STAT G8325
Gaussian Processes and Kernel Methods
§09: Basic Kernel Methods

John P. Cunningham

Department of Statistics
Columbia University

Outline

Outline

Progress...

Week	Lectures	Content
8	Nov 9, 11	Reproducing kernel Hilbert spaces • [Wah90, ch. 1] (intentionally light reading; work on projects)
9	Nov 16, 18	Basic kernel methods • [SSM98] (intentionally light reading; work on projects)

- HW4 due this Friday.
- The final project will look like:
  - Final project presentation (3-5 minutes, with slides) on Monday 14 December.
  - Final project written submission (8-16 pages, typeset in LaTeX) on Friday 18 December at noon (sharp).

Outline

Reminder: ridge regression

- Consider ℓ_2 penalized least squares regression:

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta^\top x_i)^2 \quad + \quad \rho \|\beta\|_2^2$$

  where $\beta \in \mathbb{R}^d$ is the parameter coefficient.

- We shrink β with Tikhonov regularization to avoid overfitting (large d).
- Regularization seems sensible (and necessary) when in an rkhs $\mathcal{H}$ (vs $\mathbb{R}^d$).
- This choice corresponds to nonlinear or kernel ridge regression:

$$\arg\min_{f} \sum_{i=1}^{n} (y_i - f(x_i))^2 \quad + \quad \rho \|f\|_{\mathcal{H}}^2.$$

Kernel ridge regression?

- For an rkhs $\mathcal{H}$ with rk k, we know:

$$\mathcal{H} = \left\{ f | f = \sum_{i \in \mathbb{N}} \alpha_i k_{x_i} \quad \text{where } \alpha_i \in \mathbb{R}, x_i \in \mathcal{X} \right\},$$

i.e., $\mathcal{H}$ is the span of the representers of evaluation. We also defined inner product:

$$\left\langle \sum_{i \in \mathbb{N}} \alpha_i k_{x_i}, \sum_{j \in \mathbb{N}} \alpha_j k_{x_j} \right\rangle_{\mathcal{H}} = \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{N}} \alpha_i \alpha_j k(x_i, x_j).$$

- Because k is a reproducing kernel, we have $\langle f, k_x \rangle_{\mathcal{H}} = f(x) \quad \forall f \in \mathcal{H}$.
- ...all of which suggests (remember $f(x_i) = \delta_{x_i} f = \langle f, k_{x_i} \rangle_{\mathcal{H}}$):

$$\arg\min_f \quad \sum_{i=1}^n (y_i - f(x_i))^2 \quad + \quad \rho \|f\|_{\mathcal{H}}^2 \qquad \Leftrightarrow$$

$$\arg\min_\alpha \quad \sum_{i=1}^n \left(y_i - \sum_{j \in \mathbb{N}} \alpha_j k(x_j, x_i) \right)^2 \quad + \quad \rho \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{N}} \alpha_i \alpha_j k(x_i, x_j).$$

- That expression parses, but is an infinite dimensional optimization...

Representer theorem (intuitively)

- In the usual ridge case, consider the solution for data $X \in \mathbb{R}^{d\times n}$:

$$
\begin{aligned}
\beta^* &= \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta^\top x_i)^2 \quad + \quad \rho\|\beta\|_2^2 \\
&= (XX^\top + \rho I_d)^{-1} X y \\
& \quad \text{(...and let } X = UDV^\top \text{ be the full svd, i.e. } D \in \mathbb{R}^{n\times n}) \\
&= (UD^2U^\top + \rho I_d)^{-1} UDV^\top y \\
&= UD_\rho^{-2} U^\top UDV^\top y \\
&= U(D_\rho^{-2} D) V^\top y \\
&= U(DD_\rho^{-2}) V^\top y \\
&= UDV^\top V D_\rho^{-2} V^\top y \\
&= X(X^\top X + \rho I_n)^{-1} y \\
&= \sum_{i=1}^{n} \alpha_i x_i, \quad \text{where } \alpha = (X^\top X + \rho I_n)^{-1} y \in \mathbb{R}^n.
\end{aligned}
$$

where w_i are scalar weights on each data point $X = [x_1, ..., x_n]$.

- The ridge regression solution *lives in the span* of the data points.

Representer theorem (intuitively)

- We had: $\beta^* = \arg\min_\beta \sum_{i=1}^n (y_i - \beta^\top x_i)^2 + \rho\|\beta\|_2^2 = X(X^\top X + \rho I_n)^{-1} y.$
- Consider $\mathcal{H}$; replace X with representer 'matrix' $\Phi = [k_{x_1}, ..., k_{x_n}] \in \mathcal{H}^n$.
- An intuitively appealing solution to kernel ridge regression might be:

$$\begin{aligned} f^* &= \arg\min_f \sum_{i=1}^n (y_i - f(x_i))^2 + \rho\|f\|_{\mathcal{H}}^2 \\ &= \Phi(\Phi^\top\Phi + \rho I_n)^{-1} y \\ &= \Phi(K + \rho I_n)^{-1} y \\ &= \sum_{i=1}^n \alpha_i k_{x_i}. \end{aligned}$$

where we have defined $^\top$ as in $\Phi^\top\Phi = \left\{ \left\langle k_{x_i}, k_{x_j} \right\rangle_{\mathcal{H}} \right\}_{i,j \in 1,\ldots,n}$.

