Vol.2012-MPS-91 No.6
Vol.2012-BIO-32 No.6
2012/12/6

カーネル層別逆回帰のためのモデル選択手法

日野 英逸[1,a)]　越島 健介[1,b)]　村田 昇[1,c)]

概要：データの次元を適切に削減することは，計算コスト及び記憶領域の削減，データの本質的構造の把握に繋がる．特に回帰問題においては，データが拘束されている部分空間を推定する sufficient dimension reduction の問題が盛んに研究されている．その代表例である層別逆回帰は，説明変数が楕円分布に従う場合は，応答変数の値によって層別した説明変数の中心ベクトルを用いて線型次元削減行列を推定することの正当性が理論的に保証されている．層別逆回帰の拡張として，カーネルトリックを利用することで非線型次元削減部分空間を推定する手法も提案されているが，利用するカーネル関数を適切に選択する必要がある．本研究では，カーネル層別逆回帰に利用するカーネル関数の選択手法を提案する．カーネル関数を定めることで，付随する特徴空間における説明変数の分布が決定されることに着目し，説明変数が特徴空間で正規分布に従うようにカーネルを選択する．正規分布は楕円分布の一種であり，これにより層別逆回帰における仮定が満たされる．特徴空間における分布の正規性を特性関数によって評価し，カーネル関数の凸結合の結合係数を最適化するアルゴリズムを導出する．幾つかの実データを用いた実験から，提案手法の有用性を示す．

キーワード：層別逆回帰，カーネル法，Multiple Kernel Learning

1. はじめに

次元削減は統計，機械学習を初め多くの分野で重要な問題であり，データ量の削減，データに潜む本質的な情報の抽出，或いは次元削減の後に行うデータ処理の効率化などを目的として多くの方法が提案されている．通常，次元削減は観測データとしてベクトルデータを考え，与えられた観測データ集合から何らかの基準に従い，データが有する本質的な情報を失うことなくそのデータの次元を削減するための写像を求める問題として定式化される．特に回帰問題のための教師付き線型次元削減手法として，Sliced Inverse Regression (SIR [1]) が統計の分野で開発され多くの関連した研究がある．SIR の特徴は，Linear Design Condition (LDC) という形で，応答関数やノイズにではなく，説明変数に分布を仮定する点にある．SIR では説明変数の集合を応答変数の値に応じて層別し，線型次元削減をした部分空間を求める問題を固有値問題に帰着している．説明変数が楕円分布に従う時，SIR によって求められた部分空間は，真の部分空間と一致することが示されている．SIR はカーネル関数を用いて非線型化がなされており，可視化や判別などの問題に応用されている [2].

SIR の最適性は説明変数に対する LDC と回帰関数の形状に関する緩い条件に依存している．SIR の制約を緩和するために種々の拡張がなされており [3–7] ，回帰のための線型次元削減手法は現在も盛んに研究されている．しかし，こうした SIR の拡張手法の多くはやはり説明変数の分布に何らかの仮定をおいている．また，これらの拡張は線型次元削減の枠組みで提案されているため，線型写像によって表現される部分空間の同定しか出来ない．カーネル法を用いた非線型次元削減手法である KSIR も，カーネル関数に付随する特徴空間における説明変数の分布に対する LDC の成立を仮定しており，カーネルの選択によっては LDC が満たされず，削減をしたデータを用いた回帰が良好な結果を与えない．

LDC は，説明変数の分布が楕円分布に従っている場合には満たされることが知られている [1,8]. 本研究では正規分布が楕円分布の一種であることと，説明変数が楕円分布に従っていれば SIR の想定する理想的な状況が得られるということに着目し，カーネル関数によって写像される特徴空間においてデータが正規分布に従うようにカーネル関数を設計する手法を提案する．提案手法は特徴空間における正規性の評価に経験特性関数を利用し，Multiple Kernel Learning (MKL) の枠組みでカーネル関数の最適化を行う．

[1] 早稲田大学理工学術院　〒 169–8555 東京都新宿区大久保 3–4–1
[a)] hideitsu.hino@gmail.com
[b)] jazzaphysical@ruri.waseda.jp
[c)] noboru.murata@eb.waseda.ac.jp

