再生核ヒルベルト空間を用いた回帰問題における次元削減法

福水 健次
統計数理研究所
2003年9月26日 日本数学会

M.I. Jordan, F.R. Bach (UC Berkeley) との共同研究
Outline

- Introduction
  - Dimensionality reduction for regression

- Conditional Independence and RKHS
  - Dimensionality reduction and conditional independence
  - Reproducing kernel Hilbert space
  - Conditional covariance operator

- Kernel Dimensionality Reduction for Regression
  - Algorithm and experimental results

- Extension to Variable Selection

- Summary
Dimensionality reduction

- Dimensionality reduction is important for high-dimensional data such as gene expression, text, image, etc.

- Dimensionality reduction
  - feature selection – linear or nonlinear combination of variables.
  - variable selection – subset of variables.

- Purposes of dimensionality reduction
  - Compact and readable explanation of statistical relationship
  - Computational efficiency
  - Accuracy of estimation
Dimensionality reduction for regression

- Regression: analysis of statistical dependence of $Y$ on $X$,

  Conditional probability density $p(Y \mid X)$

  $Y$: response variable,
  $X = (X_1, ..., X_m) \in \mathbb{R}^m$ explanatory variables

- Goal of dimensionality reduction
  = Find an effective subspace to select feature vector for regression.

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

  $B = (b_1, ..., b_d)$: $m \times d$ matrix $d$ ($< m$) is fixed.

Feature vector as a linear combination of $X_1, ..., X_m$.
Effective subspace contains all the information in $X$ to explain $Y$.  


Example

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

Effective subspace = direction of \( X_1 \)
Introduction

**Semiparametric problem**

Assume existence of an effective subspace

\[ p_{Y|X}(Y \mid X) = \tilde{p}(Y \mid B_0^T X) \]

\[ B_0: m \times d \quad \text{matrix} \]

i.i.d. sample \((X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})\) given.

---

**PROBLEM**

Find the effective subspace \(B_0\)

- without assuming any models about the conditional probability \(p_{Y|X}\) or the marginal distributions \(p_X\) and \(p_Y\).

- There is the infinite degree of freedom on unestimated \(p_{Y|X}\).

  \[ \rightarrow \quad \text{Semiparametric problem.} \]

- Once the effective subspace is obtained, any type of regressor can be built on that space.
Conditional Independence

- **Dimensionality reduction and conditional independence**

\[ (U, V) = (B^T X, C^T X) \quad \text{for} \quad (B, C) \in O(m) \quad (m\text{-dim. orthogonal matrix}) \]

- \( B \) gives the projector onto the effective subspace

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

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

  \[ \Leftrightarrow \quad \text{Conditional independence} \quad Y \perp V \mid U \]

- **Characterization of conditional independence**

  - Reproducing kernel Hilbert space (RKHS)
Reproducing Kernel Hilbert Space

**Definition**

\( \Omega : \text{set.} \quad H : \text{Hilbert space} \subset \{f : \Omega \to \mathbb{R}\} \)

\( H : \text{reproducing kernel Hilbert space (RKHS)} \)

\[\iff \exists k : \Omega \times \Omega \to \mathbb{R} \quad \text{symmetric function (reproducing kernel) s.t.} \]

1) \( k(\cdot, x) \in H \) for all \( x \in \Omega \).

2) \( \langle k(\cdot, x), f \rangle_H = f(x) \) for \( \forall f \in H, \ x \in \Omega \). reproducing property

**Example: Gaussian kernel**

\[ k : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}, \quad k(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma^2}\right) \]

\[ \iff \text{There is a RKHS on } \mathbb{R}^m \text{ with reproducing kernel } k. \]
Reproducing Kernel Hilbert Space

Properties of RKHS

- Condition of RKHS (Mercer)
  If a symmetric function \( k: \Omega \times \Omega \rightarrow \mathbb{R} \) is positive definite, i.e. for any \( x_1, \ldots, x_n \in \Omega \),
  \[
  (k(x_i, x_j)) \geq 0
  \]
  then, there uniquely exists a RKHS with \( k \) its reproducing kernel.

