Kernel-Based Contrast Functions for Sufficient Dimension Reduction

Michael I. Jordan
Departments of Statistics and EECS
University of California, Berkeley

Joint work with Kenji Fukumizu and Francis Bach

Outline

- Introduction
  – dimension reduction and conditional independence
- Conditional covariance operators on RKHS
- Kernel Dimensionality Reduction for regression
- Manifold KDR
- Summary

Sufficient Dimension Reduction

- Regression setting: observe *(X,Y)* pairs, where the covariate *X* is high-dimensional
- Find a (hopefully small) subspace *S* of the covariate space that retains the information pertinent to the response *Y*
- *Semiparametric formulation*: treat the conditional distribution *p(Y | X)* nonparametrically, and estimate the parameter *S*

Perspectives

- Classically the covariate vector X has been treated as ancillary in regression
- The sufficient dimension reduction (SDR) literature has aimed at making use of the randomness in X (in settings where this is reasonable)
- This has generally been achieved via inverse regression
  - at the cost of introducing strong assumptions on the distribution of the covariate X
- We'll make use of the randomness in X without employing inverse regression

Dimension Reduction for Regression

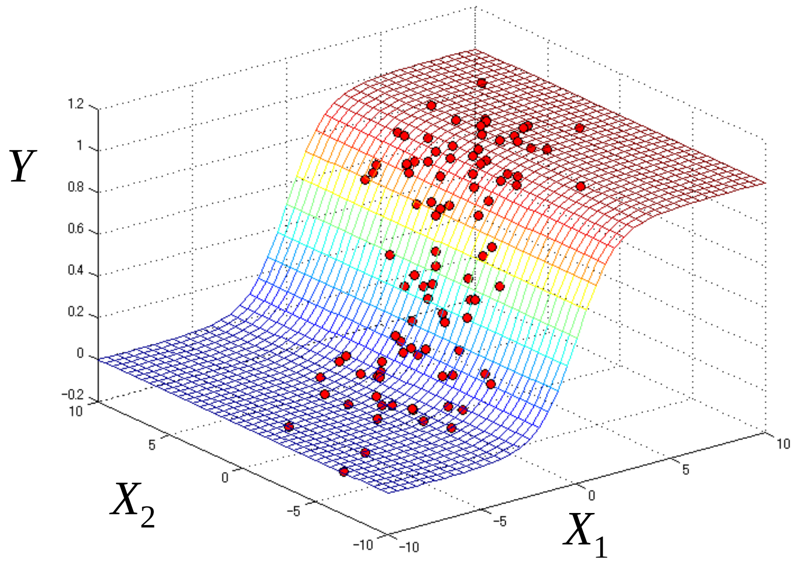
- Regression: $p(Y \mid X)$

  Y : response variable,

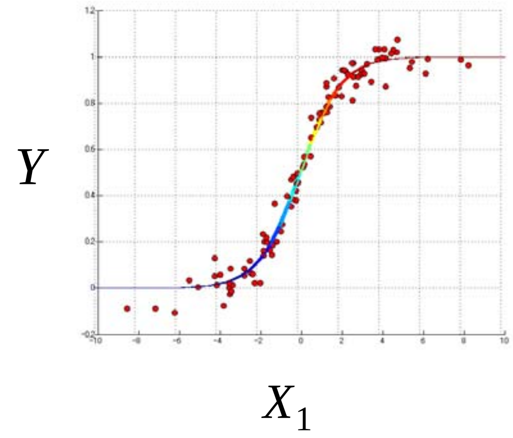
  $X = (X_1, \ldots, X_m)$: m-dimensional covariate

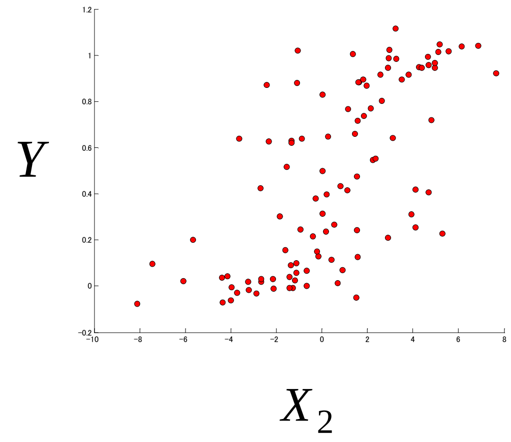
- Goal: Find the central subspace, which is defined via:

$$p(Y \mid X) = \tilde{p}(Y \mid b_1^T X, \ldots, b_d^T X) \quad \left(= \tilde{p}(Y \mid B^T X) \right)$$