2. Sliced Inverse Regression

本節では, 層別逆回帰 (Sliced Inverse Regression, SIR [1]) を概観し，次元削減部分空間の同定のための最適性条件を説明する．

2.1 Problem Formulation

X を p 次元の説明変数, Y を応答変数とし，X, Y の実現値をそれぞれ $\boldsymbol{x}, y$ で表す．X の線型次元削減を，次元削減行列 $\mathbb{R}^{p\times q}$ $(q \leq p)$ の作用 $B^\top X$ で表す．次元削減行列の列ベクトルを $\boldsymbol{b}_i \in \mathbb{R}^p, i = 1, \ldots, q$ で表す．ここで，応答変数 Y の説明変数 X に対する依存性が $\{\boldsymbol{b}_1^\top X, \cdots, \boldsymbol{b}_q^\top X\}$ のみによって表される時，この B による次元削減を，Sufficient Dimension Reduction (SDR) と呼ぶ．形式的には，

$$Y \perp\!\!\!\perp X | B^\top X \tag{1}$$

なる B を求める問題を SDR と呼ぶ．上式 (1) を満たす B の列空間を，次元削減空間と呼ぶ．通常は B の列空間の同定のみに関心があるため，B は Stiefel 多様体 $S_q^p(\mathbb{R}) = \{B \in \mathbb{R}^{p\times q} | B^\top B = I_q\}$ の元であると考える．ただし，I_q は $q \times q$ の単位行列である．本稿ではこの行列 B を，次元削減行列と呼ぶ．SDR の主な目的は，この次元削減空間を与える次元削減行列 B の推定である．つまり，通常の回帰問題と同様に，ϵ を適当な分布に従うノイズとして

$$y = r(B^\top \boldsymbol{x}, \epsilon) \tag{2}$$

のような説明変数，応答変数の関係を想定しているが，特定の回帰関数 r を仮定していないことに注意する．

2.2 SIR Procedure

層別逆回帰の特徴は，応答変数 Y で条件付けた説明変数 X に分布を導入する点にある．つまり，SIR の推定は逆回帰平均関数 $E[X|Y]$ に基づいて行う．説明変数の観測値集合は，その経験平均 $\boldsymbol{\mu}$，経験共分散行列 Σ_{xx} を用いて $\tilde{\boldsymbol{x}}_i = \Sigma_{xx}^{-1/2}(\boldsymbol{x}_i - \boldsymbol{\mu})$ のように正規化されているものとする．応答変数の値の範囲を L 個の重ならない層に分割し，各層で観測データがほぼ同数含まれるようにする．このスライスを，内部に含まれるデータの添字集合 $\mathcal{S}_l, l = 1, \ldots, L$ で表現する．ここで，$\cup_{l=1}^L \mathcal{S}_l = \{1, \ldots, n\}$ 及び $\mathcal{S}_l \cap \mathcal{S}_h = \emptyset$, $l \neq h$ が成り立つ．記述の簡単のため，記号 $\mathcal{S}_l$ で添字集合とデータの部分集合の両方を文脈に応じて表現することにする．また，集合 S の要素数を $|S|$ で表す．

今，各層におけるデータの経験平均ベクトル $\boldsymbol{\mu}_l = \frac{1}{|\mathcal{S}_l|}\sum_{i\in\mathcal{S}_l} \tilde{\boldsymbol{x}}_i$, $l = 1, \ldots, L$ を用いて，重み付き共分散行列 $M = \frac{1}{L}\sum_{l=1}^L |\mathcal{S}_l| \boldsymbol{\mu}_l \boldsymbol{\mu}_l^\top$ を推定する．前述のように観測データの平均は $\mathbf{0}$ で共分散行列は単位行列になるように変換されていることに注意すると，直観的には行列 M の上位固有値に対応する固有ベクトルは，説明変数 X と応答変数 Y の関係を記述するような部分空間に含まれることが分かる．従って，次元削減行列の推定量は行列 M の固有ベクトル $\hat{\boldsymbol{b}}_j$ を対応する固有値の大きいものから q 個計算し，$\hat{B} = (\hat{\boldsymbol{b}}_1, \ldots, \hat{\boldsymbol{b}}_q) \in \mathbb{R}^{p\times q}$ として得られる．

2.3 Linear Design Condition

次元削減空間の次元 q は既知であるとする．

Condition1 (Linear Design Condition:LDC)
$B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q)$ が linear design condition を満たすとは， 任意の $\boldsymbol{w} \in \mathbb{R}^p$ に対して条件付き期待値 $E[\boldsymbol{w}^\top X | \boldsymbol{b}_1^\top X, \ldots, \boldsymbol{b}_q^\top X]$ が $\{1, \boldsymbol{b}_1^\top X, \ldots, \boldsymbol{b}_q^\top X\}$ の線型結合で表すことが出来ることをいう．

回帰モデル (2) と LDC の下で，行列 M の q 個の主要固有ベクトルによって次元削減空間が張られることが示されている [1]．一方，Eaton [8] により，説明変数が楕円分布に従っていることが，LDC が成立するための必要条件であることが示されている．ここで楕円分布とは密度関数が，中央値ベクトル $\boldsymbol{\nu}$，正定値対称行列 Σ 及びスカラー値関数 f を用いて

$$p(\boldsymbol{x}) = |\Sigma|^{-1/2} f((\boldsymbol{x} - \boldsymbol{\nu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\nu})) \tag{3}$$

