Kernel Dimension Reduction in Regression

K Fukumizu
F R. Bach
M I. Jordan

April 18, 2014

Outline

- Sufficient Dimension Reduction (SDR)
- Reproducing Kernel Hilbert Space (RKHS)
- Conditional covariance operators on RKHS
- Population criterion
- KDR procedure (optimization)
- Consistency
- Numerical results
- Summary

Sufficient Dimension Reduction

- Regression setting: observe $(\mathbf{X}, \mathbf{Y})$ pairs, where the covariate $\mathbf{X}$ is p-dimensional (p is large).
- Want to find a subspace $\mathcal{S} \subset \mathbb{R}^p$ that retains the information pertinent to the response $\mathbf{Y}$. More formally,

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | B_{\mathcal{S}}^T \mathbf{X}$$

  where $B_{\mathcal{S}}$ denotes the orthogonal projection of $\mathbf{X}$ onto $\mathcal{S}$ (so $\mathcal{S}$ is the column space of $B_{\mathcal{S}}$).
- The SDR literature treats $\mathbf{X}$ random

Central Space

Usually, $\mathcal{S}$ is not unique. Under some weak conditions, it can be shown

$$\mathcal{S}_{Y|X} = \left\{ \bigcap_i \mathcal{S}_i : \mathbf{X} \perp\!\!\!\perp \mathbf{Y} | B_{\mathcal{S}_i}^T \mathbf{X} \right\}$$

is also a dimension reduction subspace, called **central space**, which is unique and of the minimum dimensionality of the dimension-reduction spaces.

Once the central space is identified, a conditional distribution $p(\mathbf{Y}|\mathbf{X})$ or a regression function $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$ can be inferred using low dimensional coordinates $B_{\mathcal{S}_{Y|X}}^T \mathbf{X}$.

$\{\mathcal{S} : \mathbf{Y} \perp\!\!\!\perp \mathbb{E}[\mathbf{Y}|\mathbf{X}] | B_{\mathcal{S}} X\}^T$ is called **central mean space**, which is a subspace of $\mathcal{S}_{Y|X}$.

Existing Methods

- *Strong assumption models (on $p(Y|X)$ or $\mathbb{E}[Y|X]$): OLS, PLS, CCA, ACE, Projection pursuit regression, neural networks and LASSO.
- Inverse regression: SIR (Li 1991), SAVE, pHd

SIR

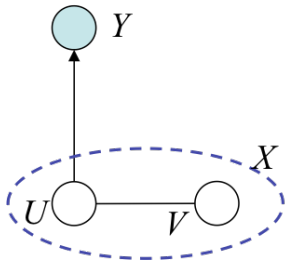
- PCA on Cov($\mathbb{E}[\mathbf{X}|\mathbf{Y}]$).
- Pros: Semiparametric with specific inferential methodology.
- Cons: Require the linearity assumption in the subspace, i.e. $\mathbb{E}\left[\mathbf{b^T X}|B_S^T\mathbf{X}\right] = c_0 + c^T B_S^T \mathbf{X}$. Lack of exhaustiveness is another issue. The estimator may converge to a set of vectors that are in S but do not span S. e.g. $Y = \left(B_S^T X\right)^2$.

- Forward methods: MAVE.
- Gradient-based methods.
- KDR: characterize conditional independence directly in a Hilbert space mapped from the data space.

Conditioinal Independence

- $(U, V) = (B^T X, C^T X)$,
  - B: projector onto a d-dimensional subspace
  - C: projector onto its complement space
  - $[B, C]$: complete orthonormal matrix
- B projects X onto the central space

$$\Leftrightarrow p_{Y|X}(y|x) = p_{Y|U}(y|B^T x)$$
$$\Leftrightarrow p_{Y|U,V}(y|u, v) = p_{Y|U}(y|u)$$
$$\Leftrightarrow Y \perp\!\!\!\perp V|U$$



- This paper characterizes conditional independence of Y and X using RKHS.

RKHS (Kernel → Associated Hilbert space)

Given a pd kernel k, the **reproducing kernel feature map** is defined to be

$$\Phi : \Omega \to H, x \to k(\cdot, x).$$

You can consider x as an index. Then, consider the (unique) vector space:

$$\text{span}\{\Phi(x) : x \in \Omega\} = \left\{ f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) : n \in \mathbb{N}, x_i \in \Omega, \alpha_i \in \mathbb{R} \right\}.$$

In this space, for $f = \sum_{i=1}^{n} \alpha_i k(\cdot, u_i)$ and $g = \sum_{i=1}^{n} \beta_i k(\cdot, v_i)$, define the inner product as

$$\langle f, g \rangle = \sum_{i,j=1}^{n} \alpha_i \beta_j k(u_i, v_j),$$

so that k has the **reproducing property**

$$\langle f, k(\cdot, x) \rangle = f(x).$$

Complete this inner product space to be a Hilbert space by adding the limits of all Cauchy sequences.

Why RKHS

- RKHS provides rich enough functions for modelling purpose.
- By the reproducing property, computation is easy:

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

and

$$f = \sum_{i=1}^{n} a_i k(\cdot, x_i)$$

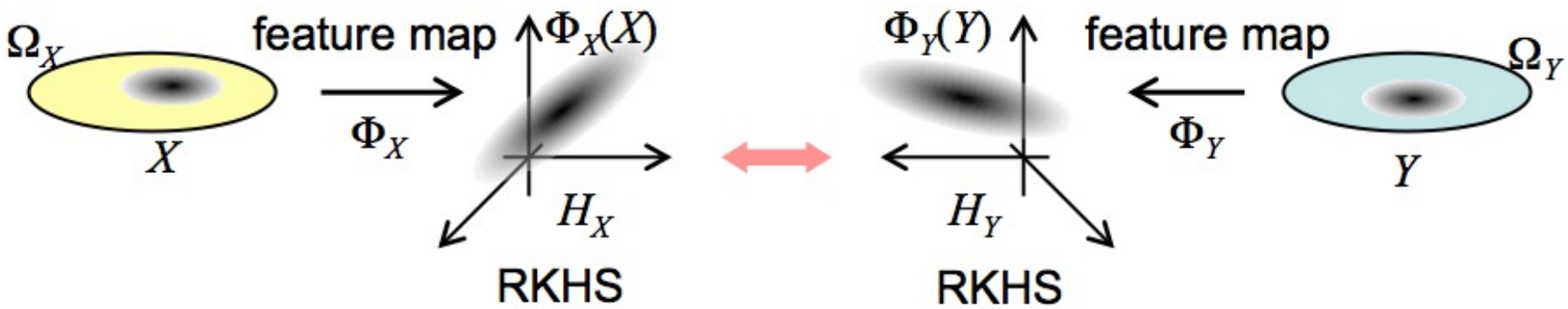
$$g = \sum_{i=1}^{n} b_j k(\cdot, x_j)$$

$$\Rightarrow \langle f, g \rangle = \sum_{i,j=1}^{n} a_i b_j k(x_i, x_j).$$

The computational cost essentially depends on the sample size n.

Kernel Methods

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



Setup:

- X, Y : random variables(vectors) on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$, resp.
- RKHSs $(\mathcal{H}_{\mathcal{X}}, k_{\mathcal{X}})$ and $(\mathcal{H}_{\mathcal{Y}}, k_{\mathcal{Y}})$ defined on $\mathcal{X}$ and $\mathcal{Y}$, resp. It is assumed that k's are pd kernels satisfying

$$\mathbb{E}_X \left[k_{\mathcal{X}}(X, X)\right] < \infty \text{ and } \mathbb{E}_Y \left[k_{\mathcal{Y}}(Y, Y)\right] < \infty.$$

So

$$\mathcal{H}_{\mathcal{X}} \subset L^2_{P_X} \text{ and } \mathcal{H}_{\mathcal{Y}} \subset L^2_{P_Y}$$

$$\mathbb{E}_X \left[f(X)^2\right] < \infty \text{ and } \mathbb{E}_Y \left[g(Y)^2\right] < \infty$$

Three representations of cross-covariance operators:

1. For $f \in \mathcal{H}_{\mathcal{X}}$ and $g \in \mathcal{H}_{\mathcal{Y}}$, **cross-covariance operator** Σ_{YX} is defined as

$$\begin{aligned}\langle g, \Sigma_{YX} f \rangle &= \mathbb{E}_{XY}\left[\left(f(X) - \mathbb{E}_X\left[f(X)\right]\right)\left(g(Y) - \mathbb{E}_Y\left[g(Y)\right]\right)\right] \\ &= \mathrm{Cov}(f(X), g(Y)).\end{aligned}$$

2. Define $m_X \in \mathcal{H}_\mathcal{X}$ (**mean element**) as

$$m_X : \langle f, m_X \rangle = \mathbb{E}_X \left[f(X) \right], \quad \forall f \in \mathcal{H}_\mathcal{X}.$$

The existence of m_X is due to **Riesz Representation Theorem** and it can be shown $m_X = \mathbb{E}_X \left[k_\mathcal{X}(\cdot, X) \right]$. Then

$$\langle g, \Sigma_{YX} f \rangle = \mathbb{E}_{XY} \left[\langle f, k_\mathcal{X}(\cdot, X) - m_X \rangle_{\mathcal{H}_\mathcal{X}} \langle g, k_\mathcal{Y}(\cdot, Y) - m_Y \rangle_{\mathcal{H}_\mathcal{Y}} \right].$$

3. A cross-covariance operator on an RKHS can be represented as an integral operator. Let S_{YX} be the integral operator $\forall \varphi \in L^2_{P_X}$

$$G_\varphi(y) := \int_{\mathcal{X}\times\mathcal{Y}} k_{\mathcal{Y}}(y, \tilde{y}) \left(\varphi(\tilde{x}) - \mathbb{E}_X \left[\varphi(X)\right]\right) dP_{XY}(\tilde{x}, \tilde{y})$$

Then,

$$(\Sigma_{YX} f)(y) = G_f(y), \quad \forall f \in \mathcal{H}_{\mathcal{X}}.$$

In other words, $\Sigma_{YX} = S_{YX}$ on $\mathcal{H}_{\mathcal{X}} \subset L^2_{P_X}$.

Remarks:

If $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ are Euclidean spaces $\mathbb{R}^p$ and $\mathbb{R}^q$ resp., then

$$\Sigma_{YX} = \mathbb{E}\left[YX^T\right] - \mathbb{E}\left[Y\right]\mathbb{E}\left[X\right]^T \text{ (covariance matrix)},$$
$$\Rightarrow \langle b, \Sigma_{YX} a \rangle = \mathrm{Cov}(b^T Y, a^T X).$$

Properties of cross-covariance operators:

- $\Sigma_{YX} = \Sigma_{XY}^*$, where "*" denotes adjoint operator.
- If $Y = X$, Σ_{XX} is self-adjoint and called **covariance operator**.
- $\Sigma_{YX} = \Sigma_{YY}^{1/2} V_{YX} \Sigma_{XX}^{1/2}$ (Cov $\leftrightarrow$ Corr), where

  $V_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ is a unique bounded operator s.t.
  $V_{YX} = Q_Y V_{YX} Q_X$,

  Q_X and Q_Y are the orthogonal projections $\mathcal{H}_{\mathcal{X}} \to \overline{\mathcal{R}(\Sigma_{XX})}$ and $\mathcal{H}_{\mathcal{Y}} \to \overline{\mathcal{R}(\Sigma_{YY})}$, resp.