$$Y = \frac{1}{1+\exp(-X_1)} + N(0;\,0.1^2)$$





central subspace = X_1 axis

Some Existing Methods

- Sliced Inverse Regression (SIR, Li 1991)
  - PCA of $\mathrm{E}[X|Y]$ → use slice of Y
  - Elliptic assumption on the distribution of X
- Principal Hessian Directions (pHd, Li 1992)
  - Average Hessian $\Sigma_{yxx} \equiv E[(Y-\bar{Y})(X-\bar{X})(X-\bar{X})^T]$ is used
  - If X is Gaussian, eigenvectors gives the central subspace
  - Gaussian assumption on X. Y must be one-dimensional
- Projection pursuit approach (e.g., Friedman et al. 1981)
  - Additive model $E[Y|X] = g_1(b_1^T X) + \ldots + g_d(b_\mathrm{d}^T X)$ is used
- Canonical Correlation Analysis (CCA) / Partial Least Squares (PLS)
  - Linear assumption on the regression
- Contour Regression (Li, Zha & Chiaromonte, 2004)
  - Elliptic assumption on the distribution of X

Dimension Reduction and Conditional Independence

- $(U, V) = (B^T X, C^T X)$

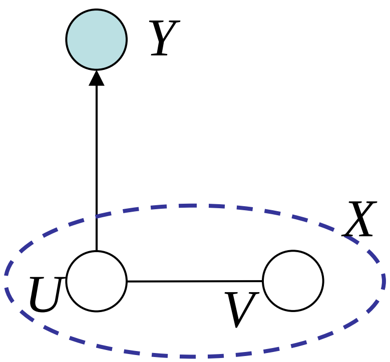
  where C: m x $(m-d)$ with columns orthogonal to B

- B gives the projector onto the central subspace

  $\Leftrightarrow \quad p_{Y|X}(y \mid x) = p_{Y|U}(y \mid B^T x)$

  $\Leftrightarrow \quad p_{Y|U,V}(y \mid u, v) = p_{Y|U}(y \mid u)$ for all y, u, v

  $\Leftrightarrow$ Conditional independence $Y \perp\!\!\!\perp V \mid U$



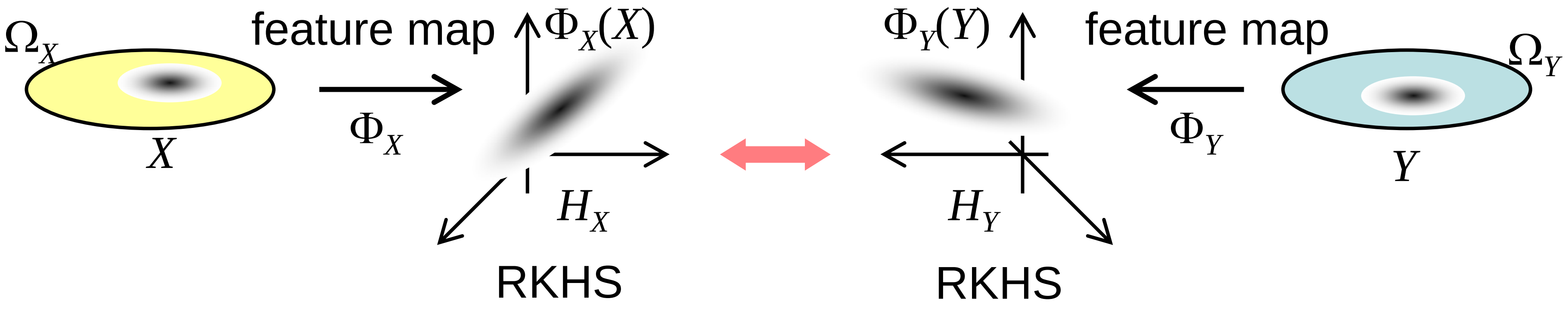
- Our approach: *Characterize conditional independence*

Outline

- Introduction
  – dimension reduction and conditional independence
- Conditional covariance operators on RKHS
- Kernel Dimensionality Reduction for regression
- Manifold KDR
- Summary

Reproducing Kernel Hilbert Spaces

- "Kernel methods"
  - RKHS's have generally been used to provide basis expansions for regression and classification (*e.g.*, support vector machine)
  - *Kernelization*: map data into the RKHS and apply linear or second-order methods in the RKHS
  - But RKHS's can also be used to characterize independence and conditional independence



Positive Definite Kernels and RKHS

- **Positive definite kernel (p.d. kernel)**

  $k : \Omega \times \Omega \to \mathbf{R}$

  k is positive definite if $k(x,y) = k(y,x)$ and for any $n \in \mathbf{N},\ x_1, \ldots x_n \in \Omega$ the matrix $\left(k(x_i, x_j)\right)_{i,j}$ (Gram matrix) is positive semidefinite.