- Reducing optimization of $\{\alpha_i\}_{i\in\mathbb{N}} \to$ optimization over $\alpha \in \mathbb{R}^n$.
- Representer theorem in a nutshell: under certain conditions, the solution f^* lives in the n-dimensional linear span of the data's representers of evaluation!

...hence *representer* theorem.

Representer theorem (properly)

- *(Representer theorem) For $f \in \mathcal{H}$, $\mathcal{H}$ an rkhs with rk k, an arbitrary loss function ℓ, a monotonically increasing regularizer g, and data $\{(x_i, y_i)\}_{i \in 1,\ldots,n}$, the program*

$$\arg\min_f \ell\left((y_1, f(x_1)), ..., (y_n, f(x_n))\right) + \rho g\left(\|f\|_{\mathcal{H}}\right)$$

  has minimizer f^* with the form $f^* = \sum_{i=1}^n \alpha_i k_{x_i}$, where $k_{x_i} \in \mathcal{H}$ are the representers of evaluation for k. (Common to define $\phi : \mathcal{X} \to \mathcal{H}, x \to k_x$.)
- This is variously written in a few ways:

$$f = \sum_{i=1}^n \alpha_i k_{x_i} \quad = \quad \sum_{i=1}^n \alpha_i \phi(x_i) \quad = \quad \Phi\alpha \qquad \text{or} \qquad f(x) = k(x, \{x_i\}_i)\alpha$$

.

- Original: [KW71]; generalization: [SHS01]; recently interesting: [DS12].
- We will prove it by considering ℓ and g in turn. For both, consider the orthogonal complement of the span of the data representers:

$$\mathcal{H}_X^\perp = \left\{ f^\perp \in \mathcal{H} \;\middle|\; \left\langle f^\perp, \sum_{i=1}^n \alpha_i k_{x_i} \right\rangle_{\mathcal{H}} = 0 \ , \ \forall \alpha_i \in \mathbb{R} \right\}.$$

Representer theorem proof (loss function ℓ)

- Assume that the solution $f^* \in \mathcal{H}$ is arbitrary. Then:

$$f^* = \sum_{i=1}^{n} \alpha_i k_{x_i} + f^{\perp}, \qquad f^{\perp} \in \mathcal{H}_X^{\perp}, \alpha_i \in \mathbb{R}.$$

- Noting that ℓ depends only on $f(x_j)$ for $j \in 1, ..., n$, we see:

$$\begin{aligned} f^*(x_j) &= \langle f, k_{x_j} \rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^{n} \alpha_i k_{x_i} + f^{\perp}, k_{x_j} \right\rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^{n} \alpha_i k_{x_i}, k_{x_j} \right\rangle_{\mathcal{H}} + \left\langle f^{\perp}, k_{x_j} \right\rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^{n} \alpha_i k_{x_i}, k_{x_j} \right\rangle_{\mathcal{H}} \\ &= \sum_{i=1}^{n} \alpha_i k(x_i, x_j). \end{aligned}$$

- Thus, the loss function ℓ is invariant to any part of $f^* \notin \operatorname{span}(k_{x_1}, ..., k_{x_n})$.

Representer theorem proof (regularizer g)

- Again assume that the solution $f^* \in \mathcal{H}$ is arbitrary. Then:

$$f^* = \sum_{i=1}^{n} \alpha_i k_{x_i} + f^{\perp}, \qquad f^{\perp} \in \mathcal{H}_X^{\perp}, \alpha_i \in \mathbb{R}.$$

- Then:

$$\begin{aligned} g\left(\|f\|_{\mathcal{H}}\right) &= g\left(\left\|\sum_{i=1}^{n} \alpha_i k_{x_i} + f^{\perp}\right\|_{\mathcal{H}}\right) \\ &= g\left(\left(\left\|\sum_{i=1}^{n} \alpha_i k_{x_i}\right\|_{\mathcal{H}}^2 + \left\|f^{\perp}\right\|_{\mathcal{H}}^2\right)^{\frac{1}{2}}\right) \\ &\geq g\left(\left\|\sum_{i=1}^{n} \alpha_i k_{x_i}\right\|_{\mathcal{H}}\right). \end{aligned}$$

- Since ℓ does not depend on $f^{\perp}$ and g is monotonically increasing, the minimizer must have $f^{\perp} = 0$.
- Thus it is proven that $f^* \in \text{span}(k_{x_1}, ..., k_{x_n})$; that is, $f^* = \sum_{i=1}^{n} \alpha_i k_{x_i}$.

