分类号：O212　　单位代码：10422
密　级：　　学　　号：200720228

山东大学

博士学位论文

论文题目：高维线性模型和部分线性模型的相合统计推断

Consistent Statistical Inferences for High-dimensional Linear Models and Partially Linear Models

作者姓名：曾云辉
学院名称：数学学院
专业名称：概率论与数理统计
指导教师：林路 教授
合作导师：

2013 年 5 月 12 日

分类号：　　单位代码：10422

密　级：　　学　　号：200720228

山东大学

博士学位论文

论文题目：高维线性模型和部分线性模型的拟合统计推断

作者姓名　曾云辉

学院名称　数学学院

专业名称　概率论与数理统计

指导教师　林路　教授

合作导师

2013　年 5 月 12 日

原 创 性 声 明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律责任由本人承担。

论文作者签名：曾云辉　　日　期：2013.5.23

关于学位论文使用授权的声明

本人完全了解山东大学有关保留、使用学位论文的规定，同意学校保留或向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权山东大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

(保密论文在解密后应遵守此规定)

论文作者签名：曾云辉　导师签名：　日　期：2013.5.23

目 录

高维线性模型和部分线性模型的相合统计推断

曾云辉

（山东大学数学学院, 济南, 250100）

摘　　要

在模型建模过程的初始阶段, 大量的预测变量被引入从而形成全模型. 但在实际的应用过程中, 过多的预测变量和较大的模型将需要大量的计算, 而且会增加参数估计的方差(variance)和均方误差(MSE), 甚至导致算法不稳定或异常退出, 直接影响参数估计的结果和模型的预测效果. 为了简化模型和增强可预测性, 将一些不重要的预测变量从全模型中剔除, 从而形成简化模型(或限制模型,restricted model). 在一定的正则条件下,简化模型中参数的估计是相合的. 然而, 当采用变量选择的模型特别需要参数估计的相合性时, 稀疏性和oracle 性质(如SCAD估计)仅在逐点意义下成立, 这意味着它不具有一个好的全局性质. 而且, 对于简化模型的估计或模型选择后(post-model-selection)估计, 如果一些重要变量不幸被剔除, 其系数虽然近似为零但对响应变量仍然有影响, 那么简化模型(亦称子模型)是误定(misspecified) 的, 从而对于子模型中的参数采用常用方法得到的估计将收敛于伪参数而不是真实参数值. 即使选出的子模型只是局部误定,采用常用的参数估计方法得出的估计也仍然是不相合的.

另一方面, 在某些应用场景下, 可能只对其中的部分变量及其系数感兴趣, 因为这些变量是可以适当控制的. 反之, 可能对其余的变量并不感兴趣, 或者这些剩余变量本身不容易被精确控制, 或者剩余变量对响应变量的影响结构还不是很清楚. 在这种情况下, 如果采用全模型同样会面临上述的问题, 而且可能会出现模型误定的情形. 如果只将响应变量对这些感兴趣的变量进行回归分析, 显然模型是有偏的, 对应的参数估计也将是有偏的; 如果在回归结构中, 将感兴趣的变量作为参数项而将剩余变量作为非参项处理, 虽然这样得到的半参回归模型相对较为合理, 但是这时非参项将可能面临维数祸根问题(curse of dimensionality). 特别在剩余变量的维数很高时, 维数祸根问题将直接导致半参回归模型的估计结果和预测效果的损失. 由于人们认识的局限性, 在实际应用中更多地将会面临到这种半参情形.

以上表明, 在包含变量选择(variable selection)或者变量指定(variable specified) 的建模过程中相合估计和置信域构造仍然是一个困难. 本论文将主要研究这种情形下感兴趣参数的估计方法以及相应模型的预测. 对于线性回归模型

$$Y = \beta^T X + \gamma^T Z + \varepsilon,$$

假设仅对参数β及协变量X感兴趣, 本论文在第二章将提出一个半参方法以调整有偏子模型

$$Y = \beta^T X + \eta,$$

从而得到参数β的相合估计, 进而进行模型预测. 详细地, 通过寻找一个方向$\tau \in \mathscr{T}$, 构造出调整模型

$$Y = \beta^T X + g(\tau^T Z) + \xi,$$

其中$g(\tau^T Z) = E(Y - \beta^T X|\tau^T Z) = \gamma^T E(Z|\tau^T Z)$, 将有偏子模型调整为部分条件无偏, 即满足

$$E(\xi|\tau^T Z, X) = 0, \quad \forall \tau \in \mathscr{T}, \forall (X, Z) \in \mathscr{W} \subset (\mathscr{X}, \mathscr{Z}).$$

采用一维非参核方法对非参项$g(\tau^T Z)$进行估计如下

$$\hat{g}(\tau^T Z) = \frac{\sum_{j=1}^n (Y_j - \beta^T X_j) K\{\tau^T (Z_j - Z)/h\}}{\sum_{j=1}^n K\{\tau^T (Z_j - Z)/h\}},$$

其中$K(\cdot)$是核函数, h是依赖于n的窗宽. 代入调整模型后, 得到参数β的估计如下

$$\tilde{\beta}_A = (\frac{1}{n}\sum_{i=1}^n \tilde{X}_i \tilde{X}_i^T)^{-1} \frac{1}{n}\sum_{i=1}^n \tilde{X}_i \tilde{Y}_i,$$

其中, $\tilde{Y}_i = Y_i - \frac{\sum_{j=1}^n Y_j K\{\tau^T (Z_j - Z_i)/h\}}{\sum_{j=1}^n K\{\tau^T (Z_j - Z_i)/h\}}$, $\tilde{X}_i = X_i - \frac{\sum_{j=1}^n X_j K\{\tau^T (Z_j - Z_i)/h\}}{\sum_{j=1}^n K\{\tau^T (Z_j - Z_i)/h\}}$. 随后, 证明了由该调整模型导出的参数估计$\tilde{\beta}_A$在子集区域$\mathscr{W}$上是$\sqrt{n}$相合的. 进一步, 基于F检验的PT估计思想, 得到了还依赖于全模型的调整PT估计$\tilde{\beta}_{APT}$. 在第二章第三节中基于该调整模型, 采用经验似然方法构造了参数β的置信域估计. 采用第二章的方法, 通过一个一维的非参函数可以成功地避免维数祸根问题, 而且该推断的性质对非参函数中的方向τ不敏感, 所以新方法对变量选择是稳健的. 无论子模型的偏大小, 数值模拟结果表明在相合意义下新的参数估计和置信域优于现有方法.

对于线性回归模型, 第二章中提出的非参调整方法能够大幅度地减小子模型的估计偏差. 但是, 这个方法的理论结果仅在协变量的一个子集区域$\mathscr{W}$上成立. 为此, 在第三章中基于子模型$Y = \beta^T X + \eta$和感兴趣参数β, 我们构造了一个全局无偏的工作模型

$$Y = \beta^T X + g_1(\tilde{Z}^{(1)}) + \cdots + g_K(\tilde{Z}^{(K)}) + \zeta_K,$$

假设$E(Z) = 0$. 主要的思路是首先将协变量Z分解成相互独立的分量$\tilde{Z}^{(1)}, \cdots, \tilde{Z}^{(q)}$, 然后利用了协变量$X$和协变量$Z$的独立分量$\tilde{Z}^{(l)}, l = 1, \cdots, q$之间的相关性信息, 对于与协变量$X$相关的独立分量$\tilde{Z}^{(l)}$,向子模型中增加一个一维的非参调整项$g_l(\tilde{Z}^{(l)}) = E(Y - \beta^T X|\tilde{Z}^{(l)}) = \gamma^T E(Z|\tilde{Z}^{(l)})$, 从而通过多步的非参调整来减小子模型的偏差.

(1)当协变量Z为正态分布时, 将采用主成分回归(PCR)方法来构造全局无偏的调整模型, 此时$g_l(\tilde{Z}^{(l)}) = \alpha_l \tilde{Z}^{(l)}$, 调整模型实际上是一个线性模型

$$Y = \beta^T X + \alpha^T \tilde{\mathbf{Z}}_K + \zeta_K,$$

其中$\tilde{Z}^{(l)}$是Z的第l个主成分;
(2)否则, 将采用独立成分分析(ICA)方法, 其中$\tilde{Z}^{(l)}$是Z 的第l 个独立成分.
基于这个调整模型, 采用一维非参核方法对非参项$g_l(\tilde{Z}^{(l)})$ 进行估计如下

$$\hat{g}_l(\tilde{Z}^{(l)}) = \frac{\sum_{j=1}^n (Y_j - \beta^T X_j) K((\tilde{Z}_j^{(l)} - \tilde{Z}^{(l)})/h_l)}{\sum_{j=1}^n K((\tilde{Z}_j^{(l)} - \tilde{Z}^{(l)})/h_l)}, l = 1, \cdots, K,$$

其中$K(\cdot)$是核函数, h_l是依赖于n的窗宽. 代入调整模型后, 得到参数β的估计如下

$$\hat{\beta}_A = (\frac{1}{n}\sum_{i=1}^n \breve{X}_i \breve{X}_i^T)^{-1} \frac{1}{n}\sum_{i=1}^n \breve{X}_i \breve{Y}_i,$$

其中, $\breve{Y}_i = Y_i - \sum_{l=1}^K \frac{\sum_{j=1}^n Y_j K((\tilde{Z}_j^{(l)} - \tilde{Z}_i^{(l)})/h_l)}{\sum_{j=1}^n K((\tilde{Z}_j^{(l)} - \tilde{Z}_i^{(l)})/h_l)}$, $\breve{X}_i = X_i - \sum_{l=1}^K \frac{\sum_{j=1}^n X_j K((\tilde{Z}_j^{(l)} - \tilde{Z}_i^{(l)})/h_l)}{\sum_{j=1}^n K((\tilde{Z}_j^{(l)} - \tilde{Z}_i^{(l)})/h_l)}$. 证明了参数$\beta$的这个点估计$\hat{\beta}_A$在协变量$X$和$Z$的全空间上是相合的, 而且它是渐近正态分布的. 同时, 由于增加的非参调整项$g_l(\tilde{Z}^{(l)})$之间相互独立, 因此避免了一般可加模型采用后拟(backfitting)方法时所产生的大量计算, 而且当调整的项数K不大时新算法的计算误差将很小. 当协变量Z服从正态分布时, 可以对线性调整模型直接采用最小二乘得到相应结果.

当增加的非参调整项数K较大时, 第三章提出的算法将会产生较大的计算误差, 该方法将失去其优势. 因此在第四章, 更一般地对于稀疏部分线性模型

$$Y = \beta^T X + \gamma^T Z + f(U) + \varepsilon,$$

提出了一个两阶段的重建模和参数估计方法. 其中, 参数β是我们感兴趣的, 参数γ是稀疏的. 为简化起见, 本章中假设U是1-维变量, $E(Z) = 0$. 实际上, $f(\cdot)$ 可推广到U是多维变量时的可加结构. 详细地, 在第一阶段中利用协变量Z的独立分量$\tilde{Z}^{(j)}$与协变量(X, U) 之间的相关性, 按照第三章的方法将首先构造一个多步调整的全局无偏模型

$$Y = \beta^T X + \sum_{j=1}^{K_0} g_j(\tilde{Z}^{(j)}) + f(U) + \zeta_{K_0}.$$

在第二阶段, 利用Zhao和Xue(2009)提出的半参变量选择方法, 在稀疏性条件下进一步对上述调整模型进行简化. 具体地, 对每一个非参项$g_j(\tilde{Z}^{(j)})$和非参项$f(U)$, 采用非参正交级数方法展开进行近似

$$g_j(\tilde{Z}^{(j)}) \approx \theta_j^T \mathbf{\Psi}_j(\tilde{Z}^{(j)}), \quad f(U) \approx \nu^T \mathbf{\Psi}_0(U).$$

然后通过组SCAD(group SCAD)方法对参数β,参数θ_j和ν进行估计, 即

$$(\hat{\beta},\hat{\theta},\hat{\nu}) = \arg\min_{\beta\in R^p,\theta_j\in R^L,\nu\in R^L} F(\beta,\theta.\nu),$$

其中,

$$F(\beta,\theta,\nu) = \sum_{i=1}^{n}\{Y_i - \beta^T X_i - \theta^T \mathbf{\Psi_i} - \nu^T \mathbf{\Psi_{0i}}\}^2 + n\sum_{j=1}^{K_0} p_{\lambda_j}(\|\theta_j\|_2),$$

$p_\lambda(\cdot)$是SCAD惩罚函数, 定义如下

$$p'_\lambda(w) = \lambda\{I(w\le\lambda) + \frac{(a\lambda - w)_+}{(a-1)\lambda} I(w>\lambda)\},$$

满足$a>2$, $w>0$, $p_\lambda(0)=0$. 令$\hat{\mathcal{M}}_n = \{1\le j\le K_0 : \hat{\theta}_j \neq 0\}$, 记$K_n = |\hat{\mathcal{M}}_n|$. 为简化起见, 假设$\hat{\mathcal{M}}_n = \{1,2,\cdots,K_n\}$. 记$g_j(\tilde{Z}^{(j)}) = E(\gamma^T Z|\tilde{Z}^{(j)}), j=1,\cdots,K_n$, $\zeta_{K_n} = Y - \beta^T X - g_1(\tilde{Z}^{(1)}) - \cdots - g_{K_n}(\tilde{Z}^{(K_n)}) - f(U)$, 从而得到简化模型

$$Y = \beta^T X + \sum_{j=1}^{K_n} g_j(\tilde{Z}^{(j)}) + f(U) + \zeta_{K_n}.$$

经过两阶段重新建模, 最终得到的模型是全局条件无偏的而且是充分化简的. 在理论结果中, 证明了基于简化模型的参数估计$\hat{\beta}$及非参估计$\hat{g}_l$ 和$\hat{f}$ 的收敛速度, 并得到了估计$\hat{\beta}$的渐近正态性. 因为变量选择主要依赖于参数的稀疏性, 当直接对部分线性模型采用变量选择方法时, 一些系数非零但与X不相关的变量可能会被选入模型, 这样可能会影响参数β 估计的有效性和稳定性.

在本论文的第五章, 对于协变量和误差项均服从正态分布的高维线性模型$Y = \beta^T X + \gamma^T Z + \varepsilon$, 基于包含感兴趣参数$\beta$的有偏子模型$Y = \beta^T X + \eta$, 将Cho和Fryzlewicz (2012) 提出的倾斜变量(tilted variables)方法和Zhang和Zhang(2012)提出的松弛投影(relaxed projection) 方法有机结合从而进行重新建模. 如果$\gamma^T E(Z|X)\neq 0$, 则$E(\eta|X = x)$是一个非零函数. 所以, 首先对有偏子模型, 采用Cho和Fryzlewicz(2012)的方法, 将协变量Z中与X相关的分量(记为$Z_{\mathcal{C}_X}$)扩充到模型中, 得到一个调整模型

$$Y = \beta^T X + \sum_{k\in\mathcal{C}_X}\gamma_k Z^{(k)} + \zeta,$$

其中$\zeta = \varepsilon + \sum_{k\in\mathcal{J}\setminus\mathcal{C}_X}\gamma_k Z^{(k)}$, $\mathcal{J} = \{1,2,\cdots,q\}$. 然后, 计算样本矩阵$\mathbb{X}$对应的倾斜变量

$$\mathbf{U}_0 \equiv (\mathbf{I}_n - \Pi_{Z_X})\mathbb{X},$$

其中, Π_{Z_X}为到由$\mathbf{Z}_{\mathcal{C}_X}$生成的空间的投影.

(1)如果倾斜变量$\mathbf{U}_0$的长度不太小, 可直接基于倾斜变量和调整模型, 得到感兴趣参数β的估计

$$\hat{\beta}_T = (\mathbf{U}_0^T\mathbb{X})^{-1}\mathbf{U}_0^T\mathbb{Y},$$

并证明了在一定条件下该估计是相合的;
(2)如果存在长度很小的倾斜变量, 则需要按照Zhang和Zhang(2012)的方法对投影进行放松. 具体地, 定义松弛投影后的倾斜变量

$$\mathbf{U} \equiv \mathbb{X} - \mathbf{Z}_{\mathcal{C}_X}\hat{\theta},$$

其中, $d = |\mathcal{C}_X|$, $\mathrm{tr}(V)$表示矩阵V的迹, λ为惩罚参数, $\hat{\theta}$ 满足

$$\hat{\theta} = \arg\min_{B\in\mathbb{R}^{d\times p}}\Big\{\frac{\mathrm{tr}((\mathbb{X} - \mathbf{Z}_{\mathcal{C}_X}B)^T(\mathbb{X} - \mathbf{Z}_{\mathcal{C}_X}B))}{2n} + \lambda\|B\|_1\Big\}.$$

利用倾斜变量$\mathbf{U}$, 即可得到参数β的一个线性估计

$$\hat{\beta}_L = (\mathbf{U}^T\mathbb{X})^{-1}\mathbf{U}^T\mathbb{Y}.$$

由于对投影进行了放松, 从而需要对估计$\hat{\beta}_L$进行纠偏. 假设$(\hat{\beta}^{(init)}, \hat{\gamma}^{(init)})$为模型参数$(\beta, \gamma)$的一个初值, 满足

$$\{\hat{\beta}^{(init)}, \hat{\gamma}^{(init)}, \hat{\sigma}\} = \arg\min_{b\in\mathbb{R}^p, r\in\mathbb{R}^q, \sigma\in\mathbb{R}^1_+}\Big\{\frac{\|\mathbb{Y} - \mathbb{X}b - \mathbb{Z}r\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0\|r\|_1\Big\}.$$

进而, 构造参数β的一个新的纠偏估计如下

$$\hat{\beta}_U = \hat{\beta}_L - (\mathbf{U}^T\mathbb{X})^{-1}\mathbf{U}^T\mathbb{Z}\hat{\gamma}^{(init)}.$$

最后, 通过该估计还可构造参数β的置信区间估计. 理论结果表明, 这样得到的参数β的点估计是相合的, 而且其置信区间估计的覆盖率是有保证的.

关键词: 部分线性模型; 有偏子模型; 半参回归; 相合估计; 主成分回归; 独立成分分析; 非参调整方法; 变量选择; 倾斜变量; 松弛投影; 惩罚最小二乘; 经验似然; 置信区域.

Consistent Statistical Inferences for High-dimensional Linear Models and Partially Linear Models

Yunhui Zeng

School of Mathematics, Shandong University,
Jinan, Shandong, 250100, P. R. China

ABSTRACT

In modeling procedure a large number of predictors usually are introduced at initial stage as a full model. But when it is applied to real problems, lots of predictors and large model will need heavy computation load and resource occupation, which may increase variance and mean squared error of the estimator, and even make the numerical algorithm unstable or abort abnormally. This will directly influence estimation of parameter and model prediction. In order to simplify the model and enhance predictability, some less significant variables are removed. So a reduced model (or restricted model) is formed. Under certain regularity conditions, the estimation of the remaining parameters in the reduced model is consistent. However, when the consistency of parameter estimation is specially taken into account for modeling with variable selection, the sparsity and oracle property (e.g. SCAD estimator) hold only in the pointwise sense, which means that such an estimator does not have a good global property. Furthermore, for the case of restricted-model (RM) estimation or post-model-selection estimation, if some significant variables are unfortunately removed, even if their coefficients are nearly zero, the restricted model (or submodel) is misspecified and the popularly used estimators for remaining parameters converge to pseudo parameters rather than the true ones. Even if the selected submodel is only locally misspecified, the commonly used estimators for the parameters in the submodel are still inconsistent.

