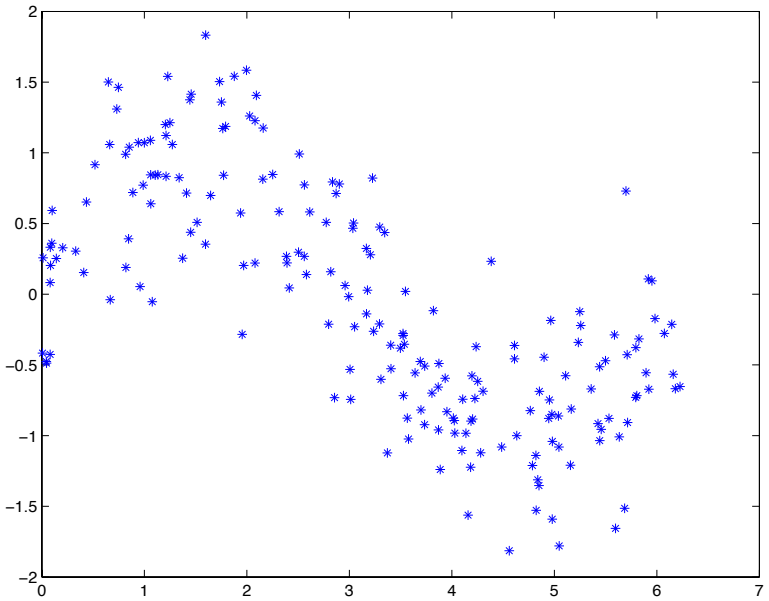
- **normality, and**
- **elliptical symmetry.**

Example 1 *We show that kernel map can bring the data distribution to better elliptical symmetry. Consider a random sample of size 200 consisting of* $\{\mathbf{x}_i = (x_{i1}, \ldots, x_{i5})\}_{i=1}^{200}$, *where*

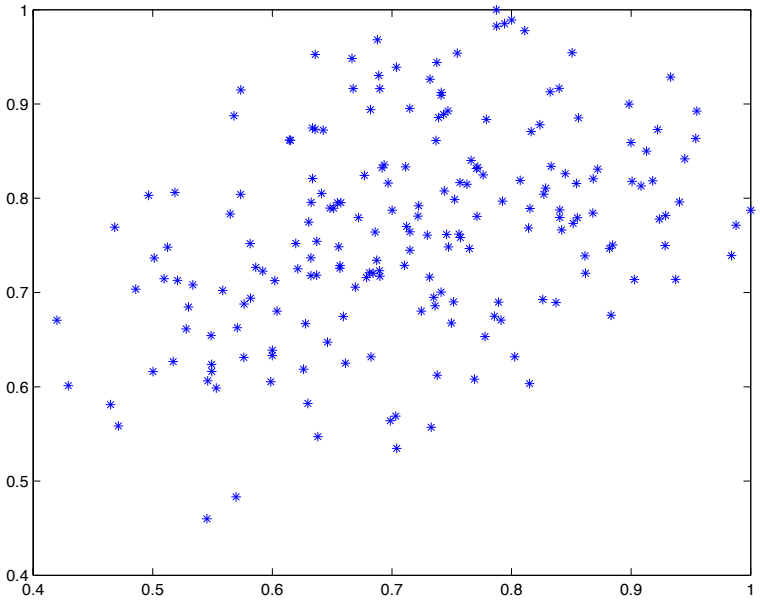
$$x_{i1}, x_{i3}, x_{i4}, x_{i5} \overset{iid}{\sim} \text{uniform}(0, 2\pi)$$

and

$$x_{i2} = \sin(x_{i1}) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \tau^2) \quad with \quad \tau = 0.4$$



Scatter plot (x_1, x_2).



"*Kernel data*" scatter along 2 random directions.