Conditional Covariance Operator

$$\begin{aligned}\Sigma_{YY|X} &:= \Sigma_Y Y - \Sigma_{YY}^{1/2} V_{YX} V_{XY} \Sigma_{YY}^{1/2} \\ &= \Sigma_Y Y - \Sigma_{YX}^{1/2} \Sigma_{XX}^{-1} \Sigma_{XY}^{1/2}.\end{aligned}$$

The first proposition relates the operator to the residual error of regression:

Proposition 2

$\forall g \in \mathcal{H}_{\mathcal{Y}}$,

$$\langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_{\mathcal{Y}}} = \inf_{f \in \mathcal{H}_{\mathcal{X}}} \mathbb{E}_{XY} \left[(g(Y) - \mathbb{E}_Y [g(Y)]) - (f(X) - \mathbb{E}_X [f(X)]) \right]^2.$$

The second proposition expresses the residual error in terms of the conditional variance:

Proposition 3

If $\mathcal{H}_{\mathcal{X}} + \mathbb{R}$ is dense in $L^2_{P_X}$ **(AS)**, then

$$\langle g, \Sigma_{YY|X} g \rangle_{\mathcal{H}_{\mathcal{Y}}} = \mathbb{E}_X \left[\mathrm{Var}_{Y|X}(g(Y)|X) \right].$$

Remarks:

- It means the space $\mathcal{H}_{\mathcal{X}}$ is rich enough to approximate any $L^2_{P_X}$ regression function $\mathbb{E}_{Y|X}[g(Y)|X]$.

Theorem 4 (Part 1)

Suppose $\overline{\mathcal{H}_{\mathcal{X}}^{B}} \subseteq \overline{\mathcal{H}_{\mathcal{X}}} \subset L^2_{P_X} \quad \forall B \in \mathbb{S}_d^m$. Then

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

Remark:
So one might be attempted to think if we can find a B such that $\Sigma_{YY|X}^{B} = \Sigma_{YY|X}$, then we don't need X any more. Instead $B^T X$ contains the same amount of information of X to explain Y. Thus $Y \perp\!\!\!\perp X | B^T X$.

This is not true unless the RKHS is **characteristic**.

Characteristic Kernel

Definition

Let $\mathcal{P}$ be the set of all probability measures on $(\Omega, \mathcal{B})$ and $(\mathcal{H}, k)$ be an RKHS over Ω with the kernel measureable and bounded. The RKHS $\mathcal{H}$ is called **characteristic** (w.r.t $\mathcal{B}$) if the map

$$\mathcal{P} \to \left\{ m_P = \mathbb{E}_P \left[k(\cdot, X) \right] \right\}$$

is 1-1.

Note: k is called characteristic kernel if $\mathcal{H}$ is characteristic.

A concrete example

Other examples: $k =$ Gaussian RBF or Laplacian
$\exp\{-\alpha \sum_{i=1}^{m} |x_i - y_i|\}$ are characteristic on $\mathbb{R}^m$ w.r.t $\mathcal{B}$.

Theorem 4 (Part 2)

If further $(\mathcal{H}_{\mathcal{X}}, P_X)$ and $(\mathcal{H}_{\mathcal{X}}^B, P_B)$ satisfy **(AS)** $\forall B \in \mathbb{S}_d^m$ and $\mathcal{H}_{\mathcal{Y}}$ is characteristic, then

$$\Sigma_{YY|X}^B = \Sigma_{YY|X} \iff Y \perp\!\!\!\perp X | B^T X.$$

Remark: The spaces of $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{X}}^B$ should be rich enough to include the regression function. Since $\mathcal{H}_{\mathcal{Y}}$ is characteristic, the (conditional) distribution of Y is fully determined by its (conditional) mean.

(AS) and the notion of **characteristic** kernel are closely related.

Proposition 5

Let $(\Omega, \mathcal{B})$ be a measurable space, k is a bounded pd kernel and $\mathcal{H}$ is the induced RKHS. Then

$$k \text{ is characteristic } \iff \mathcal{H} + \mathbb{R} \text{ is dense in } L_P^2 \quad \forall P.$$

Remark: If $\mathbb{E}_P[f(X)] = \mathbb{E}_Q[f(X)] \quad \forall f \in \mathcal{H}$ and $\mathcal{H}$ is rich enough, then the two measures are the same.

Notations

- d: dimension of the central space.
- m: dimension of the original space.
- $\mathbb{S}_d^m := \{B \in \mathbb{R}^{m \times d} : B^T B = I_d\}$ **(Stiefel manifold)**.
- $\mathbb{B}_d^m \subseteq \mathbb{S}_d^m$: subset of matrices whose columns span a dimension-reduction subspace. (not central space?)
- $B_0 \in \mathbb{B}_d^m$, the true subspace.
- $k_d(\cdot, \cdot)$: a pd kernel on $\mathbb{R}^d$.
- $k_{\mathcal{X}}^B(\mathbf{x}_1, \mathbf{x}_2) := k_d(B^T \mathbf{x}_1, B^T \mathbf{x}_2) \quad \forall B \in \mathbb{S}_d^m$.
- $\mathcal{H}_{\mathcal{X}}^B := \left\{ f : \mathcal{X} \to \mathbb{R} \mid \exists g \in \mathcal{H}_{k_d} \text{ s.t. } f(x) = g(B^T x) \right\}$. It is the RKHS associated with $k_{\mathcal{X}}^B(\cdot, \cdot)$.

Population DR Criterion

Due to Theorem 4, we should minimize the conditional covariance operator $\Sigma^B_{YY|X}$ w.r.t. B.