Outline

Kernel ridge regression

- For $f \in \mathcal{H}$, the rkhs with rk k, we seek the nonlinear regressor:

$$f^* = \arg\min_f \sum_{i=1}^n (y_i - f(x_i))^2 \quad + \quad \rho\|f\|_{\mathcal{H}}^2$$

- For $\rho > 0$, the representer theorem holds. Thus $f = \sum_{j=1}^n \alpha_j k_{x_j}$, so:

$$\begin{aligned}
f^* &= \arg\min_f \sum_{i=1}^n (y_i - f(x_i))^2 \quad + \quad \rho\|f\|_{\mathcal{H}}^2 \\
&= \arg\min_\alpha \sum_{i=1}^n \left(y_i - \left\langle \sum_{j=1}^n \alpha_j k_{x_j}, k_{x_i} \right\rangle_{\mathcal{H}} \right)^2 \quad + \quad \rho \left\| \sum_{j=1}^n \alpha_j k_{x_j} \right\|_{\mathcal{H}}^2 \\
&= \arg\min_\alpha \sum_{i=1}^n \left(y_i - \sum_{j=1}^n \alpha_j k(x_i, x_j) \right)^2 \quad + \quad \rho \sum_{i=1} \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j) \\
&= \arg\min_\alpha \|y - K\alpha\|_2^2 + \rho\alpha^\top K\alpha. \\
&= \arg\min_\alpha \alpha^\top (K^2 + \rho K)\alpha - 2\alpha^\top K y. \\
\Rightarrow &\quad \alpha^* = (K + \rho I)^{-1} y.
\end{aligned}$$

- Thus $f^* = \sum_{j=1}^n \alpha_j^* k_{x_j} = \Phi(K + \rho I)^{-1} y$, as intuitively expected.

Kernel ridge regression: a familiar form

- For $f \in \mathcal{H}$, the rkhs with rk k, we have that the nonlinear regressor:

$$f^* = \arg\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 \quad + \quad \rho \|f\|_{\mathcal{H}}^2$$

  has form $f^* = \sum_{j=1}^{n} \alpha_j^* k_{x_j} = \Phi(K + \rho I)^{-1} y$.

- Prediction at x is $f^*(x) = \left\langle \sum_{j=1}^{n} \alpha_j^* k_{x_j}, k_x \right\rangle_{\mathcal{H}} = K_{xf}(K_{ff} + \rho I)^{-1} y$.

- This is precisely our usual form for the gp posterior mean:

$$E(f(x)|X, y) = K_{xf}(K_{ff} + \rho I)^{-1} y = K_{xf} K_{yy}^{-1} y.$$

- Thus gp inference is kernel ridge regression with a bayesian interpretation.
- Kernel ridge regression is very widely used. Often no mention of gp at all.
- Differences between kernel methods and gp methods seem largely cultural.
- While true, there is a surprising difference (lest we get too comfortable).

Outline

Kernel ridge regression and gp

- We just saw that krr and gp regression give the same results in the sense that, given data $X = [x_1, ..., x_n]$ and a rkhs $\mathcal{H}$ with rk k, the krr prediction and gp posterior mean of a point x are the same:

$$f_{\alpha^*}(x) = E(f(x)|X, y) = K_{xf}(K_{ff} + \rho I)^{-1}y = K_{xf}K_{yy}^{-1}y.$$

- Kernel ridge regression optimizes over all functions $f \in \mathcal{H}$, by definition.
- It is then tempting to think that a draw f from a gp with kernel k will be a point in k's rkhs $\mathcal{H}$, i.e., $f \sim \mathcal{GP}(0, k) \in \mathcal{H}$.
- The intuition:
  - Riesz $\Rightarrow f(x) = \langle f, k_x \rangle_{\mathcal{H}}$, so a gp seems to be an iid gaussian weighted sum (with weights f^i) of basis elements k_x^i.
  - Take linear regression: $f(x) = \beta^\top x$, where $\beta \sim \mathcal{N}(0, \rho I)$.
  - or some arbitrary polynomial on $\mathbb{R}$: $f(x) = \sum_{k=1}^K \beta_k x^k$, again with $\beta \sim \mathcal{N}(0, \rho I)$.
- This intuition is **false** when $\mathcal{H}$ is infinite dimensional (sadly).