- Advantage of RKHS
  • Reproducing property makes computation easy and feasible.
    e.g.) For \( f = \sum_{i=1}^{n} a_i k(\cdot, X_i) \), \( g = \sum_{i=1}^{n} b_i k(\cdot, X_i) \)
    \[
    \langle f, g \rangle_H = \sum_{ij} a_i b_j k(X_i, X_j)
    \]
    \[
    (k(X_i, X_j))_{ij} : \text{Gram matrix}
    \]
  • RKHS is much smaller than \( L^2(\mathbb{R}^m) \)
    If \( k \) is continuous, all the functions in \( H \) are continuous, and \( H \subset C(\Omega) \) is a continuous embedding.
RKHS and Independence

Independence and characteristic functions

Random variables $X$ and $Y$ are independent

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

$e^{-\omega^T X}$ and $e^{-\eta^T Y}$ work as test functions which account for the infinite degree of freedom ($L^2$).

RKHS characterization

$H_X$ and $H_Y$ are RKHS on $\Omega_X$ and $\Omega_Y$, respectively.

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

$$\iff E_{XY}[f(X)g(Y)] = E_X[f(X)] E_Y[g(Y)] \quad \text{for all } f \in H_X, g \in H_Y$$

This is true if $H_X$ and $H_Y$ are RKHS for Gaussian kernels.

(Bach & Jordan 2002)
Cross-covariance Operator

Definition

$X$ and $Y$: random variable on $\Omega_X$ and $\Omega_Y$, respectively.

$H_X$ and $H_Y$: RKHS on $\Omega_X$ and $\Omega_Y$, respectively, with bounded kernels.

We can define a bounded operator $\Sigma_{YX} : H_X \rightarrow H_Y$ by

$$\langle g, \Sigma_{YX} f \rangle_{H_Y} = E_{XY} [f(X)g(Y)] - E_X [f(X)] E_Y [g(Y)] \quad (= \text{Cov}[f(X), g(Y)])$$

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

(Riesz’s theorem)

$\Sigma_{YX}$ is called cross-covariance operator.

Cross-covariance operator and independence

Theorem

$H_X$ and $H_Y$: RKHS with Gaussian kernel.

$X$ and $Y$ are independent $\iff \Sigma_{YX} = 0$
RKHS and Conditional Independence

**Conditional covariance**

$X$ and $Y$ are random vectors. $H_X$, $H_Y$ : RKHS with kernel $k_X$, $k_Y$, resp.

Assumption: $\exists \Sigma_{XX}^{-1}$, $E_{Y|X}[g(Y) \mid X] \in H_X$ for all $g \in H_Y$.

$$\left\langle f, \left(\Sigma_{YY} - \Sigma_{YX} \Sigma_{XX}^{-1} \Sigma_{XY}\right)g \right\rangle = E_X\left[\text{Cov}_{Y|X}[f(Y), g(Y) \mid X]\right]$$

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

\textit{c.f.} For Gaussian r.v., $\text{Cov}_{Y|X}[a^T Y, b^T Y \mid X = x] = a^T \left(V_{YY} - V_{YX} V_{XX}^{-1} V_{XY}\right)b$

- **Monotonicity** of conditional covariance operators

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

$$\Sigma_{YY|U} \geq \Sigma_{YY|X} \quad \geq : \text{in the sense of self-adjoint operators}$$
RKHS and Conditional Independence

**Conditional independence**

Theorem

\[ X = (U, V) \text{ and } Y \text{ are random vectors.} \]

\[ H_X, H_U, H_Y : \text{RKHS with Gaussian kernel } k_X, k_U, k_Y, \text{ resp.} \]

\[ E_{Y|X}[g(Y) \mid X] \in H_X \text{ and } E_{Y|U}[g(Y) \mid U] \in H_U \text{ for all } g \in H_Y. \]

\[ \iff \quad Y \perp V \mid U \iff \Sigma_{YY|U} = \Sigma_{YY|X} \]

