基于系数正则的同时回归估计

陈珩¹, 黄尉², 陈迪荣^{3*}

1. 首都经济贸易大学统计学院, 北京 100070;

2. 合肥工业大学数学学院, 合肥 230026;

3. 北京航空航天大学数学与系统科学学院, 北京 100191

E-mail: hchenbuaa@sina.cn, whuang@hfut.edu.cn, drchen@buaa.edu.cn

收稿日期: 2018-08-14; 接受日期: 2018-10-10; 网络出版日期: 2019-05-15; * 通信作者

国家自然科学基金 (批准号: 11501380, 11571267 和 91538112) 资助项目

摘要 提取两个随机向量 X 与 Y 之间的相关性是非常重要的问题. 核方法被用来提取非线性的相关性. 本文通过极小化方差 $\text{Var}[f(X) - g(Y)]$ 得到最大相关性, 称为同时回归, 其中 $f(X)$ 和 $g(Y)$ 分别是两个不同的再生核空间中的函数. 本文利用正则经验方差极小化得到估计. 为了所得的估计函数具有稀疏性, 本文采用系数的 l_1 范数作为惩罚项, 在一些常规条件下建立学习率. 同时回归问题与典型相关分析、切片逆回归等密切相关.

关键词 相关性 系数正则 同时回归 学习率

MSC (2010) 主题分类 62G08, 62J02

1 引言

近年来, 为了提取随机变量之间的非线性特征, 核方法得到了研究, 其中有核典型相关分析 (kernel canonical correlation analysis, KCCA)、核主成分分析 (kernel principal component analysis, KPCA) 和核判别分析 (kernel discriminant analysis, KDA) 等, 广泛被应用于相关性和数据降维中.

由 Li^[1] 和 Duan 和 Li^[2] 提出的切片逆回归 (sliced inverse regression, SIR) 是降维的一种方法, SIR 可以被用来寻找数据的紧凑表示, 从而发现高维数据中的内在结构. 设 $X \in \mathcal{X} \subset \mathbb{R}^p$ 和 $Y \in \mathcal{Y} \subset \mathbb{R}^q$ 是服从联合分布 P_{XY} 的两个随机向量. SIR 的核心想法是: 将 p 维数据 X 投影到某个 s 维的子空间 $\text{span}\{b_1, \dots, b_s\}$, 通常 $s \ll p$, 这些投影获取了 X 中与 Y 相关的所有信息.

文献 [3] 将 SIR 解释为基于变换的投影追踪, 从而说明了 SIR 与相关性的密切关系. 基于变换的投影追踪的目标为, 找到独立随机变量的组合, 使得变换后的随机变量有最大相关系数. 下面详细介绍文献 [3] 的解释. 对于一个方向 $b \in \mathbb{R}^p$, 定义 $R(b)$ 为最大的相关系数 $\rho(T(Y), b^T X)$, 其中 T 跑遍所

英文引用格式: Chen H, Huang W, Chen D R. Coefficient regularizer based simultaneous regression estimation (in Chinese). Sci Sin Math, 2019, 49: 1251–1260, doi: 10.1360/N012018-00200

有的变换, b^T 表示 b 的转置. 文献 [3] 证明了 SIR 中的方向是下述逐次问题的解:

$$\begin{aligned} \max_{b \in \mathbb{R}^p} \quad & R^2(b) \\ \text{s.t.} \quad & \text{Cov}[b^T X, b_k^T X] = 0, \quad k = 1, \dots, j-1, \end{aligned} \quad (1.1)$$

这里当 $j = 1$ 时没有约束条件. 第 j 个最优变换为 $T_j(y) = E(b_j^T X | Y = y)$, 它是下述问题的解:

$$\min \text{Var}[T(Y) - b_j^T X], \quad T \in L_2(\mathcal{Y}). \quad (1.2)$$

SIR 已经被成功推广和应用到各个领域. 作为一个线性方法, SIR 无法获取非线性结构. 将核方法应用到 SIR, 文献 [4] 在再生核空间中提出了核 SIR (kernel SIR, KSIR), 这是 SIR 的一种非线性推广.

对 $\mathcal{X}$ 上的 Mercer 核 $k_{\mathcal{X}}(x, x')$, 由它所导出的积分算子具有非负特征值 λ_i , 以及对应的正交特征函数 $\varphi_i(x)$, $i = 1, \dots, d$, 其中 $d \leq \infty$. 于是, $k_{\mathcal{X}}(x, x')$ 有如下特征谱分解 (参见文献 [5, 定理 4.10]):

$$k_{\mathcal{X}}(x, x') = \sum_{i=1}^d \lambda_i \varphi_i(x) \varphi_i(x'), \quad x, x' \in \mathcal{X}, \quad (1.3)$$

相应的特征映射定义为 $x \mapsto \Phi(x) := (\sqrt{\lambda_1} \varphi_1(x), \dots, \sqrt{\lambda_d} \varphi_d(x)) \in \mathcal{F}$, 其中 $\mathcal{F}$ 为特征空间, 其内积可以通过 $\langle \Phi(x), \Phi(x') \rangle = k_{\mathcal{X}}(x, x')$ 得到. 这使得我们可以只与核打交道, 而不需要知道特征映射的显示表示. 再生核 Hilbert 空间 $\mathcal{H}_{\mathcal{X}}$ 是由下述函数 h 组成的集合: 存在 $b = \{b_i\} \in \ell^2$, 使得 $h = \sum_{i=1}^d b_i \sqrt{\lambda_i} \varphi_i$. 换句话说, $h(x) = \langle b, \Phi(x) \rangle = \langle h, k_{\mathcal{X}}(x, \cdot) \rangle_{\mathcal{H}_{\mathcal{X}}}$, 其中 $\mathcal{H}_{\mathcal{X}}$ 上的内积为

$$\langle h_1, h_2 \rangle_{\mathcal{H}_{\mathcal{X}}} = \sum_{i=1}^d b_i^{(1)} b_i^{(2)}, \quad h_1 = \sum_{i=1}^d b_i^{(1)} \sqrt{\lambda_i} \varphi_i, \quad h_2 = \sum_{i=1}^d b_i^{(2)} \sqrt{\lambda_i} \varphi_i. \quad (1.4)$$

由上式可知, 再生核 Hilbert 空间中的 K 范数为 $\|h\|_{\mathcal{H}_{\mathcal{X}}} = \sqrt{\langle h, h \rangle_{\mathcal{H}_{\mathcal{X}}}}$. 对 $\mathcal{Y}$ 上的 Mercer 核 $k_{\mathcal{Y}}(y, y')$, 类似地可以定义再生核 Hilbert 空间 $\mathcal{H}_{\mathcal{Y}}$, 以及 K 范数 $\|\cdot\|_{\mathcal{H}_{\mathcal{Y}}}$.