の形で書ける分布をいう [9]. 正規分布は楕円分布の一例であることは明らかである．

3. Kernel Sliced Inverse Regresion

本節では，カーネルトリックによる SIR の非線型化手法である，カーネル層別逆回帰（Kernel Sliced Inverse Regression, KSIR [2]）を導入する．

正定値カーネル関数 $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$ が与えられたとする．KSIR のアイディアは，データを変換

$$\phi : \boldsymbol{x} \to \phi(\boldsymbol{x}) \in \mathcal{H} \tag{4}$$

によって特徴空間 $\mathcal{H}$ に写像し，$\mathcal{H}$ 上で SIR を実行することである．以下では誤解の恐れがない場合には特徴空間における次元削減行列も，原空間における次元削減行列と同様に $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_q)$ で表すものとする．特徴空間における回帰モデルを，

$$y = f(\langle \boldsymbol{b}_1, \phi(\boldsymbol{x}) \rangle, \ldots, \langle \boldsymbol{b}_q, \phi(\boldsymbol{x}) \rangle; \epsilon) \tag{5}$$

とする．本稿では，特徴空間 $\mathcal{H}$ における内積を $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ で表し，混乱の恐れがない場合には内積記号の添字 $\mathcal{H}$ は省略する．また，学習用のデータを特徴空間に写像して得られる特徴ベクトルは $\mathcal{H}$ において $\sum_{i=1}^n \phi(\boldsymbol{x}_i) = 0$ を満たすように中心化されているとする．この中心化操作は，カーネル行列のみを用いて

$$\tilde{K}_{ij} = [Q_n K Q_n]_{ij}, \quad Q_n = I_n - \frac{1}{n} 1_n 1_n^\top \tag{6}$$

で実現できる．ただし，$1_n = (1, \ldots, 1)^\top$ である．スライスは通常の SIR と同様に応答変数 Y の値に応じて作成する．各スライス内での特徴ベクトルの平均ベクトルを

$$\psi_l = \frac{1}{|S_l|} \sum_{i \in S_l} \phi(\boldsymbol{x}_i), \quad l = 1, \ldots, L \tag{7}$$

で定義すると，

$$M_{\mathcal{H}} = \frac{1}{L} \sum_{l=1}^{L} |S_l| \psi_l \psi_l^\top \tag{8}$$

で特徴空間における重み付き共分散行列が定義できる．形式的には，特徴空間の部分空間としての次元削減空間は固有値問題

$$M_{\mathcal{H}} \boldsymbol{b} = \lambda M_{\mathcal{H}} \boldsymbol{b} \tag{9}$$

を解くことで得られるが，この問題を直接解くためには特徴ベクトルが陽に計算できる形でなければならない．そこで，KSIR におけるモデル (5) の計算には特徴ベクトルの B による射影 $\langle \boldsymbol{b}_j, \phi(\boldsymbol{x}) \rangle, j = 1, \ldots, q$ が得られれば十分であることに注目する．今，$\tilde{K}$ を全データを用いて計算した中心化カーネル行列，H を

$$H_{ij} = \begin{cases} 1/|S_l|, & \boldsymbol{x}_i \in S_l \text{ かつ } \boldsymbol{x}_j \in S_l \\ 0, & otherwise \end{cases} \tag{10}$$

として，一般化固有値問題

$$\tilde{K} H \tilde{K} \boldsymbol{c} = \lambda (\tilde{K}^2 + s I_n) \boldsymbol{c} \tag{11}$$

を考える．ここで $s > 0$ は一般化固有値問題の条件数の問題を回避するための正則化パラメタである．この一般化固有値問題の解 $\boldsymbol{c}_j \in \mathbb{R}^n, j = 1, \ldots, q$ を用いて，特徴空間上の点 $\phi(\boldsymbol{x})$ の次元削減空間への射影は

$$\langle \boldsymbol{b}_j, \phi(\boldsymbol{x}) \rangle = (k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_n)) \boldsymbol{c}_j, j = 1, \ldots, q$$

で得られる．以上より，一般化固有値問題 (11) を解くことで，特徴ベクトルを陽に計算することなく，カーネル関数のみを用いて回帰モデル (5) の評価が可能となる．例えば KSIR によって次元削減をした後に線型回帰を行う場合，次のような手続きとなる．まず，与えられたデータに KSIR を適用し，特徴空間における次元削減空間を張るベクトル $\boldsymbol{c}_j, j = 1, \ldots, q$ を得る．次にデータ $\boldsymbol{x}_i$ を，各次元が

$$[\boldsymbol{v}_i]_j = \langle \boldsymbol{b}_j, \phi(\boldsymbol{x}) \rangle = (k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_n)) \boldsymbol{c}_j \tag{12}$$

で定まる q 次元ベクトルに変換する．q 次元データ $\{\boldsymbol{v}_i\}_{i=1}^n$ に対して，線型回帰モデル $y = \boldsymbol{w}^\top \boldsymbol{v} + b$ を当てはめる．

なお，一般化固有値問題 (11) を解くことで得られる射影ベクトルは，一般に直交していない．そこで，得られた射影ベクトルが互いに正規直交系をなすように，準直交化変換を施す [10].