**Minimization of conditional covariance operator**

\[
\min_{B: U = B^TX} \Sigma_{YY|U} \quad \Rightarrow \quad \text{matrix } B \text{ gives the effective subspace}
\]

- Evaluation
  - Operator norm -- maximum eigenvalue.
  - Trace norm -- sum of eigenvalues
  - **Determinant** -- product of eigenvalues
Kernel Dimensionality Reduction

**Estimation of conditional covariance operator**

\((X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})\) : i.i.d. sample from the true joint probability.

Restrict the spaces to the linear hull of
\[ \left\{ k(\cdot, X^{(i)}) - \frac{1}{n} \sum_{j=1}^{n} k(\cdot, X^{(j)}) \mid 1 \leq i \leq n \right\} \]

and
\[ \left\{ k(\cdot, Y^{(i)}) - \frac{1}{n} \sum_{j=1}^{n} k(\cdot, Y^{(j)}) \mid 1 \leq i \leq n \right\} \]

Replace \(\Sigma_{YY\mid U}\) by \(n \times n\) matrix
\[
\hat{\Sigma}_{YY\mid U} = \hat{\Sigma}_{YY} - \hat{\Sigma}_{YU} \hat{\Sigma}_{UU}^{-1} \hat{\Sigma}_{UY}
\]

where
\[
\hat{\Sigma}_{UU} = (G_U + \varepsilon I_n)^2, \quad \hat{\Sigma}_{YY} = (G_{YY} + \varepsilon I_n)^2, \quad \hat{\Sigma}_{UY} = G_U G_Y
\]

\(\varepsilon\) : regularization coefficient

\[
G_U = \left( I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) \left( k_U(U^{(i)}, U^{(j)}) \right) \left( I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right)
\]

\[
G_Y = \left( I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) \left( k_Y(Y^{(i)}, Y^{(j)}) \right) \left( I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right)
\]

reproducing property and empirical average
Kernel Dimensionality Reduction

**Kernel dimensionality reduction (KDR)**

$$\min_B \hat{\Sigma}_{YY|U} \equiv \hat{\Sigma}_{YY} - \hat{\Sigma}_{YU} \hat{\Sigma}_{UU}^{-1} \hat{\Sigma}_{UY}$$

$$\Leftrightarrow \min_B \det \left[ I_n - \hat{\Sigma}_{YY}^{-1/2} \hat{\Sigma}_{YU} \hat{\Sigma}_{UU}^{-1} \hat{\Sigma}_{UY} \hat{\Sigma}_{YY}^{-1/2} \right]$$

$$\Leftrightarrow \min_B \frac{\det \hat{\Sigma}_{[YU][YU]}}{\det \hat{\Sigma}_{YY} \det \hat{\Sigma}_{UU}}$$

where \( \hat{\Sigma}_{[YU][YU]} = \begin{pmatrix} \hat{\Sigma}_{YY} & \hat{\Sigma}_{YU} \\ \hat{\Sigma}_{UY} & \hat{\Sigma}_{UU} \end{pmatrix} \)

c.f. mutual information of Gaussian variables.

**Method of KDR**

Kernel Dimensionality Reduction (KDR)

\[
= \text{minimization of } \frac{\det \hat{\Sigma}_{[YU][YU]}}{\det \hat{\Sigma}_{YY} \det \hat{\Sigma}_{UU}}
\]

gradient-based method is used for the minimization.
**Kernel Dimensionality Reduction**

- **Wide applicability of KDR**
  - The most general approach for dimensionality reduction: no model for $p(Y|X)$.
  - KDR needs no strong assumption on the distribution of $X$, $Y$ and dimensionality of $Y$.
  - c.f. other method; SIR, pHd, CCA, PLS, etc.