On the other hand, in some application cases only part of the variables are interested, because they can be easily controlled. While for the other variables, maybe they are not emphasized or cannot be properly controlled, or their influences on the response

variable are not clear. If the full model is applied, the similar problems will be faced and the model may be misspecified. The submodel is highly biased if only regressed on the variables of interest, and so is the corresponding estimator. It is much reasonable to form a semiparametric model in which the variables of interest are parametric part and the other variables are nonparametric part, but the nonparametric part will be faced with the curse of dimensionality. Especially, when the dimension of the other variables are high the curse of dimensionality will ruin the estimation and prediction. In application fields more and more semiparametric models are to be faced.

All the above indicate that, it is still a problem to consistently estimate parameters and construct confidence regions in the modeling process with variable selection or some variables specified. In this thesis, we shall mainly study some estimation methods for parameter of interest and the corresponding model prediction in the above cases. For linear regression model

$$Y = \beta^T X + \gamma^T Z + \varepsilon,$$

suppose that parameter β and covariate X are only interested. In chapter two, a semiparametric method is firstly proposed to adjust the biased submodel

$$Y = \beta^T X + \eta.$$

By finding a proper direction $\tau \in \mathscr{T}$, a partially unbiased adjusted model

$$Y = \beta^T X + g(\tau^T Z) + \xi$$

is constructed, where $g(\tau^T Z) = E(Y - \beta^T X|\tau^T Z) = \gamma^T E(Z|\tau^T Z)$. Such an adjusted model is partially locally unbiased in the sense that

$$E(\xi|\tau^T Z, X) = 0, \quad \forall \tau \in \mathscr{T}, \forall (X, Z) \in \mathscr{W} \subset (\mathscr{X}, \mathscr{Z}).$$

Applying univariate nonparametric kernel method to estimate the nonparametric term $g(\tau^T Z)$ as following

$$\hat{g}(\tau^T Z) = \frac{\sum_{j=1}^n (Y_j - \beta^T X_j) K\{\tau^T (Z_j - Z)/h\}}{\sum_{j=1}^n K\{\tau^T (Z_j - Z)/h\}},$$

in which $K(\cdot)$ is a kernel function, h is bandwidth depending on the sample size n. Substituting $\hat{g}(\tau^T Z)$ into the adjusted model, an estimator of β is obtained as

$$\tilde{\beta}_A = (\frac{1}{n}\sum_{i=1}^n \tilde{X}_i \tilde{X}_i^T)^{-1} \frac{1}{n}\sum_{i=1}^n \tilde{X}_i \tilde{Y}_i,$$

where $\tilde{Y}_i = Y_i - \frac{\sum_{j=1}^{n} Y_j K\{\tau^T(Z_j - Z_i)/h\}}{\sum_{j=1}^{n} K\{\tau^T(Z_j - Z_i)/h\}}$, $\tilde{X}_i = X_i - \frac{\sum_{j=1}^{n} X_j K\{\tau^T(Z_j - Z_i)/h\}}{\sum_{j=1}^{n} K\{\tau^T(Z_j - Z_i)/h\}}$. It is proved that the estimator $\tilde{\beta}_A$ is $\sqrt{n}$-consistent on the subset $\mathscr{W}$. Furthermore, based on the idea of PT estimation with F-test, another estimator $\tilde{\beta}_{APT}$ which is also relied on the full model is obtained. In the third section, a confidence region estimation is constructed by empirical likelihood method. Upon the method proposed in chapter two, the curse of dimensionality is avoided by a univariate nonparametric function. Furthermore, the property of inference is insensible to the direction τ, so the new method is robust. No matter how large the bias of the submodel, simulation results show that parameter estimation and confidence region of the new method are better than those of existing methods.

Although the adjusted method proposed in chapter two can markedly reduce the estimation bias of the submodel, it only holds on a subset $\mathscr{W}$ of the covariates' support region. So in chapter three, based on the submodel $Y = \beta^T X + \eta$ and the parameter of interest β, a globally unbiased working model

$$Y = \beta^T X + g_1(\tilde{Z}^{(1)}) + \cdots + g_K(\tilde{Z}^{(K)}) + \zeta_K$$

is constructed with $E(Z) = 0$. The main idea is to firstly decompose covariate Z into independent components $\tilde{Z}^{(1)}, \cdots, \tilde{Z}^{(q)}$, then make use of the independent components correlated with covariate X. For each component $\tilde{Z}^{(l)}$, a univariate nonparametric term $g_l(\tilde{Z}^{(l)}) = E(Y - \beta^T X|\tilde{Z}^{(l)}) = \gamma^T E(Z|\tilde{Z}^{(l)})$, is added into the submodel to reduce the bias of the submodel.

(1) When the covariate Z is normally distributed, the principal component regression (PCR) is applied, $g_l(\tilde{Z}^{(l)}) = \alpha_l \tilde{Z}^{(l)}$ and the adjusted model is really a linear model

$$Y = \beta^T X + \alpha^T \tilde{\mathbf{Z}}_K + \zeta_K,$$

where $\tilde{Z}^{(l)}$ is the l-th principal component of Z;

(2) Otherwise, we adopt independent component analysis (ICA) method and $\tilde{Z}^{(l)}$ is the l-th independent component of Z.

Based on this adjusted model, applying univariate nonparametric kernel method to estimate each nonparametric term $g_l(\tilde{Z}^{(l)})$ as following

$$\hat{g}_l(\tilde{Z}^{(l)}) = \frac{\sum_{j=1}^{n}(Y_j - \beta^T X_j)K((\tilde{Z}_j^{(l)} - \tilde{Z}^{(l)})/h_l)}{\sum_{j=1}^{n} K((\tilde{Z}_j^{(l)} - \tilde{Z}^{(l)})/h_l)}, l = 1, \cdots, K,$$

in which $K(\cdot)$ is a kernel function, h_l is bandwidth depending on the sample size n. Substituting $\hat{g}_l(\tilde{Z}^{(l)})$ into the adjusted model, an estimator of β is obtained as

$$\hat{\beta}_A = (\frac{1}{n}\sum_{i=1}^{n} \breve{X}_i \breve{X}_i^T)^{-1} \frac{1}{n}\sum_{i=1}^{n} \breve{X}_i \breve{Y}_i,$$

where $\breve{Y}_i = Y_i - \sum_{l=1}^{K} \frac{\sum_{j=1}^{n} Y_j K((\tilde{Z}_j^{(l)} - \tilde{Z}_i^{(l)})/h_l)}{\sum_{j=1}^{n} K((\tilde{Z}_j^{(l)} - \tilde{Z}_i^{(l)})/h_l)}$, $\breve{X}_i = X_i - \sum_{l=1}^{K} \frac{\sum_{j=1}^{n} X_j K((\tilde{Z}_j^{(l)} - \tilde{Z}_i^{(l)})/h_l)}{\sum_{j=1}^{n} K((\tilde{Z}_j^{(l)} - \tilde{Z}_i^{(l)})/h_l)}$. It is proved that the estimator $\hat{\beta}_A$ is globally consistent on the whole space of covariates X and Z, and it is also asymptotically normal. Because the added nonparametric parts $g_l(\tilde{Z}^{(l)})$ are independent of each other, large computation load is avoided which is faced in the backfitting method for general additive model. Furthermore, the computation error is fairly small if the number K of adjusted parts is not large. When Z is normally distributed, the corresponding result can be obtained by applying least squares method to the linear adjusted model.

When the adjusted steps K is large or even close to the original dimension q of covariate Z, the method proposed in chapter three will result in large computation error and will lose its superiority. In chapter four, more generally, a two-stage remodeling and estimation method is constructed for sparse partially linear model

$$Y = \beta^T X + \gamma^T Z + f(U) + \varepsilon,$$

in which parameter β is interested and parameter γ is sparse. For simplicity, we assume that U is univariate and $E(Z) = 0$. In fact, $f(\cdot)$ can be extended to additive structure with multidimensional U. In details, in the first stage, making use of the independent components $\tilde{Z}^{(j)}$ correlated with covariate (X, U) and by a multi-step adjustment as that in chapter three, a globally unbiased model

$$Y = \beta^T X + \sum_{j=1}^{K_0} g_j(\tilde{Z}^{(j)}) + f(U) + \zeta_{K_0}$$

is reconstructed. In the second stage, we further reduce the adjusted model by a semi-parametric variable selection method proposed by Zhao and Xue(2009). In details, by a truncated expansion of orthogonal series method, each nonparametric term $g_j(\tilde{Z}^{(j)})$ and nonparametric term $f(U)$ are approximated with

$$g_j(\tilde{Z}^{(j)}) \approx \theta_j^T \Psi_j(\tilde{Z}^{(j)}), \quad f(U) \approx \nu^T \Psi_0(U).$$

Then we estimate parameters β, θ_j and ν with group SCAD method as following

$$(\hat{\beta}, \hat{\theta}, \hat{\nu}) = \arg\min_{\beta \in R^p, \theta_j \in R^{L}, \nu \in R^{L}} F(\beta, \theta, \nu),$$

where

$$F(\beta,\theta,\nu)=\sum_{i=1}^{n}\{Y_i-\beta^TX_i-\theta^T\mathbf{\Psi_i}-\nu^T\mathbf{\Psi_{0i}}\}^2+n\sum_{j=1}^{K_0}p_{\lambda_j}(\|\theta_j\|_2),$$

$p_\lambda(\cdot)$ is the SCAD penalty function defined as

$$p'_\lambda(w)=\lambda\{I(w\leq\lambda)+\frac{(a\lambda-w)_+}{(a-1)\lambda}I(w>\lambda)\},$$

satisfying $a>2$, $w>0$, $p_\lambda(0)=0$. Let $\hat{\mathcal{M}}_n=\{1\leq j\leq K_0:\hat{\theta}_j\neq 0\}$, $K_n=|\hat{\mathcal{M}}_n|$. For simplicity, we suppose that $\hat{\mathcal{M}}_n=\{1,2,\cdots,K_n\}$. Denote $g_j(\tilde{Z}^{(j)})=E(\gamma^TZ|\tilde{Z}^{(j)}),j=1,\cdots,K_n$, $\zeta_{K_n}=Y-\beta^TX-g_1(\tilde{Z}^{(1)})-\cdots-g_{K_n}(\tilde{Z}^{(K_n)})-f(U)$, finally, we get model

$$Y=\beta^TX+\sum_{j=1}^{K_n}g_j(\tilde{Z}^{(j)})+f(U)+\zeta_{K_n}.$$

After two-stage remodeling the final model is sufficiently simplified and globally conditionally unbiased. In the theoretical results, the convergence rate of parametric estimator $\hat{\beta}$ and nonparametric estimators $\hat{g}_l$ and $\hat{f}$ are proved, and the estimator $\hat{\beta}$ is proved to be asymptotically normal. Because variable selection much relies on the sparsity of the parameter, if we directly consider the partially linear model, some irrelevant variables with X but with nonzero coefficients may be selected into model. This may affect the estimation of the parameter β on its efficiency and stability.

In chapter five, for high dimensional linear model $Y=\beta^TX+\gamma^TZ+\varepsilon$ with normal covariates and error term, we combine the tilted variable method proposed by Cho and Fryzlewicz(2012) with the relaxed projection method proposed by Zhang and Zhang(2012) to remodel the biased submodel $Y=\beta^TX+\eta$ with parameter β of interest. If $\gamma^TE(Z|X)\neq 0$, then $E(\eta|X=x)$ is a nonzero function. So we shall firstly adjust the biased submodel with the method of Cho and Fryzlewicz(2012). After the components of Z correlated with X, which are denoted as $Z_{\mathcal{C}_X}$, are added into the submodel, an adjusted model

$$Y=\beta^TX+\sum_{k\in\mathcal{C}_X}\gamma_kZ^{(k)}+\zeta$$

is obtained, where $\zeta=\varepsilon+\sum_{k\in\jmath\setminus\mathcal{C}_X}\gamma_kZ^{(k)}$ and $\jmath=\{1,2,\cdots,q\}$. Then we compute the tilted variables corresponding to design matrix $\mathbb{X}$ as

$$\mathbf{U}_0\equiv(\mathbf{I}_n-\Pi_{Z_X})\mathbb{X},$$

where Π_{Z_X} is the projection onto space spanned by $\mathbf{Z}_{\mathcal{C}_X}$.

(1) If all the lengths of the tilted variables are not too short, based on the tilted variables $\mathbf{U}_0$ and the above adjusted model, parameter β can be directly estimated as

$$\hat{\beta}_T = (\mathbf{U}_0^T\mathbb{X})^{-1}\mathbf{U}_0^T\mathbb{Y}.$$

It is proved that $\hat{\beta}_T$ is consistent under mild conditions.

(2) Otherwise, it need to relax the projection by the method of Zhang and Zhang(2012). In details, the tilted variables with relaxed projection are defined as

$$\mathbf{U} \equiv \mathbb{X} - \mathbf{Z}_{\mathcal{C}_X}\hat{\theta},$$

where $d = |\mathcal{C}_X|$, $\mathrm{tr}(V)$ is the trace of matrix V, λ is the penalty parameter, $\hat{\theta}$ satisfying

$$\hat{\theta} = \arg\min_{B\in\mathbb{R}^{d\times p}}\left\{\frac{\mathrm{tr}((\mathbb{X} - \mathbf{Z}_{\mathcal{C}_X}B)^T(\mathbb{X} - \mathbf{Z}_{\mathcal{C}_X}B))}{2n} + \lambda\|B\|_1\right\}.$$

Applying the tilted variables $\mathbf{U}$ and based on the adjusted model, we can get a linear estimator

$$\hat{\beta}_L = (\mathbf{U}^T\mathbb{X})^{-1}\mathbf{U}^T\mathbb{Y}.$$

Because the projection is relaxed, so it need to reduce the bias of $\hat{\beta}_L$. Suppose that $(\hat{\beta}^{(init)}, \hat{\gamma}^{(init)})$ is an initial estimator of parameter (β, γ) of the full model, satisfying

$$\{\hat{\beta}^{(init)}, \hat{\gamma}^{(init)}, \hat{\sigma}\} = \arg\min_{b\in\mathbb{R}^p, r\in\mathbb{R}^q, \sigma\in\mathbb{R}_+^1}\left\{\frac{\|\mathbb{Y} - \mathbb{X}b - \mathbb{Z}r\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0\|r\|_1\right\},$$

a new unbiased estimator of parameter β is constructed as following

$$\hat{\beta}_U = \hat{\beta}_L - (\mathbf{U}^T\mathbb{X})^{-1}\mathbf{U}^T\mathbb{Z}\hat{\gamma}^{(init)}.$$

Finally, based on this point estimator, confidence region of β can also be estimated. It is proved that such a point estimator $\hat{\beta}_U$ of parameter β is consistent and coverage rate of the confidence region is guaranteed.

Key words: Partially linear model; biased submodel; semiparametric regression; consistent estimation; principal component regression; independent component analysis; nonparametric adjusted method; variable selection; tilted variable; relaxed projection; penalized least squares; empirical likelihood; confidence region.

第一章　绪论

本章主要介绍与论文主体相关的一些基础内容, 主要包括常用统计模型及其估计方法, 高维统计模型及其估计方法, 面临的问题和困难等.

§1.1　常用统计模型及估计方法

回归模型是研究两个（随机）变量之间关系的一种重要的统计方法. 一般地, 回归模型包括参数回归模型,非参数回归模型和半参数回归模型等. 其中, 参数回归模型假定回归模型的函数结构已知, 但是包含若干个未知参数; 非参数回归模型对回归模型的函数结构不做假定, 也不设置参数, 函数在每一点的值由样本数据决定; 当回归模型的部分函数结构已知时, 可采用半参数回归模型. 上述几种回归模型有着各自的优点和缺点, 在生物,医学,经济,工业,金融以及工程技术等领域有着广泛的应用.

参数回归模型一般包括线性回归模型和非线性回归模型等, 其中, 线性回归模型的结构简单, 是应用最广泛的模型之一, 见文献Montgomery等(2001), Kutner等(2004), Draper和Smith(1998). 线性回归模型的一般形式为$Y = X^T\beta + \varepsilon$, 其中, Y为响应变量, X为p维预测变量, β为p维未知参数向量, ε为随机误差, 假设满足$E(\varepsilon|X) = 0$, $Var(\varepsilon|X) = \sigma^2$. 在线性回归模型的参数估计理论与方法中, 最小二乘法占有非常重要的地位, 其基本思想是参数β的真值应该使得均方误差$(\mathbb{Y} - \mathbb{X}^T\beta)^T(\mathbb{Y} - \mathbb{X}^T\beta)$达到最小, 即$\hat{\beta} = (\mathbb{X}\mathbb{X}^T)^{-1}(\mathbb{X}\mathbb{Y})$. 有关最小二乘法的详细介绍及其理论结果, 可参考文献Pfeffermann(1984), Harville(1976), Lewis 和Odell(1966). 一般地, 当研究人员建立响应变量与预测变量之间的数学表达式时, 这些模型中的参数往往是非线性的. 非线性回归模型的形式为$Y = r(X;\beta) + \varepsilon$, 其中, Y为响应变量, $X = (X^{(1)}, \cdots, X^{(p)})^T$为$p$维预测变量, $\beta = (\beta_1, \cdots, \beta_p)^T$为$p$维未知参数向量, 回归结构$r(x;\beta)$已知但是依赖于未知参数$\beta$, ε为随机误差. 对于非线性回归模型来说, 至少存在一个期望函数关于参数的导数, 该导数至少依赖一个参数. 关于非线性回归模型的估计方法及相关性质, 参见文献Bates和Watts(1988), Seber和Wild(2003). 参数回归模型最显著的优点是简单直观, 而且对于感兴趣的因素而言, 模型的可解释性强. 参数回归模型使得对回归函数的估计可以划归为对几个参数的估计, 并且当参数模型假设被证实是合理的情况下, 参数方法要比非参数方法更加有效, 见文献Härdle和Mammen(1993), Hart(1997).

但是, 如果假设的参数模型结构不正确, 将会导致错误的结论. 为了减少模型的偏差, 这时人们采用非参数的方法来估计模型内在的函数结构. 非参数回归模型的一般形式为$Y = r(X) + \varepsilon$, 其中, Y为响应变量, X为p维预测变量, $r(x)$为结构未知的p维函数, ε为随机误差. 非参数回归模型既不预先设定函数的结构形式, 也不引入未知参数, 函数在每一点的值都由所观测到的样本数据所决定, 因此非参数模型方法具有较大的

适应性, 在降低建模偏差方面具有较大优势. 非参数回归模型的理论与应用, 详见文献Hart(1997), Efronmovich(1999).

常用的非参数估计方法主要有核平滑, 正交级数和样条等方法. 几种非参数估计方法在本质上是一样的, 只是在处理具体问题时存在不同的技术上的便利性差异. 本文主要用到核平滑方法和正交级数方法. (1)核平滑方法. 核平滑方法简单易用, 在应用时便于处理. 传统的核平滑方法主要有Nadaraya-Watson估计, Priestly-Chao估计, Gasser-Muller估计. 局部多项式方法由于具有良好的渐近性质和边界适应性, 近年来受到越来越多的关注, 详细的算法和理论性质见文献Fan和Gijbels(1996). (2)正交级数方法. 正交级数方法的主要思想是用一些标准函数基的线性组合去逼近回归函数, 其基本做法就是用一个有限截断的正交级数(也称为部分和)去近似回归函数. 常用的正交级数有三角余弦级数, 幂级数, Harr级数等.