RKHS of a GP draw [Wah90, ch. 1]

- Let k be a Mercer kernel, so that $k(x, x') = \sum_{i \in \mathbb{N}} \lambda_i \phi_i(x) \phi_i(x')$, where $\{\phi_i\}$ forms an orthonormal basis of L_2.
- The Karhunen-Loeve transform tells us $f \sim \mathcal{GP}(0, k)$ has expansion: $f(x) = \sum_{i \in \mathbb{N}} z_i \phi_i(x)$, where the variables z_i are independent and normal.
- The z_i are the projection onto that eigenfunction $z_i = \int f(x) \phi_i(x) dx$; thus:

$$
\begin{aligned}
E(z_i) &= E\left(\int f(x) \phi_i(x) dx \right) = \int E(f(x)) \phi_i(x) dx = 0. \\
E(z_i z_j) &= E\left(\int \int f(x) f(x') \phi_i(x) \phi_j(x') dx dx' \right) \\
&= \int \int E(f(x) f(x')) \phi_i(x) \phi_j(x') dx dx' \\
&= \int \int k(x, x') \phi_i(x) \phi_j(x') dx dx' \\
&= \lambda_i \mathbb{1}(i = j).
\end{aligned}
$$

- Here again is that tempting intuition: a gp is just a sequence of weighted independent $\mathcal{N}(0, \lambda_i)$ variables, which looks like (but is not) it is in $\mathcal{H}$.

RKHS of a GP draw [Wah90, ch. 1]

- Consider $f_M(x) = \sum_{i=1}^{M} z_i \phi_i(x)$. $f_M \to f$ in quadratic mean:

$$
\begin{aligned}
E(|f_M(x) - f(x)|^2) &= E\left(\left|\sum_{i=M+1}^{\infty} z_i \phi_i(x)\right|^2\right) \\
&= \sum_{i=M+1}^{\infty} \lambda_i \phi_i^2(x) \\
&\to 0.
\end{aligned}
$$

- However, $\mathcal{H}$ does not contain the limit of this sequence, which is hinted at by the fact that its expectation:

$$
\begin{aligned}
E\left(\|f_M\|_{\mathcal{H}}^2\right) &= E\left(\sum_{i=1}^{M} \frac{z_i^2}{\lambda_i}\right) \\
&= M \\
&\to \infty.
\end{aligned}
$$

- The above is not a proof; see [Kal70, Dri73, LP$^+$73, Háj62, LB01].
- Nonetheless, it is the case that for a rkhs $\mathcal{H}$ with rk k, a gp draw $f \sim \mathcal{GP}(0, k)$ is not (a.s.) a member of $\mathcal{H}$.

Outline

Kernel mean estimation

- Consider an rkhs $\mathcal{H}$ with rk $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and representers $\phi : \mathcal{X} \to \mathbb{R}$.
- (can be a bit more generic: the rk machinery is not explicitly needed)
- $\mathcal{H}$ offers a sensible notion of (squared) distance between points:

$$\begin{aligned} d^2_{\mathcal{H}}(x, x') &= \|\phi(x) - \phi(x')\|^2_{\mathcal{H}} \\ &= \langle \phi(x), \phi(x) \rangle_{\mathcal{H}} - 2 \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} + \langle \phi(x'), \phi(x') \rangle_{\mathcal{H}} \\ &= k(x, x) - 2k(x, x') + k(x', x'). \end{aligned}$$

- Given a probability distribution P, an object of regular interest is:

$$\mu_P = \arg\min_{\mu} \int_{\mathcal{X}} \|\phi(x) - \mu\|^2_{\mathcal{H}} \; dP(x)$$

...e.g. in kernel PCA (upcoming).

- This object looks like the usual expected value/mean...

Kernel mean estimation

- This object looks like the usual expected value/mean:

$$\begin{aligned}
\mu_P &= \arg\min_{\mu} \int_{\mathcal{X}} \|\phi(x) - \mu\|^2_{\mathcal{H}} \, dP(x) \\
&= \arg\min_{\mu} \langle \mu, \mu \rangle_{\mathcal{H}} - 2E_P\left(\langle \mu, \phi(x) \rangle_{\mathcal{H}}\right). \\
&= \arg\min_{\mu} \langle \mu, \mu \rangle_{\mathcal{H}} - 2\left\langle \mu, E_P(\phi(x)) \right\rangle_{\mathcal{H}}. \\
&\Rightarrow \quad \mu_P = E_P(\phi(x)).
\end{aligned}$$

- Similarly we have the finite case $\hat{\mu}_P = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$.
- Notice that both $\mu_P, \hat{\mu}_P \in \mathcal{H}$. Sometimes useful, sometimes not useful...
- What point $x_\mu \in \mathcal{X}$ is the pre-image of μ_P, i.e. $\mu_P = \phi(x_\mu)$?
- This is called the *pre-image problem* (for kernel mean estimation).