4. Convex Combination of Kernel Functions

多くの多変量解析手法は，カーネル法により特徴空間における解析手法に拡張が可能である．カーネル法の詳細は [11] に詳しい．カーネル関数を選択することは，特徴空間への説明変数の写像を定める．すなわち，カーネル関数を定めることは，説明変数 X の特徴空間における分布を定めることに他ならない．本研究では，観測した説明変数値が特徴空間において正規分布に従うようにカーネル関数を選択する方法を提案する．正規分布は楕円分布の一つであり，SIR が説明変数の分布に課す条件が特徴空間において満たされることになる．これにより，KSIR による非線型次元削減の性能向上が期待される．

説明変数の実現値 $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1,\ldots,n}$ を考える．$\boldsymbol{x}_i$ は入力空間 $\mathcal{X}$ に属するとする．Multiple Kernel Learning (MKL) の枠組みでは，入力空間 $\mathcal{X}$ から S 個の異なる特徴空間 $\mathcal{H}_s$ への特徴写像 $\phi_1, \ldots, \phi_S$, $\phi_s : \mathcal{X} \to \mathcal{H}_s$, $s = 1, \ldots, S$ が与えられていると考える．この特徴空間の次元は任意であり，関数空間でも良い．ここでは，各写像 ϕ_s に，$k_s(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi_s(\boldsymbol{x}), \phi_s(\boldsymbol{x}') \rangle_{\mathcal{H}_s}$ なる再生核 k_s が対応する状況を考える．以下では，ϕ_s と k_s の両方を適宜利用する．また，$K_s = [k_s(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1,\ldots,n} = [\langle \phi_s(\boldsymbol{x}_i), \phi_s(\boldsymbol{x}_j) \rangle]_{i,j=1,\ldots,n}$ でデータ $\mathcal{D}$ から生成したカーネル行列を表すものとする．カーネル関数の凸結合の結合係数が属する空間を $S-1$ 単体 Δ_S として，

$$k_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{s=1}^{S} \beta_s k_s(\boldsymbol{x}, \boldsymbol{x}'), \ \boldsymbol{\beta} \in \Delta_S \tag{13}$$

の形のカーネル関数の凸結合を考える．この結合されたカーネルに対して，$\phi_{\boldsymbol{\beta}} = (\sqrt{\beta_1}\phi_1, \ldots, \sqrt{\beta_S}\phi_S)$ と置けば $k_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi_{\boldsymbol{\beta}}(\boldsymbol{x}), \phi_{\boldsymbol{\beta}}(\boldsymbol{x}') \rangle$ が成り立つことに注意する．また，カーネル行列を用いた表現では

$$K_{\boldsymbol{\beta}} = \sum_{s=1}^{S} \beta_s K_s, \ \boldsymbol{\beta} \in \Delta_S, \tag{14}$$

のように書くことができる．この凸結合係数 $\boldsymbol{\beta}$ が変化することで，結合後のカーネル関数に付随する特徴写像が変化し，説明変数の特徴空間における分布が変化する．

MKL をカーネル関数の凸結合により定まる特徴空間における分布を調整する方法として捉えるアイディアは [12] において提案され，LDA による判別を前提として，特性関数に基づく最適化の目的関数が提案されている．本稿では，[12] に基づき，特性関数に基づく最適化問題としてカーネル最適化問題を定式化する．

5. 経験特性関数

分布関数 $F(\boldsymbol{x})$ を持つ確率変数 X の特性関数は

$$c(\boldsymbol{t}) = \int_{-\infty}^{\infty} e^{i\boldsymbol{t}^\top \boldsymbol{x}} \mathrm{d}F(\boldsymbol{x}), \quad \boldsymbol{t} \in \mathcal{X} \tag{15}$$

で定義される．なお，$c(\boldsymbol{t})$ は X の確率密度関数の Fourier 変換に他ならない．独立同一分布に従うサンプル $\mathcal{D} = \{\boldsymbol{x}_j\}_{j=1,\ldots,n}$ に対し，経験特性関数は

$$c_{\mathcal{D}}(\boldsymbol{t}) = \frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} e^{i\boldsymbol{t}^\top \boldsymbol{x}_j} \tag{16}$$