在 KSIR 中, 特征 $\Phi(x)$ 被投影到 s 维子空间

$$(\langle b_1, \Phi(X) \rangle, \dots, \langle b_s, \Phi(X) \rangle), \quad b_j \in \mathcal{F}, \quad j = 1, \dots, s, \quad (1.5)$$

其中 b_j ($j = 1, \dots, s$) 是 KSIR 的方向. 一个重要的问题是, 如何估计 $\langle b_j, \Phi(X) \rangle$, $1 \leq j \leq s$.

给定独立取自于 P_{XY} 的样本 $\mathbf{z} = \{Z_i\}_{i=1}^n$, $Z_i = (X_i, Y_i)$, $i = 1, 2, \dots, n$, 文献 [4] 通过求解一个广义特征值问题, 对子空间 $(\langle b_1, \Phi(X) \rangle, \dots, \langle b_s, \Phi(X) \rangle)$ 进行估计. 一般地, 利用经验数据来估计方向 b_j ($j = 1, \dots, s$) 是不可能的. 文献 [6] 考虑了正则 KSIR, 并得到下述相容性结果:

$$|\langle \hat{b}_j, \Phi(x) \rangle - \langle b_j, \Phi(x) \rangle| = o_p(1), \quad j = 1, \dots, d_{\Gamma}, \quad (1.6)$$

其中 d_{Γ} 是 $\Gamma = \text{Cov}[E[\Phi(X) | Y = y]]$ 的秩.

如同前面所述的 SIR 与 (1.2) 的关系, 我们可以知道 KSIR 与下述称之为同时回归的问题等价. 对 j ($j = 1, \dots, s$) 依次考虑第 j 个最小化问题 [7]:

$$\begin{aligned} (f_j, g_j) = \min \quad & \text{Var}[f(X) - g(Y)] \\ \text{s.t.} \quad & (f, g) \in \mathcal{H}_{\mathcal{X}} \times \mathcal{H}_{\mathcal{Y}}, \quad \text{Var}[f(X)] = 1, \\ & \text{Cov}[f(X), f_k(X)] = 0, \quad k = 1, \dots, j-1. \end{aligned} \quad (1.7)$$

文献 [7] 提出了以 K 范数为正则项的经验方差最小化算法, 并在一些温和的条件下建立了学习效率 $O_p(n^{-\theta})$, 其中 $\forall \theta < \frac{1}{2}$, 并且利用该结果改进了文献 [8] 中的学习率. 为了所得的估计函数具有稀疏性, 系数正则化的方法发挥了重要的作用 (参见文献 [9]). 本文提出并考虑系数正则化方差极小算法.

为了给出算法, 首先引入假设空间, 记 $\mathbf{x} = \{X_i\}_{i=1}^n$, $\mathbf{y} = \{Y_i\}_{i=1}^n$,

$$\mathcal{H}_{\mathbf{x}} = \left\{ \sum_{i=1}^n \alpha_i k_{\mathcal{X}}(X_i, \cdot) \mid \alpha_i \in \mathbb{R}, 1 \leq i \leq n \right\}, \quad \mathcal{H}_{\mathbf{y}} = \left\{ \sum_{i=1}^n \beta_i k_{\mathcal{Y}}(Y_i, \cdot) \mid \beta_i \in \mathbb{R}, 1 \leq i \leq n \right\}, \quad (1.8)$$

并定义正则项

$$\Omega_{\mathbf{x}}(f) = \inf \left\{ \sum_{i=1}^n |\alpha_i| \mid f = \sum_{i=1}^n \alpha_i k_{\mathcal{X}}(X_i, \cdot) \right\}, \quad \Omega_{\mathbf{y}}(g) = \inf \left\{ \sum_{i=1}^n |\beta_i| \mid g = \sum_{i=1}^n \beta_i k_{\mathcal{Y}}(Y_i, \cdot) \right\}. \quad (1.9)$$

本文对 (f_j, g_j) ($j = 1, \dots, s$) 的估计通过下述依次正则经验极小化问题得到:

$$\begin{aligned} (f_{j,\mathbf{z}}, g_{j,\mathbf{z}}) &= \arg \min \widehat{\text{Var}}[f(X) - g(Y)] + \lambda(\Omega_{\mathbf{x}}(f) + \Omega_{\mathbf{y}}(g)) \\ \text{s.t. } (f, g) &\in \mathcal{H}_{\mathbf{x}} \times \mathcal{H}_{\mathbf{y}}, \quad \widehat{\text{Var}}[f(X)] = 1, \\ \widehat{\text{Cov}}[f_{k,\mathbf{z}}(X), f(X)] &= 0, \quad k = 1, \dots, j-1, \end{aligned} \quad (1.10)$$

其中 $\lambda > 0$ 是正则化参数, $\widehat{\text{Var}}[h(Z)] = \widehat{\text{Cov}}[h(Z), h(Z)]$, $\widehat{\text{Cov}}[h_1(Z), h_2(Z)]$ 是 $h_1(Z)$ 和 $h_2(Z)$ 的经验协方差, 由下式定义:

$$\widehat{\text{Cov}}[h_1(Z), h_2(Z)] = \frac{1}{n} \sum_{i=1}^n (h_1(Z_i) - \hat{E}(h_1(Z)))(h_2(Z_i) - \hat{E}(h_2(Z))), \quad \hat{E}(h(Z)) = \frac{1}{n} \sum_{i=1}^n h(Z_i). \quad (1.11)$$

该极小化问题可以通过交替最小平方算法来求解 (参见文献 [10]).