Pre-image problem

- The pre-image problem, for finite data and some statistic S, is:

$$x_\mu = \{x \in \mathcal{X} : \phi(x) = S\left(\phi(x_1), ..., \phi(x_n)\right)\}.$$

- The pre-image problem is useful:
  - Consider the mean $S\left(\phi(x_1), ..., \phi(x_n)\right) = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$.
  - We have seen kernels on interesting spaces (graphs, rankings, etc.).
  - We do not know how to sensibly average n graphs.
  - Kernel $\rightarrow$ a distance metric; pre-image $\rightarrow$ the mean under that distance.
  - Also useful in simple spaces ($\mathbb{R}^d$) $\rightarrow$ consider distance in $\mathcal{H}$ rather than $\mathbb{R}^d$.

- The pre-image problem is a problem:
  - $\phi : \mathcal{X} \rightarrow \mathcal{H}$ is not injective. Example: $\phi : \mathbb{R} \rightarrow \mathbb{R}_+, x \rightarrow x^2$.
  - $\phi : \mathcal{X} \rightarrow \mathcal{H}$ is not surjective. Note $\Phi = \{\phi(x) \in \mathcal{H} \ \forall x \in \mathcal{X}\} \subset \mathcal{H}$.
  - Thus x_μ such that $\mu_P = \phi(x_\mu)$ generally exists only in trivial circumstances.

Pre-image problem

- Common approach: optimize over $\mathcal{X}$ through the mapping ϕ.
- That is, apply the constraint set $\Phi = \{\phi(x) \in \mathcal{H} \ \forall x \in \mathcal{X}\} \subset \mathcal{H}$.

Convex? No.

Bounded? Yes.

$$\begin{aligned} \mu_P &= \arg\min_{\mu} \frac{1}{n} \sum_{i=1}^{n} \|\phi(x_i) - \mu\|_{\mathcal{H}}^2 \qquad \text{for } x_i \sim P \\ &\rightarrow \\ x_\mu &= \arg\min_{x} \frac{1}{n} \sum_{i=1}^{n} \|\phi(x_i) - \phi(x)\|_{\mathcal{H}}^2 \\ x_\mu &= \arg\min_{x} \frac{1}{n} \sum_{i=1}^{n} k(x_i, x_i) - 2k(x_i, x) + k(x, x). \end{aligned}$$

- Not the unconstrained optimum in most cases (restating ϕ is not invertible).
- Optimization over $\mathcal{X}$ is also often difficult (gradients on ranking space?).
- The pre-image problem is not solved in any satisfactory way...

Outline

Principal components analysis

- PCA produces an r dimensional orthogonal projection by:

$$
\begin{aligned}
\begin{bmatrix} v_1 & \dots & v_r \end{bmatrix} &= \arg \min_{v_\ell^\top v_j = \mathbb{1}(\ell = j)} \sum_{i=1}^{n} \left\| \bar{x}_i - \sum_{j=1}^{r} v_j v_j^\top \bar{x}_i \right\|_2^2 \\
&= \arg \max_{v_\ell^\top v_j = \mathbb{1}(\ell = j)} \sum_{j=1}^{r} v_j^\top \bar{X} \bar{X}^\top v_j
\end{aligned}
$$

  where $\bar{X} = X - \frac{1}{n} X 1 1^\top \in \mathbb{R}^{d \times n}$ is the centered data matrix.
- Solution is the first r eigenvectors v_j of $\bar{X}\bar{X}^\top$: $\bar{X}\bar{X}^\top v_j = \lambda_j v_j$.
- Observations:
  - The loss ℓ operates only on inner products $\bar{X}^\top v_j$.
  - The constraint is equivalent (up to a normalizer) with any increasing $g(\|v_j\|)$.
  - $\bar{X}\bar{X}^\top v_j = \lambda_j v_j$ means $v_j \in \mathrm{span}(x_1, \dots, x_n)$.
- Thus we suspect that PCA can be readily 'kernelized', and that the representer theorem will hold. (cache this remark...)

Kernel eigenvalue problem

- [SSM98] calls kernel PCA (kpca) the solution to $\lambda_j v_j = \bar{C} v_j$, where

$$\bar{C} = \frac{1}{n} \sum_{i=1}^{n} \left(\phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j) \right) \left(\phi(x_i) - \frac{1}{n} \sum_{j=1}^{n} \phi(x_j) \right)^{\top},$$

  the covariance 'matrix' in $\mathcal{H}$.

  This 'outer product' notation is frustrating but common. All steps are legitimate, just loosely written.

- Now the eigenvector $v_j \in \mathcal{H}$, and we know (assert) that v_j obeys the representer theorem (lies in the span); thus: $v_j = \sum_{i=1}^{n} \alpha_j^i \bar{\phi}(x_i)$.

- We now use this representation of v_j in the quadratic form $v_j^\top \bar{C} v_j$, the Rayleigh quotient whose solutions form the eigenvector basis.