  – Example: Gaussian RBF kernel $k(x,y) = \exp\left(-\|x-y\|^2 / \sigma^2\right)$

- **Reproducing kernel Hilbert space (RKHS)**

  k: p.d. kernel on Ω

  ⇨ $\exists H$: reproducing kernel Hilbert space (RKHS)

  1) $k(\cdot, x) \in H$ for all $x \in \Omega$.
  2) $\mathrm{Span}\left\{k(\cdot,x) \mid x \in \Omega\right\}$ is dense in H.
  3) $\left\langle k(\cdot, x), f \right\rangle_H = f(x)$ (reproducing property)

Functional data

$$\Phi : \Omega \to H, \quad x \mapsto k(\cdot, x) \qquad i.e. \quad \Phi(x) = k(\cdot, x)$$

Data: $X_1, \ldots, X_N$ → $\Phi_X(X_1), \ldots, \Phi_X(X_N)$: functional data

Why RKHS?

- By the reproducing property, computing the inner product on RKHS is easy:

$$\langle \Phi(x), \Phi(y) \rangle = k(x, y)$$

$$f = \sum_{i=1}^{N} a_i \Phi(x_i) = \sum_i a_i k(\cdot, x_i), \qquad g = \sum_{j=1}^{N} b_j \Phi(x_j) = \sum_j b_j k(\cdot, x_j)$$

$$\Rightarrow \quad \langle f, g \rangle = \sum_{i,j} a_i b_j k(x_i, x_j)$$

- The computational cost essentially depends on the sample size. Advantageous for high-dimensional data of small sample size.

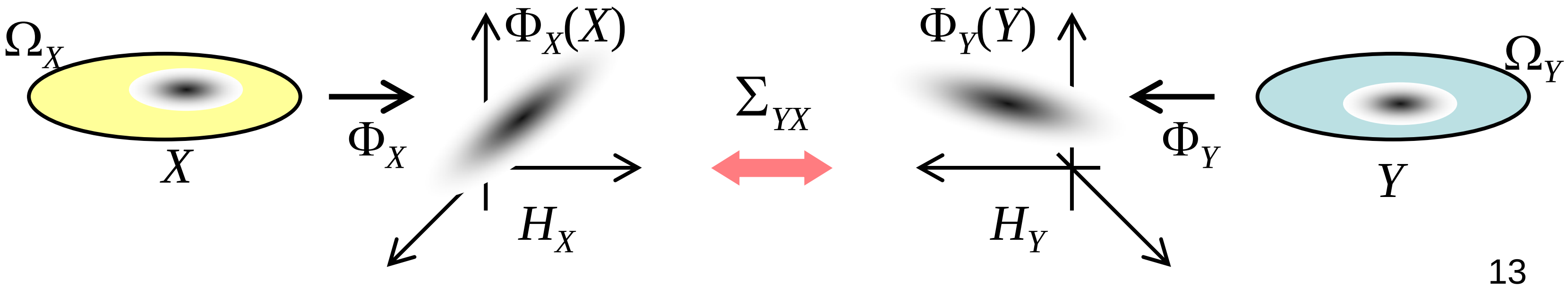
Covariance Operators on RKHS

- X , Y : random variables on Ω_X and Ω_Y , resp.
- Prepare RKHS (H_X, k_X) and (H_Y, k_Y) defined on Ω_X and Ω_Y, resp.
- Define random variables on the RKHS H_X and H_Y by

$$\Phi_X(X) = k_X(\cdot, X) \qquad \Phi_Y(Y) = k_Y(\cdot, Y)$$

- Define the covariance operator Σ_{YX}

$$\Sigma_{YX} = E[\Phi_Y(Y)\langle \Phi_X(X), \cdot \rangle] - E[\Phi_Y(Y)]E[\langle \Phi_X(X), \cdot \rangle]$$



Covariance Operators on RKHS

- Definition

$$\Sigma_{YX} = E[\Phi_Y(Y)\langle \Phi_X(X), \cdot \rangle] - E[\Phi_Y(Y)]E[\langle \Phi_X(X), \cdot \rangle]$$

Σ_{YX} is an operator from H_X to H_Y such that

$$\langle g, \Sigma_{YX} f \rangle = E[g(Y)f(X)] - E[g(Y)]E[f(X)] \quad (= \mathrm{Cov}[f(X), g(Y)])$$

for all $f \in H_X, g \in H_Y$

- *cf.* Euclidean case

$V_{YX} = \mathrm{E}[YX^T] - \mathrm{E}[Y]\mathrm{E}[X]^T$: covariance matrix

$$(b, V_{YX} a) = Cov[(b, Y), (a, X)]$$

Characterization of Independence

- Independence and cross-covariance operators

If the RKHS's are "rich enough":

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad \Sigma_{XY} = O$$

$$\Updownarrow$$

$\Rightarrow$ is always true
$\Leftarrow$ requires an assumption on the kernel (universality)

$$\mathrm{Cov}[f(X), g(Y)] = 0$$

or

$$E[g(Y)f(X)] = E[g(Y)]E[f(X)]$$

for all $f \in H_X, g \in H_Y$

e.g., Gaussian RBF kernels are universal

$$k(x,y) = \exp\left(-\|x-y\|^2/\sigma^2\right)$$

– *cf*. for Gaussian variables,

X and Y are independent $\Leftrightarrow$ $V_{XY} = O$ i.e. uncorrelated

- Independence and characteristic functions

  Random variables X and Y are independent

$$\Leftrightarrow \quad E_{XY}\left[e^{i\omega^T X} e^{i\eta^T Y}\right] = E_X\left[e^{i\omega^T X}\right] E_Y\left[e^{i\eta^T Y}\right] \qquad \text{for all } \omega \text{ and } \eta$$

