2 相关变量线性回归模型

2.1 模型建立

在这里我们先建立模型, 并通过极大似然估计得到目标函数, 关于模型的优化和求解我们将在下一章进行详细讨论.

我们有 $m \times p$ 维的设计矩阵 X_1 和 $n \times p$ 维的设计矩阵 X_2, 同时将 X_1, X_2 拼在一起得到的大设计矩阵记为 X. 已知的 m 维的因变量 y_{11} 和 n 维的因变量 y_{22}, 需要估计的部分分别为 $y_{21} \in R^{n\times 1}$, $y_{12} \in R^{m\times 1}$.

将 $(y_{11}, y_{21})^T \hat{=} \tilde{Y_1}$ 看作是随机变量 Y_1 的 $(m+n)$ 维样本，将 $(y_{12}, y_{22})^T \hat{=} \tilde{Y_2}$ 看作是随机变量 Y_2 的 $(m+n)$ 维样本.

我们建立如下线性回归模型：

$$\begin{aligned} \tilde{Y_1} &= X\beta_1 + \epsilon_1. \\ \tilde{Y_2} &= X\beta_2 + \epsilon_2. \end{aligned} \tag{2.1}$$

其中 $\epsilon_1 \sim N(0, \sigma_1^2), \epsilon_2 \sim N(0, \sigma_2^2)$

即

$$\begin{aligned} y_{11} &= X_1\beta_1 + \epsilon_{11}. \qquad y_{12} = X_1\beta_2 + \epsilon_{12}. \\ y_{21} &= X_2\beta_1 + \epsilon_{21}. \qquad y_{22} = X_2\beta_2 + \epsilon_{22}. \end{aligned} \tag{2.2}$$

其中 $\epsilon_{11}, \epsilon_{21} \sim N(0, \sigma_1^2), \epsilon_{12}, \epsilon_{22} \sim N(0, \sigma_2^2)$

我们从而得到 $\hat{y_{12}} = X_1\beta_2, \hat{y_{21}} = X_2\beta_1$, 我们不妨假设 Y_1, Y_2 的期望是 0,

$$E\{Y_k\} = \sum_{i=1}^{p} E\{x_i\}\beta + E\{\epsilon\}, k = 1, 2$$

其中 x_i 是每个属性变量的随机变量, 且随机方差的期望是 0, 所以实际操作时只需要将 X 的每一列,y_{11} 和 y_{22} 进行中心化即可, 从而得到随机变量 Y_1, Y_2 方差及协方

差的估计:

$$\hat{D}(Y_1) = \hat{D}(\tilde{Y_1}) = (y_{11}^T y_{11} + \beta_1^T X_2^T X_2 \beta_1)/(m+n) \hat{=} \sigma_{11} \tag{2.3}$$

$$\hat{D}(Y_2) = \hat{D}(\tilde{Y_2}) = (y_{22}^T y_{22} + \beta_2^T X_1^T X_1 \beta_2)/(m+n) \hat{=} \sigma_{22} \tag{2.4}$$

$$\hat{cov}(Y_1, Y_2) = \hat{cov}(\tilde{Y_1}, \tilde{Y_2}) = (y_{11}^T X_1 \beta_2 + y_{22}^T X_2 \beta_1)/(m+n) \hat{=} \sigma_{12} \tag{2.5}$$

我们取 Σ 为 $\tilde{Y_1}, \tilde{Y_2}$ 理论上的协方差矩阵, 有

$$\hat{\Sigma} = \begin{bmatrix} \sigma_{11} I_{m+n} & \sigma_{12} I_{m+n} \\ \sigma_{12} I_{m+n} & \sigma_{22} I_{m+n} \end{bmatrix}$$

其中 I_{m+n} 为 $m+n$ 阶标准矩阵.