で定義される．特に正規分布を扱うときには，以下の定理から分かるように，特性関数の絶対値の二乗のみを考察すれば良い:

Theorem1　ある分布の特性関数 $c(\boldsymbol{t})$ が与えられた時，その分布が正規分布であるための必要十分条件は $-\log|c(\boldsymbol{t})|^2$ が

$$-\log|c(\boldsymbol{t})|^2 = \boldsymbol{t}^\top \Sigma \boldsymbol{t}$$

の形をしていることである．ここで，Σ は正定値対称行列である．

平均ベクトル $\boldsymbol{\mu}$，共分散行列 Σ で定まる多次元正規分布の特性関数は

$$c^*(\boldsymbol{t}) = \exp(i\boldsymbol{t}^\top \boldsymbol{\mu}) \exp(-\frac{1}{2}\boldsymbol{t}^\top \Sigma \boldsymbol{t}) \tag{17}$$

であり，その絶対値の二乗は $|c^*(\boldsymbol{t})|^2 = \exp(-\boldsymbol{t}^\top \Sigma \boldsymbol{t})$ となる．データ分布が正規分布であるという仮定の下で $c^*(\boldsymbol{t})$ の経験特性関数による近似 $c^*_{\mathcal{D}}(\boldsymbol{t})$ は，平均と共分散行列の推定量を式 (17) に代入することで得られる．

関数 (16) 及び (17) は内積のみを用いて記述されているので, 写像 $\phi_{\boldsymbol{\beta}}$ によって特徴空間に写像された場合も，カーネル関数 $k_{\boldsymbol{\beta}}$ の値のみを用いて計算することが出来る．つまり，特徴空間におけるデータ $\phi_{\boldsymbol{\beta}}(\boldsymbol{x}_j)$ と，特徴空間における一点 $\phi_{\boldsymbol{\beta}}(\boldsymbol{t})$ を考えた時，特徴空間における分布の経験特性関数は

$$c_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta}) = \frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \exp\left(i k_{\boldsymbol{\beta}}(\boldsymbol{t}, \boldsymbol{x}_j)\right), \tag{18}$$

で計算出来て，その絶対値の二乗は

$$\begin{aligned} |c_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta})|^2 &= \left\{ \frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \cos\left(\sum_{s=1}^{S} \beta_s k_s(\boldsymbol{t}, \boldsymbol{x}_j) \right) \right\}^2 \\ &+ \left\{ \frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \sin\left(\sum_{s=1}^{S} \beta_s k_s(\boldsymbol{t}, \boldsymbol{x}_j) \right) \right\}^2 \end{aligned} \tag{19}$$

となる．

正規分布の経験特性関数の絶対値二乗 $|c^*_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta})|$ の定義には共分散行列が現れるが，この量を特徴付ける二次形式が陽に共分散行列を計算することなく評価可能である:

$$\begin{aligned} S^*_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta}) &\stackrel{\mathrm{def}}{=} \langle \phi_{\boldsymbol{\beta}}(\boldsymbol{t}), \hat{\Sigma}_{\mathcal{D}} \phi_{\boldsymbol{\beta}}(\boldsymbol{t}) \rangle & (20) \\ &= \frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \left(k_{\boldsymbol{\beta}}(\boldsymbol{t}, \boldsymbol{x}_j) - \hat{\mu}_{\boldsymbol{\beta}}(\boldsymbol{t}) \right)^2 \\ &= \boldsymbol{\beta}^\top \left(\frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j) \boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j)^\top \right) \boldsymbol{\beta} \\ &= \boldsymbol{\beta}^\top V_{\mathcal{D}} \boldsymbol{\beta}. & (21) \end{aligned}$$

ここで，$\hat{\mu}_{\boldsymbol{\beta}}(\boldsymbol{t}) = \sum_{s=1}^{S} \beta_s \hat{\mu}_s(\boldsymbol{t})$ は点 $\phi_{\boldsymbol{\beta}}(\boldsymbol{t})$ において評価された特徴空間における平均ベクトルであり，その要素は $\hat{\mu}_s(\boldsymbol{t}) = (1/n) \sum_{\boldsymbol{x}_i \in \mathcal{D}} k_s(\boldsymbol{t}, \boldsymbol{x}_i)$ である．ベクトル $\boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j)$ と行列 $V_{\mathcal{D}}$ はそれぞれ

$$\begin{aligned} \boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j) &= \begin{pmatrix} k_1(\boldsymbol{t}, \boldsymbol{x}_j) - \hat{\mu}_1(\boldsymbol{t}) \\ \vdots \\ k_S(\boldsymbol{t}, \boldsymbol{x}_j) - \hat{\mu}_S(\boldsymbol{t}) \end{pmatrix}, \\ V_{\mathcal{D}} &= \frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j) \boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j)^\top \end{aligned}$$