  I.e., $e^{i\omega^T x}$ and $e^{i\eta^T y}$ work as test functions

- RKHS characterization

  Random variables $X \in \Omega_X$ and $Y \in \Omega_Y$ are independent

$$\Leftrightarrow \quad E_{XY}\left[f(X)g(Y)\right] = E_X\left[f(X)\right] E_Y\left[g(Y)\right] \qquad \text{for all } f \in \mathcal{H}_X,\ g \in \mathcal{H}_Y$$

  – RKHS approach is a generalization of the characteristic-function approach

RKHS and Conditional Independence

- Conditional covariance operator

X and Y are random vectors. $\mathcal{H}_X$, $\mathcal{H}_Y$: RKHS with kernel k_X, k_Y, resp.

Def. $\Sigma_{YY|X} \equiv \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$: conditional covariance operator

– Under a universality assumption on the kernel

$$\langle g, \Sigma_{YY|X} g \rangle = E\left[\mathrm{Var}[g(Y) \mid X]\right]$$

cf. For Gaussian $\mathrm{Var}_{Y|X}[a^T Y \mid X = x] = a^T \left(V_{YY} - V_{YX} V_{XX}^{-1} V_{XY}\right) a$

– Monotonicity of conditional covariance operators

$X = (U,V)$: random vectors

$$\Sigma_{YY|U} \geq \Sigma_{YY|X}$$

$\geq$: in the sense of self-adjoint operators

- Conditional independence

Theorem

$X = (U,V)$ and Y are random vectors.

$\mathcal{H}_X$, $\mathcal{H}_U$, $\mathcal{H}_Y$: RKHS with Gaussian kernel k_X, k_U, k_Y, resp.

$$\Rightarrow \quad Y \perp\!\!\!\perp V \mid U \quad \Leftrightarrow \quad \Sigma_{YY|U} = \Sigma_{YY|X}$$

This theorem provides a new methodology for solving the sufficient dimension reduction problem

Outline

- Introduction
  – dimension reduction and conditional independence
- Conditional covariance operators on RKHS
- Kernel Dimensionality Reduction for regression
- Manifold KDR
- Summary

Kernel Dimension Reduction

- Use a universal kernel for $B^T X$ and Y

$$\Sigma_{YY|B^T X} \geq \Sigma_{YY|X}$$

($\geq$: the partial order of self-adjoint operators)

$$\Sigma_{YY|B^T X} = \Sigma_{YY|X} \quad \Leftrightarrow \quad X \perp\!\!\!\perp Y \mid B^T X$$

- KDR objective function:

$$\min_{B:\, B^T B = I_d} \mathrm{Tr}\left[\Sigma_{YY|B^T X} \right]$$

which is an optimization over the Stiefel manifold

Estimator

- Empirical cross-covariance operator

$$\hat{\Sigma}_{YX}^{(N)} = \frac{1}{N}\sum_{i=1}^{N}\{k_Y(\cdot,Y_i)-\hat{m}_Y\}\otimes\{k_X(\cdot,X_i)-\hat{m}_X\}$$

$$\hat{m}_X = \frac{1}{N}\sum_{i=1}^{N}k_X(\cdot,X_i) \qquad \hat{m}_Y = \frac{1}{N}\sum_{i=1}^{N}k_Y(\cdot,Y_i)$$

$\hat{\Sigma}_{YX}^{(N)}$ gives the empirical covariance:

$$\langle g, \hat{\Sigma}_{YX}^{(N)} f\rangle = \frac{1}{N}\sum_{i=1}^{N} f(X_i)g(Y_i) - \frac{1}{N}\sum_{i=1}^{N} f(X_i)\frac{1}{N}\sum_{i=1}^{N} g(Y_i)$$

- Empirical conditional covariance operator

$$\hat{\Sigma}_{YY|X}^{(N)} = \hat{\Sigma}_{YY}^{(N)} - \hat{\Sigma}_{YX}^{(N)}\left(\hat{\Sigma}_{XX}^{(N)}+\varepsilon_N I\right)^{-1}\hat{\Sigma}_{XY}^{(N)}$$