If $\mathcal{X} \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}^q$, and $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are Gaussian RBF, i.e. $k(x, y) = e^{-\|x-y\|^2/\sigma^2}$, then

$$\mathbb{B}^m_d = \arg\min_{B \in \mathbb{S}^m_d} \Sigma^B_{YY|X}$$

The minimization refers to the self-adjoint operator with the minimum partial order.

$$\Updownarrow$$

$$\arg\min_{B \in \mathbb{S}_d^m} \mathbf{Tr}\left(\Sigma_{YY|U}\right).$$

Remarks:

1. This is the population contrast function.
2. Fukumizu et al.(2004) used determinant to evaluate the operator size. Another option is to use the largest eigen-value.

$\Updownarrow$ (due to Proposition 2)

Minimizing sum of residual errors for predicting Y using U (under some conditions), i.e.

$$\arg\min_{B \in \mathbb{S}_d^m} \sum_{a=1}^{\infty} \min_{f \in \mathcal{H}_{\mathcal{X}}^B} \mathbb{E}_{XY} \left[\left| (\xi_a(Y) - \mathbb{E}_Y [\xi_a(Y)]) - (f(X) - \mathbb{E}_X [f(X)]) \right|^2 \right],$$

where $\{\xi_a\}_{a=1}^{\infty}$ is a complete orthonormal system of $\mathcal{H}_{\mathcal{Y}}$.

Empirical cross-covariance operator:

$$\langle g, \widehat{\Sigma}_{YX}^{B(n)} f \rangle = \frac{1}{n} \sum_{i}^{n} \langle f, k_{\mathcal{X}}^{B}(\cdot, \mathbf{x}_i) - \hat{m}_X^B \rangle_{\mathcal{H}_{\mathcal{X}}^B} \langle g, k_{\mathcal{Y}}(\cdot, \mathbf{y}_i) - \hat{m}_Y \rangle_{\mathcal{H}_{\mathcal{Y}}}$$

where

$$\hat{m}_X^B = \frac{1}{n} \sum_{i}^{n} k_{\mathcal{X}}^{B}(\cdot, \mathbf{x}_i), \quad \hat{m}_Y = \frac{1}{n} \sum_{i}^{n} k_{\mathcal{Y}}(\cdot, \mathbf{y}_i).$$

Besides, since $f \in \mathcal{H}_{\mathcal{X}}^B$ and $g \in \mathcal{Y}$, we have

Empirical conditional covariance operator:

$$\hat{\Sigma}^{B(n)}_{YY|X} = \hat{\Sigma}^{(n)}_{YY} - \hat{\Sigma}^{B(n)}_{YX}\left(\hat{\Sigma}^{B(n)}_{XX} + \epsilon_n I_n\right)^{-1}\hat{\Sigma}^{B(n)}_{XY}$$

ϵ_n : regularization coefficient

Let

$$\hat{K}_Y := \left(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right) G_Y \left(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right),$$

$$\hat{K}_X^B := \left(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right) G_X^B \left(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right),$$

where

$$[G_Y]_{ij} := k_{\mathcal{Y}}(\mathbf{y}_i, \mathbf{y}_j),$$

$$[G_X^B]_{ij} := k_{\mathcal{X}}^B(\mathbf{x}_i, \mathbf{x}_j).$$

Then

$$\hat{\Sigma}_{YY|X}^{B(n)} = \frac{1}{n}\left[\hat{K}_Y - \hat{K}_X^B\left(\hat{K}_X^B + n\epsilon_n I_n\right)^{-1}\hat{K}_Y\right]$$

Evaluating the size by trace (empirical contrast function):

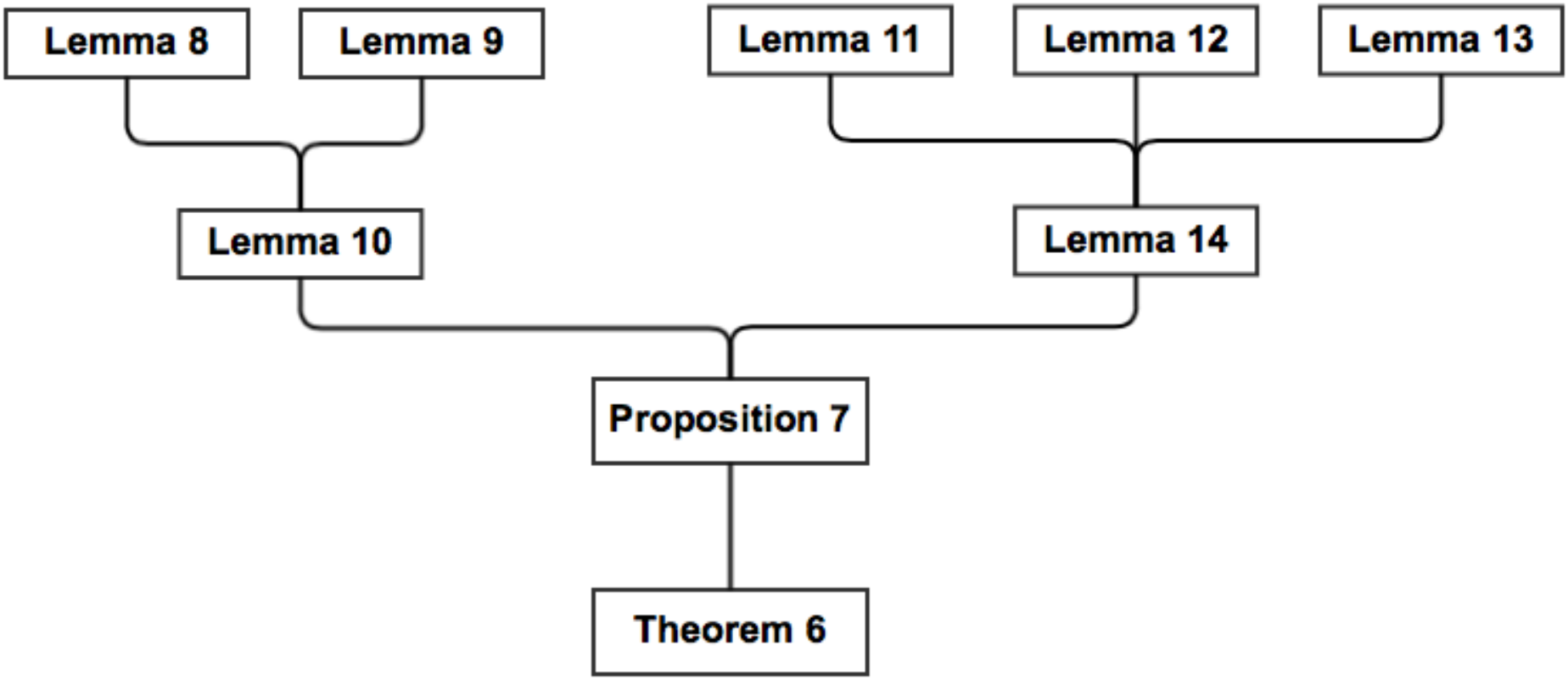
$$\begin{aligned}\mathbf{Tr}\left(\hat{\Sigma}^{B(n)}_{YY|X}\right) &= \frac{1}{n}\mathbf{Tr}\left(\hat{K}_Y - \hat{K}^B_X(\hat{K}^B_X + n\epsilon_n I_n)^{-1}\hat{K}_Y\right) \\ &= \epsilon_n \mathbf{Tr}\left(\hat{K}_Y\left(\hat{K}^B_X + n\epsilon_n I_n\right)^{-1}\right).\end{aligned}$$