- **Computational cost**
  - Multiplication of $n \times n$ matrices is computationally hard.
    → Incomplete Cholesky decomposition
  - Local minimum → annealing is used in gradient method.
Existing Methods

- **Sliced Inverse Regression (SIR, Li 1991)**
  - PCA of $E[X|Y] \rightarrow$ use slice of $Y$.
  - Semiparametric method: no assumption on $p(Y|X)$.
  - Elliptic assumption on the distribution of $X$ is necessary.

- **Principle Hessian Direction (pHd, Li 1992)**
  - Average Hessian $\Sigma_{jxx} \equiv E[(Y - \bar{Y})(X - \bar{X})(X - \bar{X})^T]$ is used.
  - If $X$ is Gaussian, eigenvectors gives the effective directions.
  - Gaussian assumption on $X$. $Y$ must be one-dimensional.

- **Projection pursuit approach (e.g. Friedman et al. 1981)**
  - Additive model $E[Y|X] = g_1(b_1^T X) + ... + g_d(b_d^T X)$ is used.

- ** Canonical Correlation Analysis (CCA) / Partial Least Square (PLS)**
  - Linear assumption on the regression.

- **Nonparametric approach**
Experiments

**Synthesized data**

- **Data**

  \[ X: 2 \text{ dim}, \ Y: 1 \text{ dim} \]

  100 data

  \[ Y \sim 2 \exp(-X_1^2) + N(0; 0.1^2) \]

- **Problem**: find one-dim. effective subspace

- **Results**

<table>
<thead>
<tr>
<th></th>
<th>SIR</th>
<th>pHd</th>
<th>CCA</th>
<th>PLS</th>
<th>KDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angle (deg.)</td>
<td>-86.522</td>
<td>57.015</td>
<td>-10.416</td>
<td>-26.093</td>
<td>0.298</td>
</tr>
</tbody>
</table>
Wine data (from UCI Machine Learning Repository)

- Data
  13 dim. 178 data.
  3 classes
  2 dim. projection

KDR

CCA

PLS

SIR

pHd
Experiments

Experiments on classification accuracy

– Purpose:
  to see how much information on \( Y \) is maintained in the low-dimensional subspace of \( X \).


– Data sets for binary classification from UCI repository.

– Comparison with pHd.
  Many methods are NOT applicable for binary classification tasks.
Experiments

- Results

Breast-cancer-Wisconsin

X: 30 dim.
# training data=200
# test data=369

Heart-disease

X: 13 dim.
# training data=149,
# test data=148

Ionosphere

X: 34 dim.
# training data=151
# test data=200

![Graphs showing classification rates for different datasets and feature reduction methods.](attachment:image.png)
Extension to Variable Selection

Variable selection by KDR

- Select a subset $(X_{i_1}, \ldots, X_{i_d})$ from $\{X_1, \ldots, X_m\}$.
- Principle
  \[ Y \perp V \mid U \iff \Sigma_{YY|U} = \Sigma_{YY|X} \]
- Objective function for variable selection.

\[
\min_U \frac{\det \hat{\Sigma}_{[YU][YU]}}{\det \hat{\Sigma}_{YY} \det \hat{\Sigma}_{UU}}
\]

\(\min\) is taken over all the subsets

\[ U = (X_{i_1}, \ldots, X_{i_d}) \text{ where } 1 \leq i_1 < \cdots < i_d \leq m \]

- Problem: combinatorial explosion for computation
  
  Evaluation of all the combinations is intractable for large $m$ and $d$.
  
  \(\rightarrow\) some approximate optimization methods are needed.
Experiments of Variable Selection

■ Small data set

- *Boston Housing:*
  
  X : 13 dim.,
  
  Y = house price,
  
  506 data.

- 4 variables are selected.

  \[ \binom{13}{4} = 715. \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>ACE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRIM</td>
<td></td>
<td>O</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INDUS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHAS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOX</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RM</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>AGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIS</td>
<td></td>
<td>O</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAX</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>PTRATIO</td>
<td>O</td>
<td>O</td>
<td></td>
<td>O</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LSTAT</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>O</td>
</tr>
</tbody>
</table>

ACE: Breiman & Friedman (1985)
Application to Gene Selection

AML/ALL classification (Golub et al. Science 1999)
- Microarray data: 7129 dim. 38 training data.
- Binary classification: AML / ALL.
- Golub et al. show 50 effective genes using neighborhood analysis.