ε_N: regularization coefficient

- Estimating function for KDR:

$$\mathrm{Tr}\left[\hat{\Sigma}_{YY|U}^{(N)}\right] = \mathrm{Tr}\left[\hat{\Sigma}_{YY}^{(N)} - \hat{\Sigma}_{YU}^{(N)}\left(\hat{\Sigma}_{UU}^{(N)} + \varepsilon_N I\right)^{-1}\hat{\Sigma}_{UY}^{(N)}\right] \qquad U = B^T X$$

$$= \mathrm{Tr}\left[G_Y - G_Y G_U \left(G_U + N\varepsilon_N I_N\right)^{-1}\right]$$

where

$$G_U = \left(I_N - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right)K_U\left(I_N - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right)$$: centered Gram matrix

$$K_U = k(B^T X_i, B^T X_j)$$

- Optimization problem:

$$\min_{B:B^T B = I_d} \mathrm{Tr}\left[G_Y\left(G_{B^T X} + N\varepsilon_N I_N\right)^{-1}\right]$$

Experiments with KDR

- Wine data

  Data

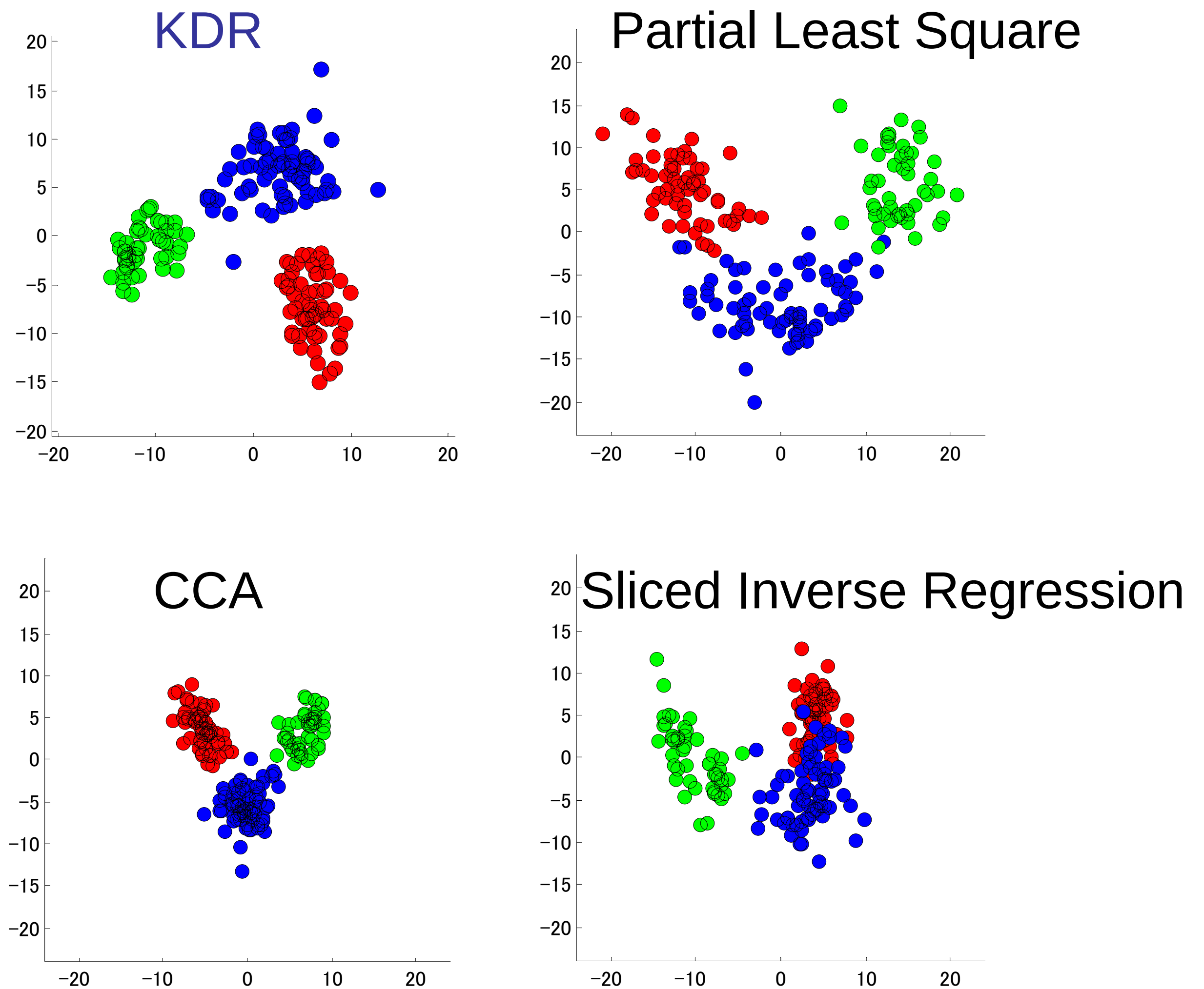
  13 dim. 178 data

  3 classes

  2 dim. projection

$$k(z_1, z_2) = \exp\left(-\|z_1 - z_2\|^2 / \sigma^2\right)$$

$$\sigma = 30$$



Consistency of KDR

<u>Theorem</u>

Suppose k_d is bounded and continuous, and

$$\varepsilon_N \to 0, \;\; N^{1/2}\varepsilon_N \to \infty \;\; (N \to \infty).$$