である．特徴空間においてデータが正規分布に従っている時，その経験特性関数は

$$\begin{aligned} &c^*_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta}) = \\ &\exp\left(i \hat{\mu}_{\boldsymbol{\beta}}(\boldsymbol{t})\right) \exp\left(-\frac{1}{2n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \left(k_{\boldsymbol{\beta}}(\boldsymbol{t}, \boldsymbol{x}_j) - \hat{\mu}_{\boldsymbol{\beta}}(\boldsymbol{t}) \right)^2 \right), \end{aligned}$$

であり，その絶対値の二乗は

$$\begin{aligned} &|c^*_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta})|^2 \\ &= \exp\left\{ -\frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \left(k_{\boldsymbol{\beta}}(\boldsymbol{t}, \boldsymbol{x}_j) - \hat{\mu}_{\boldsymbol{\beta}}(\boldsymbol{t}) \right)^2 \right\} & (22) \\ &= \exp\left\{ -\boldsymbol{\beta}^\top V_{\mathcal{D}} \boldsymbol{\beta} \right\} = \exp\{ -S^*_{\mathcal{D}}(\boldsymbol{t}, \boldsymbol{\beta}) \} & (23) \end{aligned}$$

となる．上式から，$-\log|c^*_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta})|^2 = S^*_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta}) = \boldsymbol{\beta}^\top V_{\mathcal{D}} \boldsymbol{\beta}$ が成り立つことが分かり，正規分布の絶対値の二乗は対数変換により単純な二次形式として表せることが分かる．そこで，正規分布に限らず一般に経験特性関数の絶対値の二乗も (負の) 対数関数で変換したものを考え,

$$S_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta}) = -\log|c_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta})|^2 \tag{24}$$

と記す.

特性関数を利用する際，評価点 $\phi_{\boldsymbol{\beta}}(\boldsymbol{t})$ の選択が問題となる．本研究では，n 個の学習データからランダムに $\lfloor n/10 \rfloor$ 点をサンプリングして評価点として利用し，それらの平均で特性関数値を評価することにする．記述の簡単のため，以下では $\boldsymbol{t}$ における特性関数を評価する場合はこの平均操作が取られているものと仮定する．

任意に評価点 $\boldsymbol{t}$ を選択した上で，正規性の評価尺度を

$$J(\boldsymbol{\beta};\mathcal{D}) = \left(S^*_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta}) - S_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta}) \right)^2 \tag{25}$$

で定義する．KSIR のためのカーネル最適化問題は,

$$\min_{\boldsymbol{\beta}\in\Delta_S} J(\boldsymbol{\beta};\mathcal{D}) \tag{26}$$

のように定式化される．本研究ではこの最適化問題を，拡張ラグランジュ法と逐次二次計画法を利用して解いた [13].

6. Experimental Result

本節ではまず，人工データを用いて提案するアルゴリズムがガウス性を最大化するようなカーネルを選択できることを確認する．次に，種々のデータセットを用いた実験により，提案手法の有用性を示す.

6.1 人工データによる実験

まず，2 次元ユークリッド空間において正規分布に従うデータ集合 $\{\boldsymbol{x}_i\}_{i=1}^n$ を生成し，このデータから以下の 5 つのカーネル関数を用いてカーネル行列を作成する:

(1) Linear kernel: $k_1(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{x}_i^\top\boldsymbol{x}_j$
(2) Gaussian kernel: $k_2(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-0.1\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2)$
(3) Laplace kernel: $k_3(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-0.1\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2)$
(4) Bessel kernel: $k_4(\boldsymbol{x}_i,\boldsymbol{x}_j)=\frac{J_2(2\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2)}{(\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2)^{-2}}$
(5) Polynomial kernel: $k_5(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i^\top\boldsymbol{x}_j+1)^2$

ここで, J_2 は次数 2 の第一種ベッセル関数である．上記のカーネルのうち，linear kernel はユークリッド空間における通常の内積に対応するカーネルであり，このカーネルに対応する特徴空間は元のユークリッド空間そのものである．従って，元の空間でデータが正規分布に従っている場合は，linear kernel を選択することで特徴空間におけるデータ分布は正規性に従うことになる．データ集合 $\{\boldsymbol{x}_i\}_{i=1}^n$ から生成した k_1 から k_5 に対応するカーネル行列を $K_1,\ldots,K_5$ として, $K=\sum_{s=1}^5\beta_sK_s$ の結合係数を提案手法により最適化する．図 1 に，最適化アルゴリズムの第 $1,3,45,50$ ステップにおける結合係数 $\beta_1,\ldots,\beta_5$ をプロットしたものを示す．この結果から，特徴空間においてデータが正規分布に従うようなカーネル関数が選択されていることが分かる.