本文的主要目的是建立该算法在 $j = 1$ 时的学习率. 我们在证明了估计误差为 $O_p(n^{-1/6})$ 的基础上, 采用迭代的技巧建立 $O_p(n^{-1/5})$ 的学习率. 与文献 [7] 中 K 范数做正则项不同, $\Omega_{\mathbf{x}}(f)$ 、 $\Omega_{\mathbf{y}}(g)$ 并不是对所有的 $f \in \mathcal{H}_{\mathbf{x}}$ 、 $g \in \mathcal{H}_{\mathbf{y}}$ 都有定义, 为了克服这个困难, 我们需要构造合适的 Banach 空间, 将 $\Omega_{\mathbf{x}}(f)$ 、 $\Omega_{\mathbf{y}}(g)$ 与 Banach 空间的范数联系起来.

本文的结构如下: 第 2 节给出过误差估计; 第 3 节利用迭代技巧给出学习率.

2 过误差分析

本节估计过误差 $\mathcal{E}(f, g) - \mathcal{E}(f_1, g_1)$, 其中, 对任意满足 $\text{Var}[f(X)] < \infty$ 、 $\text{Var}[g(Y)] < \infty$ 的 f 和 g , 误差 $\mathcal{E}(f, g)$ 由下式定义:

$$\mathcal{E}(f, g) = \text{Var}[f(X) - g(Y)]. \quad (2.1)$$

对上述 f 和 g , 经验误差定义为 $\mathcal{E}_n(f, g) = \widehat{\text{Var}}[f(X) - g(Y)]$. 首先将正则过误差分解成下面两种不同的误差:

$$\mathcal{E}(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) - \mathcal{E}(f_1, g_1) + \lambda(\Omega_{\mathbf{x}}(f_{1,\mathbf{z}}) + \Omega_{\mathbf{y}}(g_{1,\mathbf{z}})) = S(\mathbf{z}) + P(\mathbf{z}) + \lambda(\|f_1\| + \|g_1\|), \quad (2.2)$$

其中

$$S(\mathbf{z}) = \{\mathcal{E}(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) - \mathcal{E}_n(f_{1,\mathbf{z}}, g_{1,\mathbf{z}})\} + \{\mathcal{E}_n(f_1, g_1) - \mathcal{E}(f_1, g_1)\} \quad (2.3)$$

称为样本误差;

$$P(\mathbf{z}) = \{\mathcal{E}_n(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) + \lambda(\Omega_{\mathbf{x}}(f_{1,\mathbf{z}}) + \Omega_{\mathbf{y}}(g_{1,\mathbf{z}}))\} - \{\mathcal{E}_n(f_1, g_1) + \lambda(\|f_1\| + \|g_1\|)\} \quad (2.4)$$

称为假设误差.

为了估计误差, 需要引入 Banach 空间, 它包含了算法所需要的假设空间 $\mathcal{H}_{\mathbf{x}}$. Banach 空间 $C_{k_{\mathcal{X}}}$ 是如下形式的所有函数:

$$f = \sum_{i=1}^{\infty} a_i k_{\mathcal{X}}(x_i, \cdot), \quad \{a_i\}_{i=1}^{\infty} \in \ell^1, \quad \{x_i\} \subseteq \mathcal{X} \quad (2.5)$$

构成的集合, 其范数定义为

$$\|f\| := \inf \left\{ \sum_{i=1}^{\infty} |a_i| \mid f = \sum_{i=1}^{\infty} a_i k_{\mathcal{X}}(x_i, \cdot) \right\}, \quad (2.6)$$

类似地可以定义 Banach 空间 $C_{k_{\mathcal{Y}}}$ 以及其相应的范数.

对任意有限集 $\mathbf{x} = \{x_i\}_{i=1}^n$, $\mathbf{y} = \{y_i\}_{i=1}^n$, 由定义可知, $\|f\| \leq \Omega_{\mathbf{x}}(f)$, $\|g\| \leq \Omega_{\mathbf{y}}(g)$. 注意到, 对任意 Mercer 核 $k_{\mathcal{X}}$ 和 $k_{\mathcal{Y}}$, 有

$$\sup_{x \in \mathcal{X}} k_{\mathcal{X}}(x, x) = \sup_{x, x' \in \mathcal{X}} k_{\mathcal{X}}(x, x'), \quad \sup_{y \in \mathcal{Y}} k_{\mathcal{Y}}(y, y) = \sup_{y, y' \in \mathcal{Y}} k_{\mathcal{Y}}(y, y'). \quad (2.7)$$

可知 $\forall f \in C_{k_{\mathcal{X}}}, \forall g \in C_{k_{\mathcal{Y}}}$, 有 $\|f\|_{\infty} \leq \sup_{x \in \mathcal{X}} k_{\mathcal{X}}(x, x) \|f\|$, $\|g\|_{\infty} \leq \sup_{y \in \mathcal{Y}} k_{\mathcal{Y}}(y, y) \|g\|$. 此外, 由

$$\left\| \sum_{i=N_1}^{N_2} a_i k_{\mathcal{X}}(x_i, \cdot) \right\|_{\mathcal{H}_{\mathcal{X}}}^2 = \sum_{i,j=N_1}^{N_2} a_i a_j k_{\mathcal{X}}(x_i, x_j) \leq \sup_{x \in \mathcal{X}} k_{\mathcal{X}}(x, x) \left(\sum_{i=N_1}^{N_2} |a_i| \right)^2 \quad (2.8)$$

可得, 序列 $\sum_{i=1}^{\infty} a_i k_{\mathcal{X}}(x_i, \cdot)$ 在 $\mathcal{H}_{\mathcal{X}}$ 中收敛. 因此, $f \in \mathcal{H}_{\mathcal{X}}$. 类似的结果对 $g \in C_{k_{\mathcal{Y}}}$ 成立, 进一步有

$$\|f\|_{\mathcal{H}_{\mathcal{X}}} \leq \sqrt{\sup_{x \in \mathcal{X}} k_{\mathcal{X}}(x, x)} \|f\|, \quad \|g\|_{\mathcal{H}_{\mathcal{Y}}} \leq \sqrt{\sup_{y \in \mathcal{Y}} k_{\mathcal{Y}}(y, y)} \|g\|. \quad (2.9)$$

下面先估计样本误差.