在利用核平滑方法进行非参数估计时涉及到窗宽的选择, 对于正交级数方法涉及到截断点的选择, 均称为非参平滑参数. 非参平滑参数的选取将直接影响非参估计方法的好坏, 常用的平滑参数选择方法包括数据驱动方法, 如交叉验证(Cross-Validation, Stone (1974))和广义交叉验证(Generalized Cross-Validation, Craven和Wahba(1979)), 还有插入法(Plug-in method, Ruppert, Sheather和Wand(1995))等.

一维非参回归方法可以很自然地推广到多维非参回归, 然而, 由于“维数祸根”(curse of dimensionality)问题, 在多维变量的情况下未知函数的估计效果会很不好. 因此, 人们又提出了新的模型来进行改进, 如Buja,Hastie和Tibshirani(1989)提出的可加模型, Friedman和Stuetzle(1981)提出的投影追踪模型, Hastie和Tibshirani(1993)提出的变系数模型等. 一些最新的估计方法, 见文献Fan,Härdle和Mammen(1998), Fan和Jiang(2005), Lin,Cui和Zhu(2008), Lin和Li(2008)等.

由于参数模型和非参数模型存在如上的固有问题, 介于参数模型和非参模型之间的半参数模型应运而生. 半参模型既包含参数部分又包含非参部分, 因此这类模型具有参数模型的可解释性和非参模型的灵活性, 在实际中有很强的适应性. 本学位论文将涉及其中的部分线性模型和可加部分线性模型.

(1) 部分线性模型可表示为$Y = X^T\beta + f(U) + \varepsilon$, 其中, Y为响应变量, X和U为相关的协变量, 通常U取单变量, 误差项ε与X和U独立且期望为0方差为σ^2, β为未知参数向量, $f(\cdot)$为未知的光滑函数. Speckman(1988)介绍了部分线性模型截面最小二乘估计的方法, 并证明了参数估计的$\sqrt{n}$相合性及渐近正态性, 同时在一些适当的条件下得到了非参数项估计的收敛速度, 更多结果见文献Härdle,Liang和Gao(2000).

(2) 可加部分线性模型具有形式$Y = X^T\beta + \sum_{l=1}^{d} f_l(U^{(l)}) + \varepsilon$, 其中, Y 是响应变量, X和$U = (U^{(1)}, \cdots, U^{(d)})^T$分别是$R^p$和$R^d$上的协变量, $f_1, \cdots, f_d$是未知的光滑函数, β是p-维的未知参数向量, ε是随机误差. 相对于部分线性模型, 对可加部分线性模型的研究要更困难. 原因在于可加部分线性模型中有更多的非参项, 其计算由

于用到后拟算法(backfitting algorithm)而变得更加复杂. 后拟方法(backfitting procedures)由Breiman和Friedman(1985), Buja,Hastie和Tibshirani(1989)提出, Opsomer和Ruppert(1999)提出了一个$\sqrt{n}$相合的后拟估计.

在半参模型中, 人们主要对参数部分感兴趣, 把非参部分看作无穷维的讨厌参数. 因此, 人们所期望的是更加有效的参数估计, 而获得半参有效估计的方法通常采用截面方法. Li(2000) 提出用一般的级数方法去估计可加部分线性模型, 并证明如果误差是条件同方差的, 用该方法得到的参数估计能够达到可加部分线性模型的半参有效界. 由于核平滑方法的简单易用, 所以Fan和Li(2003)对可加部分线性模型又提出了核平滑方法, 该方法在估计过程中充分利用到非参的可加结构. 关于这个模型最新的进展, 可以参考文献Deng和Liang(2010), Zhou,Jiang和Qian(2011)等.

§1.2 高维统计模型及估计方法

高维数据分析与建模是目前统计领域研究的热点之一. 数值计算技术和存取技术的快速发展为人们存储数据带来了极大的便利, 现场所收集数据的维数也呈几何级数速度增长. 但海量数据在为我们提供了更多信息的同时, 也为如何进行有效的和及时的数据分析带来了极大的挑战. 传统的统计分析主要考虑协变量的维数p远远比样本容量n小的情形. 近年来, 一类"变量维数p相当或大于样本容量n"的高维数据分析问题在诸多领域内出现, 从生物学,基因生命学,经济学, 到航空航天,气候气象,模式识别,文字挖掘等领域, 参考专著Bühlmann和van de Geer(2011).

1. 高维统计模型

在高维数据模型分析中, 目前常用的是线性模型$Y = \theta^T W + \varepsilon$, 其中, 协变量$W$包含了所有可能对预测变量$Y$有影响的变量, 而且$W$是外生的即满足$E(\varepsilon|W) = 0$. 因此$W$的维数$r$是很大的, 甚至可能超过样本数据的容量$n$. 此外, 还有高维可加模型,高维部分线性模型,高维变系数模型等. 基于维数r与样本容量n的大小关系, 高维模型主要有如下几类:

(1) 高维模型(high dimensional model): $r = O(n^\delta)$, 其中$\delta > 0$, 即模型维数随着样本容量的增长呈多项式级增长. 特别, 当$0 < \delta < 1$时, 维数r随着样本容量n的增长而增长, 但是满足$r < n$, 在这种情况下模型维数小于样本容量.

(2) 超高维模型(ultra high dimensional model): $r = O(\exp(n^\alpha))$, 其中$\alpha > 0$, 即模型维数随着样本容量的增长呈指数级增长.

在高维统计模型中, 通常假设重要的协变量的个数是比较少的, 比如生物基因维数很高, 但是导致某类疾病发生的一般只有少数几个或者几十个基因. 所以, 在做高维模型的统计推断时一般假定模型是稀疏的. 在实际研究过程中, 目前存在如下两种稀疏性的定义.

(1) 稀疏性定义1: 记$S_0 = \{1 \le j \le r : \theta_j \ne 0\}$, 令$s_0 = |S_0|$表示集合的元素个数, 满足$s_0 \ll r$. 根据定义1, 协变量$W$中仅有一小部分分量对响应变量$Y$有影响, 即对应系数不为零, 而其余分量的对应系数则严格为零.

(2) 稀疏性定义2: 直观定义为, 仅有一小部分W分量对应系数是大的, 而其余分量的系数很小但是未必为零. 常见定义如, 记$s_1 = \|\theta\|_1$, 或$s_1 = \|\theta\|_2$, 或$s_1 = \sum_{j=1}^{r} \min\{|\theta_j|/\delta, 1\}$,$\delta > 0$, 满足$s_1 = o(n)$即相对样本容量$n$ 而言是小量. 显然, 定义2更广泛, 在最近的研究中更多的学者采用定义2的稀疏性定义.

2. 常用降维方法

对于高维数据的处理, 一个基本的分析思路是提出有效的降维方法, 在不损失信息的前提下将变量的维数降低到合理的程度, 然后再利用低维的统计分析方法进行合理的统计推断. 一方面由于高维数据分析十分困难, 采用全模型进行统计推断其结果基本不可用, 另一方面高维数据中的信息往往主要包含在一个或几个低维结构中, 因此降维是分析高维数据的一个重要手段. 传统的主成分分析,充分降维和近年来比较流行的变量选择等都属于降维方法, 其它相关的降维方法参考李根,邹国华和张新雨(2012)的综述性文献.

2.1 主成分分析(PCA)

主成分分析方法的初衷是通过投影的方法, 将多维数据以尽可能少的信息损失投影到较低维的空间, 使数据降维从而达到简化数据结构的目的. 同时, 它也是将多个相关变量以尽可能少的信息损失为原则, 进行综合化为少数几个不相关变量的方法. 设$\mathbb{X} = (X_1, \cdots, X_n)^T = (\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(p)})$为来自正态分布总体$N(\mu, \Sigma)$的一个样本, 其中$n > p$, $\Sigma > 0$. 设$\nu_1 \ge \nu_2 \ge \cdots \ge \nu_p \ge 0$为$\mathbb{X}^T\mathbb{X}$的特征值, $\gamma_1, \cdots, \gamma_p$为其对应的标准正交特征向量. 称$\gamma_j$为$\mathbb{X}$在$\mathbf{R}^p$中的第$j$个主轴向量(简称主轴), $j = 1, \cdots, p$; $X_i^T\gamma_1, \cdots, X_i^T\gamma_p$为$X_i$的主坐标, $i = 1, \cdots, n$; n个样本点的第j个主坐标形成的向量$U_j = \mathbb{X}\gamma_j = (X_1^T\gamma_j, \cdots, X_n^T\gamma_j)^T$ 为$\mathbb{X}$的第j个主成分, $j = 1, \cdots, p$. 对样本点x, 称值$x^T\gamma_j$为该样本关于第j个主成分的得分. 样本主成分具有如下一些性质:

(1)主成分向量长度平方等于对应的特征值, 主成分向量之间相互正交. 特别, 对于正态分布变量, 则主成分向量之间也是相互独立的.

(2)称$\mathbb{U} = \mathbb{X}\Gamma$为对$\mathbb{X}$的主成分变换, 称$\mathbb{X} = \mathbb{U}\Gamma^T$为用主成分对$\mathbb{X}$的数据恢复变换, 其中$\Gamma = (\gamma_1, \cdots, \gamma_p)$.

(3)称$(\frac{1}{\nu_j}\mathbf{X}^{(j)T}U_j)^2 = \nu_j\gamma_{ij}^2$为主成分$U_j$对$\mathbf{X}^{(j)}$的变差贡献, 则主成分$U_j$对$\mathbb{X}$的变差贡献等于它对$\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(p)}$的变差贡献和, $\mathbf{X}^{(j)}$的总变差等于各主成分对它的变差贡献和.

(4)如果用总变差$\mathrm{tr}(\mathbb{X}^T\mathbb{X})$反映样本$\mathbb{X}$的差异信息, 则$\nu_j$为主成分$U_j$对差异信息的贡献大小, 即有$\mathrm{tr}(\mathbb{X}^T\mathbb{X}) = \Sigma_{i=1}^{n}\|X_i\|^2 = \Sigma_{j=1}^{p}\|U_j\|^2 = \Sigma_{j=1}^{p}\nu_j$.

关于主成分分析的最新研究成果, 主要有如下两类. 一类是主成分本身的推广, 包

括Gnanadesikan提出的广义主成分(generalized principal components), Hastie和Stuetzle提出的主曲线(principal curves), Webb提出的核主成分(kernel principal components), Besse和Ferraty提出的曲线性主成分分析(curvilinear PCA), Dauxois和Pousse提出的半线性主成分(semi-linear PC), Donnel等提出的可加主成分(additive principal components, APC), Salinelli定义的非线性主成分(nonlinear principal components, NLPCs), 详细可参见文献Salinelli(2009). 另一类是针对高维情形的推广, 研究的主要内容包括主成分分析的相合性,稀疏性以及主成分回归等, 详细可参考文献Johnstone和Lu(2009), Shen,Shen和Marronx(2011), Barrios和Lansangan(2010).

2.2 充分降维(sufficient dimension reduction)

考虑一个回归模型, 其中Y为响应变量, X为p维预测变量, 假设Y和X有联合分布$F(Y,X)$, 回归的主要目标就是在已知$X=x$的条件下估计分布$F(Y|X=x)$. 充分降维的基本想法是, 寻找一个矩阵$B\in\mathbb{R}^{d\times p}$满足$d<p$, 使得$F(Y|X=x)=F(Y|Bx)$, 即给定$Bx$的条件下$Y$与$x$独立. 当$B$的维数$d$比较小时, 就可通过$BX$代替$X$建立对响应变量$Y$的回归方程来实现降维. 充分降维的思想是由Cook于1994年提出的, 并经许多统计学者不断发展和完善; 可参考文献Cook和Ni(2005), Adragni和Cook(2009).

定义一个降维运算$S:\mathbb{R}^p\to\mathbb{R}^d$, $d<p$, 称该运算为充分的, 如果满足如下三个条件之一:

1) 逆降维条件: $X|(Y,S(X))\sim X|S(X)$;

2) 正向降维条件: $Y|X\sim Y|S(X)$;

3) 联合降维条件: $X\perp Y|S(X)$.

其中, $\perp$表示独立, $\sim$表示同分布. 三个条件都意味着低维空间$S(X)$包含了X中关于Y的全部信息, 从而有$E(Y|X)=E(Y|S(X))$.

在实际使用中, 最有用的是最小充分降维空间, 即所有充分降维空间的交集, 称为中心子空间, 记作$S_{Y|X}$. 估计中心子空间的方法主要分为非参数方法和参数方法. 常用的非参数方法包括切片逆回归(SIR),切片平均方差估计(SAVE),图回归等. 切片逆回归利用了样本估计的一阶条件矩$E(X|Y)$来建立对$S_{Y|X}$的估计, 在一定条件下得到的估计只是$S_{Y|X}$的一个子集; 而切片平均方差估计利用了样本估计的前两阶矩$E(XX^T|Y)$, 它尽可能地去估计$S_{Y|X}$, 但是在检测线性性方面不如切片逆回归. Cook和Forzani(2009)在条件正态性的假设下采用最大似然法估计$S_{Y|X}$, 结果表明基于模型的新的估计量优于非参方法, 并且对偏离正态性有较好的稳健性.

Li(2007)研究发现很多基于矩的充分降维方法可以归结为如下的广义特征值问题

$$M_n v_{nj}=\lambda_{nj}N_n v_{nj},\ \forall j=1,2,\cdots,p$$

其中, $M_n\geq 0$是一个依赖于特定降维方法的非负定对称核矩阵, $N_n>0$为正定对称矩阵, v_{nj}为M_n关于N_n的第j个特征向量, λ_{nj}为对应的广义特征值. 采用记号M_n, N_n和λ_{nj}

表示是对应总体M, N和λ_j的样本形式. 在一定的边际分布条件下, 对应于前d个非零特征值$\lambda_{n1} \geq \cdots \geq \lambda_{nd}$的特征向量$\{v_{n1}, \cdots, v_{nd}\}$将组成中心子空间的一个基底的相合估计. 在具体计算时可以将该广义特征值问题转化为一个约束最小二乘目标函数的最小化问题进行求解. 记

$$\hat{V} = \arg\min_{V \in \mathbb{R}^{p \times d}} \sum_{j=1}^{p} \|N_n^{-1} m_j - VV^T m_j\|_{N_n}^2 \ s.t.\ V^T N_n V = \mathbf{I}_d,$$

其中, m_j表示$M_n^{1/2}$的第j列,$j = 1, \cdots, p$. 从而, span$(\hat{V})$便是中心子空间的一个估计. 对于切片逆回归(SIR), 矩阵$M = Cov[E\{X - E(X)|Y\}]$, $N = \Sigma$为X的协方差阵; 对于切片平均方差估计(SAVE), 矩阵$M = \Sigma^{1/2} E[\{\mathbf{I}_p - Cov(z|Y)\}^2]\Sigma^{1/2}$, $N = \Sigma$, 其中$z = \Sigma^{-1/2}\{X - E(X)\}$.

最初的充分降维方法在估计结果的线性组合中采用了所有的预测变量, 所以其解释性不好, 难以识别重要的变量, 而且可能影响到降维方法的效果. 因此, 逐渐地提出了稀疏的充分降维方法以及分量独立的稀疏充分降维方法等, 见文献Li(2007),Chen,Zou和Cook(2010)等.

2.3 变量选择(variable selection)

变量选择在高维数据分析中发挥着重要的作用. 传统的变量选择方法, 例如Akaike(1973)提出的AIC信息准则(Akaike' s information criterion), Mallows(1973)提出的C_p准则, Schwarz(1978)提出的BIC信息准则(Bayesian information criterion), 采用逐步选择子集的变量选择方法. 当预测变量的维数为p时, 子集的数量则有2^p个, 因此当p较大的时候计算量将会变得特别大, 同时估计量的渐近性质很难得到. 而且, 选择的最优变量子集还缺乏稳定性, 见文献Breiman(1996), Fan和Li(2001), Miller(2002).

近年来, 一些新的变量选择方法应运而生, 如Lasso方法(Tibshirani(1996)), 桥回归(bridge regression, Fu(1998)), Elastic Net (Zou和Hastie(2005)), SCAD (smoothly clipped absolute deviation, Fan和Li(2001), Fan和Peng(2004)), SIS(sure independence screening, Fan和Lv(2008)), NIS (nonparametric independence screening, Fan,Feng和Song(2011)), DC-SIS方法(Li,Zhong和Zhu(2012)), Dantzig selector(Candès和Tao(2007)), 广义Dantzig selector方法(James和Radchenko (2009)), 适应的Dantzig selector方法(Dicker和Lin(2009)). 这些方法可以分为三类: (1)惩罚类方法, 有Lasso方法, 桥回归方法, Elastic Net方法, SCAD方法, 以及相应的group Lasso方法和group SCAD方法等; (2)独立性方法, 包括SIS方法, NIS方法, DC-SIS方法等; (3)线性优化方法, 包括Dantzig selector方法, 广义Dantzig selector方法, 适应的Dantzig selector方法等. 关于高维变量选择方法的最新研究进展, 参见综述性文献Fan和Lv(2010).

此处介绍几种与论文主体有关的变量选择方法, 包括Lasso方法,SCAD方法和group SCAD方法等. 对于线性模型$Y = X^T\beta + \epsilon$, 基于惩罚最小二乘的变量选择就是使下式

达到最小：

$$F(\beta) := \frac{1}{2n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \sum_{l=1}^{p} p_{\lambda_l}(|\beta_l|),$$

其中, 其中p是协变量X和参数β的维数, $p_{\lambda_l}(\cdot)$是惩罚函数, 惩罚函数的形式决定了惩罚最小二乘的变量选择方法的表现. 基于惩罚最小二乘的变量选择方法能够同时实现变量选择和参数估计. 当$p_\lambda(|t|) = \lambda|t|^s$时, 我们称它为$L_s$惩罚. 特别地, 当$0 < s < 2$时, L_s惩罚就是桥回归(Fu(1998)); L_1惩罚的最小二乘估计就是Lasso(Tibshirani(1996)); L_2惩罚的情况就是岭回归. Fan和Li(2001)提出了SCAD惩罚, 其定义如下：

$$p_\lambda(|t|) = \begin{cases} \lambda|t|, & 0 \le |t| < \lambda; \\ -(|t|^2 - 2a\lambda|t| + \lambda^2)/2(a-1), & \lambda \le |t| < a\lambda; \\ (a+1)\lambda^2/2, & |t| \ge a\lambda, \end{cases}$$

满足$a > 2$, $\lambda > 0$. 而且, Fan和Li(2001)证明了SCAD惩罚估计量可以达到oracle性质, 也就是该估计量的性质跟潜在的真模型是预先给定了的一样好.

通过将最小二乘惩罚改成似然惩罚, 便可以直接将惩罚最小二乘的方法推广到处理广义线性模型. 与传统的非连续的变量选择方法相比, 惩罚最小二乘和惩罚似然方法具有连续性, 并且比传统方法更稳健. 然而, 对于不同的惩罚函数, 相应的估计量的表现也截然不同. 例如, 回归系数的岭估计全是非零的, 因此就无法给出一个易于解释的模型, 而Lasso方法能将某些很小的系数压缩成零. 当设计阵列正交时, Lasso方法的参数估计具有显式表示

$$\hat{\beta}_j^{Lasso} = \mathrm{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \lambda/2)_+,$$

其中, $\hat{\beta}^0$为基于全模型(包含所有预测变量的模型)的参数β的最小二乘估计. 当设计阵非列正交时, 可给定λ然后使用二次算法求解参数估计. 正是由于Lasso方法的简单性和压缩性, 同时为了改善其自身存在的一些缺陷, 因此出现了它的很多衍生版本, 如Elastic Net (Zou和Hastie(2005)), 用于群组数据的group Lasso (Yuan和Lin(2006)), LAD-Lasso(Wang,Li和Jiang(2007)), Scaled Lasso(Sun和Zhang(2012))等等. 关于Lasso方法的更详细性质, 参考文献Zhang和Huang(2008)及Bühlmann和van de Geer(2011).