Let S_0 be the set of optimal parameters:

$$S_0 = \left\{ B \mid B^T B = I_d, \mathrm{Tr}\left[\Sigma_{YY|X}^{B} \right] = \min_{B'} \mathrm{Tr}\left[\Sigma_{YY|X}^{B'} \right] \right\}$$

Then, under some conditions, for any open set $U \supset S_0$

$$\Pr\left(\hat{B}^{(N)} \in U \right) \to 1 \quad (N \to \infty).$$

Lemma

Suppose k_d is bounded and continuous, and

$$\varepsilon_N \to 0, \;\; N^{1/2}\varepsilon_N \to \infty \;\; (N \to \infty).$$

Then, under some conditions,

$$\sup_{B:B^T B = I_d} \left| \mathrm{Tr}\left[\hat{\Sigma}_{YY|X}^{B(N)} \right] - \mathrm{Tr}\left[\Sigma_{YY|X}^{B} \right] \right| \to 0 \;\; (N \to \infty)$$

in probability.

Conclusions

- Introduction
  - – dimension reduction and conditional independence
- Conditional covariance operators on RKHS
- Kernel Dimensionality Reduction for regression
- Technical report available at:

www.cs.berkeley.edu/~jordan/papers/kdr.pdf

Regression on Manifolds using Kernel Dimension Reduction

Jens Nilsson[1] Fei Sha[2] Michael I. Jordan[3]

[1]Centre for Mathematical Sciences
Lund University

[2]Computer Science Division
University of California, Berkeley

[3]Computer Science Division and Department of Statistics
University of California, Berkeley

June 28, 2008

Dimensionality Reduction in Regression

- Find a lower-dimensional subspace $\mathcal{Z} \subset \mathcal{X}$ and a mapping

$$\mathcal{X} \ni x_i \mapsto z_i \in \mathcal{Z}, \quad i = 1, \ldots, m$$

  such that $\{z_i\}$ retains maximal predictive power w.r.t. $\{y_i\}$.
- "Supervised dimensionality reduction"

Sufficient Dimension Reduction

- Parameterize $\mathcal{Z}$ by $\boldsymbol{B} \in \mathbb{R}^{D \times d}$ where $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} = \boldsymbol{I}$.
- Find $\boldsymbol{B}$ such that

$$Y \perp\!\!\!\perp X \mid \boldsymbol{B}^{\mathrm{T}} X \qquad (1)$$

- Under weak conditions, the intersection of all such $\mathcal{Z}_{\boldsymbol{B}}$ defines the central subspace, $\mathcal{S}$.

Kernel Dimension Reduction

Measure conditional independence in RKHS

- Map X and Y to reproducing kernel Hilbert spaces $\boldsymbol{H}_X, \boldsymbol{H}_Y$.

$$X \mapsto f \in \boldsymbol{H}_X, \quad Y \mapsto g \in \boldsymbol{H}_Y$$

- Cross-covariance $\boldsymbol{C}_{fg}$ between f and g can be represented by an operator $\boldsymbol{\Sigma}_{YX} : \boldsymbol{H}_X \to \boldsymbol{H}_Y$ such that

$$\langle g, \boldsymbol{\Sigma}_{YX} f \rangle_{\mathcal{H}_{\mathcal{Y}}} = \boldsymbol{C}_{fg}, \quad \forall f, g \tag{2}$$

- Conditional covariance operator

$$\boldsymbol{\Sigma}_{YY|X} = \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY}. \tag{3}$$

KDR Theorem

1. $\mathbf{\Sigma}_{YY|X} \prec \mathbf{\Sigma}_{YY|\boldsymbol{B}^{\mathrm{T}}X}$
2. $\mathbf{\Sigma}_{YY|X} = \mathbf{\Sigma}_{YY|\boldsymbol{B}^{\mathrm{T}}X} \quad \Longleftrightarrow \quad Y \perp\!\!\!\perp X|\boldsymbol{B}^{\mathrm{T}}X$

- The central space can be found by minimizing $\mathbf{\Sigma}_{YY|\boldsymbol{B}^{\mathrm{T}}X}$.

KDR Algorithm

- The minimization of $\boldsymbol{\Sigma}_{YY|\boldsymbol{B}^{\mathrm{T}}X}$ can be formulated as

$$\begin{array}{ll} \min & \mathrm{Tr}\left[\!\left[\boldsymbol{K}_Y^c(\boldsymbol{K}_{\boldsymbol{B}^{\mathrm{T}}X}^c + N\epsilon\boldsymbol{I})^{-1}\right]\!\right] \\ \text{such that} & \boldsymbol{B}^{\mathrm{T}}\boldsymbol{B} = \boldsymbol{I} \end{array} \qquad (4)$$

where $\boldsymbol{K}_Y^c$ and $\boldsymbol{K}_{\boldsymbol{B}^{\mathrm{T}}X}^c$ are centered Gram matrices.