Optimization problem

$$\hat{B}^{(n)} = \arg\min_{B \in \mathbb{S}_d^m} \mathbf{Tr}\left(\hat{K}_Y \left(\hat{K}_X^B + n\epsilon_n I_n\right)^{-1}\right).$$

Non-convex, steepest descent method with line search.

Consistency



Assumptions

A-1 For any bounded continuous function g on $\mathcal{Y}$, the function

$$B \to \mathbb{E}_X \left[\mathbb{E}_{Y|B^T X} \left[g(Y) | B^T X \right]^2 \right]$$

is continuous on $\mathbb{S}_d^m$.

A-2 Let P_B be the probability distribution of $BB^T X$. The Hilbert space $\mathcal{H}_{\mathcal{X}}^B$ satisfies **(AS)**.

A-3 There exists a measurable function $\phi : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}\left[\phi(X)\right]^2 < \infty$ and the Lipschitz condition

$$\| k_d(B^T X, \cdot) - k_d(\tilde{B}^T X, \cdot) \|_{\mathcal{H}_d} \leq \phi(x) D(B, \tilde{B}),$$

where $D()$ is the distance defined on Stiefel manifold.

Lemma 8:

$$\left|\mathbf{Tr}\left(\widehat{\Sigma}^{(n)}_{YY|X}\right)\right| - \mathbf{Tr}\left(\Sigma_{YY} - \Sigma_{YX}\left(\Sigma_{XX} + \epsilon_n I\right)^{-1}\Sigma_{XY}\right)$$
$$\leq \frac{1}{\epsilon_n}\left\{\left(\|\widehat{\Sigma}^{(n)}_{YX}\|_{HS} + \|\Sigma_{YX}\|_{HS}\right)\|\widehat{\Sigma}^{(n)}_{YX} - \Sigma_{YX}\|_{HS} + \|\Sigma_{YY}\|_{tr}\|\widehat{\Sigma}^{(n)}_{XX} - \Sigma_{XX}\|\right\}$$
$$+ \left|\mathbf{Tr}\left(\widehat{\Sigma}^{(n)}_{YY} - \Sigma_{YY}\right)\right|$$

Lemma 9:

$$\sup_{B\in\mathbb{S}^m_d}\|\widehat{\Sigma}^{B(n)}_{XX} - \Sigma^B_{XX}\|_{HS} = O_p(1/\sqrt{n})$$
$$\sup_{B\in\mathbb{S}^m_d}\|\widehat{\Sigma}^{B(n)}_{XY} - \Sigma^B_{XY}\|_{HS} = O_p(1/\sqrt{n})$$
$$\sup_{B\in\mathbb{S}^m_d}\left|\mathbf{Tr}\left(\widehat{\Sigma}^{B(n)}_{YY} - \Sigma^B_{YY}\right)\right| = O_p(1/\sqrt{n})$$

Lemma 10:

$$\sup_{B\in\mathbb{S}^m_d}\left|\mathbf{Tr}\left(\widehat{\Sigma}^{B(n)}_{YY|X}\right) - \mathbf{Tr}\left(\Sigma_{YY} - \Sigma^B_{YX}\left(\Sigma^B_{XX} + \epsilon_n I\right)^{-1}\Sigma^B_{XY}\right)\right| = O_p(1/(\epsilon_n\sqrt{n}))$$

if $\epsilon_n = O(1/\sqrt{n})$ as $n \to \infty$

Lemma 11:

$$\mathbf{Tr}\left(\Sigma_{YX}\left(\Sigma_{XX}+\epsilon I\right)^{-1}\Sigma_{XY}\right) \to \mathbf{Tr}\left(\Sigma_{YY}^{1/2}V_{YX}V_{XY}\Sigma_{YY}^{1/2}\right)$$

Lemma 12:

$$L_\epsilon(B) = \mathbf{Tr}\left(\Sigma_{YX}^{B}\left(\Sigma_{XX}^{B}+\epsilon I\right)^{-1}\Sigma_{XY}^{B}\right) \text{ is continuous } \forall \epsilon > 0$$

Lemma 13:

$$L_0(B) \text{ is continuous}$$

Lemma 14:

$$\sup_{B\in\mathbb{S}_d^m} \mathbf{Tr}\left(\Sigma_{YY|X}^{B} - \left\{\Sigma_{YY} - \Sigma_{YX}^{B}\left(\Sigma_{XX}^{B}+\epsilon_n I\right)^{-1}\Sigma_{XY}^{B}\right\}\right)$$

$$\text{as } n\to\infty, \epsilon_n \to 0$$

Proposition 7

Under (A-1), (A-2) and (A-3), $\mathbf{Tr}\left(\hat{\Sigma}^{B(n)}_{YY|X}\right)$ and $\mathbf{Tr}\left(\Sigma^{B}_{YY|X}\right)$ are continuous, and

$$\sup_{B\in\mathbb{S}_d^m}\left|\mathbf{Tr}\left(\hat{\Sigma}^{B(n)}_{YY|X}\right)-\mathbf{Tr}\left(\Sigma^{B}_{YY|X}\right)\right|\to 0$$

in probability.