Kernel eigenvalue problem

- The kernel Rayleigh quotient $v_j^\top \bar{C} v_j$:

$$\begin{aligned} v_j^\top \bar{C} v_j &= \left(\sum_{i=1}^{n} \alpha_j^i \bar{\phi}(x_i)\right)^\top \left(\frac{1}{n}\sum_{i=1}^{n} \bar{\phi}(x_i)\bar{\phi}(x_i)^\top\right)\left(\sum_{i=1}^{n} \alpha_j^i \bar{\phi}(x_i)\right) \\ &= \frac{1}{n}\alpha^\top \left(\left\{\langle \bar{\phi}(x_i), \bar{\phi}(x_j)\rangle_{\mathcal{H}}\right\}_{i,j\in 1,\ldots,n} \left\{\langle \bar{\phi}(x_i), \bar{\phi}(x_j)\rangle_{\mathcal{H}}\right\}_{i,j\in 1,\ldots,n}\right)\alpha \\ &= \frac{1}{n}\alpha_j^\top \bar{K}^2 \alpha_j . \end{aligned}$$

...where this last equation is a properly defined quadratic form of a finite dimensional matrix.

- Centering operations in $\mathcal{H}$ behave as expected:

$$\begin{aligned} \bar{C} v_j &= \left(\frac{1}{n}\bar{\Phi}\bar{\Phi}^\top\right)\bar{\Phi}\alpha_j \\ &= \frac{1}{n}\bar{\Phi}\left(\Phi - \frac{1}{n}\Phi 1 1^\top\right)^\top \left(\Phi - \frac{1}{n}\Phi 1 1^\top\right)\alpha_j \\ &= \frac{1}{n}\bar{\Phi}\left(K - \frac{1}{n}1 1^\top K - \frac{1}{n}K 1 1^\top + \frac{1}{n^2}1 1^\top K 1 1^\top\right)\alpha_j \\ &= \frac{1}{n}\bar{\Phi}\bar{K}\alpha_j , \end{aligned}$$

...and thus $\bar{K}$ is often called the centered kernel matrix.

Kernel PCA

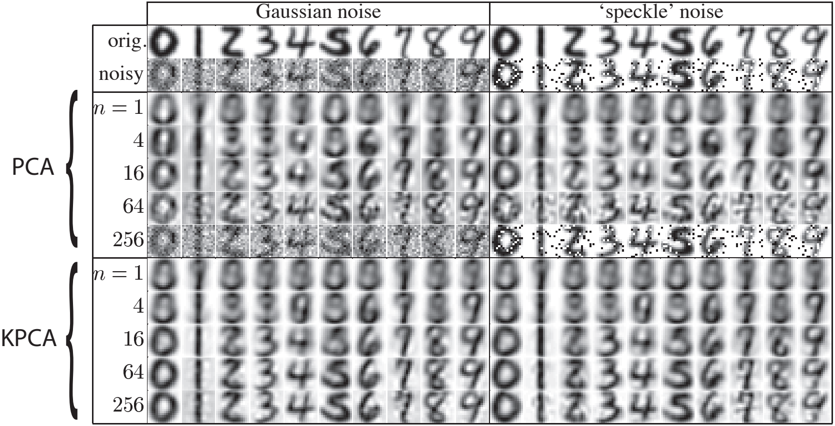
- We know the eigenvectors (functions) $v_j = \bar{\Phi}\alpha_j = \sum_{i=1}^{n} \alpha_j^i \bar{\phi}(x_i) \in \mathcal{H}$.
- We know $\lambda_j v_j = \bar{C} v_j = \frac{1}{n} \bar{\Phi} \bar{K} \alpha_j \quad \in \mathcal{H}$.
- $v_j \in$ span $\{\phi(x_1), ..., \phi(x_n)\} \rightarrow$ equivalently consider projection onto $\bar{\Phi}$:

$$\begin{aligned} \bar{\Phi}^\top \lambda_j v_j &= \bar{\Phi}^\top \bar{C} v_j \\ \bar{\Phi}^\top \left(\lambda_j \bar{\Phi}\alpha\right) &= \bar{\Phi}^\top \left(\frac{1}{n} \bar{\Phi} \bar{K} \alpha_j\right) \\ \lambda_j \bar{K} \alpha_j &= \frac{1}{n} \bar{K}^2 \alpha_j. \end{aligned}$$

- Thus $\alpha_j \in \mathbb{R}^n$ an eigenvector of $\bar{K} \Rightarrow v = \bar{\Phi}\alpha_j$ is an eigenfunction in $\mathcal{H}$.
- $\alpha_i^\top \alpha_j = \mathbb{1}(i = j)$ because they are eigenvectors of $\bar{K}$ (symmetric and real).
- Accordingly, $\langle v_i, v_j \rangle_{\mathcal{H}} = \alpha_i \bar{K} \alpha_j = \mathbb{1}(i = j)$ also, so v_i are orthogonal.
- For v_j to be orthonormal, $\|\alpha_j\| = \frac{1}{\sqrt{n\lambda_j}}$, since we had $n\lambda_j \alpha_j = \bar{K}\alpha_j$.
- KPCA then projects x' onto v_j as $\langle v_j, \phi(x') \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \alpha_j^i k(x_i, x')$.