Reduction to Manifolds?

- Large literature on “manifold learning”
- The goal is to uncover the intrinsic geometry underlying a data set
  - often the goal is visualization
- This is usually done without taking into account a response variable
- As before, we’re motivated to find a way to estimate manifolds while taking into account a response
  - e.g., can help provide guidance for visualization
- We’ll combine (normalized) graph Laplacian technology with KDR

Eigenvectors of the Graph Laplacian

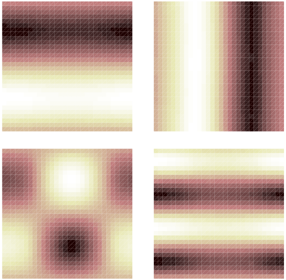
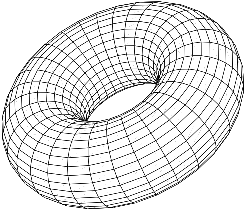


Figure: First four (non-constant) eigenvectors of the graph Laplacian on a torus.



- Harmonics on the manifold
- Reflect intrinsic coordinates

mKDR Formulation

Nilsson, Sha, and Jordan (2007)

- Compute the eigenvectors $\boldsymbol{v}_i,\ i = 1, \ldots, M$ of the normalized graph Laplacian
- Define an RKHS explicitly as the span of these eigenvectors
- Approximate the image of central subspace with a linear transformation $\boldsymbol{\Phi}\boldsymbol{V}^{\mathrm{T}}$

mKDR Algorithm

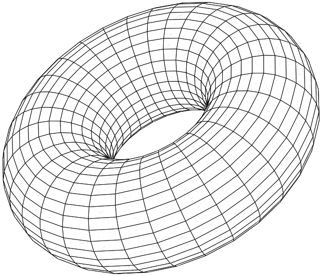
- mKDR minimization problem:

$$\begin{aligned} \min \quad & \mathrm{Tr}\left[\!\left[\boldsymbol{K}_Y^c (\boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{V}^{\mathrm{T}} + N\epsilon \boldsymbol{I})^{-1} \right]\!\right] \\ \text{such that} \quad & \boldsymbol{\Omega} \succeq 0 \\ & \mathrm{Tr}(\boldsymbol{\Omega}) = 1 \end{aligned} \tag{5}$$

- $\boldsymbol{\Phi} = \sqrt{\boldsymbol{\Omega}}$

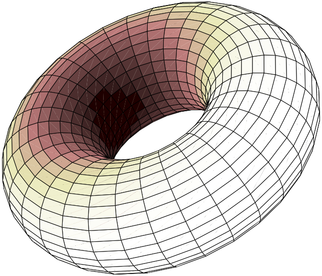
Regression on a Torus

- $\{x_i\}$ have intrinsic coordinates $[\theta_i, \phi_i] \in \mathbf{S}^1 \times \mathbf{S}^1$
- y is a logistic function of $||(\theta, \phi)||$



Regression on a Torus

- $\{x_i\}$ have intrinsic coordinates $[\theta_i, \phi_i] \in \mathbf{S}^1 \times \mathbf{S}^1$
- y is a logistic function of $||(\theta, \phi)||$



mKDR finds the Central Subspace

- $\boldsymbol{\Omega}$ nearly rank 1 $\Rightarrow \boldsymbol{\Omega} \approx \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$;   Project onto $\boldsymbol{a}$

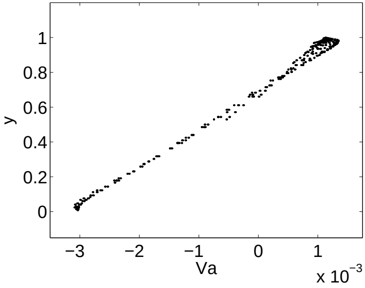


Figure: Uniform grid sampling

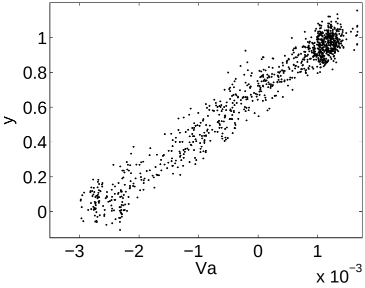


Figure: Uniformly random sampling with additive noise

mKDR can be used to guide visualization

- Map $\{x_i\}$ onto the eigenvectors $\{v_i\}$ with largest weight in $\mathbf{\Phi}$.
- "Predictive eigenvectors" in contrast to principal eigenvectors used in e.g. Laplacian eigenmaps.