对于SCAD方法, Fan和Li(2001)研究了在有限维情况下参数估计的oracle性质. 当设计阵列正交时, SCAD方法的参数估计具有显式表示

$$\hat{\beta}_j^{SCAD} = \begin{cases} \mathrm{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \lambda)_+, & |\hat{\beta}_j^0| \le 2\lambda; \\ ((a-1)\hat{\beta}_j^0 - \mathrm{sign}(\hat{\beta}_j^0)a\lambda)/(a-2), & 2\lambda < |\hat{\beta}_j^0| < a\lambda; \\ \hat{\beta}_j^0, & |\hat{\beta}_j^0| \ge a\lambda. \end{cases}$$

由于其具有良好的oracle性质, 同时还具有稀疏性,无偏性和连续性, 因此受到广泛的青睐. Fan和Peng(2004)将这些结果推广到了高维(即$p = o(n^{1/3})$)的情形. Wang,Chen和Li

(2007)将SCAD方法推广, 提出了group SCAD方法. Wang,Li和Tsai(2007)研究了SCAD惩罚函数下惩罚参数的选择, 文中指出基于SCAD惩罚函数采用GCV方法选择惩罚参数时得到的选模型将包含所有的重要变量, 可能有一些不重要的变量会被选入模型. 但是, 这样能够避免模型的误定. Kim,Choi和Oh(2008), Kim和Kwon(2012)进一步研究了SCAD方法的估计性质. Lv和Fan(2009), Fan和Lv(2011)研究了在超高维(即$p_n = O(\exp(cn))$)情形下SCAD方法的估计性质.

3. 本论文主要研究内容

对于进行降维之后的统计推断问题, 在其余变量系数严格为零的稀疏性定义下并在一定的正则条件下, 采用传统的Lasso方法和SCAD方法等可以得到相合的参数估计. 然而, 当采用变量选择的模型特别需要参数估计的相合性时, 稀疏性和oracle性质(如SCAD 估计)仅在逐点意义下成立. Leeb和Pöstscher(2008),Leeb和Pöstscher(2005)表明, 在样本容量趋于无穷时稀疏估计的最大均方误差(maximal scaled mean squared error)也将趋于无穷. 这意味着基于选模型的参数估计不具有一个好的全局性质. 而且, 对于简化模型的估计或模型选择后(post-model-selection)估计, 如果一些重要变量不幸被剔除, 虽然其系数近似为零但对响应变量仍然有影响, 那么简化模型(亦称子模型)是有偏的, 从而对于子模型中参数采用常用方法得到的估计将收敛于伪参数而不是真实参数值. 即使选出的子模型只是局部有偏, 采用常用的估计方法得出的参数估计也仍然是不相合的. 况且, 在高维模型情形下检验稀疏性和非稀疏性是一件困难的事.

另一方面, 在某些应用场景下, 可能只对其中的部分变量及其系数感兴趣, 因为这些变量是可以适当控制的. 反之, 对其余的变量并不感兴趣, 或者这些剩余变量本身不容易被精确控制, 或者剩余变量对响应变量的影响结构还不是很清楚. 在这种情况下, 如果采用全模型同样会面临上述的问题, 而且可能会出现模型误定的情形; 如果只将响应变量对这些感兴趣变量进行回归分析, 显然模型是有偏的, 对应的参数估计也将是有偏的; 如果在回归结构中, 将感兴趣变量作为参数项而将剩余变量作为非参项处理, 固然这样得到的半参回归模型相对较为合理, 但是这时非参项将可能面临维数祸根问题(curse of dimensionality). 特别在剩余变量的维数很高时, 维数祸根问题将直接导致半参回归模型的估计结果和预测效果的损失.

以上表明, 在包含变量选择(variable selection)或者变量指定(variable specified)的建模过程中相合估计和置信域构造仍然是一个困难, 关于这方面的参考文献不多. Zhang (1992)对于线性模型研究了基于广义预测误差准则(generalized final prediction error criterion)的选模型下参数估计的性质, 发现经过变量选择后参数的推断效果会变差, 特别是点估计存在偏差从而不是相合的, 而且置信区间会变大. 在有感兴趣参数情形下, Shen,Huang和Ye(2004)研究了对线性模型的讨厌参数部分按广义预测误差准则进行模型选择后基于选模型的兴趣参数估计的性质, 结果表明兴趣参数的估计是渐近相合的而方差比真模型有所增加. 但是由于其变量选择所用准则实际是C_p准则和BIC准

则的推广, 所以在讨厌变量维数较大时不可避免地会遇到变量选择中固有的相关问题. 对于高维线性模型, Leeb(2009)对于协变量和误差项服从联合正态分布的情形下基于变量选择后子模型的预测进行了研究, 其结果表明, 就覆盖率和区间长度而言, 在一定条件下得到的预测区间是近似有效的. 对于包含感兴趣参数的一类部分线性单指标模型, Li,Zhu和Zhu(2011)利用充分降维方法导出了兴趣参数的一种估计方法, 并得到了相应的渐近正态性结果, 但在单指标中变量维数较大时该方法将受到非参方法固有缺陷的影响.

本论文将主要研究这种情形下包含在变量选择后子模型中的参数或变量指定模型中的感兴趣参数的估计方法以及相应模型的预测. 对于线性模型$Y = \theta^T W + \varepsilon$, 假设$W$的前$p$个分量(记作$X$)为相对重要的, 而后$q = r - p$个分量(记作$Z$)经过变量选择后被剔除, 对应的参数$\theta$被分解为$\theta = (\beta^T, \gamma^T)^T$, 从而得到变量选择后的工作模型为$Y = \beta^T X + \eta$, η是新的误差项. 由于$\eta = \gamma^T Z + \varepsilon$, 当系数$\gamma$非零而且协变量$Z$与$X$相关时, 一般地有$\gamma^T E(Z|X) \neq 0$, 则$E(\eta|X = x)$ 是一个非零函数. 因此, 一个很自然的想法就是将协变量Z中与X相关的部分分量引入该工作子模型, 从而实现对该有偏工作子模型进行纠偏. 类似地, 适用于高维部分线性模型.

本论文采用的方法主要分为两类. 一类是第二,三,四章采用的半参方法, 通过寻找方向τ, 从而增加一个或多个一维的非参项来调整有偏子模型. 随后, 采用截面最小二乘方法, 得到了由该纠偏模型导出的参数β的估计及其相合性和渐近正态性. 当采用多步调整时, 为了避免一般部分可加模型采用后拟算法(backfitting)带来的巨大运算量, 在第三章中我们采用了主成分分析或独立成分分析方法, 以保证增加的非参调整项之间相互独立. 为了防止调整项过多, 在第四章中我们采用了组变量选择方法对调整模型进行化简. 当调整的项数K不大时新算法的计算误差将很小. 另一类方法是第五章的松弛投影方法, 适用于协变量和误差项均服从正态分布的高维线性模型. 其中, 首先采用Cho和Fryzlewicz(2012)的倾斜变量方法, 将协变量Z中与X相关的分量扩充到有偏子模型中, 从而得到一个调整模型. 如果倾斜变量的长度不太小, 可直接基于倾斜变量和调整模型进行兴趣参数的估计; 如果存在长度很小的倾斜变量, 则按照Zhang和Zhang(2012)的方法对投影进行放松后再对感兴趣参数进行估计, 包括点估计和置信区间估计.

第二章　高维线性回归模型的有偏子模型的半参估计方法

§2.1　引言

在线性模型建模的初始阶段, 大量的预测变量被引入从而形成全模型, 但是额外变量的引入将会增加参数估计的方差(variance)和均方误差(MSE), 甚至导致数值算法不稳定, 参考文献Wu,Harris和McAuley(2007). 为了简化模型和增强可预测性, 将一些不重要的预测变量从全模型中剔除, 形成简化模型(稀疏模型). 在一定的正则条件下, 简化模型中参数的估计是相合的. 例如, Fan和Li(2001), Fan和Peng(2004)提出的SCAD估计(smoothly clipped absolute deviation)已经具有稀疏性和oracle性质, 该性质定义如下: 将参数向量真值中的零分量和近似为零的分量估计为零, 而非零分量的估计仍然是相合的. 也就是说, 当零分量和很小的分量从模型中剔除时,保留在模型中的分量仍然可被相合估计.

然而, 当采用变量选择的模型特别需要参数估计的相合性时, 自然而然地将会产生以下的问题. 首先, SCAD估计的上述稀疏性和oracle性质仅在逐点意义下成立. 文献Leeb和Pöstscher(2008), Leeb和Pöstscher(2005)表明, 在样本容量趋于无穷时一个稀疏估计的最大均方误差(maximal scaled mean squared error)也将趋于无穷. 这意味着这样一个估计不具有一个好的全局性质. 其二, 对于限制模型(RM)估计或模型选择后(post-model-selection)估计, 如果一些重要变量不幸被剔除, 那么限制模型(亦称子模型)是误定的, 从而对于子模型中参数采用常用方法得到的估计将收敛于伪参数, 而不是真实参数值; 见文献Severini(1998)、Hall和Inoue(2003)、Schennach(2007). 即使选出的子模型只是局部误定,常用的RM参数估计也仍然是不相合的; 见文献Claeskens和Carroll(2007), Hjort和Claeskens(2003). 其三, PT(preliminary test)估计虽然本质上区别于RM(restricted model)估计, 但是它依赖于检验过程; 参见文献Ehsanes Saleh(2006), Sen和Ehsanes Saleh(1987)以及Sen(1979). 尽管这样的估计过程只是估计子模型中的参数而且具有一些好的渐近性质, 但是它依赖于全模型的参数估计, 从而当全模型的参数是高维时这样的估计是不可得到的. 再有, 基于不相合估计的置信域必然是无效的, 因为它的中心不是参数向量的真值.

以上表明, 在包含变量选择的建模过程中相合估计和置信域构造仍然是一个困难. 对于线性模型, 本章将提出一个半参调整方法, 使得有偏子模型变为部分无偏. 从而, 对于工作子模型中的参数构造出两类调整估计. 进一步对工作子模型中的参数, 基于调整的估计函数, 采用经验似然(Owen 1988,1990)构造出一个有效的置信域. 采用本章方法, 通过一个一维的非参函数将成功地避免维数祸根(curse of dimensionality), 而且该推断的性质对非参函数中变量的方向不敏感甚至无关. 所以新方法对变量选择是稳健

的. 无论子模型的偏大小, 在相合意义下新的参数估计和置信域优于现有方法. 对于该新方法，一个自然的疑问就是采用半参方法而不是参数方法去调整线性模型, 因为经常地, 如果可行的话, 参数方法会更有效, 所以自然而然会选择参数方法. 本章第§2.2节将会对这个问题进行解释.

本章主要基于文献Lin,Zeng和Zhu(2008), 剩余部分内容组织如下. 在第§2.2节对于线性回归模型的子模型, 将提出一个纠偏方法, 从而构造出一个调整的RM估计和调整的PT估计. 无论变量选择过程如何, 参数估计的相合性和有效性都可以得到. 在第§2.3节, 采用调整的估计函数结合经验似然用于构造置信域. 第§2.4节给出了一些数值模拟结果, 以表明调整方法的计算效果并与传统方法进行比较. 第§2.5节是结束语. 第§2.6节提供了一些定理的证明过程.

§2.2 线性子模型的相合估计

本节将首先定义一个工作子模型, 然后将其调整为部分无偏. 基于这个调整后的新模型, 引入了一个调整RM估计和一个调整PT估计. 无论全模型的参数维数高低, 都可以证明估计的相合性和有效性.

§2.2.1 模型和现有估计方法

本章中我们将考虑以下的线性回归模型

$$Y_i = \beta^T X_i + \gamma^T Z_i + \varepsilon_i,\ i = 1, \cdots, n, \tag{2.2.1}$$

其中, Y_i是响应变量Y的独立观测, X_i和Z_i分别是p-维协变量X和q-维协变量Z的独立观测, $\beta \in \mathscr{B}$和$\gamma \in \mathscr{G}$分别是p-维和q-维参数向量, $\varepsilon_1, \cdots, \varepsilon_n$假设满足

$$E(\varepsilon_i|X_i, Z_i) = 0,\ D(\varepsilon_i|X_i, Z_i) = \sigma^2$$

的随机误差ε的独立观测.

对于模型(2.2.1), 参数向量$(\beta^T, \gamma^T)^T$的最小二乘(least squares,简记为LS)估计定义为

$$\begin{pmatrix} \hat{\beta}_F \\ \hat{\gamma}_F \end{pmatrix} = \begin{pmatrix} \mathbb{X}^T\mathbb{X} & \mathbb{X}^T\mathbb{Z} \\ \mathbb{Z}^T\mathbb{X} & \mathbb{Z}^T\mathbb{Z} \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{X}^T \\ \mathbb{Z}^T \end{pmatrix} \mathbb{Y}, \tag{2.2.2}$$

其中$\mathbb{X} = (X_1, \cdots, X_n)^T$, $\mathbb{Z} = (Z_1, \cdots, Z_n)^T$, $\mathbb{Y} = (Y_1, \cdots, Y_n)^T$. 在一定的条件下, 这样一个估计是相合的和有效的. 然而, 当参数向量是高维时, 这个估计是得不到的, 因为在高维情况下(2.2.2)中的逆矩阵很容易变成奇异的或近似奇异的. 即使仅对估计参数β感兴趣, 当γ的维数q是高维时, 上述估计$\hat{\beta}_F$仍然是得不到的.

但是, 这个样本观测提供了两类信息. 首先, 为了得到相合估计, 我们必须利用变量Z的信息. 其次, 由于变量X和Z之间的关系未知, 就像不清楚变量Z如何影响Y一样, 任何参数型的建模方法将不可用. 以上这些将促使我们采用半参的建模方法. 不失一般性, 本章中假设, 经过变量选择过程后, 高维协变量Z被剔除, 这样将得到如下的一个子模型作为工作模型

$$Y_i = \beta^T X_i + \eta_i, \ i = 1, \cdots, n. \tag{2.2.3}$$

这里假定参数β的维数p不是很高. 如果仅利用低维子模型(2.2.3), 可以得到参数β的最小二乘估计

$$\hat{\beta}_S = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}. \tag{2.2.4}$$

实际上该估计也是一个RM估计, 因为它仅仅是通过限制子模型(2.2.3)导出的估计. 注意到$v_1(x) = E(Z|X = x)$是一个非零函数,

$$E(\eta|X) = E(Y - \beta^T X|X) = E(\gamma^T Z + \varepsilon|X) = \gamma^T E(Z|X). \tag{2.2.5}$$

因此, 当$\gamma \neq 0$和$E(Z|X) \neq 0$时, $v_2(x) = E(\eta|X = x)$也是一个非零函数. 这表明子模型(2.2.3)是条件有偏的, 同时, 即使$|E(\eta|X)|$是小量估计$\hat{\beta}_S$也是不相合的.

另一方面, 对于全模型(2.2.1), 假定随机误差项ε_i是正态分布的. 对于子模型(2.2.3), 参数β的一个常用PT估计定义为

$$\hat{\beta}_{PT} = \begin{cases} \hat{\beta}_S, & \text{if } \mathcal{L}_n < \chi^2_q(\alpha); \\ \hat{\beta}_F, & \text{if } \mathcal{L}_n \geq \chi^2_q(\alpha). \end{cases} \tag{2.2.6}$$

其中, $\mathcal{L}_n$是用于检验$H_0 : \gamma = 0$的profile对数似然比检验统计量, $\chi^2_q(\alpha)$是自由度为q的χ^2-分布的上分位点$100\alpha\%$, $0 < \alpha < 1$. 尽管相应的概率小, 但是这个估计仍然是有偏的; 下一节中将给出一个详细的讨论.

进一步, 利用现有的非参方法难以将估计$\hat{\beta}_S$和$\hat{\beta}_{PT}$调整为相合, 因为全模型(2.2.1)和子模型(2.2.3)涉及到多维的变量和参数; 参见Hjort和Glad(1995), Hjort和Jones(1996), Naito(2004), Glad (1998), Lin,Cui和Zhu(2008)等等.

§2.2.2 调整模型

当协变量X和Z不相关时, 若协变量Z为中心化设计, 即$E(Z) = 0$, 从(2.2.5)可以发现子模型(2.2.3)已经是无偏的. 中心化条件是一个普通条件, 如果Z不是中心化的可以用$Z - E(Z)$代替Z. 在这种情况下, 常见的最小二乘估计(LS)$\hat{\beta}_S$是相合的, 尽管它的方差很大. 以下我们假定协变量X和Z是相关的.

对子模型(2.2.3), 定义如下的纠偏形式

$$Y_i = \beta^T X_i + g(\tau^T Z_i) + \xi(\tau^T Z_i), \ i = 1, \cdots, n, \tag{2.2.7}$$

其中, 方向τ给定而且$\tau \in \mathscr{T} = \{\tau \in R^q : \|\tau\| = 1, \Sigma(\tau) > 0\}$,

$$\Sigma(\tau) = E\{(X - E(X|\tau^T Z))(X - E(X|\tau^T Z))^T\},$$
$$g(\tau^T Z) = E(\eta|\tau^T Z),\ \xi(\tau^T Z) = \eta - g(\tau^T Z),\ \eta = Y - \beta^T X.$$

这里$\Sigma(\tau) > 0$意指$\Sigma(\tau)$是一个正定矩阵. 为了实现上述模型(2.2.7), 需要说明如下.

(a) 模型(2.2.7)实际是一个特殊的部分线性单指标模型. 根据文献Härdle等(2000), Chen (1988), Härdle等(1993), Zhu和Xue(2006), 条件$\tau \in \mathscr{T}$用以保证模型(2.2.7)是可识别的. 在条件$\tau \in \mathscr{T}$下, 将能证明下面构造的新估计的相合性与τ的选取无关, 而且它的有效性对τ的选取不敏感. 当然τ的一些分量可被设计为零以降低τ的维数. 以下可把模型(2.2.7)视作一个部分线性模型, 不需要去估计τ, 而是仅需要找到τ的一些合适的经验值. 详见第2.2.3节和第2.4节.

(b) 在模型(2.2.7)中, 误差项$\xi(\tau^T Z, X)$是部分条件无偏的, 即满足

$$E(\xi(\tau^T Z)|\tau^T Z, X) = 0,\ \ \forall \tau \in \mathscr{T}, \forall (X, Z) \in \mathscr{W} \subset (\mathscr{X}, \mathscr{Z}), \tag{2.2.8}$$

其中, 样本子空间$\mathscr{W}$定义为

$$\mathscr{W} = \Big\{(X, Z) \in (\mathscr{X}, \mathscr{Z}) : E(Z|X, \tau^T Z) = E(Z|\tau^T Z), \tau \in \mathscr{T}\Big\}.$$

这个无偏性可以保证相应的β估计的相合性. 但是, 这个部分条件无偏性与全模型(2.2.1)的无偏性有所差别, 在全模型中误差项是全局条件无偏的, 即$E(\varepsilon|X, Z) = 0, \forall\ X, Z$.

这样定义调整模型(2.2.7)后, 将高维协变量Z和参数γ用一个1-维非参函数$g(\tau^T Z)$替代就可以避免维数祸根问题(curse of dimensionality), 其中τ不需要估计而且它的取值对β的新估计的主要性质影响不大.