6.2 実データを用いた実験

UCI repository と, 回帰分析の教科書の実例として採用されている例から 8 個のデータセットを用いて，提案手法によって次元削減をした場合の線形回帰による予測誤差を調べる．また，NEDO の実証実験において収集した太陽光発電パネルの出力を，前日の気象庁発表の府県予報と地域時系列予報を利用して回帰する問題 (データ名"pv") にも提案手法を適用した．本実験では以下の 40 個のカーネル関数の線型結合を最適化した:

(1) Polynomial kernel: $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i^\top\boldsymbol{x}_j+o)^d$,
$o\in\{0.5,1\}, d\in\{2,3,4,5\}$
(2) Gaussian kernel: $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\lambda\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2)$,
$\lambda\in\{0.01,0.1,0.2,\ldots,0.9,1,1.5,2,3,4,5\}$
(3) Laplace kernel: $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\lambda\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2)$,
λ は Gaussian kernel と同じ範囲.

真の部分空間の次元は未知であるため，部分空間の次元を 1 からもとのデータの次元 −1 になるまで増やして実験を行ない，各次元における回帰の絶対誤差の平均を示す．絶対誤差の平均は，10-fold のクロスバリデーションによって算出した．また，単一のカーネル関数を用いた KSIR との比較のため，学習データを利用した 10-fold クロスバリデーションにより各次元に射影した場合の回帰誤差が最小となるようなカーネルを上記の候補カーネルから選択した．この単一のカーネル関数を用いた KSIR で次元削減をした場合の線型回帰，原空間での線型回帰，原空間での線型 SIR による回帰，及び提案手法により最適化したカーネル関数を用いた KSIR による次元削減をした場合の線型回帰それぞれの平均絶対誤差を図 2 に示す．KSIR は線型の SIR の性能を上回る可能性があるが，カーネル関数の選択が問題となる．実際，クロスバリデーションにより選択したカーネル関数を用いた場合，線型の SIR を上回る回帰性能を示した例はほとんど無い一方で，提案手法により最適化されたカーネルを用いた KSIR は多くのデータセットに対して良好な回帰性能を示している.

7. 終わりに

本研究では，カーネル層別逆回帰 (KSIR) におけるカーネル関数の選択手法として，データが特徴空間で正規分布に従うようにカーネル関数を選ぶアプローチを提案した．特徴空間におけるデータ分布の正規性を特性関数を用いて評価し，正規分布の特性関数と経験特性関数の二乗誤差からなる目的関数を，カーネル関数の結合係数に関して最適化した．最適化したカーネル関数を用いた KSIR により，クロスバリデーションによって選択したカーネル関数を用いた KSIR を上回る予測精度を達成した.

謝辞

本研究の一部は，新エネルギー・産業技術総合開発機構 (NEDO) 研究開発委託事業「安全・低コスト大規模蓄電システム技術開発」にて行われた.

参考文献

[1] K.-C. Li: "Sliced inverse regression for dimension reduction", Journal of the American Statistical Association, **86**, 414, pp. 316–327 (1991).
[2] H.-M. Wu: "Kernel sliced inverse regression with applications to classification", Journal of Computational and Graphical Statistics, **17**, 3, pp. 590–610 (2008).
[3] R. Cook and S.Weisberg: "Sliced inverse regression for dimension reduction: Comment", Journal of the American Statistical Association, **86**, 414, pp. 328–332 (1991).
[4] B. Li and S. Wang: "On directional regression for dimension reduction", Journal of the American Statistical

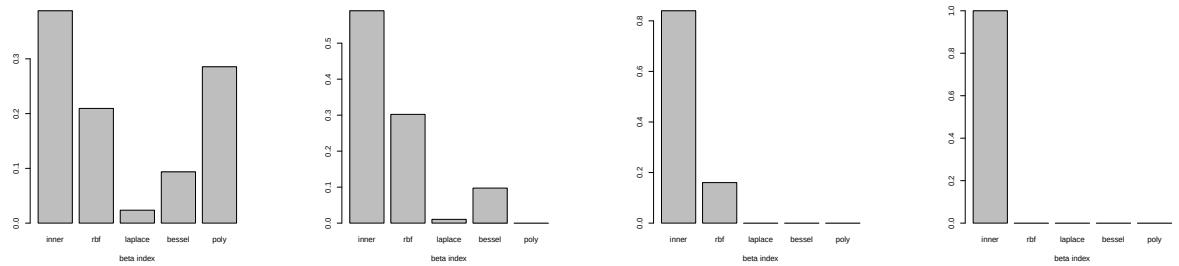

図 1 人工データを用いたカーネル選択実験結果．各カーネル関数の結合係数を棒グラフに表す．4 個のグラフは左から最適化の繰り返し第 1, 3, 45, 50 回目の結果。