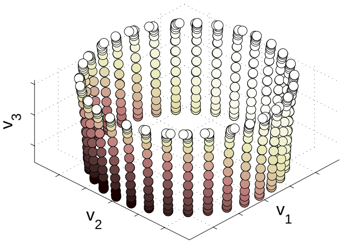


Figure: Principal eigenvectors

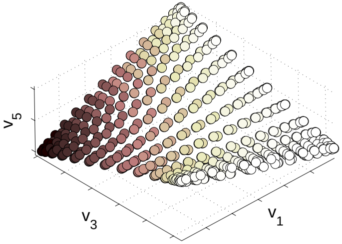
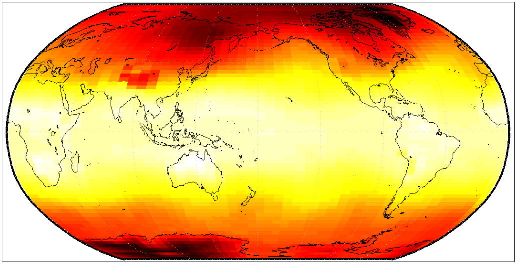


Figure: Predictive eigenvectors

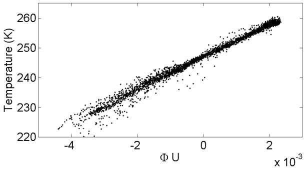
Global Temperature Data



- $\{y_i\}$ are satellite measurements of atmospherical temperatures around the globe.
- 3168 observation points
- $\{x_i\}$ lie on a spheroid in $\mathbb{R}^3$
- Regress the temperature y on x.

Regression Model of Temperature Distribution

Compute the central space $\mathbf{\Phi} \boldsymbol{V}^{\mathrm{T}}$ and use linear regression to model $\mathbf{E}[Y|\mathbf{\Phi} \boldsymbol{V}^{\mathrm{T}}]$.



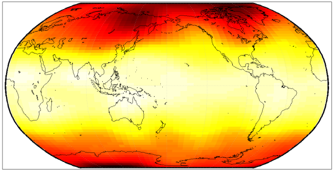


Figure: Predicted temperature

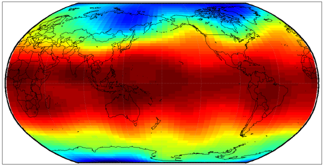


Figure: Central space coordinate

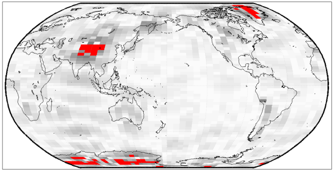
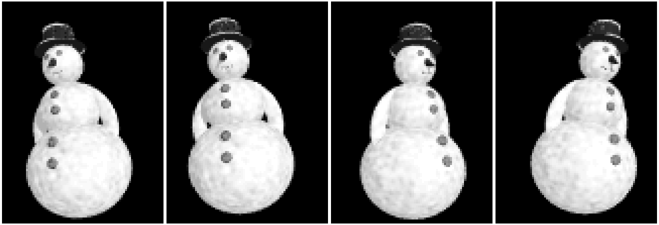


Figure: Prediction error

Visualization of an Image Data Manifold



- $\{x_i\}$ are a set of 1000 grayscale images of size 100×80 pixels
- 4 degrees of freedom: rotation angle, tilt angle and translations in the image plane
- Data lie on a 4-dimensional manifold in $\mathbb{R}^{100 \cdot 80}$
- Create a lower-dimensional embedding that captures the variation in rotation angle

Unsupervised Embedding

- Project onto the principal eigenvectors, i.e. Laplacian Eigenmaps

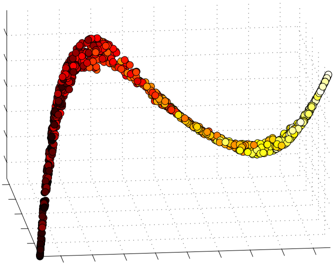


Figure: Principal eigenvectors. Color by tilt angle.

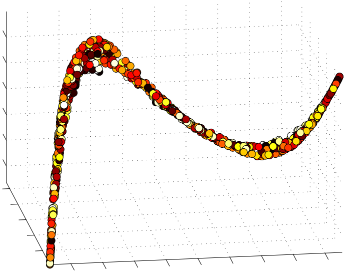


Figure: Principal eigenvectors. Color by rotation angle.

Predictive embedding guided by mKDR

- Apply mKDR with rotation angle as response
- Map data onto predictive eigenvectors of the Graph Laplacian

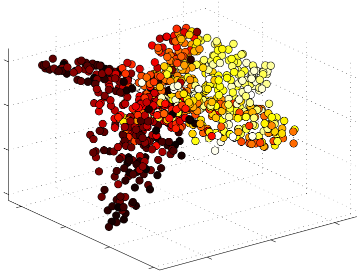


Figure: Predictive eigenvectors

Summary

- mKDR discovers manifolds that optimally preserve predictive power w.r.t response variables.
- mKDR enables:
  - flexible regression modeling
  - supervised exploration of nonlinear data manifolds
- mKDR extends:
  - sufficient dimension reduction to nonlinear manifolds.
  - manifold learning to the supervised setting.