于是我们得到 $\tilde{Y_1}, \tilde{Y_2}$ 的联合对数似然函数

$$logL(\tilde{Y_1}, \tilde{Y_2}) = -\frac{1}{2} \begin{pmatrix} \tilde{Y_1} - X\beta_1 \\ \tilde{Y_2} - X\beta_2 \end{pmatrix}^T \hat{\Sigma}^{-1} \begin{pmatrix} \tilde{Y_1} - X\beta_1 \\ \tilde{Y_2} - X\beta_2 \end{pmatrix} - \frac{1}{2} log|\hat{\Sigma}| + const$$

取 $t = \sigma_{11}\sigma_{22} - \sigma_{12}^2$, 我们将 $\hat{\Sigma}^{-1} = \frac{1}{t} \begin{bmatrix} \sigma_{22} I_{m+n} & -\sigma_{12} I_{m+n} \\ -\sigma_{12} I_{m+n} & \sigma_{11} I_{m+n} \end{bmatrix}$,$|\hat{\Sigma}| = t^{m+n}$ 代入, 舍去末尾的常数项，最大化联合对数似然函数函数，化简得到目标优化函数:

$min\ f(\beta_1, \beta_2)$

$= \frac{1}{t}\{\sigma_{22}(y_{11} - X_1\beta_1)^T(y_{11} - X_1\beta_1 + \sigma_{11}(y_{22} - X_2\beta_2)^T(y_{22} - X_2\beta_2)\} + (m+n)logt$

其中

同时我们又已知了 Y_1, Y_2 的高度相关性, 故我们还需要添加一个约束条件, 在 $\tilde{Y_1}, \tilde{Y_2}$ 标准化的前提下:

$$||\tilde{Y_1} - \tilde{Y_2}||_2 \leq \epsilon$$

2.2 对回归参数的 LASSO 约束

首先, 我们给出有界约束二次规划问题的一种标准形式：

$$\begin{cases} \min\limits_{x} \frac{1}{2}||y - Ax||_2^2 \\ s.t. \ \ ||x||_1 \leq \lambda \end{cases} \tag{2.6}$$

随后通过拉格朗日乘子法将其转变为无约束凸优化问题:

$$\min_{x} \frac{1}{2}||y - Ax||_2^2 + \lambda||x||_1$$

此处的理解是: 把这个问题想象成为一个处理设计矩阵 A 条件数过大, 或者奇异的情形. 通过添加一个 ℓ_1 正则项,x 的分量中比较小的部分会收缩到零.

关于分量较小部分收缩到到零的解释, 实际上, 在设计矩阵 A 有性质 $A^TA = I$ 时我们可以给出其显式解:

$$x_j = sgn(a_i^T x)(|a_i^T x| - \lambda)_+ \tag{2.7}$$

其中 A=$(a_1, a_2, \ldots, a_n)$, 每个 a_i 是一个列向量, 且 $(\ )_+$ 代表取正部. 每个分量上形式如图:

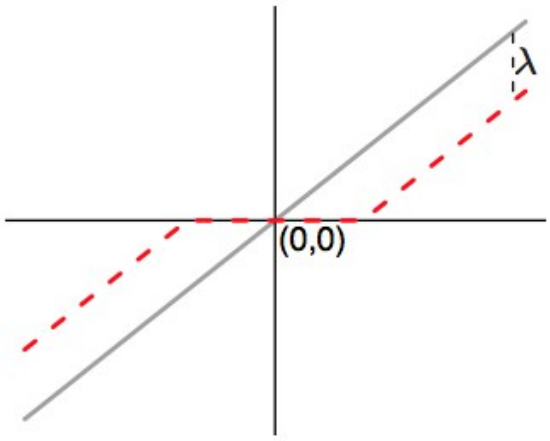

图 2.1　soft-thresholding 算子

当 λ 值较小时, 易看出该分量取值 (红色部分) 收缩到 0. 由于当使用 l_1 约束的时

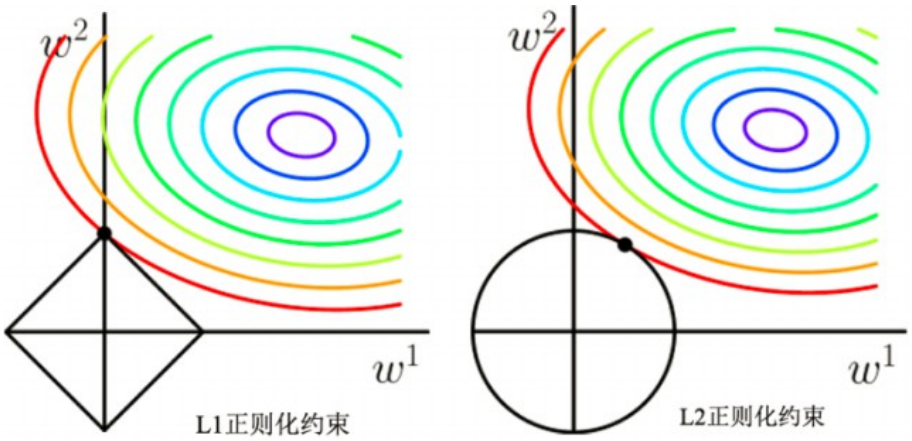

图 2.2　直观看 Lasso 回归结果的稀疏性

候, 得到的约束是一个多边形区域, 求得解往往都是落在顶点上, 故此时得到的解会在一些分量上取 0 值, 这使得 LASSO 问题的解具有了稀疏性.

从 Bayes 的角度看, 我们可以直接取 β_1, β_2 的先验分布为 Laplace 分布, 即

$$f_1(\beta_1) = (\frac{1}{2}\lambda_1)^p \exp\{-\lambda_1 \|\beta_1\|_{l_1}\}$$

$$f_2(\beta_2) = (\frac{1}{2}\lambda_2)^p \exp\{-\lambda_2 \|\beta_2\|_{l_1}\}$$

取 $\lambda_1 = \lambda_2 = (2\sigma^2)^{-1}\lambda$, 舍去常数项得到我们的目标函数:

$$\begin{cases} \min f(\beta_1, \beta_2) + \lambda_1 ||\beta_1||_1 + \lambda_2 ||\beta_2||_1 \\ s.t.\ \ ||\tilde{Y_1} - \tilde{Y_2}||_2 \leq \epsilon \end{cases} \tag{2.8}$$

3 对于约束条件的转化

可以看到目前的约束条件

$$||\tilde{Y_1} - \tilde{Y_2}||_2 \leq \epsilon$$

相当复杂, 其中 $\tilde{Y_1}, \tilde{Y_2}$ 中不仅包含已有的数据, 同时还包含了待优化的参数 β_1, β_2. 但是直观上看, 当两组标准化之后的数据很接近时, 它们的回归系数应该理所当然的很接近

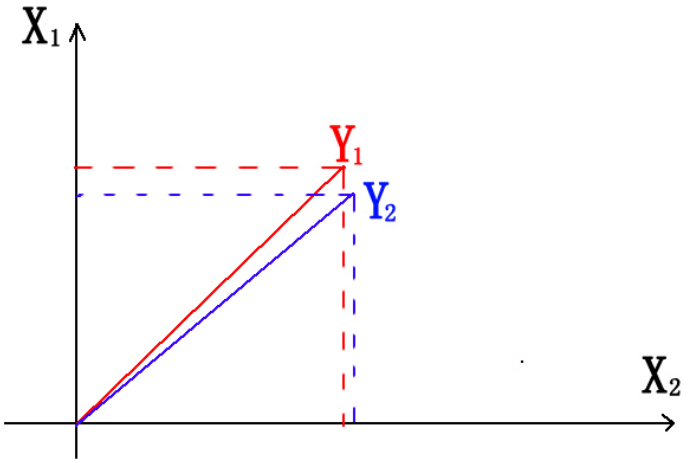

图 3.1 标准化之后向量相近

从而我们通过下面的一系列的证明来简化约束条件, 这里只做 $m+n \geq p$ 时的证明.

引理 3.0.1 若 A 为对称矩阵, 则有 $||A||_2 = \rho(A)$, 其中 $\rho(A)$ 为 A 的谱范数.

证明

$$||A||_2 = \sqrt{\lambda_{max}(A^TA)} = \sqrt{\lambda_{max}(A^2)}$$

若 λ 是 A 的一个特征根, 则 λ^2 必是 A^2 的特征根.

又对称矩阵的特征根 λ 为实数, 故

$$\lambda_{max}(A^2) = \max_{1\le i\le n}(\lambda_i(A))^2 = (\max_{1\le i\le n}|\lambda_i(A)|)^2$$

从而

$$||A||_2 = \sqrt{\lambda_{max}(A^2)} = \max_{1\le i\le n}|\lambda_i(A)| = \rho(A) \qquad \square$$

引理 3.0.2 $A \in R^{m\times n}$, $m > n$ 且 A 的行秩大于 n, 对于任意 $\epsilon > 0$, 存在 $\delta > 0$, 对于任意 $||Ax||_2 \le \epsilon$, 使得 $||x||_2 \le \delta$.