Remark: Lemmas 12 and 13 show the continuity of the two traces. Lemmas 10 and 14 prove the uniform convergence.

Theorem 6

Suppose k_d is bounded and continuous, and

$$\epsilon_n = O\left(\frac{1}{\sqrt{n}}\right).$$

Let $\mathbb{B}_d^m$ be the set of optimum parameters

$$\mathbb{B}_d^m = \arg\min_{B \in \mathbb{S}_d^m} \mathbf{Tr}\left(\Sigma_{YY|X}^B\right)$$

Then, under (A-1), (A-2) and (A-3), $\mathbb{B}_d^m$ is non-empty and for any open set $U \supset \mathbb{B}_d^m$ in $\mathbb{S}_d^m$

$$P(\hat{B}^{(n)} \in U) \to 1.$$

Numerical Simulation

Compare SIR, SAVE, pHd and KDR based on $\|\hat{B} - B_0\|_F$.

Models for simulation:

A

$$Y = \frac{X_1}{0.5 + (X_2 + 1.5)^2} + (1 + X_2)^2 + \sigma\epsilon$$

where $X \sim \mathsf{N}(0, I_4)$.

So $B_0^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$.

TABLE 1
Comparison of KDR and other methods for regression (A)

	KDR		SIR		SAVE		pHd	
σ	NORM	SD	NORM	SD	NORM	SD	NORM	SD
0.1	0.11	0.07	0.55	0.28	0.77	0.35	1.04	0.34
0.4	0.17	0.09	0.60	0.27	0.82	0.34	1.03	0.33
0.8	0.34	0.22	0.69	0.25	0.94	0.35	1.06	0.33

Each row used 100 random replications with 100 samples each.

B

$$Y = \sin^2(\pi X_2 + 1) + \sigma\epsilon$$

where $X \sim$ Unif on $[0,1]^4 \setminus \left\{x \in \mathbb{R}^4 | x_i \leq 0.7, i = 1,2,3,4\right\}$.

So $B_0^T = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix}$.

TABLE 2
Comparison of KDR and other methods for regression (B)

	KDR		SIR		SAVE		pHd	
σ	NORM	SD	NORM	SD	NORM	SD	NORM	SD
0.1	0.05	0.02	0.24	0.10	0.23	0.13	0.43	0.19
0.2	0.11	0.06	0.32	0.15	0.29	0.16	0.51	0.23
0.3	0.13	0.07	0.41	0.19	0.41	0.21	0.63	0.29

C

$$Y = \frac{1}{2}(X_1 - a)^2 \epsilon$$

where $X \sim N(0, I_{10})$.

TABLE 3
Comparison of KDR and other methods for regression (C)

	KDR		SIR		SAVE		pHd	
a	NORM	SD	NORM	SD	NORM	SD	NORM	SD
0.0	0.17	0.05	1.83	0.22	0.30	0.07	1.48	0.27
0.5	0.17	0.04	0.58	0.19	0.35	0.08	1.52	0.28
1.0	0.18	0.05	0.30	0.08	0.57	0.20	1.58	0.28

Each row used 100 random replications with 500 samples each.