Greedy optimization algorithm
1. Start from one variable.
2. For already chosen $t$ variables $S_t = \{X_i, \ldots, X_{i_t}\}$, evaluate the objective function values for $S_t \cup \{X_j, X_k\} - \{X_i\}$ for all combinations of $X_j, X_k$ not in $S_t$ and $X_i$ in $S_t$, and select the best one.
3. Repeat this up to $d$ variables.

Results
- 50 genes are selected by KDR method
<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Golub99</th>
<th>Lee03</th>
<th>Szabo02</th>
<th>Li02</th>
<th>Fuj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukotriene C4 synthase (LTC4S)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Zyxin</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>FAH</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>V-yes-1 Yamaguchi sarcoma</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Leptin receptor</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>CD33 CD33 antigen (differentiati)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Liver mRNA for interferon-gamma</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>PRG1 Protelytic 1, secretory</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>DF D component of complement (ad</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>INTERLEUKIN-8 PRECURSOR</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>MYELOID LEUKEMIA</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Peptidyl-prolyl cis-trans</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Cathepsin D lysosomal aspa</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>TGAX Integrin, alpha X (antige</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>GAL33 Lectin, galactose-bind</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Epb72 gene exon 1</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>MAJOR HISTOCOMPATIBILITY</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Lysozyme</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Azurocidin gene</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>PFC Properdin P factor, comple</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Lysophospholipase homolog (HU-K5)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>PPGB Protective protein for beta</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Catalase (EC 1.11.1.6) 5flank</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Ferritin heavy chain</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>CD36 CD36 antigen (collagen typ</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>EUKARYOTIC PEPTIDE CHAIN</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>GB DEF = CD36 gene exon 15</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Colony-stimulating factor 1</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Carbonic anhydrase II</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Hepatocyte growth factor-like pr</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>MPO Myeloperoxidase</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>CHRNA7 Cholinergic receptor, ni</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>AFFX-HUMTRRM11507_M_at</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Complement component 1 Inh</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Glycoporphin Sta (type</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Glycophorin E</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>AFFX-HUMTRRM11507_3_at</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Metabotropic glutamate receptor</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>GB DEF = Neutrophil elastase ge</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>ELA2 Elastase 2, neutrophil)</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>GB DEF = Kazal-type serine prote</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>LCAT Leptin-cholesterol acyltr</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Aldehyde dehydrogenase 2,</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Annexin VIII</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
<tr>
<td>Protease, serine, 3 (tryp</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
<td>o</td>
</tr>
</tbody>
</table>

#agree/#selected 25/50 10/28 4/9 8/10 29/50
Application to Gene Selection

- Classification accuracy
  - Evaluation of classification rate for 34 independent test data.
  - Application of SVM using the selected genes.

- Results
  - KDR + SVM: 32 correct / 34 samples
  - Golub et al: 29 correct + 5 rejected / 34 samples
  - 50 genes in Golub et al + SVM: 32 correct / 34 samples

- KDR method provides effective genes.
- KDR method accounts for the combination of genes.
  Many methods use relation between only $X_j$ and $Y$ for selection.
Summary

- Kernel method is suitable for semiparametric problems
  - Dimensionality reduction for regression = conditional independence.
  - Conditional covariance operators gives the criterion for the conditional independence.

- Kernel dimensionality reduction (KDR)
  - The most general approach for dimensionality reduction.
  - KDR has wide applicability to feature / variable selection.
  - K.f. other methods have some restrictions.
  - KDR finds effective features / variables in practical problems.

- Future/ongoing studies
  - Theoretical analysis of the estimator: consistency etc.
  - Extension to more general Bayesian networks.
Tech report etc.

http://www.ism.ac.jp/~fukumizu/