证明 我们可以从 A 中抽取线性无关的 n 行组成可逆的 $n \times n$ 的方阵 $\bar{A}$, 从而

$$x^T\bar{A}^T\bar{A}x = ||\bar{A}x||_2 \le ||Ax||_2 \le \epsilon$$

由于 $\bar{A}$ 可逆, 从而 $\bar{A}^T\bar{A}$ 为正定矩阵, 故对任意 $x > 0$,$x^T\bar{A}^T\bar{A}x > 0$, 当 $\epsilon \to 0$, 则 $x \to 0$, 从而存在 $\delta > 0$

$$||x||_2 \le \delta \qquad \square$$

定理 3.0.1 在 §2.2 的模型 (2.1) 中, 设 β_1, β_2 的估计值分别为 b_1, b_2, 当 $m+n \ge p$ 时, 对于任意 $\epsilon > 0$, 存在 $\delta > 0$, 对于任意 $||\tilde{Y_1} - \tilde{Y_2}||_2 \le \epsilon$, $||b_1 - b_2||_2 \le \delta$.

证明 由于 $m + n \ge p$, 从而 X^TX 可逆, 我们记 $H = X(X^TX)^{-1}X^T$, 由于矩阵二范数为诱导范数, 即

$$||H||_2 = \max_{||x||_2\neq 0}\frac{||Hx||_2}{||x||_2} \tag{3.1}$$

从而有

$$\frac{||H(\tilde{Y_1} - \tilde{Y_2})||_2}{||\tilde{Y_1} - \tilde{Y_2}||_2} \le ||H||_2$$

$$||X(b_1 - b_2)||_2 = ||H(\tilde{Y_1} - \tilde{Y_2})||_2 \le ||H||_2||\tilde{Y_1} - \tilde{Y_2}||_2$$

H 的特征值不全为 0, 又 H 为投影矩阵, 其特征值为 0 或 1, 有从而

$$||H||_2 = \sqrt{\rho(H^T H)} = \sqrt{\rho(H)} = 1$$

故

$$||X(b_1 - b_2)||_2 \leq ||\tilde{Y_1} - \tilde{Y_2}|| \leq \epsilon$$

由引理 3.0.2, 所以

$$\exists \delta > 0, \ \ ||b_1 - b_2||_2 \leq \delta \qquad \square$$

于是原本的约束条件

$$||\tilde{Y_1} - \tilde{Y_2}||_2 \leq \epsilon$$

就可以转化成

$$||b_1 - b_2|| \leq \delta$$

故最终我们寻求优化的目标函数为

$$\begin{cases} \min f(\beta_1, \beta_2) + \lambda_1 ||\beta_1||_1 + \lambda_2 ||\beta_2||_1 \\ s.t. \ \ ||\beta_1 - \beta_2||_2 \leq \epsilon \end{cases} \tag{3.2}$$

4 优化步骤的实现

由于优化的目标函数中,t 作为 β_1, β_2 的二次函数存在于分母当中, 这个我们直接进行优化求解带了很大的困难. 于是开始尝试将 t 和 β_1, β_2 进行迭代下降的进行优化.

4.1 数据的预处理

将 X, y_{11}, y_{22} 进行中心化处理, 再将 y_{11}, y_{22} 进行标准化处理, 这时 Y_1, Y_2 都看做均值为 0 前方差为 1, 给出待估回归系数 β_1, β_2 的初始值 $\beta_1^{(0)}, \beta_2^{(0)}$

4.2 计算 $\sigma_{11}^{(k)}, \sigma_{22}^{(k)}$

根据 (2.3)(2.4)(2.5), 我们计算 $\sigma_{11}^{(k)}, \sigma_{22}^{(k)}$

$$\sigma_{11}^{(k)} = (y_{11}^T y_{11} + \beta_1^{(k-1)T} X_2^T X_2 \beta_1^{(k-1)})/(m+n) \tag{4.1}$$

$$\sigma_{22}^{(k)} = (y_{22}^T y_{22} + \beta_2^{(k-1)T} X_1^T X_1 \beta_2^{(k-1)})/(m+n) \tag{4.2}$$

$$\sigma_{12}^{(k)} = (y_{11}^T X_1 \beta_2^{(k-1)} + y_{22}^T X_2 \beta_1^{(k-1)})/(m+n) \tag{4.3}$$