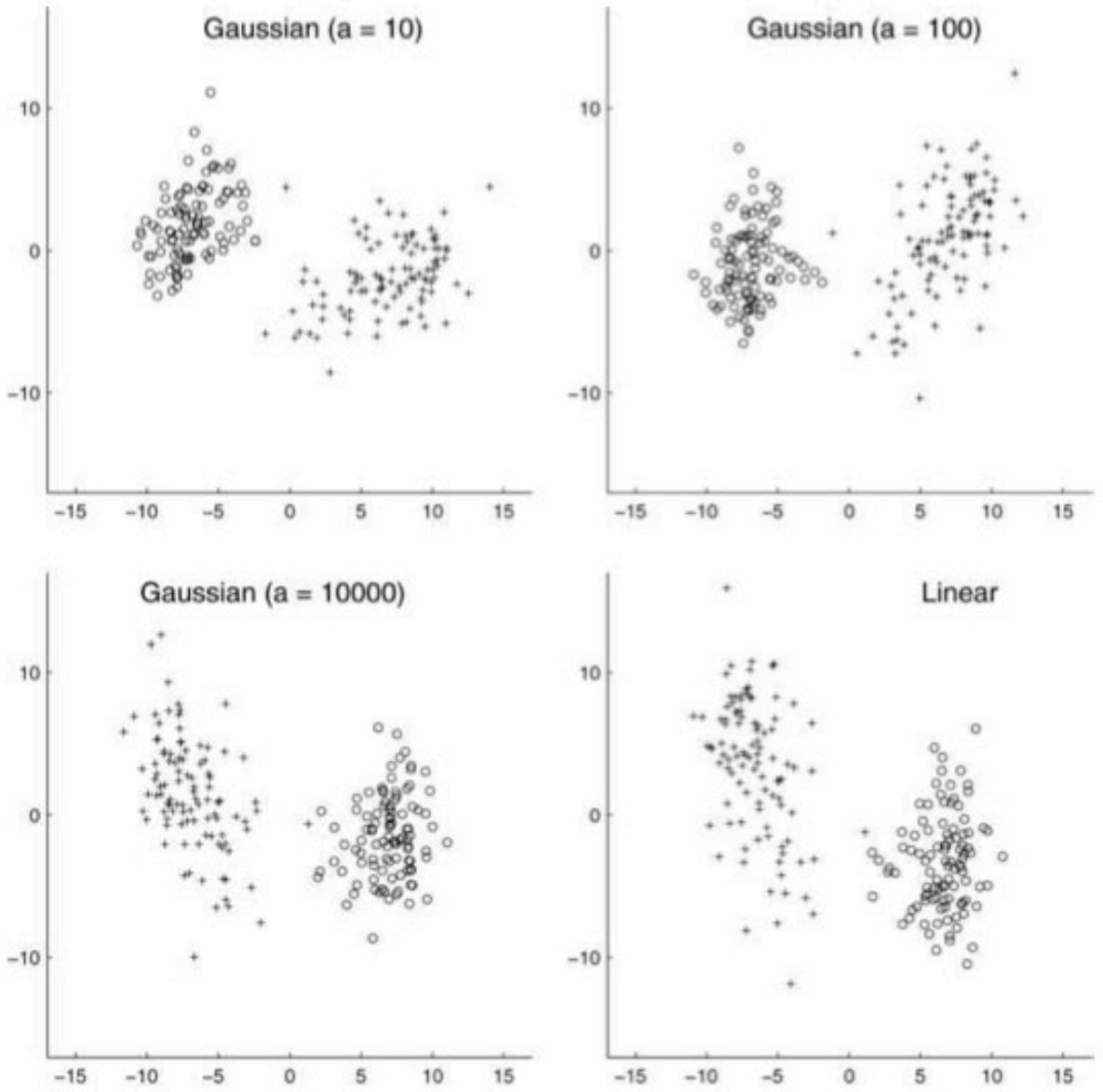
Comment: pHd failed for the model because the estimating direction only appears in the variance.

Real Data

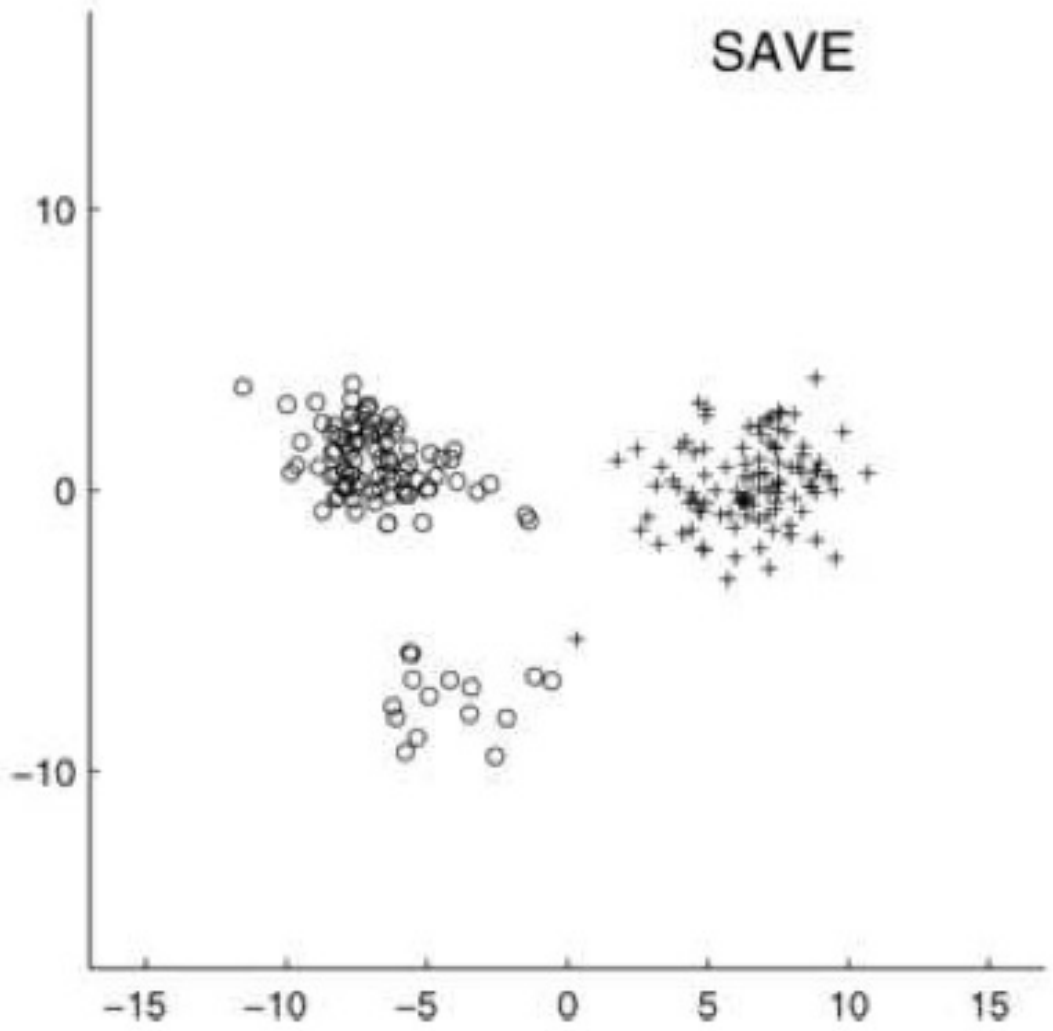
A Swiss Bank Notes

Y: counterfeit or genuine (binary)
X_1: length
X_2: left height
X_3: right height
X_4: distance of inner frame to the lower border
X_5: distance of inner frame to the upper border
X_6: diagonal length