对于调整模型(2.2.7), 给定β, 可以通过非参方法来估计$g(\tau^T Z)$. 为了表示方便, 对$g(\tau^T Z)$采用核估计如下

$$\hat{g}(\tau^T Z) = \frac{\sum_{j=1}^n (Y_j - \beta^T X_j) K\{\tau^T (Z_j - Z)/h\}}{\sum_{j=1}^n K\{\tau^T (Z_j - Z)/h\}},$$

其中$K(\cdot)$是核函数, h是依赖于n的窗宽. 这样, 调整模型(2.2.7)可被近似表示为

$$Y_i \approx \beta^T X_i + \hat{g}(\tau^T Z_i) + \xi(\tau^T Z_i),\ i = 1, \cdots, n.$$

上述模型可以等价地表示为

$$\tilde{Y}_i \approx \beta^T \tilde{X}_i + \xi(\tau^T Z_i),\ i = 1, \cdots, n, \tag{2.2.9}$$

其中

$$\tilde{Y}_i = Y_i - \frac{\sum_{j=1}^{n} Y_j K\{\tau^T(Z_j - Z_i)/h\}}{\sum_{j=1}^{n} K\{\tau^T(Z_j - Z_i)/h\}},\ \tilde{X}_i = X_i - \frac{\sum_{j=1}^{n} X_j K\{\tau^T(Z_j - Z_i)/h\}}{\sum_{j=1}^{n} K\{\tau^T(Z_j - Z_i)/h\}}.$$

从形式上看, 分别用$(\tilde{X}_i, \tilde{Y}_i)$和$\xi(\tau^T Z_i)$ 替换子模型(2.2.3)中的变量(X_i, Y_i)和误差项η_i就可以得到新的模型(2.2.9). 称模型(2.2.9)是子模型(2.2.3)的一个调整, 而且将其作为一个可实现的工作模型.

§2.2.3 调整估计

通过调整模型(2.2.9) 可以得到参数β的调整最小二乘估计如下

$$\tilde{\beta}_A = S_n^{-1} \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i \tilde{Y}_i, \tag{2.2.10}$$

其中, $S_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{X}_i \tilde{X}_i^T$. 这个估计可看作一个RM估计, 因为它主要依赖于(2.2.9)中的数据. 作为一个特殊的部分线性模型, 假设调整模型(2.2.7)满足如下的正则条件:

(C2.1) $\sup_{\tau \in \mathscr{T}, Z \in \mathscr{Z}} E(\|X\|^3 \mid \tau^T Z) < \infty$.

(C2.2) $g(\cdot)$的一阶导数和二阶导数是一阶Lipschitz连续的.

进一步, 假设核函数和窗宽满足如下条件:

(C2.3) $\int K(u)du = 1$, $\int uK(u)du = 0$, $h = O(n^{-1/5})$.

类似于部分线性模型的结果(见Härdle等(2000)), 以下定理成立.

定理2.1 对于全模型(2.2.1), 若满足条件(C2.1)-(C2.3), 同时满足以下条件:

(C2.4) $\xi(\tau^T Z_i)$为同方差的, 而且方差为$\sigma^2(\tau)$.

则, $\forall\ (X, Z) \in \mathscr{W}$,

$$\sqrt{n}\,(\tilde{\beta}_A - \beta) \longrightarrow_d N\left(0, \sigma^2(\tau)\Sigma^{-1}(\tau)\right),$$

其中, "$\longrightarrow_d$"表示依分布收敛.

这个定理表明, 对于任意的$(X, Z) \in \mathscr{W}$, 调整估计$\tilde{\beta}_A$是$\sqrt{n}$-相合的, 而且调整估计的相合性与γ和τ无关, 也就是说, 相合性不依赖于变量选择和方向τ的选取, 即使大量的协变量或一些重要变量在变量选择时被剔除. 进一步, 给定方向τ, 对于调整模型(2.2.7), 估计$\tilde{\beta}_A$是渐近半参有效的; 见文献Chamberlain(1992). 而在全模型(2.2.1)下, 估计的效依赖于方向τ.

下面讨论如何选择方向τ. 从$\Sigma(\tau)$的结构可以发现, 如果X和Z之间的相关性弱, 则$\Sigma(\tau)$对τ的选取不敏感. 这样, 在选取时可仅关注$\sigma^2(\tau)$. 在实际中可对τ进行经验选取. 由于γ是一个高维参数向量, 其部分分量经常被估计为0, 这样τ的一些分量可被选取为0. 如果要从理论上解决方向τ的选取问题, 可参考文献Lin,Zhu和Gai(2012), 此处不再进行描述. 第2.5节的模拟结果表明, 估计的效果对τ的选取不敏感. 这意味着估计对τ的选取和变量选择过程是稳健的.

另外, 对估计的渐近正态性, 条件(C2.4)中的等方差不是必需的. 对于$\xi(\tau^T Z_i)$异方差的情形, 记$\xi(\tau^T Z_i)$的方差为$\sigma_i^2(\tau)$. 基于(2.2.10), 构造加权的最小二乘估计如下

$$\tilde{\beta}_A(W) = S_n^{-1}(W)\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2(\tau)}\tilde{X}_i\tilde{Y}_i, \tag{2.2.11}$$

其中$S_n(W) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2(\tau)}\tilde{X}_i\tilde{X}_i^T$. 这时，以下定理成立.

定理2.2 对于全模型(2.2.1), 若满足条件(C2.1)-(C2.3), 同时满足以下条件:

(C2.5) $\sigma_i^2(\tau)$是已知函数, 存在常数C_1和C_2, 使得

$$0 < C_1 \le \min_{i\le n}\frac{1}{\sigma_i^2(\tau)} \le \max_{i\le n}\frac{1}{\sigma_i^2(\tau)} \le C_2 < \infty,$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2(\tau)} < \infty,$$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2(\tau)}(X_i - E(X|\tau^T Z))(X_i - E(X|\tau^T Z))^T \longrightarrow_p B(\tau), \quad (n\to\infty),$$

其中, $B(\tau)$是一个正定矩阵, "$\longrightarrow_p$"表示依概率收敛.

那么, $\forall\ (X,Z)\in\mathscr{W}$,

$$\sqrt{n}\left(\tilde{\beta}_A(W) - \beta\right) \longrightarrow_d N\left(0, B^{-1}(\tau)\Sigma(\tau)B^{-1}(\tau)\right).$$

这个定理也表明新估计$\tilde{\beta}_A(W)$的$\sqrt{n}$-相合性与γ和τ无关. 而且, 当$\sigma_i^2(\tau)$未知时, 若$\hat{\sigma}_i^2(\tau)$是$\sigma_i^2(\tau)$的一个相合估计使得

$$\max_{1\le i\le n}\left|\frac{1}{\hat{\sigma}_i^2} - \frac{1}{\sigma_i^2}\right| = o_p(n^{-\delta}),\ \delta \ge 1/4,$$

在估计(2.2.11)中σ_i^2用其相合估计$\hat{\sigma}_i^2(\tau)$替换, 定理2.2的结果仍然成立; 详见文献Härdle等(2000).

§2.2.4 调整的PT估计

在全模型(2.2.1)的条件下, 假设ε_i是正态分布的. 采用调整估计$\tilde{\beta}_A$, 定义子模型(2.2.3)中参数β的调整PT估计如下

$$\tilde{\beta}_{APT} = \begin{cases} \tilde{\beta}_A, & \text{if } \mathcal{L}_n < \chi_q^2(\alpha); \\ \hat{\beta}_F, & \text{if } \mathcal{L}_n \ge \chi_q^2(\alpha), \end{cases} \tag{2.2.12}$$

其中, $\chi_q^2(\alpha)$是自由度为q的χ^2-分布的$100\alpha\%$上分位点, $\mathcal{L}_n$是针对零假设$H_0:\gamma = 0$的profile对数似然比检验, 其定义如下

$$\mathcal{L}_n = 2\log\frac{\sup_{\beta\in\mathscr{B},\gamma\in\mathscr{G},\sigma>0} L_n(\beta,\gamma,\sigma)}{\sup_{\beta\in\mathscr{B},\gamma=0,\sigma>0} L_n(\beta,\gamma,\sigma)} \tag{2.2.13}$$

其中$L_n(\beta,\gamma,\sigma)$是似然函数. 可以验证

$$\mathcal{L}_n = n\log\frac{\hat{\sigma}_S^2}{\tilde{\sigma}_F^2},$$

其中

$$\hat{\sigma}_S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\hat{\beta}_S)^2,\ \tilde{\sigma}_F^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\hat{\beta}_F - Z_i^T\hat{\gamma}_F)^2.$$

对于γ而言其中心点取为向量0, 考虑向量0的一个收缩邻域(shrinking neighborhood), 特殊地考虑$\{K_n\}$序列, 其中

$$K_n : \gamma = \gamma_n = n^{-1/2}\xi,\ \ \xi \in R^q.$$

记$G_m(t;\mu,\Sigma)$为一个期望向量为μ、方差矩阵为Σ的m-维多元正态分布, $H_r(t;\delta)(t>0)$为一个非中心化的自由度为r、非中心化参数为$\delta \geq 0$的χ^2-分布. 进一步, 记

$$E = \{u : \mathcal{L}_n(u) \geq \chi_q^2(\alpha)\},$$

$$D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} = \begin{pmatrix} E(XX^T) & E(XZ^T) \\ E(ZX^T) & E(ZZ^T) \end{pmatrix}^{-1},$$

$$D_{11.2} = D_{11} - D_{12}D_{22}^{-1}D_{21},$$

$$D_{22.1} = D_{22} - D_{21}D_{11}^{-1}D_{12},$$

$$\delta = \sigma^{-2}\xi^T D_{22.1}\xi.$$

则下述定理成立.

定理2.3 对于全模型(2.2.1), 如果定理2.1的条件满足, 而且误差项ε服从正态分布, 则, $\forall\ (X,Z) \in \mathscr{W}$,

$$\begin{aligned}&\lim_{n\to\infty} P\{\sqrt{n}(\tilde{\beta}_{APT} - \beta) \leq t\,|K_n\}\\ &= H_p(\chi_p^2(\alpha);\delta)G_p(t;0,\sigma^2(\tau)\Sigma^{-1}(\tau))\\ &\quad + \int_E G_p(t + D_{12}D_{22}^{-1}u;0,\sigma^2 D_{11.2})dG_q(u;0,\sigma^2 D_{22}).\end{aligned}$$

对于异方差情形, 可以得到类似的结果. 采用上述定理可以对调整PT估计$\tilde{\beta}_{APT}$和(2.2.6)定义的初始PT估计$\hat{\beta}_{PT}$进行比较. 根据文献Sen等(1987), $\hat{\beta}_{PT}$满足

$$\begin{aligned}&\lim_{n\to\infty} P\{\sqrt{n}(\hat{\beta}_{PT} - \beta) \leq t\,|K_n\}\\ &= H_p(\chi_p^2(\alpha);\delta)G_p(t - (E(XX^T))^{-1}E(XZ^T)\xi;0,\sigma^2(E(XX^T))^{-1})\\ &\quad + \int_E G_p(t + D_{12}D_{22}^{-1}u;0,\sigma^2 D_{11.2})dG_q(u;0,\sigma^2 D_{22}).\end{aligned}$$

这样, 大的$\|\xi\|$将可能导致$\hat{\beta}_{PT}$有大的渐近偏, 尽管这样的概率很小. 然而, 定理2.3表明, 新的调整PT估计的渐近偏与ξ无关. 从而, 在相合性和变量选择的稳健性方面, 调整PT估计比初始PT估计有明显改进. 进一步, 如果$D_{12}D_{22}^{-1}=0$, 新的调整PT估计是$\sqrt{n}$-相合的, 而且与$\|\xi\|$的大小无关.

§2.3 线性子模型的置信域

本节将构造子模型(2.2.3)中参数向量β的置信域. 虽然可以通过定理2.1或定理2.2来构造, 但是构造过程不方便, 因为需要首先估计渐近协方差矩阵. 特别地, 如果仅用子模型(2.2.3)中的数据, 渐近协方差矩阵是不可估计的. 下面将利用经验似然方法(empirical likelihood)来构造置信域, 见文献Owen(1988, 1990).

基于调整的部分无偏模型(2.2.9), 能够得到估计函数$Z_{ni}(\beta)=\tilde{X}_i(\tilde{Y}_i-\beta^T\tilde{X}_i)$. 由于$E(Z_{ni}(\beta)|X_j,Z_j,j\geq 1)=o_p(1)$, 定义对数似然比如下

$$L(\beta)=-2\sup\Big\{\sum_{i=1}^{n}\log(np_i):p_i\geq 0,\sum_{i=1}^{n}p_i=1,\sum_{i=1}^{n}p_iZ_{ni}(\beta)=0\Big\}.$$

利用拉格朗日乘子(Lagrange multipliers), 上述经验似然比可改写为

$$L(\beta)=2\sum_{i=1}^{n}\log\{1+\lambda^T(\beta)Z_{ni}(\beta)\},$$

其中, 拉格朗日乘子$\lambda(\beta)$由下式决定

$$\frac{1}{n}\sum_{i=1}^{n}\frac{Z_{ni}(\beta)}{1+\lambda^T(\beta)Z_{ni}(\beta)}=0. \tag{2.3.1}$$

定理2.4 如果定理2.1或定理2.2的条件满足, 则对于子模型(2.2.3), 有

$$(X,Z)\in\mathscr{W},L(\beta)\longrightarrow_d\chi^2(p).$$

利用定理2.4, 参数β的一个近似置信域可按如下构造

$$\{\beta:L(\beta)\leq\chi^2_\alpha(p)\}. \tag{2.3.2}$$

无论变量选择方法和方向τ如何选取, 这个置信域估计总是可以获得的. 尽管上述的得分函数定义依赖于τ, 根据文献Kitamura, Tripathi和Ahn(2004), 用于构造经验似然的得分不需要选为最优. 但是, 类似于文献Owen(1991),有

$$L(\beta)=(\tilde{\beta}_A-\beta)^T\Big(S_n^{-1}\frac{1}{n}\sum_{i=1}^{n}(\tilde{Y}_i-\beta^T\tilde{X}_i)^2\tilde{X}_i\tilde{X}_i^TS_n^{-1}\Big)^{-1}(\tilde{\beta}_A-\beta)+O_p(n^{-1}).$$

这样, 置信域(2.3.2)的准确性依赖于τ. 然而置信域的准确性对τ的选取不敏感. 这一点在下一节的模拟结果中将会得到验证. 同样地, 参数β的经验似然估计$\tilde{\beta}_E$依赖于得分函数, 从而依赖于τ; 更准确地, 类似于文献Qin和Lawless(1994), 有

$$\sqrt{n}(\tilde{\beta}_E - \beta) \longrightarrow_d N(0, \sigma^2(\tau)\Sigma^{-1}(\tau)).$$

但是, 像上面所述, 其渐近方差$\sigma^2(\tau)\Sigma^{-1}(\tau)$对$\tau$的选取不敏感.

§2.4 模拟研究

本节通过模拟来展现新方法的估计效果, 将调整RM估计$\tilde{\beta}_A$与初始RM估计$\hat{\beta}_S$和全模型估计$\hat{\beta}_F$进行比较. 而且, 还比较了基于纠偏模型(2.2.7)、全模型(2.2.1) 和有偏子模型(2.2.3)的置信域. 对点估计, 将比较其估计的点图、平均值和最大最小值. 对置信域, 将按分量比较其置信区间长度和覆盖率. 这些模拟结果和第2节中的理论结果是相吻合的.

§2.4.1 点估计

在多维线性模型(2.2.1)中, β和γ分别按如下取7-维和20-维向量:

$$\beta = (1.0, -2.5, 1.5, -0.5, 1.2, 2.0, -1.5)^T$$
$$\gamma = (\gamma_I^T, \gamma_{II}^T)^T,$$

其中, γ_I和γ_{II}是按如下取值的10-维向量

$$\gamma_I = (0.2, 0.1, 0.15, 0.01, 0.02, 0.1, 0.25, 0.16, 0.15, 0.3)^T,$$
$$\gamma_{II} = (0.2, 0.1, 0.15, 0.2, 0.16, 0.1, 0.02, 0.3, 0.15, 0.01)^T.$$

协变量X服从正态分布$N_7(0, \mathbf{I})$, Z的前10个分量(记为Z_I)服从正态分布$N_{10}(2.0, 0.5^2\mathbf{I})$, Z的后10个分量(记为Z_{II})服从正态分布$N_{10}(1.0, 0.8^2\mathbf{I})$, 误差项$\varepsilon$服从正态分布$N(0, 0.2^2)$. 核函数选为$K(u) = \frac{3}{4}(1-u^2)I_{(-1,1)}(u)$. 虽然基于数据驱动的方法可以用于选取窗宽$h$, 但是调整估计对$h$的选取不敏感. 为了简化起见, 窗宽取为$h = n^{-1/5}$. 这里我们用不同的经验方法来选取$\tau$以验证调整估计对$\tau$的选取是稳健的. 分别记$X(j)$和$Z(j)$分别表示$X$和$Z$的第$j$个分量. 对于点估计情形, 不同样本容量和重复估计次数对模拟结果影响不明显, 因此不妨取样本容量$n = 200$、重复估计次数$m = 50$.

情形1: X和Z不相关. 取$\tau = \mathbf{1}/\|\mathbf{1}\|$, 其中$\mathbf{1}$表示所有分量都为1的向量.

图2.1是对应的模拟结果. 通过图2.1发现$\tilde{\beta}_A$和$\hat{\beta}_F$都有好的估计效果, 虽然$\hat{\beta}_F$比$\tilde{\beta}_A$较好一点. 注意到, 当观测数据不相关时, 基于全模型的估计$\hat{\beta}_F$仍然有好的表现, 但是后面的模拟结果将表明当数据相关时基于全模型的估计将会崩溃. 另一方面, 图2.1表明

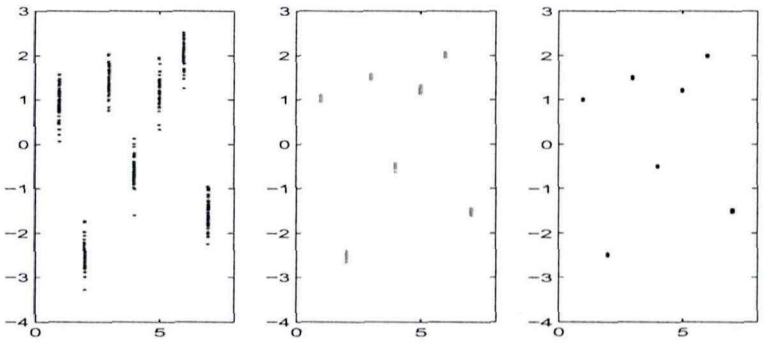

图 2.1: 情形1的点估计模拟结果, 左边图表示$\hat{\beta}_S$, 中间图表示$\tilde{\beta}_A$, 右边图表示$\hat{\beta}_F$.

初始RM估计$\hat{\beta}_S$是有偏的和分散的. 根据第2.2节, 当数据不相关而且Z中心化时, $\hat{\beta}_S$具有好的估计效果. 但是这里Z不是中心化的, 所以$\hat{\beta}_S$的表现较差.

情形2: X和Z先按情形1中设计, 然后取分量$Z^{(12)} = X^{(2)} + 5.0 \times 10^{-5}$. 取$\tau = \mathbf{1}/\|\mathbf{1}\|$.

图2.2是对应的模拟结果. 在这个情形下, 调整估计$\tilde{\beta}_A$仍然效果较好. 与情形1一样, $\hat{\beta}_S$是有偏的和分散的. 然而, 基于全模型的估计$\hat{\beta}_F$的第二个分量没有有效值, 也就是说它的估计值超出了图2.2的表示范围. 实际上, 对于第二个分量的估计值, 其$Max = 8433.3, Min = -9556.9, Avg = 377.7712$, 远远偏离了真实值$-2.5$. 这样与情形1相比, 当观测数据是部分相关时, 基于全模型的估计$\hat{\beta}_F$将变坏.