记 $\kappa = \max\{k_{\mathcal{X}}(x, x), k_{\mathcal{Y}}(y, y) \mid x \in \mathcal{X}, y \in \mathcal{Y}\}$, 由文献 [7] 可知, $\forall t > 0$, 存在 $\mathbf{Z}_1 \subseteq (\mathcal{X} \times \mathcal{Y})^n$, 满足 $P_{\mathbf{X}\mathbf{Y}}^n(\mathbf{Z}_1) \geq 1 - 18e^{-t}$, 以及

$$|\mathcal{E}(f, g) - \mathcal{E}_n(f, g)| \leq 64\kappa \sqrt{\frac{t}{n}} (\|f\|_{\mathcal{H}_{\mathcal{X}}}^2 + \|g\|_{\mathcal{H}_{\mathcal{Y}}}^2), \quad f \in \mathcal{H}_{\mathcal{X}}, \quad g \in \mathcal{H}_{\mathcal{Y}}, \quad \mathbf{z} \in \mathbf{Z}_1. \quad (2.10)$$

命题 2.1 对于任意的 $\lambda < 1$, $t \leq n$ 且 $64\kappa\sqrt{t/n} \leq 1$, 以及任意的 $\mathbf{z} \in \mathbf{Z}_1$, 有

$$S(\mathbf{z}) \leq 64\kappa^2 \sqrt{\frac{t}{n}} ((\|f_{1,\mathbf{z}}\| + \|g_{1,\mathbf{z}}\|)^2 + (\|f_1\| + \|g_1\|)^2). \quad (2.11)$$

由 (2.9) 和 (2.10) 可得

$$\mathcal{E}(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) - \mathcal{E}_n(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) \leq 64\kappa^2 \sqrt{\frac{t}{n}} (\|f_{1,\mathbf{z}}\| + \|g_{1,\mathbf{z}}\|)^2, \quad (2.12)$$

$$\mathcal{E}_n(f_1, g_1) - \mathcal{E}(f_1, g_1) \leq 64\kappa^2 \sqrt{\frac{t}{n}} (\|f_1\| + \|g_1\|)^2. \quad (2.13)$$

命题 2.1 由 $S(\mathbf{z})$ 的定义而得证.

为了估计假设误差, 我们需要以下假设.

假设 2.1 (i) 核 $k_{\mathcal{X}}$ 和 $k_{\mathcal{Y}}$ 满足 γ ($0 < \gamma < 1$) 阶的 Lipschitz 条件, 即存在 $c_1 > 0$, 使得

$$\begin{aligned} |k_{\mathcal{X}}(u, x) - k_{\mathcal{X}}(u, x')| &\leq c_1 |x - x'|_2^\gamma, \quad \forall u, x, x' \in \mathcal{X}, \\ |k_{\mathcal{Y}}(v, y) - k_{\mathcal{Y}}(v, y')| &\leq c_1 |y - y'|_2^\gamma, \quad \forall v, y, y' \in \mathcal{Y}. \end{aligned} \quad (2.14)$$

(ii) X 的边缘分布 ρ_X 和 Y 的边缘分布 ρ_Y 都满足条件 L_τ ($0 < \tau < \infty$), 即对某个 $c_2 > 0$ 和任意球 $B(x, \delta) := \{u \in \mathcal{X} \mid |u - x|_2 < \delta\}$, $B(y, \delta) := \{v \in \mathcal{Y} \mid |v - y|_2 < \delta\}$, 成立

$$\rho_X(B(x, \delta)) \geq c_2 \delta^\tau, \quad \rho_Y(B(y, \delta)) \geq c_2 \delta^\tau, \quad \forall x \in \mathcal{X}, \quad \forall y \in \mathcal{Y}, \quad 0 < \delta \leq 1. \quad (2.15)$$

下面的假设考虑覆盖数问题.

假设 2.2 存在 $\alpha > 0$ 和 $c_\alpha > 0$, 使得

$$\max\{\mathcal{N}(\delta, \mathcal{X}, |\cdot|_2), \mathcal{N}(\delta, \mathcal{Y}, |\cdot|_2)\} \leq c_\alpha \left(\frac{1}{\delta}\right)^\alpha, \quad \forall \delta > 0. \quad (2.16)$$

设 $\Delta > 0$ 是一个常数, $\{s_i\}_{i=1}^n \subset \mathcal{S}$, 这里 $\mathcal{S} = \mathcal{X}$ 或 $\mathcal{S} = \mathcal{Y}$. 若对任意的 $s \in \mathcal{S}$, 存在某个 $j \in \{1, 2, \dots, n\}$, 使得 $|s - s_j|_2 < \Delta$, 则称 $\{s_i\}_{i=1}^n$ 在 $\mathcal{S}$ 是 Δ 稠密的.