Kernel PCA

- KPCA is widely used as a compression or visualization tool.



- We were fast and loose (in the common way that linear dimensionality reduction $\leftrightarrow$ eigenproblem) with the representer theorem. Let's revisit that...

Outline

KPCA through the lens of the Stiefel manifold

- PCA produces an r dimensional orthogonal projection by:

$$\begin{aligned}
\begin{bmatrix} v_1 & \dots & v_r \end{bmatrix} &= \arg \max_{v_\ell^\top v_j = \mathbb{1}(\ell = j)} \sum_{j=1}^{r} v_j^\top \bar{X} \bar{X}^\top v_j \\
V &= \arg\max \quad tr\left(V^\top \bar{X}\bar{X}^\top V\right) \\
& \quad\ \text{subject to } \ V \in \mathsf{St}(\mathbb{R}^d, r),
\end{aligned}$$

  where $\mathsf{St}(\mathbb{R}^d, r) = \left\{ M \in \mathbb{R}^{d \times r} : M^\top M = I_d \right\}$, the Stiefel manifold of orthonormal r-frames in d dimensions [CG15].

- The Stiefel manifold exists similarly in a (separable) Hilbert space:

$$\mathsf{St}(\mathcal{H}, r) = \left\{ [m_1, ..., m_r] \in \mathcal{H}^r : \langle m_i, m_j \rangle_{\mathcal{H}} = \mathbb{1}(i = j) \right\}.$$

- The underlying problem of KPCA is then:

$$\begin{aligned}
V &= \arg\max \quad tr\left(V^\top \bar{C} V\right) \\
& \quad\ \text{subject to } \ V \in \mathsf{St}(\mathcal{H}, r).
\end{aligned}$$

- This **does not** obey the representer theorem $\rightarrow$ two neat implications...

Implication 1: KPCA *asserts* the representer theorem

- Decompose $V \in \mathsf{St}(\mathcal{H}, r)$ as:

$$V = V_{\mathcal{X}} + V^{\perp} = \Phi A + \Phi^{\perp} B$$

where $\Phi^{\perp}$ is a basis of the orthogonal complement of Φ (ignoring centering), $A = \{\alpha_j^i\}_{i=1,\dots,n; j=1,\dots,r}$ as previously, and $B = \{\beta_j^i\}_{i \in \mathbb{N}; j=1,\dots,r}$ similarly.

- Then $V \in \mathsf{St}(\mathcal{H}, r) \Rightarrow V^{\top} V = I_r$, so:

$$\begin{aligned} V^{\top} V &= \left(\Phi A + \Phi^{\perp} B\right)^{\top} \left(\Phi A + \Phi^{\perp} B\right) \\ &= A^{\top} K A + \left(\Phi^{\perp} B\right)^{\top} \Phi^{\perp} B. \\ &= A^{\top} K A + \Psi^{\perp}. \end{aligned}$$

- Notice $\Psi^{\perp}$ is positive semidefinite:

$$\begin{aligned} v^{\top} \Psi^{\perp} v &= \sum_{k=1}^{r} \sum_{\ell=1}^{r} v_k v_{\ell} \Psi^{\perp}_{k\ell} \\ &= \sum_{k=1}^{r} \sum_{\ell=1}^{r} v_k v_{\ell} \left\langle \sum_{i=1}^{\infty} \beta_{i,k} \phi_i^{\perp}, \sum_{j=1}^{\infty} \beta_{j,\ell} \phi_j^{\perp} \right\rangle_{\mathcal{H}} \\ &= \left\langle \sum_{k=1}^{r} \sum_{i=1}^{\infty} v_k \beta_{i,k} \phi_i^{\perp}, \sum_{\ell=1}^{r} \sum_{j=1}^{\infty} v_{\ell} \beta_{j,\ell} \phi_j^{\perp} \right\rangle_{\mathcal{H}} \\ &= \left\| \sum_{k=1}^{r} \sum_{i=1}^{\infty} v_k \beta_{i,k} \phi_i^{\perp} \right\|_{\mathcal{H}}^{2} \geq 0. \end{aligned}$$

Implication 1: KPCA *asserts* the representer theorem

- If $V \in \mathsf{St}(\mathcal{H}, r) \Rightarrow V^\top V = I_r$ and $V^\top V = A^\top K A + \Psi^\perp$ for $\Psi^\perp \succeq 0$,
- Then the original KPCA problem:

$$\begin{array}{rcll} V & = & \arg\max & tr\left(V^\top \bar{C} V\right) \\ & & \text{subject to} & V \in \mathsf{St}(\mathcal{H}, r). \end{array}$$

  is equivalent to:

$$\begin{array}{rcll} V & = & \arg\max & tr\left(V^\top \bar{C} V\right) \\ & & \text{subject to} & \sigma_1(V) \leq 1 \\ & & & V \in \mathsf{span}(\Phi), \end{array}$$

  where $\sigma_1(V) \leq 1 \Rightarrow V \in \left\{M \in \mathcal{H}^r : M^\top M \preceq I_r\right\}$ (spectral norm unit ball).