4.3 对 $t^{(k)}$ 的优化

对 $t^{(k)}$ 的优化可以归结为以下对 t 的优化问题

$$\begin{cases} \min \quad f(t|\beta_1^{(k-1)}, \beta_2^{(k-1)}) = \\ \qquad \frac{1}{t}\{\sigma_{22}^{(k)}||y_{11} - X_1\beta_1^{(k-1)}||_2^2 + \sigma_{11}^{(k)}||y_{22} - X_2\beta_2^{(k-1)}||_2^2\} + (m+n)logt \\ s.t. \ |t - \{\sigma_{11}^{(k)}\sigma_{22}^{(k)} - (\sigma_{12}^{(k)})^2\}| \le \epsilon \end{cases} \tag{4.4}$$

我们这里采用约束优化问题的对数障碍函数的内点法进行求解, 记 $t_0^{(k)} = \sigma_{11}^{(k)}\sigma_{22}^{(k)} - (\sigma_{12}^{(k)})^2$, 则其变为如下无约束优化问题

$$min\ f(t|\beta_1^{(k-1)}, \beta_2^{(k-1)}) - \lambda_1 \log(t - t_0 + \epsilon) - \lambda_2 \log(\epsilon - t + t_0) \tag{4.5}$$

我们可以看出来直接对 (4.4) 进行求导就可以求解.

4.4 对 $\beta_1^{(k)}, \beta_2^{(k)}$ 的优化

ADMM 算法 (S. Boyd, 2011) 是一种解决凸优化问题简单而有效的方法, 其可以看做是将对偶分解法和约束优化增广拉格朗日方法的优点综合在一起的一次尝试. 它主要解决如下形式的约束优化问题:

$$\begin{aligned} &minimize\ f(x) + g(z) \\ &subject\ to\ Ax + Bz = c \end{aligned} \tag{4.6}$$

其中 $x \in R^n,\ z \in R^m,\ A \in R^{p\times n},\ B \in R^{p\times m},\ c \in R^p$, 并且假设 f, g 都是凸函数. 写出其增广拉格朗日函数

$$L_\rho(x, z, y) = f(x) + g(z) + y^T(Ax + Bz - c) + (\rho/2)||Ax + Bz - c||_2^2$$

ADMM 算法的求解迭代形式为

$$x^{k+1} := \underset{x}{argmin}\, L_\rho(x, z^k, y^k) \tag{4.7}$$

$$z^{k+1} := \underset{z}{argmin}\, L_\rho(x^{k+1}, z, y^k) \tag{4.8}$$

$$y^{k+1} := y^k + \rho(Ax^{k+1} + Bz^{k+1} - c) \tag{4.9}$$

ADMM 可以看做是是简单的高斯-赛德尔迭代, 但是对 x 和 z 是分别优化而不是同时优化求解的. 同时由于 z 的优化求解是在 x 之后的, 所以 x 和 z 的眉部优化不是完全对称的.

ADMM 还可以写成一种微微不同但是更加简洁的形式. 我们定义残差 $r = Ax + Bz - c$, 有

$$y^T r + (\rho/2)||r||_2 = (\rho/2)||r + (1/\rho)y||_2^2 - (1/2\rho)||y||_2^2$$

$$= (\rho/2)||r+u||_2^2 - (\rho/2)||u||_2^2$$

其中 $u = (1/\rho)y$ 被称为尺度化的对偶向量. 使用尺度化的对偶向量, 我们用如下式子表达 ADMM:

$$x^{k+1} := \underset{x}{argmin}(f(x) + (\rho/2)||Ax + Bz^k - c + u^k||_2^2) \tag{4.10}$$

$$z^{k+1} := \underset{z}{argmin}(g(z) + (\rho/2)||Ax^{k+1} + Bz - c + u^k||_2^2) \tag{4.11}$$