引理 2.1 ^[11] 在假设 2.1 和 2.2 条件下, 设 $\{X_j\}_{j=1}^n$ 和 $\{Y_j\}_{j=1}^n$ 是服从于 ρ_X 和 ρ_Y 的独立取样, 则对任意 $t > 1$, 在假设 2.1(ii) 和 2.2 成立时, 至少以 $1 - e^{-t}$ 的概率, $\{X_j\}_{j=1}^n$ 和 $\{Y_j\}_{j=1}^n$ 在 $\mathcal{X}$ 和 $\mathcal{Y}$ 中是 $c_{\alpha, \tau} \left(\frac{\log n + t}{n}\right)^{\frac{\gamma}{\tau}}$ -稠密的, 其中 $c_{\alpha, \tau}$ 是仅依赖于 α 和 τ 的常数.

由引理 2.1 可知, 存在 $\mathbf{Z}_2 \subseteq (\mathcal{X} \times \mathcal{Y})^n$, 满足 $P_{X,Y}^n(\mathbf{Z}_2) \geq 1 - e^{-t}$, 使得任意 $\mathbf{z} = \{(X_i, Y_i)\}_{i=1}^n \in \mathbf{Z}_2$, $\{X_j\}_{j=1}^n$ 和 $\{Y_j\}_{j=1}^n$ 是 $c_{\alpha, \tau} \left(\frac{\log n + t}{n}\right)^{\frac{\gamma}{\tau}}$ -稠密的.

假设误差 $P(\mathbf{z})$ 可以由下面的命题控制.

命题 2.2 设假设 2.1 和 2.2 成立. 存在常数 C_0 , 对任意 $t > 1$, 至少以 $1 - e^{-t}$ 的概率下式成立:

$$P(\mathbf{z}) \leq C_0 \left(\left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} + \sqrt{\frac{t}{n}} \right), \quad (2.17)$$

其中 C_0 是与 $\mathbf{z}$ 、 t 和 n 无关的常数.

证明 对任意 $\eta \in (0, \|f_1\|)$, 存在 f_1 和 g_1 的表示 $f_1 = \sum_{i=1}^{\infty} a_i k_{\mathcal{X}}(u_i, \cdot)$, $g_1 = \sum_{i=1}^{\infty} b_i k_{\mathcal{Y}}(v_i, \cdot)$, 其中 $u_i \in \mathcal{X}$, $v_i \in \mathcal{Y}$, 而且 $\{a_i\}_{i=1}^{\infty}$ 和 $\{b_i\}_{i=1}^{\infty}$ 分别满足 $\sum_{i=1}^{\infty} |a_i| \leq \|f_1\| + \eta$ 和 $\sum_{i=1}^{\infty} |b_i| \leq \|g_1\| + \eta$. 由引理 2.1 知, 以 $1 - e^{-t}$ 的概率成立: 对任意 $i = 1, 2, \dots, n$, 存在 $X_{j_i} \in \{X_i\}_{i=1}^n$ 和 $Y_{k_i} \in \{Y_i\}_{i=1}^n$, 使得下式成立:

$$\max\{|u_i - X_{j_i}|, |v_i - Y_{k_i}|\} \leq c_{\alpha, \tau} \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}}.$$

选择 N 使它满足

$$\sum_{i=N+1}^{\infty} |a_i| \leq \frac{\eta}{2}, \quad \sum_{i=N+1}^{\infty} |b_i| \leq \frac{\eta}{2},$$

定义

$$f = \sum_{i=1}^N a_i k_{\mathcal{X}}(X_{j_i}, \cdot), \quad g = \sum_{i=1}^N b_i k_{\mathcal{Y}}(Y_{k_i}, \cdot),$$

则由引理 2.1 可得

$$\begin{aligned} \|f - f_1\| &= \left\| \sum_{i=1}^N a_i k_{\mathcal{X}}(X_{j_i}, \cdot) - \sum_{i=1}^N a_i k_{\mathcal{X}}(u_i, \cdot) - \sum_{i=N+1}^{\infty} a_i k_{\mathcal{X}}(u_i, \cdot) \right\| \\ &\leq C(\|f_1\| + \eta) \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} + \eta, \end{aligned} \quad (2.18)$$

其中 $C = c_1 c_{\alpha, \tau}^{\gamma}$. 同理可得

$$\|g - g_1\| \leq C(\|g_1\| + \eta) \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} + \eta.$$

由 $\|f\|_{\infty} \leq \kappa \|f\|$ 可得

$$\|f - f_1\|_{\infty} \leq C\kappa(\|f_1\| + \eta) \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} + \kappa\eta,$$

则

$$|\text{Var}[f(X)] - 1| = |\text{Var}[f(X)] - \text{Var}[f_1(X)]| \leq 4\kappa^2(3\|f_1\| + \eta) \left(C(\|f_1\| + \eta) \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} + \eta \right). \quad (2.19)$$

由文献 [7, 引理 2] 可知, 对于任意的 $t < n$, 至少以 $1 - 18e^{-t}$ 的概率下式成立:

$$|\text{Var}[f(X)] - \widehat{\text{Var}}[f(X)]| \leq 32\kappa \|f\|_{\mathcal{H}_{\mathcal{X}}}^2 \sqrt{\frac{t}{n}}, \quad f \in \mathcal{H}_{\mathcal{X}}. \quad (2.20)$$

结合上述两个不等式可知,

$$|\widehat{\text{Var}}[f(X)] - 1| \leq 4\kappa^2(3\|f_1\| + \eta) \left(C(\|f_1\| + \eta) \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} + \eta \right) + 32\kappa \|f\|_{\mathcal{H}_{\mathcal{X}}}^2 \sqrt{\frac{t}{n}}. \quad (2.21)$$

定义

$$a = \sqrt{\widehat{\text{Var}}[f(X)]}.$$

当 η 充分小且 n 充分大时, 有 $a \geq \frac{1}{2}$, 则

$$\begin{aligned} \left| \frac{1}{a} - 1 \right| &\leq |2(1 - a^2)| \\ &\leq 2 \left(4\kappa^2(3\|f_1\| + \eta) \left(C(\|f_1\| + \eta) \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} + \eta \right) + 32\kappa \|f\|_{\mathcal{H}_{\mathcal{X}}}^2 \sqrt{\frac{t}{n}} \right) \\ &\leq 2 \left(4\kappa^2(3\|f_1\| + \eta) \left(C(\|f_1\| + \eta) \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} + \eta \right) + 32\kappa^2 \|f\|^2 \sqrt{\frac{t}{n}} \right). \end{aligned} \quad (2.22)$$