  In fact, the spectral norm unit ball is the convex hull of the corresponding Stiefel manifold.

- This is a (very) different problem than the KPCA solution.
- So what happened?

Implication 1: KPCA *asserts* the representer theorem

- Compare

$$
\begin{aligned}
V \quad &= \quad \arg\max \quad tr\left(V^\top \bar{C} V\right) \\
&\qquad \text{subject to} \;\; V \in \mathsf{St}(\mathcal{H}, r). \\
&\Leftrightarrow \\
V \quad &= \quad \arg\max \quad tr\left(V^\top \bar{C} V\right) \\
&\qquad \text{subject to} \;\; \sigma_1(V) \leq 1 \\
&\qquad\qquad\qquad\quad V \in \mathsf{span}(\Phi),
\end{aligned}
$$

...with...

$$
\begin{aligned}
V \quad &= \quad \arg\max \quad tr\left(V^\top \bar{C} V\right) \\
&\qquad \text{subject to} \;\; V \in \mathsf{St}(\mathcal{H}, r). \\
&\qquad\qquad\qquad\quad V \in \mathsf{span}(\Phi).
\end{aligned}
$$

- This latter problem *asserts* the representer theorem, and results in the familiar KPCA solution.
- An outcome of our earlier, seemingly harmless claim, "just like how the eigenvectors of $XX^\top$ are in the span of the data X, the eigenvectors (functions) of $\Phi\Phi^\top$ are also in the span of Φ."

Implication 2: KPCA projects the cholesky factors

- We'll use regular KPCA (i.e., assert the representer theorem):

$$\begin{aligned} V \quad = \quad & \arg\max \quad tr\left(V^\top \bar{C} V\right) \\ & \text{subject to} \quad V \in \mathsf{St}(\mathcal{H}, r). \\ & \qquad\qquad\quad V \in \mathsf{span}(\Phi). \end{aligned}$$

- Then, recalling that $V^\top V = I_r$, we see

$$\begin{aligned} V^\top V &= A^\top \Phi^\top \Phi A \\ &= A^\top K A \\ &= M^\top C^{-T} K C^{-1} M \\ &= I. \end{aligned}$$

  where $K = C^\top C$ is the Cholesky decomposition, and $M \in \mathsf{St}(\mathbb{R}^n, r)$.

- That is, $A^\top K A = I \Rightarrow A$ must factor as $C^{-1}M$ for some $M : M^\top M = I$.
- Then the projection $V^\top \Phi = M^\top C^{-\top} \Phi^\top \Phi = M^\top C$.
- In short, KPCA is just an orthogonal projection of the Cholesky factors C.

Outline

References

[CG15] John P Cunningham and Zoubin Ghahramani.
Linear dimensionality reduction: Survey, insights, and generalizations.
Journal of Machine Learning Research, 2015.

[Dri73] Michael F Driscoll.
The reproducing kernel hilbert space structure of the sample paths of a gaussian process.
Probability Theory and Related Fields, 26(4):309–316, 1973.

[DS12] Francesco Dinuzzo and Bernhard Schölkopf.
The representer theorem for hilbert spaces: a necessary and sufficient condition.
In *Advances in neural information processing systems*, pages 189–196, 2012.

[Háj62] Jaroslav Hájek.
On linear statistical problems in stochastic processes.
Czechoslovak Mathematical Journal, 12(3):404–444, 1962.

[Kal70] Gopinath Kallianpur.
The role of reproducing kernel hilbert spaces in the study of gaussian processes.
Advances in probability and related topics, 2:49–83, 1970.

[KW71] George Kimeldorf and Grace Wahba.
Some results on tchebycheffian spline functions.
Journal of mathematical analysis and applications, 33(1):82–95, 1971.

[LB01] Milan Lukić and Jay Beder.
Stochastic processes with sample paths in reproducing kernel hilbert spaces.
Transactions of the American Mathematical Society, 353(10):3945–3969, 2001.

[LP+73] Raoul Le Page et al.
Subgroups of paths and reproducing kernels.
The Annals of Probability, 1(2):345–347, 1973.

[SHS01] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola.
A generalized representer theorem.
In *Computational learning theory*, pages 416–426. Springer, 2001.

[SSM98] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller.
Nonlinear component analysis as a kernel eigenvalue problem.
Neural computation, 10(5):1299–1319, 1998.

[Wah90] Grace Wahba.
Spline models for observational data, volume 59.
Siam, 1990.