$$u^{k+1} := u^k + Ax^{k+1} + Bz^{k+1} - c \tag{4.12}$$

我们定义第 k 次迭代的残差为 $r^k = Ax^k + bz^k - c$, 我们可以看到

$$u^k = u^0 + \sum_{j=1}^{k} r^j$$

当 $\epsilon \to 0$ 时,$\delta \to 0$, 根据 ADMM 算法, 我们取初始值为 $\beta_1^{(k)}, \beta_2^{(k)}$ 通过如下迭代求解 $\beta_1^{(k+1)}, \beta_2^{(k+1)}$

$$\beta_1 := \underset{\beta_1}{argmin}\{\frac{\sigma_{22}^{(k)}}{2t^{(k)}}||y_{11} - X_1\beta_1||_2^2 + \lambda_1||\beta_1||_1 + \rho||\beta_1 - \beta_2||_2^2 + \Lambda^T(\beta_1 - \beta_2)\} \tag{4.13}$$

$$\beta_2 := \underset{\beta_2}{argmin}\{\frac{\sigma_{11}^{(k)}}{2t^{(k)}}||y_{22} - X_2\beta_2||_2^2 + \lambda_1||\beta_2||_1 + \rho||\beta_1 - \beta_2||_2^2 + \Lambda^T(\beta_1 - \beta_2)\} \tag{4.14}$$

$$\Lambda := \Lambda + \rho(\beta_1 - \beta_2) \tag{4.15}$$

其中 (4.11)(4.12) 式为 l_1 约束问题, 我们可以通过很多相关方法进行求解，比如临近梯度算子 (proximal gradient descent, Parikh, 2014), 同轮法 (Lars, Efron) 等等, 当然也同样可以通过 ADMM 算法进行求解.

取定 ϵ_0, 当 $|f(x^{(k+1)}) - f(x^{(k)})| \leq \epsilon_0$ 时终止迭代.

4.5 关于优化方法收敛性的证明

在这里我们将目标函数的优化看作是不可微目标函数的 block coordinate method(Tseng, 2001), 我们将证明当目标函数满足一些微弱约束条件时的我们的优化方法收敛.

首先我们将我们的目标函数看作是向量空间 $(\beta_1;\beta_2;t)$ 上的目标函数. 从而我们优化问题中的向量可以根据 block coorfinate 的框架写成如下特殊的形式:

$$f(\beta_1,\beta_2,t)=f_0(\beta_1,\beta_2,t)+\sum_{\beta_k\in\beta_1,\beta_2}f_k(\beta_k)$$

在这个形式下, 我们可以断言 f 的不可微部分是可分的. 比如 f_k 只单独的依赖于 β_1 或 β_2. 我们记 $x=(\beta_1,\beta_2,t)=(\beta_1^1,\beta_2^1,\cdots,\beta_1^{(m+n)},\beta_2^{(m+n)},t)$, 其中 $\beta_k^{(n)}$ 表示 β_k 的第 n 个分量. 接下来将证明在 f 满足如下凸性和连续性条件的前提下通过 block coordinate descent method 每一步得到的 cluster point 都是 f 的不动点.

我们先给出如下定义:

- 对于任意 $h:R^m\to R$, 将 h 的有效域记为 $dom\ h$:

$$dom\ h=\{x\in R^m|h(x)\leq\infty\}$$

- 对于任意 $x\in dom\ h$ 和任意 $d\in R^m$, 将 h 在 x 方向的下微分记为:

$$h^{'}(x;d)=liminf_{\lambda\searrow 0}[h(x+\lambda d)-h(x)]/\lambda$$

- h 是拟凸的, 如果有:

$$h(x+\lambda d)\leq max\{h(x),h(x+d)\}$$

  对于任意的 x,d,$\lambda\in[0,1]$

- h 是半变量的, 如果 h 在 $dom\ h$ 上的任何线上都不是常值.
- h 在 x_0 处下半连续, 若 x_0 附近的函数值接近 $h(x_0)$ 或大于 $h(x_0)$

容易验证我们的目标函数有如下的性质:

- f_0 在 $dom\ f_0$ 上连续
- 对于任意的 $x\in\{\beta,t\},x\to f(\beta,t)$ 是拟凸的也是半变量的.
- f_0,f_1,f_1 在其有效域上下半连续.

后面根据参考文献 [3](Tseng,2001) 中的一系列证明, 最终得证我们的优化方法是收敛的.