令 $\tilde{f}(X) = \frac{f(X)}{a}$, 于是,

$$\|\tilde{f} - f\| = \|f\| \left| \frac{1}{a} - 1 \right| \leq 2\|f_1\| \left| \frac{1}{a} - 1 \right|. \quad (2.23)$$

由 $f_{1,\mathbf{z}}$ 和 $g_{1,\mathbf{z}}$ 的定义可知,

$$\mathcal{E}_n(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) + \lambda(\Omega_{\mathbf{x}}(f_{1,\mathbf{z}}) + \Omega_{\mathbf{y}}(g_{1,\mathbf{z}}))$$

$$\begin{aligned}
&\leq \mathcal{E}_n(\tilde{f}, g) + \lambda(\Omega_{\mathbf{x}}(\tilde{f}) + \Omega_{\mathbf{y}}(g)) \\
&\leq \mathcal{E}_n(\tilde{f}, g) + \lambda\left(\frac{\sum_{i=1}^N |a_i|}{a} + \sum_{i=1}^N |b_i|\right) \\
&\leq \mathcal{E}_n(\tilde{f}, g) + \lambda\left(\frac{\|f_1\|}{a} + \|g_1\| + \eta + \frac{\eta}{a}\right). \quad (2.24)
\end{aligned}$$

再由经验方差定义和 (2.23) 可知,

$$\begin{aligned}
\mathcal{E}_n(\tilde{f}, g) &\leq \mathcal{E}_n(f, g) + 4\|f\|(3\|f\| + 2\|g\|)\left|\frac{1}{a} - 1\right| \\
&\leq \mathcal{E}_n(f, g) + 16\|f_1\|(3\|f_1\| + 2\|g_1\|)\left|\frac{1}{a} - 1\right|. \quad (2.25)
\end{aligned}$$

我们又注意到,

$$\begin{aligned}
\mathcal{E}_n(f, g) &\leq \mathcal{E}_n(f_1, g_1) + 8\kappa^2(\|f_1\| + \|g_1\| + \eta)(\|f - f_1\| + \|g - g_1\|) \\
&\leq \mathcal{E}_n(f_1, g_1) + 8\kappa^2(\|f_1\| + \|g_1\| + \eta)\left(C\left(\frac{\log n + t}{n}\right)^{\frac{\gamma}{\tau}}(\|f_1\| + \|g_1\| + 2\eta) + 2\eta\right). \quad (2.26)
\end{aligned}$$

至少以 $1 - e^{-t}$ 的概率下式成立:

$$\begin{aligned}
P(\mathbf{z}) &\leq \mathcal{E}_n(\tilde{f}, g) + \lambda\left(\frac{\|f_1\|}{a} + \|g_1\| + \eta + \frac{\eta}{a}\right) - \mathcal{E}_n(f_1, g_1) - \lambda(\|f_1\| + \|g_1\|) \\
&\leq 8\kappa^2(\|f_1\| + \|g_1\| + \eta)\left(C\left(\frac{\log n + t}{n}\right)^{\frac{\gamma}{\tau}}(\|f_1\| + \|g_1\| + 2\eta) + 2\eta\right) \\
&\quad + 16\|f_1\|(3\|f_1\| + 2\|g_1\|)\left|\frac{1}{a} - 1\right| + \lambda\left|\frac{\|f_1\|}{a} - \|f_1\| + \eta + \frac{\eta}{a}\right|. \quad (2.27)
\end{aligned}$$

令 $\eta \rightarrow 0$, 命题得证. $\square$

给定 $R > 0$, 定义 $\mathcal{B}_R = \{(f, g) \in C_{k_x} \times C_{k_y} \mid \|f\| + \|g\| \leq R\}$. 记 $\mathbf{Z} = \mathbf{Z}_1 \cap \mathbf{Z}_2$, 其中 $\mathbf{Z}_1$ 在命题 2.1 前给出, $\mathbf{Z}_2$ 在引理 2.1 后给出, 显然, $P_{X,Y}^n(\mathbf{Z}) \geq 1 - 19e^{-t}$. 根据 (2.2) 的分解, 以及命题 2.1 和 2.2 得到如下推论.

推论 2.1 在假设 2.1 和 2.2 的条件下, 假设 R 是满足 $(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) \in \mathcal{B}_R$ ($\forall \mathbf{z} \in \mathbf{Z}$) 的常数. 对任意 $1 < t \leq n$, $\lambda \leq 1$, 有下式成立:

$$\mathcal{E}(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) - \mathcal{E}(f_1, g_1) + \lambda(\Omega_{\mathbf{x}}(f_{1,\mathbf{z}}) + \Omega_{\mathbf{y}}(g_{1,\mathbf{z}})) \leq C\left(R^2\sqrt{\frac{t}{n}} + \sqrt{\frac{t}{n}} + \lambda + \left(\frac{\log n + t}{n}\right)^{\frac{\gamma}{\tau}}\right). \quad (2.28)$$

由推论 2.1, 再结合 (2.24) 和 (2.26), 可知存在 C_3 使得下式成立:

$$\Omega_{\mathbf{x}}(f_{1,\mathbf{z}}) + \Omega_{\mathbf{y}}(g_{1,\mathbf{z}}) \leq \frac{C_3}{\lambda}, \quad \forall \mathbf{z} \in \mathbf{Z}. \quad (2.29)$$

因此, $(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) \in \mathcal{B}_{C_2\lambda^{-1}}$, 将 $R = \frac{C_3}{\lambda}$ 代入推论 2.1, 不难发现 $\lambda = n^{-1/6}$ 时误差最小. 于是得到过误差的一个渐近估计.