情形3: X和Z先按情形1中设计, 然后取分量$Z^{(12)} = 1.01 \times X^{(2)}, Z^{(13)} = X^{(3)} + 5.0 \times 10^{-5}$. 方向$\tau$按情形1中选取.

图2.3是对应的模拟结果. 与情形1和情形2一样, $\hat{\beta}_S$是有偏的和分散的, 而$\tilde{\beta}_A$总是表现比较好. 进一步, 估计$\hat{\beta}_F$的第2个分量和第3个分量没有返回有效值. 实际上, 对于$\hat{\beta}_F$第二个分量的估计值, 其$Max = 280.4017, Min = -1010.124, Avg = -28.9834$, 偏离真实值$-2.5$较多; 对于$\hat{\beta}_F$第三个分量的估计值, 其$Max = 12978.41, Min = -12955.42$,

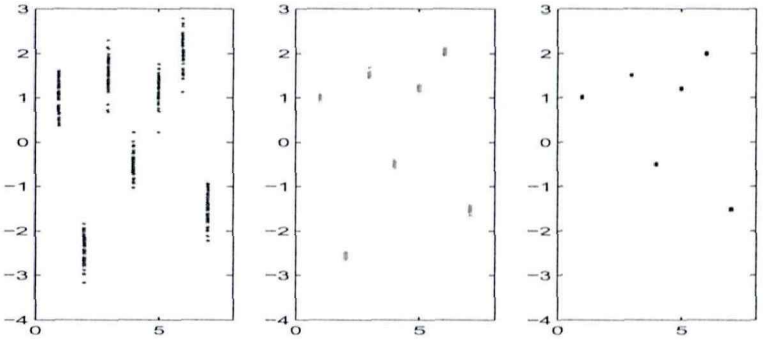

图 2.2: 情形2的点估计模拟结果, 左边图表示$\hat{\beta}_S$, 中间图表示$\tilde{\beta}_A$, 右边图表示$\hat{\beta}_F$.

$Avg = 809.7536$, 远远偏离了真实值1.5.

情形4: X和Z先按情形1中设计, 然后取分量$Z^{(12)} = 1.01 \times X^{(2)}, Z^{(13)} = X^{(3)} + 5.0 \times 10^{-5}$. τ的前10个分量和后10个分量分别取为$1/\sqrt{10}$和0.

图2.4是对应的模拟结果. 与情形3一样, $\tilde{\beta}_A$表现较好, 但是$\tilde{\beta}_S$和$\hat{\beta}_F$表现很差. 与情形3相比, 虽然τ差别较大, 但是估计$\tilde{\beta}_A$表现很稳定, 对τ的选取不敏感.

总而言之, 调整的RM估计$\tilde{\beta}_A$对变量选择、方向选择和协变量之间的相关性是稳健的.

§2.4.2 置信域

对置信域估计, 取$\alpha = 0.1$. 如果没有特殊说明, 模型和协变量按照§2.4.1节的点估计中设计. 尽管感兴趣的参数向量β是7-维的, 这里仅列出前5个分量的模拟结果.

情形1: X和Z不相关. 取$\tau = \mathbf{1}/\|\mathbf{1}\|$.

表2.1和表2.2列举了不同样本容量和不同重复估计次数下的覆盖率情况. 尽管有一点差别, 但总的来说基于子模型和纠偏模型的置信域的覆盖率接近0.9. 随着样本容量的增大, 例如$n = 500$, 这个差别基本上消失. 表2.3表明, 基于纠偏模型的置信域明显地比基于子模型的置信域小. 另外, 基于全模型的置信域覆盖率太小所以不可信. 因此本文提出的纠偏方法能够明显地改善置信域.

情形2: X和Z的分量相关$Z^{(2)} = X^{(2)} + e$, $e \sim N(0, 10^{-6})$.

表2.4和表2.5列举了取不同τ时的覆盖率结果. 其结果与X和Z不相关时的表现类似, 而且对τ的不同取值不敏感. 表2.6和表2.7列举了取不同τ时的置信域结果. 与变量不相关时的情形不同, 当数据相关时基于全模型的置信域效果较差; 例如, 分量β_2的置信区间边界远离于真实值-2.5. 但是基于纠偏模型的置信域对变量之间的相关性是稳健的. 因此本文提出的纠偏方法能够改善置信域.

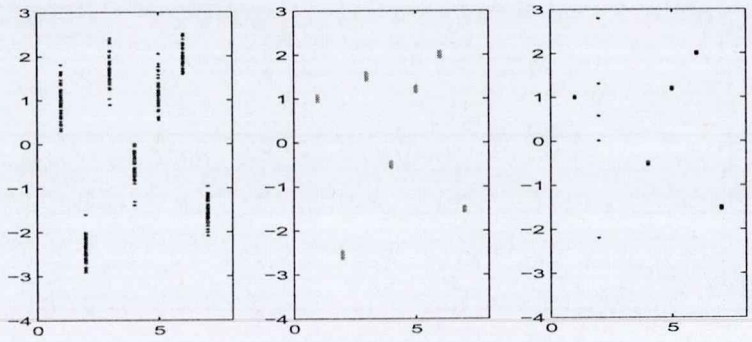

图 2.3: 情形3的点估计模拟结果, 左边图表示$\hat{\beta}_S$, 中间图表示$\tilde{\beta}_A$, 右边图表示$\hat{\beta}_F$.

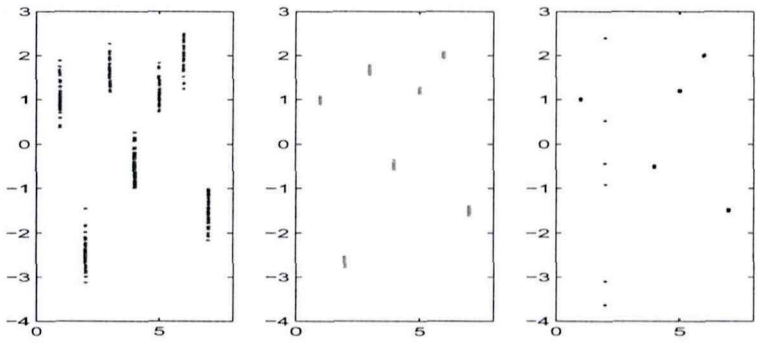

图 2.4: 情形4的点估计模拟结果, 左边图表示$\hat{\beta}_S$, 中间图表示$\tilde{\beta}_A$, 右边图表示$\hat{\beta}_F$.

表 2.1: 样本容量$n = 200$、重复估计次数$m = 200$时参数β各分量的置信区间覆盖率

elements of β	sub-model	bias-corrected model	full model
β_1	0.900	0.825	0.555
β_2	0.895	0.825	0.535
β_3	0.855	0.825	0.550
β_4	0.880	0.835	0.530
β_5	0.925	0.865	0.525

表 2.2: 样本容量$n = 500$、重复估计次数$m = 100$时参数β各分量的置信区间覆盖率

elements of β	sub-model	bias-corrected model	full model
β_1	0.87	0.93	0.80
β_2	0.86	0.86	0.74
β_3	0.93	0.89	0.85
β_4	0.86	0.88	0.84
β_5	0.92	0.90	0.79

§2.5 结论和讨论

在前面各节中引入了一种针对子模型的纠偏模型, 即使当子模型的偏相当大时该模型是部分无偏的, 证明了由该纠偏模型导出的参数估计是$\sqrt{n}$相合的. 进一步, 采用经验似然方法并结合调整后的估计函数, 构造出了参数向量的有效置信域. 与调整估计的表现类似, 置信域的精确度与γ无关, 因此对变量选择也是稳健的. 新方法的本质思想是对于变量选择过程中剔除掉的高维部分采用一个一维的非参函数进行补偿. 从而避免了维数祸根问题, 相应的调整模型是部分无偏的, 相应的推断对方向参数τ的选

表 2.3: 样本容量$n = 200$、重复估计次数$m = 50$时参数β各分量的置信区间

elements of β	sub-model	bias-corrected model	full model
β_1	[0.5148,1.4836]	[0.9651,1.0381]	[0.9777,1.0164]
β_2	[-2.9406,-1.9646]	[-2.5309,-2.4588]	[-2.5147,-2.4761]
β_3	[1.0074,1.9753]	[1.4625,1.5345]	[1.4815,1.5195]
β_4	[-0.9538,0.0167]	[-0.5357,-0.4633]	[-0.5147,-0.4758]
β_5	[0.7558,1.7303]	[1.1634,1.2352]	[1.1805,1.2186]

表 2.4: 样本容量$n = 200$、重复估计次数$m = 50$、$\tau_1 = \tau_2 = \tau_{12} = \tau_{16} = 1/\sqrt{4}$ 其余τ分量为0时参数β各分量的置信区间覆盖率

elements of β	sub-model	bias-corrected model	full model
β_1	0.875	0.855	0.550
β_2	0.940	0.865	0.475
β_3	0.875	0.860	0.500
β_4	0.935	0.880	0.525
β_5	0.895	0.830	0.510

表 2.5: 样本容量$n = 200$、重复估计次数$m = 50$、$\tau_1 = 1/\sqrt{1.24}, \tau_2 = \tau_{12} = 0.5/\sqrt{1.24}, \tau_{16} = 0.8/\sqrt{1.24}$ 其余τ分量为0时参数β各分量的置信区间覆盖率

elements of β	sub-model	bias-corrected model	full model
β_1	0.905	0.820	0.540
β_2	0.845	0.810	0.495
β_3	0.920	0.840	0.520
β_4	0.845	0.835	0.525
β_5	0.865	0.855	0.555

取也是稳健的. 本章的调整方法不仅用于线性模型, 还可应用于其它模型如部分线性模型、非线性模型等, Gai等(2011)已研究了部分线性模型情形, 非线性模型情形需要进一步研究. 另外, 本章的相合性结果仅在协变量的一个子集区域上成立, 这也有待于参考Lin,Zhu和Gai(2012)进一步进行研究.

表 2.6: 样本容量$n=200$、重复估计次数$m=50$、$\tau_1=\tau_2=\tau_{12}=\tau_{16}=1/\sqrt{4}$ 其余τ分量为0时参数β各分量的置信区间

elements of β	sub-model	bias-corrected model	full model
β_1	[0.4772,1.4105]	[0.9431,1.0583]	[0.9832,1.0220]
β_2	[-2.8286,-1.8909]	[-2.5620,-2.4133]	[-5.4867,-5.4142]
β_3	[1.0021,1.9420]	[1.4440,1.5585]	[1.4838,1.5230]
β_4	[-0.9612, -0.0196]	[-0.5655,-0.4485]	[-0.5216,-0.4824]
β_5	[0.7808,1.7143]	[1.1402,1.2556]	[1.1820,1.2213]

表 2.7: 样本容量$n=200$、重复估计次数$m=50$、$\tau_1=1/\sqrt{1.24}, \tau_2=\tau_{12}=0.5/\sqrt{1.24}, \tau_{16}=0.8/\sqrt{1.24}$ 其余τ分量为0时参数β各分量的置信区间

elements of β	sub-model	bias-corrected model	full model
β_1	[0.4817,1.4121]	[0.9413,1.0566]	[0.9839,1.0227]
β_2	[-2.8058,-1.8682]	[-2.6290,-2.4452]	[-6.2070,-6.1344]
β_3	[0.9900,1.9337]	[1.4430,1.5582]	[1.4828,1.5221]
β_4	[-0.9731, -0.0341]	[-0.5646,-0.4469]	[-0.5211,-0.4819]
β_5	[0.7903,1.7271]	[1.1404,1.2572]	[1.1813,1.2208]

§2.6 附录

§2.6.1 定理2.1和定理2.2的证明

令$\tau^T Z=U$. 在这种情况下模型(2.2.7)是一个标准的部分线性回归模型. 这时可以直接利用现有的结果得到定理2.1和定理2.2, 例如, 文献Härdle等(2000)中的定理2.11和定理2.12. □

§2.6.2 定理2.3的证明

根据现有文献, 如Sen等(1987), 有

$$\begin{aligned}
&\lim_{n\to\infty} P\{\sqrt{n}(\hat{\beta}_{PT}-\beta)\le t\,|K_n\}\\
&=\lim_{n\to\infty} P\{\sqrt{n}(\hat{\beta}_S-\beta)\le t\,|K_n, \mathcal{L}_n(u)<\chi^2_q(\alpha)\}\\
&\quad+\lim_{n\to\infty} P\{\sqrt{n}(\hat{\beta}_F-\beta)\le t\,|K_n, \mathcal{L}_n(u)\ge\chi^2_q(\alpha)\}\\
&=H_p(\chi^2_p(\alpha);\delta)G_p(t-(E(XX^T))^{-1}E(XZ^T)\xi;0,\sigma^2(E(XX^T))^{-1})\\
&\quad+\int_E G_p(t+D_{12}D_{22}^{-1}u;0,\sigma^2D_{11.2})dG_q(u;0,\sigma^2D_{22}).
\end{aligned}$$

这样, 只需要考虑$\mathcal{L}_n(u) < \chi_q^2(\alpha)$的情形. 进而, 由定理2.1可得到

$$U_n = \Gamma(\tau)\sqrt{n}(\tilde{\beta}_A - \beta) \longrightarrow_d N(0, \sigma^2(E(XX^T))^{-1}),$$

其中$\Gamma(\tau) = \sigma\sigma^{-1}(\tau)E^{-1/2}(XX^T)\Sigma^{1/2}(\tau)$. 另一方面, 容易证明

$$V_n = \sqrt{n}(\hat{\beta}_S - \beta) - (E(XX^T))^{-1}E(XZ^T)\xi \longrightarrow_d N(0, \sigma^2(E(XX^T))^{-1}).$$

上述结果表明U_n和V_n同分布的. 结合以上结果, 得到

$$\begin{aligned}
&\lim_{n\to\infty} P\{\sqrt{n}(\tilde{\beta}_A - \beta) \le t \,|K_n, \mathcal{L}_n(u) < \chi_q^2(\alpha)\} \\
&= \lim_{n\to\infty} P\{U_n \le \Gamma(\tau)t \,|K_n, \mathcal{L}_n(u) < \chi_q^2(\alpha)\} \\
&= \lim_{n\to\infty} P\{V_n \le \Gamma(\tau)t \,|K_n, \mathcal{L}_n(u) < \chi_q^2(\alpha)\} \\
&= \lim_{n\to\infty} P\{\sqrt{n}(\hat{\beta}_S - \beta) \le (E(XX^T))^{-1}E(XZ^T)\xi + \Gamma(\tau)t \,|K_n, \mathcal{L}_n(u) < \chi_q^2(\alpha)\} \\
&= H_p(\chi_p^2(\alpha); \delta)G_p(\Gamma(\tau)t; 0, \sigma^2(E(XX^T))^{-1}) \\
&= H_p(\chi_p^2(\alpha); \delta)G_p(t; 0, \sigma^2(\tau)\Sigma^{-1}(\tau)),
\end{aligned}$$

即证. □

§2.6.3 定理2.4的证明

与定理2.1和定理2.2的证明过程类似, 令$\tau^T Z = U$, 则模型(2.2.7)是一个标准的部分线性回归模型. 定理2.4可以直接由现有文献结果得到, 例如, Wang和Jing(2003), Shi和Lau(2000). □

第三章　多维线性回归有偏子模型的多步调整相合推断

§3.1　引言

将回归分析方法应用于实际问题时,在最初建模阶段会引入大量的相关变量,从而形成全模型. 尽管全模型很自然,而且模型的偏很小甚至无偏,但是它将需要很大的计算量, 同时由于全模型的复杂性其参数估计和模型预测的精度将降低. 因此需要简化全模型、提高模型预测精度.在现有方法中变量选择是典型的方法. 对于线性回归模型,有很多变量选择的方法,可以分为传统方法和基于惩罚的方法两类. 传统方法易于实现而且也很实用, 如逐步回归、最优子集选择、AIC和BIC信息准则方法, 见Miller(2002). 近年来比较流行的是基于惩罚的方法, 包括Tibshirani(1996)的LASSO方法、Fan和Li(2001)的SCAD方法及Candès和Tao(2007)的Dantzig方法,等等.

在很多情形,下面的多维线性回归模型特别与实际应用相关

$$Y = \beta^T X + \gamma^T Z + \varepsilon, \tag{3.1.1}$$

其中Y是1维的响应变量, X和Z分别是p维和q维取值连续的协变量, 对应的参数为β和γ, ε是随机误差. 在该模型中, 我们主要对参数β感兴趣. 例如,在临床试验中X由治疗方法组成而且易于控制,Z代表许多临床因素如病人的年龄和体重. Z中的变量可能对Y有影响, 但不是主要感兴趣的. 为了弥补对Y有潜在的不可忽略的影响,在模型(3.1.1)中引入了讨厌变量Z, 见Shen等(2004).

我们的主要目标是估计参数β. 称模型(3.1.1) 为全模型, 它包括所有相关的变量而且是无偏的模型, 即$E(\varepsilon|X,Z) = 0$. 基于上述全模型,给定(Y,X,Z)的一组独立观测$(Y_i,X_i,Z_i), i = 1,\cdots,n$. 参数向量$(\beta^T,\gamma^T)^T$的最小二乘估计定义为

$$\begin{pmatrix} \hat{\beta}_F \\ \hat{\gamma}_F \end{pmatrix} = \arg\min_{\beta,\gamma} \sum_{i=1}^{n} (Y_i - \beta^T X_i - \gamma^T Z_i)^2 = \begin{pmatrix} \mathbb{X}^T\mathbb{X} & \mathbb{X}^T\mathbb{Z} \\ \mathbb{Z}^T\mathbb{X} & \mathbb{Z}^T\mathbb{Z} \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{X}^T \\ \mathbb{Z}^T \end{pmatrix} \mathbb{Y}, \tag{3.1.2}$$

其中$\mathbb{X} = (X_1, X_2, \cdots, X_n)^T, \mathbb{Z} = (Z_1, Z_2, \cdots, Z_n)^T, \mathbb{Y} = (Y_1, Y_2, \cdots, Y_n)^T$. 由于全模型(3.1.1)是无偏的, 所以在一定的条件下估计(3.1.2)是相合的. 但是当全模型参数向量的维数很高时, 估计(3.1.2)存在很多计算方面的问题, 例如大的计算量和资源占用. 当设计阵$(\mathbb{X},\mathbb{Z})$为奇异的或近似奇异时,数值算法可能会崩溃并异常退出, 例如当协变量之间高度相关时. 因此, 基于全模型的传统方法的估计是不可信的.

如果仅对感兴趣的协变量X进行回归, 能够得到下面的子模型

$$Y = \beta^T X + \eta. \tag{3.1.3}$$

这样一个模型可作为一个工作模型. 变量X的分量可被提前给定, 因此这个子模型是非随机的低维模型.

基于子模型(3.1.3), 可以得到参数β的最小二乘估计如下

$$\hat{\beta}_S = \arg\min_\beta \sum_{i=1}^{n}(Y_i - \beta^T X_i)^2 = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}, \tag{3.1.4}$$

在子模型(3.1.3)中,有

$$E(\eta|X) = E(Y - \beta^T X|X) = E(\varepsilon + \gamma^T Z|X) = \gamma^T E(Z|X). \tag{3.1.5}$$

因此,当$\gamma^T E(Z|X = x) \neq 0$时$E(\eta|X = x)$是一个非零函数. 这意味着子模型(3.1.3)是条件有偏的, 所以估计$\hat{\beta}_S$可能是不相合的.