KDR with GRF:


Gaussian (a = 10)
Gaussian (a = 100)
Gaussian (a = 10000)
Linear
10
0
−10
−15
−10
−5
0
5
10
15

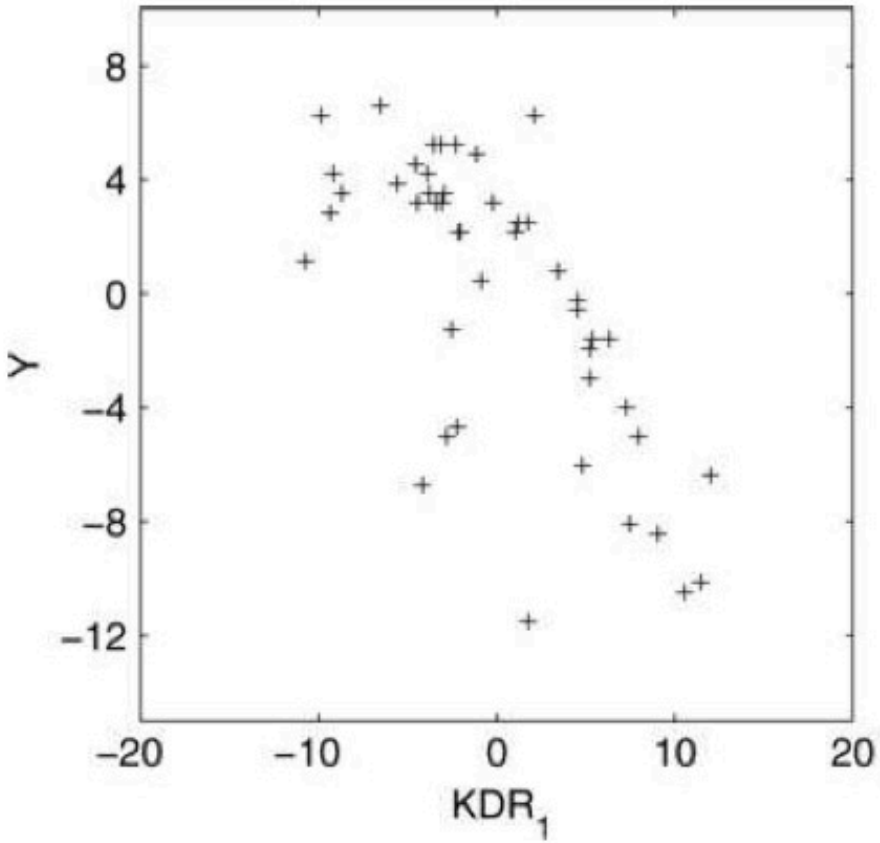
SAVE:

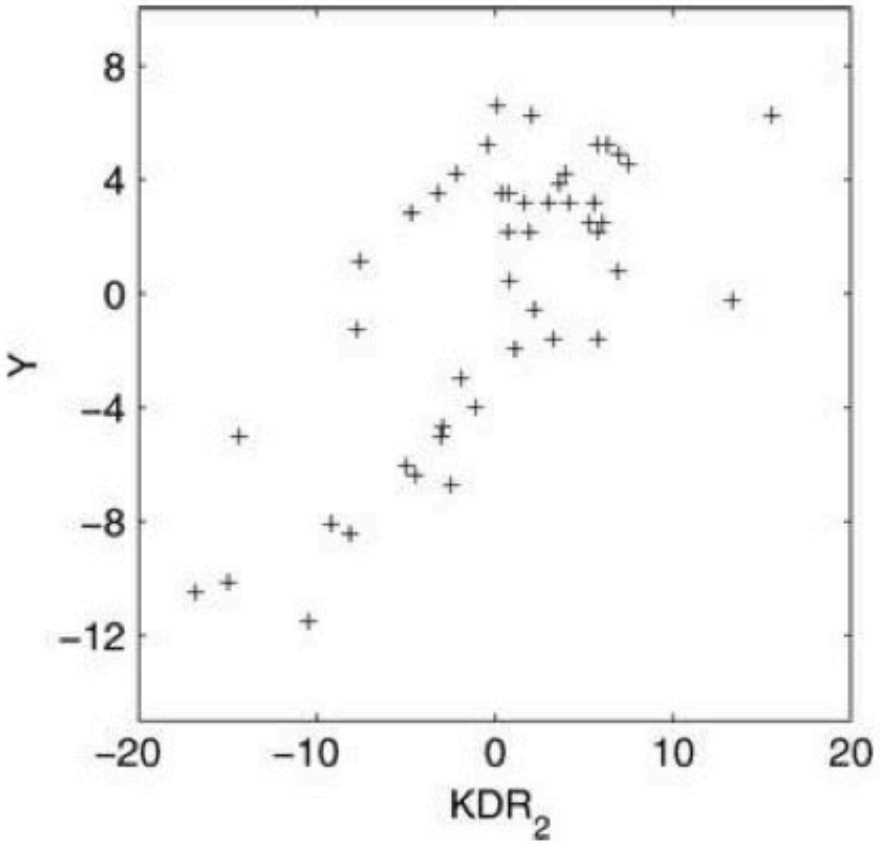


B Evaporation

Y: soil evaporation
X_1: max daily soil temperature
X_2: min daily soil temperature
X_3: area under the daily soil temperature curve
X_4: max daily air temperature
X_5: min daily air temperature
X_6: average daily air temperature
X_7: max daily humidity
X_8: min daily humidity
X_9: area under the humidity curve
X_{10}: total wind speed

Contains 46 daily observations. KDR with RBF ($\sigma^2 = 10$) yields two dimensional subspace.


8
4
0
−4
−8
−12
Y
−20
−10
0
10
20
KDR1


8
4
0
−4
−8
−12
Y
−20
−10
0
10
20
KDR2

Summary