定理 2.1 在假设 2.1 和 2.2 的条件下, 存在常数 C , 使得下式成立:

$$\mathcal{E}(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) - \mathcal{E}(f_1, g_1) + \lambda(\Omega_{\mathbf{x}}(f_{1,\mathbf{z}}) + \Omega_{\mathbf{y}}(g_{1,\mathbf{z}})) \leq C\left(n^{-1/6}\sqrt{t} + \left(\frac{\log n + t}{n}\right)^{\frac{\gamma}{\tau}}\right), \quad \forall \mathbf{z} \in \mathbf{Z}. \quad (2.30)$$

3 学习率

本节利用过误差给出学习率, 我们需要引用文献 [7] 的结果. 由 $f_1 \in C_{k_x}$ 、 $g_1 \in C_{k_y}$ 和 (2.7) 可知, $f_1 \in \mathcal{H}_X$, $g_1 \in \mathcal{H}_Y$, 于是下述命题成立.

命题 3.1 ^[7] 存在常数 C , 对任意满足 $\text{Var}[f(X)] = 1$ 的 $f \in \mathcal{H}_X$, 以及任意 $g \in \mathcal{H}_Y$, 下式成立:

$$\text{Var}[f(X) - \sigma f_1(X)] + \text{Var}[g(Y) - \sigma g_1(Y)] \leq C(\mathcal{E}(f, g) - \mathcal{E}(f_1, g_1)), \quad (3.1)$$

其中 $\sigma = \text{sign}(\text{Cov}[f(X), f_1(X)])$.

由上述命题可知, 为了得到比 $O_p(n^{-1/6})$ 更好的学习率, 我们需要改进定理 2.1 给出的过误差估计. 为此需要迭代技巧, 下面给出迭代公式.

命题 3.2 ^[7] 假设 $t \leq n$ 且 $64R^2\sqrt{t/n} \leq 1$. 设 $(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) \in \mathcal{B}_R$, $\forall \mathbf{z} \in \mathbf{Z}$, 则 $\forall \mathbf{z} \in \mathbf{Z}$, $(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) \in \mathcal{B}_{r(R)}$ 成立, 其中 $r(R)$ 由下面等式给出:

$$r(R) = C\lambda^{-1} \left(R^2 \sqrt{\frac{t}{n}} \left(1 + \left(\frac{t}{n} \right)^{1/4} \right) + \sqrt{\frac{t}{n}} + \lambda + \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} \right), \quad (3.2)$$

C 是常数.

令 $R^{(0)} := \frac{C_3}{\lambda}$, 并定义 $R^{(j)} = r(R^{(j-1)})$, $j \in \mathbb{N}$. 由 $(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) \in \mathcal{B}_{R^{(0)}}$, 反复利用命题 3.2 可得 $(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) \in \mathcal{B}_{R^{(j)}}$, $\forall j \geq 1$, $\mathbf{z} \in \mathbf{Z}$.

为了避免叙述累赘, 我们只介绍迭代一次所得到的改进收敛阶.

定理 3.1 假设 $\frac{2\gamma}{\tau} > \frac{1}{10}$, 且假设 2.1 和 2.2 成立, 那么存在常数 C_4 , 使得下式成立:

$$\begin{aligned} & \mathcal{E}(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) - \mathcal{E}(f_1, g_1) + \lambda(\Omega_{\mathbf{x}}(f_{1,\mathbf{z}}) + \Omega_{\mathbf{y}}(g_{1,\mathbf{z}})) \\ & \leq C_4 \left(n^{-1/5} t^{3/2} + \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} \right), \quad \forall \mathbf{z} \in \mathbf{Z}. \end{aligned} \quad (3.3)$$

证明 取 $\lambda = n^{-1/5}$, 根据命题 3.2, 存在常数 C , 使得 $R^{(1)} \leq C(\sqrt{tn^{1/10}} + n^{1/5}(\frac{\log n + t}{n})^{\frac{\gamma}{\tau}})$. 将 $R = R^{(1)}$ 代入推论 2.1, 并利用 $(\sqrt{tn^{1/10}} + n^{1/5}(\frac{\log n + t}{n})^{\frac{\gamma}{\tau}})^2 \leq 2(tn^{1/5} + n^{2/5}(\frac{\log n + t}{n})^{\frac{2\gamma}{\tau}})$, 可得

$$\begin{aligned} & \mathcal{E}(f_{1,\mathbf{z}}, g_{1,\mathbf{z}}) - \mathcal{E}(f_1, g_1) + \lambda(\Omega_{\mathbf{x}}(f_{1,\mathbf{z}}) + \Omega_{\mathbf{y}}(g_{1,\mathbf{z}})) \\ & \leq 2C \left(\left(tn^{1/5} + n^{2/5} \left(\frac{\log n + t}{n} \right)^{\frac{2\gamma}{\tau}} \right) \sqrt{\frac{t}{n}} + \sqrt{\frac{t}{n}} + n^{-1/5} + \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} \right) \\ & \leq C' \left(n^{-1/5} t^{3/2} + \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} \right), \quad \forall \mathbf{z} \in \mathbf{Z}, \end{aligned} \quad (3.4)$$

其中 C' 为常数. 定理成立, 只要注意到

$$\max \left\{ t^{1/2} n^{-1/10} \left(\frac{\log n + t}{n} \right)^{\frac{2\gamma}{\tau}}, t^{3/2} n^{-3/10}, \sqrt{\frac{t}{n}}, n^{-1/5} \right\} = O(n^{-1/5} t^{3/2}). \quad (3.5)$$

证毕. □

结合定理 3.1 和命题 3.1, 可以证明下述学习率.