在模型中保留变量X, 然后对变量Z应用变量选择如基于惩罚的方法, 就可以得到一个选模型. 如果全模型是稀疏的, 即只有很小一部分变量的系数远离于零点而其余系数为零或近似为零, 这个变量选择的过程总是可行的. 在很多情形下选模型可能是有偏的, 相应的参数估计也可能如此. 当特别考虑参数β的估计相合性时, 基于惩罚的方法可能会面临一些难题. 当采用了一个大的惩罚参数时, 选模型一般会过小. 如果一些重要变量不幸被剔除, 选模型就是误定的而且可能是有偏的. 相应地, 当矩阵$(\mathbb{X}, \mathbb{Z})^T(\mathbb{X}, \mathbb{Z})$不是对角阵时, 基于选模型的参数的估计将是不相合的; 见Hall 和Inoue (2003), Schennach (2007). 有时当一些不重要的变量被从全模型中剔除, 其系数虽然近似为零但对响应变量仍然有影响, 这时选模型可能也是有偏的.

在很多情形下期望参数β的估计是相合估计. 例如, 在第二章中关于线性回归模型针对子模型我们提出了一个非参调整方法. 与已有的方法相比, 这个方法仅采用了一维非参估计, 却能够大幅度地减小子模型的估计偏差. 但是, 这个方法的结果仅在协变量的一个子集区域上成立, 尽管这个子集可能相当大.

本章的目标是构造子模型中参数β的相合估计. 为此, 我们构造了一个全局无偏的工作模型, 但却没有采用多维非参估计. 主要的思路是通过多步调整来减小子模型的偏差, 利用变量X和变量Z之间的相关信息, 在每一步中将对子模型增加一个一维非参调整项. 当变量Z为正态分布变量时, 将采用主成分回归方法来构造全局无偏的调整模型; 否则, 将采用独立成分分析(ICA)方法. 基于这个调整模型, 得到了一个相合的点估计, 而且证明它是渐近正态分布的. 模拟结果表明, 新的估计是稳健的, 而且它优于已有的估计方法.

本章的后续部分组织如下. 在§3.2节, 基于子模型提出了一个全局无偏的调整模型, 然后导出了一个调整的的参数估计. 随后得到了这个新的调整估计的相合性和渐近正态性. 第§3.3节给出了一些模拟结果. 第§3.4节对全文做了一个简短总结并给出了几点备注. 第§3.5节提供了一些理论结果的证明过程.

§3.2 线性子模型的相合点估计

本章中假定q是有限的, $p \ll q$, $E(\varepsilon|X,Z) = 0$, $Var(\varepsilon|X,Z) = \sigma^2$. 而且,协变量$X$和参数$\beta$提前给定. 因此, 子模型(3.1.3)是固定模型.

我们首先将子模型调整为全局条件无偏, 然后得到感兴趣参数β的一个估计. 进一步, 我们将证明这个新的估计是全局相合而且是渐近正态分布的.

§3.2.1 调整的无偏模型

根据上述,子模型(3.1.3)可能是有偏的. 下面我们将引入两种纠偏方法来构造出一个调整的无偏模型.

(i) 正态分布采用主成分回归

如果Z为中心化的正态分布变量, 则可以对Z的总体应用主成分分析(PCA)方法. 设$\Sigma_{Z,Z}$为Z的协方差阵. 存在一个$q \times q$的正交矩阵Q,使得$Q\Sigma_{Z,Z}Q^T = \Lambda$, 其中$\Lambda$是对角阵$\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_q)$, $\Sigma_{Z,Z}$的特征值满足$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q \geq 0$. 记$Q^T = (Q_1, Q_2, \cdots, Q_q)$, 其中$Q_i$是特征值$\lambda_i$对应的$q$-维特征向量, $i = 1, 2, \cdots, q$. 令$\tilde{Z} = QZ$, 则其第i个分量$\tilde{Z}^{(i)} = Q_i^T Z$是$Z$的第$i$个主成分, 满足对于$1 \leq i, j \leq q, i \neq j$,$\tilde{Z}^{(i)}$ 和$\tilde{Z}^{(j)}$是不相关的.

取$\tau_i = Q_i, i = 1, \cdots, q$, 令$M \triangleq \{i : E(\tilde{Z}^{(i)}|X) \neq 0, 1 \leq i \leq q\}$, $K = |M|$表示集合中的元素个数. 根据以上定义,有$K \leq q$. 不失一般性,我们假定$M = \{1, \cdots, K\}$.

构造如下的纠偏模型

$$Y = \beta^T X + g_1(\tilde{Z}^{(1)}) + \cdots + g_K(\tilde{Z}^{(K)}) + \zeta_K, \tag{3.2.1}$$

其中

$$\begin{aligned}
g_1(\tilde{Z}^{(1)}) &= E(Y - \beta^T X|\tilde{Z}^{(1)}) \\
&= E(\gamma^T Z + \varepsilon|\tilde{Z}^{(1)}) \\
&= E(\gamma^T Z|\tilde{Z}^{(1)}), \\
g_i(\tilde{Z}^{(i)}) &= E(Y - \beta^T X - g_1(\tilde{Z}^{(1)}) - \cdots - g_{i-1}(\tilde{Z}^{(i-1)})|\tilde{Z}^{(i)}) \\
&= E(Y - \beta^T X|\tilde{Z}^{(i)}) \\
&= E(\gamma^T Z|\tilde{Z}^{(i)}), \quad i = 2, \cdots, K,
\end{aligned}$$

$\zeta_K = Y - \beta^T X - g_1(\tilde{Z}^{(1)}) - \cdots - g_K(\tilde{Z}^{(K)})$. 模型(3.2.1)基于$Z$总体的特征向量,依赖于$X$和$Z$的分布.

实际上, 非参部分可以进一步化简为

$$\begin{aligned}
g_i(\tilde{Z}^{(i)}) &= E(\gamma^T Z|\tilde{Z}^{(i)}) \\
&= E(\gamma^T Q^T(\tau_1^T Z, \tau_2^T Z, \cdots, \tau_q^T Z)^T|\tau_i^T Z) \\
&\triangleq \alpha_i \tilde{Z}^{(i)}, \quad i = 1, \cdots, K.
\end{aligned}$$

因此，调整模型(3.2.1)实际上是一个线性模型

$$Y = \beta^T X + \alpha^T \tilde{\mathbf{Z}}_K + \zeta_K, \tag{3.2.2}$$

其中$\alpha = (\alpha_1, \cdots, \alpha_K)^T$, $\tilde{\mathbf{Z}}_K = (\tau_1^T Z, \cdots, \tau_K^T Z)^T$.

当X和Z为联合正态分布$N\left(\mathbf{0}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{pmatrix}\right)$时, $E(\tau_i^T Z|X) = \tau_i^T \Sigma_{ZX} \Sigma_{XX}^{-1} X$. 如果$\tau_i$正交于由$\Sigma_{ZX}\Sigma_{XX}^{-1}$的$p$个列向量生成的线性空间, 则有$E(\tau_i^T Z|X) = 0$. 因此, 如果多数$\tau_i$ $(i = 1, \cdots, q)$正交于这个线性空间, 则有$K \ll q$. 如果所有τ_i $(i = 1, \cdots, q)$投影到这个线性空间均为非零, 则有$K = q$.

(ii) 非正态分布采用独立成分分析

如果Z为中心化的非正态分布变量, 上面的主成分分析将不适用, 因为主成分分析得不到独立分量. 在这种情况下, 我们将采用独立成分分析(ICA)方法来处理.

1) 假设Z 由一些独立变量$\tilde{Z}^{(i)}$的线性组合生成, 即$Z = A\tilde{Z}$, 其中A 是一个未知的方阵, $\tilde{Z}$ 是一个由独立分量组成的未知向量.

已经证明, A 和$\tilde{Z}$可被估计当且仅当变量$\tilde{Z}^{(i)}$为非正态分布变量; 见Hyvarinen等(2001). 基于峰度或负熵, 当Z分布已知时通过关系式$\tilde{Z} = QZ$其中$Q = (Q_1\ Q_2\ \cdots\ Q_q)^T$, 采用FastICA 方法可以求解独立变量$\tilde{Z}^{(i)}, i = 1, \cdots, q$; 见Hyvarinen和Oja(1997). FastICA方法基于随机变量的总体, 一般没有显式解, 而是通过迭代算法实现.

按照同样的记号, 取$\tau_i = Q_i, i = 1, \cdots, q$, 令$K = |M| \triangleq \#\{i : E(\tau_i^T Z|X) \neq 0, 1 \leq i \leq q\}$表示集合中的元素个数, 简记为$M = \{1, \cdots, K\}$. 与(i)类似, 我们能够构造调整的线性模型(3.2.2).

2) 假设Z 由一些独立变量$\tilde{Z}^{(i)}$的非线性组合生成, 即$Z = F(\tilde{Z})$, 其中$F(\cdot) : R^q \to R^q$是一个未知的非线性映射, $\tilde{Z}$是由独立分量组成的未知随机变量. 近期以来, 非线性ICA算法被应用于不同领域, 例如语音处理、图像去噪, 相关文献中也研究了非线性ICA解的理论性质; 见Simas Filho和Seixas(2007).

通过对非线性映射F或者独立变量$\tilde{Z}^{(i)}$设置一定的限制, 独立变量$\tilde{Z}^{(i)}$能够被适当地估计. 关于估计非线性独立成分的主要统计准则和一些算法, 见Simas Filho和Seixas (2007). 为了技术上的方便和简化, 本章中我们假定$Z = (Z^{(1)}, \cdots, Z^{(q)})^T$, $Z^{(l)} = \Sigma_{j=1}^q F_{lj}(\tilde{Z}^{(j)}), l = 1, \cdots, q$, 其中$F_{lj}(\cdot)$为标量值函数.

采用同样的记号, 类似于(i), 可以构造调整的部分线性模型(3.2.1), 其中

$$\begin{aligned} g_i(\tilde{Z}^{(i)}) &= E(\gamma^T Z|\tilde{Z}^{(i)}) = \Sigma_{l=1}^q \gamma_l F_{li}(\tilde{Z}^{(i)}), \quad i = 1, \cdots, K \\ \zeta_K &= Y - \beta^T X - g_1(\tilde{Z}^{(1)}) - \cdots - g_K(\tilde{Z}^{(K)}). \end{aligned}$$

当X和Z服从中心点为$\mathbf{0}$、尺度矩阵为$\Sigma \equiv \begin{pmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{pmatrix}$的联合椭球对称分布、

即$\Sigma^{-\frac{1}{2}}\begin{pmatrix} X \\ Z \end{pmatrix}$的分布在正交变换下具有不变性, 而且$Z$由一些独立变量的线性组合生成时, 有$E(\tilde{Z}^{(i)}|X) = E(\tau_i^T Z|X) = \tau_i^T \Sigma_{ZX}\Sigma_{XX}^{-1}X$. 在这个一般性条件下, (i)中联合正态分布条件下关于q和K的关系仍然成立. 见Kariya和Kurata(2004)中的命题1.19.

在(i)和(ii)中, 变量$\tilde{Z}^{(i)}$之间相互独立, 它们都是总体分布的特征. 可以证明,

$$E(\zeta_K|X, \tilde{Z}^{(i)}, i = 1, \cdots, K) = 0, \quad \forall Z, X.$$

因此, 调整的部分线性模型(3.2.1)是全局条件无偏的. 其证明过程见第§3.5节.

所以, 基于子模型(3.1.3), 得到了一个调整模型(3.2.1). 它实际上是一个特殊的可加部分线性模型, 其中$\beta^T X$是参数部分, $g_i(\tilde{Z}^{(i)}), i = 1, \cdots, K$是非参部分, ζ_K是随机误差. 在Z为中心化的条件下, 有$E(g_i(\tilde{Z}^{(i)})) = 0, i = 1, \cdots, K$. 因此, 非参项$g_1(\tilde{Z}^{(1)}), \cdots, g_K(\tilde{Z}^{(K)})$能够被识别.

与子模型(3.1.3)相比, 调整模型(3.2.1)的非参部分项$g_i(\tilde{Z}^{(i)}), i = 1, \cdots, K$可看作误差项$\eta$的纠偏. 它们都是一维的非参函数$\tilde{Z}^{(i)}$. 在这种情况下, 普通最小二乘(LS)方法不能使用, 否则将导致不相合的估计. 从而, 需要寻求新的方法去构造感兴趣参数β的相合估计.

§3.2.2 相合估计和渐近性质

一般地, 可以使用一维的非参方法去估计非参项. 为简化起见, 给定参数β, 我们使用一维核估计方法来估计$g_i(\tilde{Z}^{(i)})$如下

$$\hat{g}_i(\tilde{Z}^{(i)}) = \frac{\sum_{j=1}^n (Y_j - \beta^T X_j) K((\tilde{Z}_j^{(i)} - \tilde{Z}^{(i)})/h_i)}{\sum_{j=1}^n K((\tilde{Z}_j^{(i)} - \tilde{Z}^{(i)})/h_i)}, i = 1, \cdots, K, \tag{3.2.3}$$

其中$K(\cdot)$是核函数, h_i是依赖于样本大小n和指标i的窗宽. 模型(3.2.1)可用下面的模型来近似

$$Y = \beta^T X + \hat{g}_1(\tilde{Z}^{(1)}) + \cdots + \hat{g}_K(\tilde{Z}^{(K)}) + \zeta_K, \tag{3.2.4}$$

形式上可用下面的线性模型来等价表示

$$\breve{Y}_i = \beta^T \breve{X}_i + \zeta_{Ki}, \ i = 1, \cdots, n, \tag{3.2.5}$$

其中

$$\breve{Y}_i = Y_i - \sum_{l=1}^K \frac{\sum_{j=1}^n Y_j K((\tilde{Z}_j^{(l)} - \tilde{Z}_i^{(l)})/h_l)}{\sum_{j=1}^n K((\tilde{Z}_j^{(l)} - \tilde{Z}_i^{(l)})/h_l)}, \breve{X}_i = X_i - \sum_{l=1}^K \frac{\sum_{j=1}^n X_j K((\tilde{Z}_j^{(l)} - \tilde{Z}_i^{(l)})/h_l)}{\sum_{j=1}^n K((\tilde{Z}_j^{(l)} - \tilde{Z}_i^{(l)})/h_l)}.$$

通过模型(3.2.5), 可得到参数β的一个调整的最小二乘估计

$$\hat{\beta}_A = (\frac{1}{n}\sum_{i=1}^n \breve{X}_i \breve{X}_i^T)^{-1} \frac{1}{n}\sum_{i=1}^n \breve{X}_i \breve{Y}_i, \tag{3.2.6}$$

其中假设调整模型(3.2.1)是等方差的, 即$Var(\zeta_{Ki}|X_i,\tilde{Z}_{Ki})=\sigma^2_{\zeta_K}, i=1,\cdots,n$. 如果调整模型是异方差的, 需要将上面的估计修改如下

$$\hat{\beta}_A=(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma^2_{\zeta_{Ki}}}\check{X}_i\check{X}_i^T)^{-1}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma^2_{\zeta_{Ki}}}\check{X}_i\check{Y}_i,$$

其中$\sigma^2_{\zeta_{Ki}}=Var(\zeta_{Ki}|X_i,\tilde{Z}_{Ki})$. 这里，假定每一个$\sigma^2_{\zeta_{Ki}}$都是已知的. 如果未知，采用相合估计来代替, 见Härdle等(2000).

由于模型(3.2.1)是一个特殊的可加部分线性模型, 假设其满足如下条件:

(C3.1) $EZ=0,(X_i^T,\tilde{Z}_{Ki}^T)$具有有限支撑集, $E(\zeta_K^4|X,\tilde{Z}_K)$是有限的.

(C3.2) Z的密度函数$f(Z)$在其支撑集上下有界于一个正的常数, $E(Y|\tilde{Z}^{(l)}),E(X|\tilde{Z}^{(l)}),$ $l=1,\cdots,K$以及$f(Z)$为二阶可微.

(C3.3) $K(\cdot)$为有界、对称的二阶核函数.

(C3.4) 窗宽$h_l,l=1,\cdots,K$满足$n\to\infty, nh_l^2\to\infty, nh_l^4\to 0$.

下述定理表明调整估计$\hat{\beta}_A$ 是$\sqrt{n}$-相合的而且服从渐近正态分布.

定理3.1 对于满足$K\ll q$的调整模型(3.2.1), 在条件(C3.1)-(C3.4)下, 如果二阶矩$E((X-\sum_{l=1}^{K}E(X|\tilde{Z}^{(l)}))(X-\sum_{l=1}^{K}E(X|\tilde{Z}^{(l)}))^T)$ 正定, 则

$$n^{\frac{1}{2}}(\hat{\beta}_A-\beta)\longrightarrow_d N(0,\Sigma), \tag{3.2.7}$$

其中β 是参数向量的真值, "$\longrightarrow_d$"表示依分布收敛, $\Sigma=\Phi^{-1}\Omega\Phi^{-1}$, $\Phi=E(DD^T)$, $\Omega=E(\zeta_K^2DD^T)$, $D=X-\sum_{l=1}^{K}E(X|\tilde{Z}^{(l)})$.

注3.1 调整估计$\hat{\beta}_A$是渐近无偏的. 实际上, $\mathrm{bias}(\hat{\beta}_A)=E(\hat{\beta}_A)-\beta=K*(O(h_{max}^2)+O(n^{-1}))$, 其中$h_{max}=\max_{1\le l\le K}h_l$. 在条件(C3.4)下, $\mathrm{bias}(\hat{\beta}_A)$ 趋于0. 当$K=0$时, 估计$\hat{\beta}_A$是无偏的. 当$K\ll q$时, $\hat{\beta}_A$的偏很小. K越大, 偏$\mathrm{bias}(\hat{\beta}_A)$越大.

注3.2 通过上述定理可以发现调整估计$\hat{\beta}_A$是全局相合的, 其相合性与参数γ无关, 而且对辅助向量$\tau_l,l=1,\cdots,K$不敏感.

注3.3 定理的成立条件对模型(3.2.1)中误差项ζ_K的方差没有约束. 所以定理同样适用于误差项ζ_K异方差的情形，只是需要对渐近方差Σ的形式进行替换.

注3.4 当误差项为等方差, 即$Var(\zeta_{Ki}|X_i,\tilde{Z}_i)=\sigma^2_{\zeta_K}$, 渐近协方差的逆为

$$\Sigma^{-1}=E((X-\sum_{l=1}^{K}E(X|\tilde{Z}^{(l)}))(X-\sum_{l=1}^{K}E(X|\tilde{Z}^{(l)}))^T)/\sigma^2_{\zeta_K},$$

这便是Chamberlain (1992)中导出的半参效的界.

特别地, 当X和Z为联合正态分布时, 可以直接使用线性模型(3.2.2). 这样便可以得到β和α的一个调整的最小二乘估计

$$\begin{pmatrix}\hat{\beta}_L\\ \hat{\alpha}_L\end{pmatrix}=\begin{pmatrix}\mathbb{X}^T\mathbb{X} & \mathbb{X}^T\tilde{\mathbb{Z}}_K\\ \tilde{\mathbb{Z}}_K^T\mathbb{X} & \tilde{\mathbb{Z}}_K^T\tilde{\mathbb{Z}}_K\end{pmatrix}^{-1}\begin{pmatrix}\mathbb{X}\\ \tilde{\mathbb{Z}}_K\end{pmatrix}^T\mathbb{Y} \tag{3.2.8}$$

其中, 假定模型(3.2.2)为等方差的, 即$Var(\zeta_{Ki}|X_i, \tilde{Z}_{Ki}) = \sigma^2_{\zeta_K}$, $i = 1, \cdots, n$.