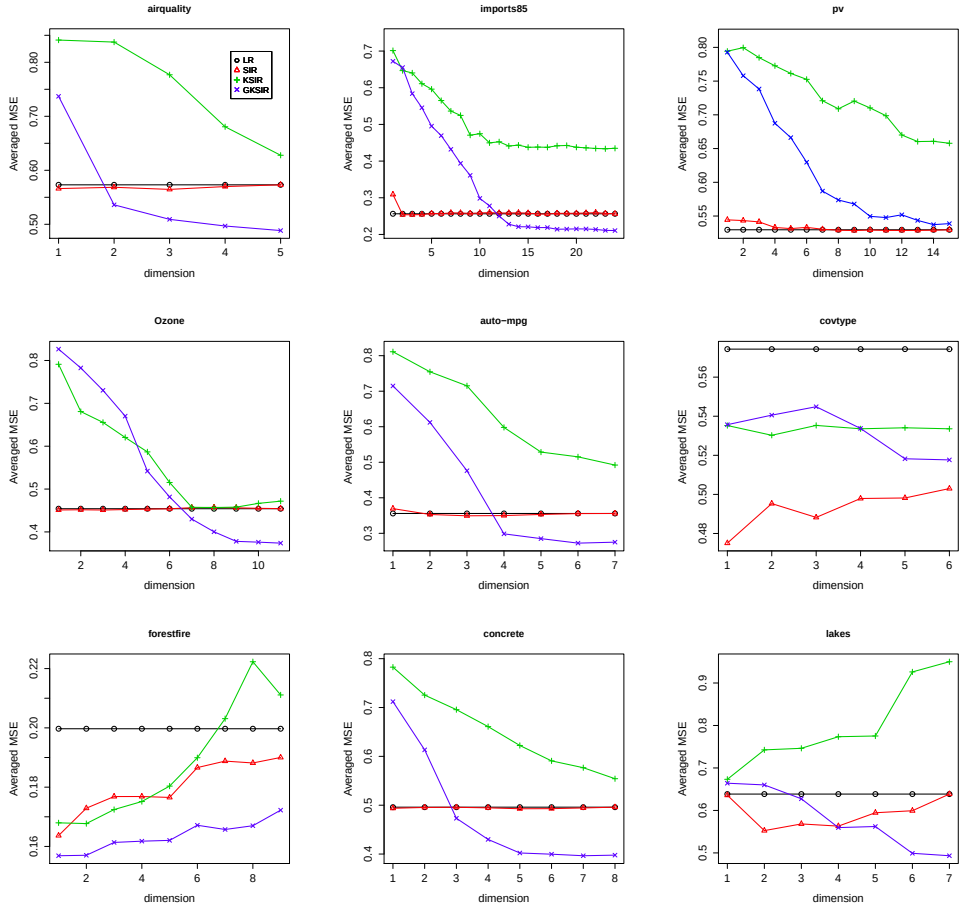

図 2 SIR, KSIR, 及び提案手法で次元削減を行ったデータを用いた線型回帰による予測の絶対誤差の平均.

Association, **102**, pp. 997–1008 (2007).

[5] R. D. Cook and L. Ni: “Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach”, Journal of the American Statistical Association, **100**, 470, pp. 410–428 (2005).

[6] L. Scrucca: “Model-based SIR for dimension reduction”, Computational Statistics & Data Analysis, **55**, 11, pp. 3010–3026 (2011).

[7] H. Hino, K. Wakayama and N. Murata: “Sliced inverse regression with conditional entropy minimization”, ICPR ’12 (2012)

[8] M. L. Eaton: “A characterization of spherical distributions”, Journal of Multivariate Analysis, **20**, 2, pp. 272–276 (1986).

[9] F. Romito: “Elliptically symmetric distributions: A review of achieved results and open issues”, New developments in classification and data analysis, Studies in Classification, Data Analysis, and Knowledge Organization, **3**, pp. 359–366 (2005).

[10] A. Hyvärinen, J. Karhunen and E. Oja: “Independent Component Analysis”, J. Wiley, New York (2001).

[11] J. Shawe-Taylor and N. Cristianini: “Kernel Methods for Pattern Analysis”, Cambridge University Press, New York, NY, USA (2004).

[12] H. Hino, N. Reyhani and N. Murata: “Multiple kernel learning with gaussianity measures”, Neural Computation, **24**, 7, pp. 1853–1881 (2012).

[13] A. Ghalanos and S. Theussl: “Rsolnp: General Nonlinear Optimization Using Augmented Lagrange Multiplier Method” (2012). R package version 1.12.