定理 3.2 在定理 3.1 的条件下, 存在常数 C , $\forall \mathbf{z} \in \mathbf{Z}$, 使得下式成立:

$$\text{Var}[f_{1,\mathbf{z}}(X) - \sigma f_1(X)] + \text{Var}[g_{1,\mathbf{z}}(Y) - \sigma g_1(Y)] \leq C \left(n^{-1/5} t^{3/2} + \left(\frac{\log n + t}{n} \right)^{\frac{\gamma}{\tau}} \right). \quad (3.6)$$

证明 对任意满足 $\text{Var}[f(X)] \neq 0$ 的 $f \in \mathcal{H}_{\mathcal{X}}$, 定义 $b_f = (\text{Var}[f(X)])^{-1/2}$. 根据文献 [7] 的方法可以证明 $|b_{f_{1,z}} - 1| \leq 64\kappa(R^{(1)})^2 \sqrt{t/n}$, $\sqrt{\text{Var}[f_{1,z}(X)]} \leq 1 + (R^{(1)})^2 \sqrt{t/n}$. 由

$$\begin{aligned} & |\text{Var}[f_{1,z}(X) - \sigma f_1(X)] - \text{Var}[b_{f_{1,z}} f_{1,z}(X) - \sigma f_1(X)]| \\ & \leq |b_{f_{1,z}} - 1| \sqrt{\text{Var}[f_{1,z}(X)]} \sqrt{\text{Var}[(b_{f_{1,z}} + 1)f_{1,z}(X) - 2\sigma f_1(X)]} \end{aligned} \quad (3.7)$$

得到常数 C , 使得下述不等式成立:

$$|\text{Var}[f_{1,z}(X) - \sigma f_1(X)] - \text{Var}[b_{f_{1,z}} f_{1,z}(X) - \sigma f_1(X)]| \leq 64C(R^{(1)})^2 \sqrt{\frac{t}{n}}. \quad (3.8)$$

另一方面, 利用文献 [7, (A.4)] 可得

$$|\mathcal{E}(b_{f_{1,z}} f_{1,z}, g_1) - \mathcal{E}(f_{1,z}, g_{1,z})| \leq C(R^{(1)})^2 \sqrt{\frac{t}{n}}. \quad (3.9)$$

再结合 $\text{Var}[b_{f_{1,z}} f_{1,z}] = 1$ 和定理 3.1 可知下述不等式成立:

$$\begin{aligned} & \text{Var}[b_{f_{1,z}} f_{1,z}(X) - \sigma f_1(X)] + \text{Var}[g_{1,z} - \sigma g_1(Y)] \\ & \leq C(\mathcal{E}(b_{f_{1,z}} f_{1,z}, g_1) - \mathcal{E}(f_{1,z}, g_{1,z})), \quad \forall \mathbf{z} \in \mathbf{Z}. \end{aligned} \quad (3.10)$$

定理由 (3.8) 和 (3.10) 得证. $\square$

在 $\frac{\gamma}{\tau} > \frac{1}{5}$ 的条件下, 由定理 3.2 给出的学习率是 $O_p(n^{-1/5})$, 而定理 2.1 只能给出 $O_p(n^{-1/6})$ 的学习率.

利用迭代后的 $R^{(J)}$ ($J \geq 2$), 可以进一步改善学习率, 为此需要用到文献 [9] 中 $R^{(J)}$ 的如下估计:

$$R^{(J)} \leq \max\{(2a)^{2^J-1}(R^{(0)})^{2^J}, 2b, (2b)(4ab)^{2^{J-1}-1}\}. \quad (3.11)$$

具体推导过程参照定理 3.1 和 3.2 的证明.

参考文献

- 1 Li K C. Sliced inverse regression for dimension reduction. J Amer Statist Assoc, 1991, 86: 316–327
- 2 Duan N, Li K C. Slicing regression: A link-free regression method. Ann Statist, 1991, 19: 505–530
- 3 Chen C H, Li K C. Can SIR be as popular as multiple linear regression? Statist Sinica, 1998, 8: 289–316
- 4 Wu H M. Kernel sliced inverse regression with applications to classification. J Comput Graph Statist, 2008, 17: 590–610
- 5 Cucker F, Zhou D X. Learning Theory: An Approximation Theory Viewpoint. New York: Cambridge University Press, 2007
- 6 Wu Q, Liang F, Mukherjee S. Kernel sliced inverse regression: Regularization and consistency. Abstr Appl Anal, 2013, 2013: 540725
- 7 Chen H, Chen D R. Statistical performance of optimal scoring in reproducing kernel Hilbert spaces. J Statist Plann Inference, 2018, 194: 122–135
- 8 Cai J, Sun H W. Convergence rate of kernel canonical correlation analysis. Sci China Math, 2011, 54: 2161–2170
- 9 Shi L, Feng Y L, Zhou D X. Concentration estimates for learning with l^1 -regularizer and data dependent hypothesis spaces. Appl Comput Harmon Anal, 2011, 31: 286–302
- 10 Lykou A, Whittaker J. Sparse CCA using a Lasso with positivity constraints. Comput Statist Data Anal, 2010, 54: 3144–3157

- 11 Tong H Z, Chen D R, Yang F H. Support vector machines regression with l^1 -regularizer. J Approx Theory, 2012, 164: 1331–1344

Coefficient regularizer based simultaneous regression estimation

Heng Chen, Wei Huang & Dirong Chen

Abstract It is very important to extract the correlation between two random vectors X and Y . The kernel method is used to extract nonlinear correlations. In this paper, the maximum correlation is obtained by minimizing variance $\text{Var}[f(X) - g(Y)]$, which is called simultaneous regression, where $f(X)$ and $g(Y)$ are respectively functions in two different reproducing kernel Hilbert spaces. We use regularized empirical variances minimization to construct estimates. In order to obtain the estimated function with sparsity, the l_1 -norm of coefficient is used as the penalty term. We establish a learning rate under some mild conditions. Simultaneous regression is associated with canonical correlation analysis and slice inverse regression, etc.

Keywords correlation, coefficient regularized, simultaneous regression, learning rate

MSC(2010) 62G08, 62J02

doi: 10.1360/N012018-00200