命题3.2 对于满足$K \ll q$的调整线性模型(3.2.2), 如果二阶矩

$$\Sigma_L \triangleq E \begin{pmatrix} X^T X & X^T \tilde{Z}_K \\ \tilde{Z}_K^T X & \tilde{Z}_K^T \tilde{Z}_K \end{pmatrix}$$

正定, 则

$$n^{\frac{1}{2}} \left(\begin{pmatrix} \hat{\beta}_L \\ \hat{\alpha}_L \end{pmatrix} - \begin{pmatrix} \beta \\ \alpha \end{pmatrix} \right) \longrightarrow_d N(0, \sigma^2_{\zeta_K} \Sigma_L^{-1}), \tag{3.2.9}$$

其中, β和α是参数向量的真值, "$\longrightarrow_d$"表示依分布收敛.

当调整线性模型(3.2.2)为异方差时, 结果类似.

§3.2.3 调整模型的实现

在定理3.1的条件(C3.1)中, 要求协变量Z为中心化的. 就象线性回归模型的最小二乘, 这个条件可被弱化. 实际上, 如果Z不是中心化的, 记$\mu_Z = EZ$, 定义$Z^* = Z - \mu_Z$, 全模型$Y = \beta^T X + \gamma^T Z + \varepsilon$可以改写为

$$Y = \theta_0^* + \beta^T X + \gamma^T Z^* + \varepsilon, \tag{3.2.10}$$

其中$\theta_0^* = \gamma^T \mu_Z$. 这时就可以对模型(3.2.10)使用调整方法.

调整模型(3.2.1)的非参项$g_i(\tilde{Z}^{(i)})$依赖于一维变量$\tilde{Z}^{(i)}$, $i = 1, \cdots, K$. 因此首先需要选取步数K. 在实现时, 首先计算$\tilde{Z}^{(l)}(l = 1, \cdots, q)$与$X$的$q$个多重相关系数, 给定$\delta > 0$选取$R = \{\tilde{Z}^{(l)} : |\text{mcorr}(\tilde{Z}^{(l)}, X)| \geq \delta,\ l = 1, \cdots, q\}$, 其中mcorr$(u, V)$表示$u$与$V$的多重相关系数的样本形式. 取$K = |R|$, 即集合$R$中的元素个数. 给定参数$\beta$, 当用一维核估计$\hat{g}_i(\tilde{Z}^{(i)})$来估计$g_i(\tilde{Z}^{(i)})$时, 窗宽$h_i$的选择比核函数$K(\cdot)$的选择更重要. 可以采用一些流行的窗宽选择方法来选择h_i, 例如最小化MISE、leave-one-out CV或GCV方法.

在调整的线性模型(3.2.2)中, 使用了$\tau_i, i = 1, \cdots, K$. 当协变量Z的分布未知时, 需要通过估计来近似选取. 1) 当Z为正态分布时, 如果协方差阵Σ_{ZZ}的特征值$\lambda_i, i = 1, \cdots, q$互不相同, 则当$n \to \infty$时序列$\sqrt{n-1}(u_i - \tau_i)$渐近于正态分布$N(\mathbf{0}, V_i)$, 其中$V_i = \sum_{j \neq i}^{q} \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \tau_j \tau_j^T$, u_i是$S = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})(Z_i - \bar{Z})^T$的第$i$个特征向量, $\bar{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i$; 见Anderson(2003). 这样, 就可以用u_i来近似τ_i. 2) 当Z为非正态分布时, 可以用Samarov和Tsybakov(2004)或者Chen和Bickel(2005,2006)提出的方法来给出$\tau_i, i = 1, \cdots, K$的一个相合估计. 将这些估计带入模型(3.2.1), 能够发现由近似产生的误差可以忽略, 而不会影响方法的渐近性质. 关于独立成分的估计, 目前已有一些成熟的主要统计原则和算法, 见文献Simas Filho和Seixas(2007),Acharya和Panda(2008).

§3.3 数值模拟

本节将通过数值模拟表明新方法的性能. 我们对新估计$\hat{\beta}_A$、基于子模型的估计$\hat{\beta}_S$、基于全模型的估计$\hat{\beta}_F$进行比较, 还将比较采用相应估计的模型预测误差. 记子模型(3.1.3)为模型(I), 调整模型(3.2.1)或(3.2.5)为模型(II), 全模型(3.1.1)为模型(III).

取全模型(3.1.1)的维数为30, 子模型(3.1.3)的维数为5. 我们将比较这些估计的均方误差(MSEs)和对应模型的均方预测误差(MSPEs).

在全模型(3.1.1)中, β和γ分别设为

$$\beta = (0.5, -3.5, 2.5, -1.5, 4.0)^T, \quad \gamma = (\gamma_a, \gamma_b)^T,$$

其中

$$\gamma_a = (0.1, 0.1, 0.15, 0.1, 0.1, 0.1, 0.15, 0.25, 0.2, 0.5, 0.1, 0.1, 0.25, 0.1, 0.3),$$
$$\gamma_b = (0.1, 0.25, 0.1, 0.1, 0.3, 0.1, 0.1, 0.25, 0.3, 0.5).$$

误差项ε服从正态分布$N(0, 0.3^2)$. 选择Epanechnikov核$K(u) = 3/4(1-u^2)I_{(-1,1)}(u)$为核函数, 非参窗宽通过GCV方法选取. 我们将通过不同方式选取τ_l, 以表明调整估计对τ_l是稳健的.

设定协变量X和Z分别按照Case I和Case II相关. 在Case I中, $Z^{(5*i)} = X^{(i)} + \omega, i = 1, \cdots, p$, 其中$\omega \sim N(0, 10^{-10})$. 在Case II中, X和Z的协方差阵为

$$\Sigma_{X,Z} = (\sigma_{ij}), \quad \sigma_{ij} = \begin{cases} 1.0, & j = 5i, \quad i = 1, \cdots, p; \\ 0, & \text{otherwise.} \end{cases}$$

样本容量$n = 500$, 算法重复次数$m = 1000$. 调整步数K取为p, 即与$X^{(i)}(i = 1, \cdots, p)$相关的$Z^{(j)}(j = 1, \cdots, q)$的变量个数.

将考虑如下情形:

1) 基于总体的主成分分析, τ_l取其真值.

假设$X \sim N(\mu_X, \mathbf{I}_5)$, $Z \sim N(\mu_Z, \mathbf{I}_{25})$ 其中$\mu_X = (1, 3, -2, 5, -1)^T$, $\mu_Z = 2(\mathbf{1}_{15}, \mathbf{0}_{10})^T + 3(\mathbf{0}_{15}, \mathbf{1}_{10})^T$, $\mathbf{1}_r$表示所有分量为1的r-维行向量, $\mathbf{0}_s$表示所有分量为0的s-维行向量, $\mathbf{I}_k$表示所有对角元素为1的$k \times k$对角阵. 由于$Var(Z) = \mathbf{I}_{25}$, τ_l取为真值

$$\tau_l^{(j)} = \begin{cases} 1.0, & j = 5 * l, \quad l = 1, \cdots, p; \\ 0, & \text{otherwise.} \end{cases}$$

显然, τ_l是标准化的, 即$\|\tau_l\| = 1$. 模拟结果见表3.1.

表 3.1: τ_l取真值时参数估计的均方误差和模型的均方预测误差.

No.	MSEs				MSPEs		
	β	$\hat{\beta}_S$	$\hat{\beta}_A$	$\hat{\beta}_F$	model (I)	model (II)	model (III)
Case I	β_1	0.0806	0.0198	1.7842×10^{12}			
n=500	β_2	1.0746	0.2642	3.9904×10^{13}			
m=1000	β_3	0.0066	0.0980	2.4380×10^{12}	1.9109	0.6099	2.1980×10^{5}
	β_4	1.4170	0.1047	2.3286×10^{13}			
	β_5	0.1062	0.2554	9.0173×10^{11}			
Case II	β_1	0.1162	0.0198	1.3660×10^{6}			
n=500	β_2	1.4334	0.2607	1.7027×10^{5}			
m=1000	β_3	0.0318	0.0959	1.7989×10^{5}	3.0619	0.7303	6.9606×10^{5}
	β_4	2.1305	0.1060	1.0980×10^{5}			
	β_5	0.0783	0.2586	2.3062×10^{4}			

2) 基于样本的主成分分析, τ_l取估计值.

假设$X \sim N(\mu_X, 0.25\,\mathbf{I}_5)$, $Z \sim N(\mu_Z, 4.0\,\mathbf{I}_{25})$ 其中$\mu_X = (1,3,-2,5,-1)^T$, $\mu_Z = 2(\mathbf{1}_{15},\mathbf{0}_{10})^T + 3(\mathbf{0}_{15},\mathbf{1}_{10})^T$.

设$(Y_i, X_i, Z_i), i = 1,\cdots,n$为$(Y,X,Z)$的一个独立观测样本, 样本协方差阵$S = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\bar{Z})(Z_i-\bar{Z})^T$, 其中$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$. 先取$\tau_l$ 为S的第l个特征向量u_l,$l = 1,\cdots,q$, 基于样本计算$\tau_l^T Z$与X的q个多重相关系数, 然后取前p个最大的多重相关系数对应的向量$\{\tau_{l_t} : t = 1,\cdots,p\}$. 表3.2列举了详细的模拟结果.

表 3.2: τ_l取估计值时参数估计的均方误差和模型的均方预测误差.

No.	MSEs				MSPEs		
	β	$\hat{\beta}_S$	$\hat{\beta}_A$	$\hat{\beta}_F$	model(I)	model(II)	model(III)
Case I	β_1	0.0714	0.0277	2.6798×10^{10}			
n=500	β_2	1.1603	0.3564	2.2090×10^{10}			
m=1000	β_3	0.0292	0.1664	4.2535×10^{10}	2.8768	2.1755	5.7280×10^{3}
	β_4	1.4788	0.1213	7.7148×10^{10}			
	β_5	0.1319	0.3031	7.0836×10^{10}			
Case II	β_1	0.1972	0.1191	2.4895×10^{6}			
n=500	β_2	5.1409	4.1299	3.8820×10^{5}			
m=1000	β_3	1.1076	1.5663	5.7254×10^{6}	2.2671	1.5255	2.2547×10^{3}
	β_4	2.7417	1.3899	2.3762×10^{6}			
	β_5	3.7741	3.9448	1.6863×10^{5}			

3) τ_l随机选取.

假设$X \sim N(\mu_X, 0.25\mathbf{I}_5)$, $Z \sim N(\mu_Z, 4.0\mathbf{I}_{25})$ 其中$\mu_X = (1,3,-2,5,-1)^T$, $\mu_Z = 2(\mathbf{1}_{15}, \mathbf{0}_{10})^T + 3(\mathbf{0}_{15}, \mathbf{1}_{10})^T$.

先按概率0.2做贝努力试验$P\{\tau_l^{(j)} = 1\} = 0.2, \quad j = 1, \cdots, q, \quad l = 1, \cdots, p$以选取$\tau_l$, 然后对$\tau_l$进行标准化. 相应的模拟结果见表3.3.

在表3.1和表3.2中, 向量τ_l分别基于总体和样本的主成分分析方法选取. 从表3.1至表3.3可以发现, 对于新估计$\hat{\beta}_A$, 至少有三个以上分量的MSEs小于估计$\hat{\beta}_S$, 而且五个分量中最大的MSE比$\hat{\beta}_S$的小很多. 估计$\hat{\beta}_A$和$\hat{\beta}_S$的MSEs远小于估计$\hat{\beta}_F$. 基于新估计的调整模型的MSPE小于子模型的MSPE, 而远小于全模型的MSPE. 因此, 就均方误差和均方预测误差而言, 新的调整模型及其参数估计表现最佳, 而且对τ_l的选取是稳健的.

§3.4 结束语

本章基于预先给定的子模型, 通过多步调整方法构造了一个条件无偏的调整模型. 其主要思想是分析子模型的残余项, 利用了协变量之间的相关信息. 新的调整模型仅包括一个低维的参数部分和多个一维的非参部分. 这样估计过程便于实现, 所导出的调整估计$\hat{\beta}_A$是渐近无偏和相合的, 而且服从渐近正态分布.

当$K \ll q$时, 基于调整模型的新估计一般优于现有的估计. 更确切地: 1) 估计$\hat{\beta}_S$的偏$(E(XX^T))^{-1}E(XZ^T)\gamma$是个常数. 新估计$\hat{\beta}_A$的偏$K*(O(h_{max}^2)+O(n^{-1}))$趋于0. 所以, 如果子模型是误定的而且$K$不大(例如$K \ll q$), 新估计$\hat{\beta}_A$的偏将比$\hat{\beta}_S$的偏小很多. 尤其, 对于调整的线性模型, 新估计$\hat{\beta}_L$是无偏的. 2) 当$K \ll q$, 调整模型(3.2.5)或(3.2.2)总

表 3.3: τ_l随机选取时参数估计的均方误差和模型的均方预测误差.

No.		MSEs			MSPEs		
	β	$\hat{\beta}_S$	$\hat{\beta}_A$	$\hat{\beta}_F$	model(I)	model(II)	model(III)
Case I	β_1	0.0935	0.0256	1.9104×10^{11}			
n=500	β_2	1.0938	0.2326	5.2779×10^{12}			
m=1000	β_3	0.0226	0.0927	3.9372×10^{11}	2.8386	1.5971	8.3777×10^{4}
	β_4	1.4791	0.0937	2.5891×10^{12}			
	β_5	0.1227	0.2326	3.5851×10^{11}			
Case II	β_1	0.2616	0.2244	3.4396×10^{6}			
n=500	β_2	5.1239	3.8396	2.2277×10^{7}			
m=1000	β_3	1.0540	1.3376	1.1387×10^{6}	2.2965	1.8749	7.5368×10^{5}
	β_4	2.6669	1.3812	4.9274×10^{6}			
	β_5	3.6827	3.8761	9.8622×10^{6}			

是可行的, 因此, 新估计$\hat{\beta}_A$或$\hat{\beta}_L$是可获得的, 除非X和Z之间的相关性很强. 当全模型(3.1.1)的维数比调整的线性模型(3.2.2)大很多时, 如果X和Z相关, 全模型的设计矩阵可能会是奇异的或者近似奇异. 因此, 通常的最小二乘估计算法可能崩溃并异常退出, 估计$\hat{\beta}_F$得不到或不可信.

当协变量Z的大多数独立成分都与协变量X相关时, 调整步数K将接近于甚至等于Z的维数q. 这时候调整模型及其估计可能会失去其有效性. 这种情形需要一些其它的技术方法来处理, 将在下一章进行研究.

§3.5 附录

§3.5.1 模型(3.2.1)无偏性的证明

证明. 在调整模型(3.2.1)中,随机误差项ζ_K 满足

$$
\begin{aligned}
&E(\zeta_K|X,\tilde{Z}^{(i)},i=1,\cdots,K)\\
&=E(Y-\beta^TX-g_1(\tilde{Z}^{(1)})-\cdots-g_K(\tilde{Z}^{(K)})|X,\tilde{Z}^{(i)},i=1,\cdots,K)\\
&=E(\gamma^TZ+\epsilon|X,\tilde{Z}^{(i)},i=1,\cdots,K)-\Sigma_{j=1}^{K}E(E(\gamma^TZ|\tilde{Z}^{(j)})|X,\tilde{Z}^{(i)},i=1,\cdots,K)\\
&=E(\gamma^TZ|X,\tilde{Z}^{(i)},i=1,\cdots,K)-\Sigma_{j=1}^{K}E(\gamma^TZ|\tilde{Z}^{(j)})\\
&=E(\Sigma_{l=1}^{q}\gamma_l\Sigma_{j=1}^{q}F_{lj}(\tilde{Z}^{(j)})|X,\tilde{Z}^{(i)},i=1,\cdots,K)-\Sigma_{j=1}^{K}E(\Sigma_{l=1}^{q}\gamma_lF_{lj}(\tilde{Z}^{(j)})|\tilde{Z}^{(j)})\\
&=\Sigma_{l=1}^{q}\gamma_l\Sigma_{j=1}^{K}F_{lj}(\tilde{Z}^{(j)})-\Sigma_{j=1}^{K}\Sigma_{l=1}^{q}\gamma_lF_{lj}(\tilde{Z}^{(j)})\\
&=0,\quad\forall Z,X
\end{aligned}
$$

其中$Z=(Z^{(1)},\cdots,Z^{(q)})^T$, $Z^{(l)}=\Sigma_{j=1}^{q}F_{lj}(\tilde{Z}^{(j)}),l=1,\cdots,q$, $\{i:E(\tilde{Z}^{(i)}|X)\neq 0,1\leq i\leq q\}=\{1,\cdots,K\}$. 所以,调整模型(3.2.1)是全局条件无偏的. □

§3.5.2 定理3.1的证明

证明. 渐近正态分布结果主要来自Fan和Li(2003)中定理2.1或者高集体和赵林诚(1992)中定理1的证明. 这里我们仅给出当调整模型(3.2.1)为等方差情形时的证明思路, 异方差情形时的证明是类似的.

在调整估计(3.2.6)中,按照模型(3.2.1)用$\beta^TX_i+g_1(\tilde{Z}_i^{(1)})+\cdots+g_K(\tilde{Z}_i^{(K)})+\zeta_{Ki}$代替$Y_i$, 将得到

$$
\hat{\beta}_A-\beta=S_n^{-1}\frac{1}{n}(\sum_{i=1}^{n}\check{X}_i\sum_{l=1}^{K}\tilde{g}_{li}+\sum_{i=1}^{n}\check{X}_i\zeta_{Ki}-\sum_{i=1}^{n}\check{X}_i\bar{\zeta}_{Ki}),\tag{3.5.1}
$$

其中$S_n=\frac{1}{n}\sum_{i=1}^{n}\check{X}_i\check{X}_i^T$, $g_{li}=E(Y_i-\beta^TX_i|\tilde{Z}_i^{(l)})$,

$$
\check{X}_i=X_i-\sum_{l=1}^{K}\frac{\sum_{j=1}^{n}X_jK((\tilde{Z}_j^{(l)}-\tilde{Z}_i^{(l)})/h_l)}{\sum_{j=1}^{n}K((\tilde{Z}_j^{(l)}-\tilde{Z}_i^{(l)})/h_l)}\triangleq X_i-\sum_{l=1}^{K}\hat{X}_{li},
$$

$\tilde{g}_{li} = g_{li} - \sum_{k=1}^{K}\sum_{j=1}^{n}\frac{g_{lj}K_{h_k}(\tilde{Z}_j^{(k)}-\tilde{Z}_i^{(k)})}{\sum_{j=1}^{n}K_{h_k}(\tilde{Z}_j^{(k)}-\tilde{Z}_i^{(k)})} \triangleq g_{li} - \sum_{k=1}^{K}\hat{g}_{kli}$,

$\tilde{\zeta}_{Ki} = \sum_{l=1}^{K}\sum_{j=1}^{n}\frac{\zeta_{Kj}K_{h_l}(\tilde{Z}_j^{(l)}-\tilde{Z}_i^{(l)})}{\sum_{j=1}^{n}K_{h_l}(\tilde{Z}_j^{(l)}-\tilde{Z}_i^{(l)})} \triangleq \sum_{l=1}^{K}\hat{\zeta}_{li}$.

首先, 根据$\check{X}$的定义、非参估计理论和大数定律, 容易得出

$$S_n \xrightarrow{P} E((X - \sum_{l=1}^{K}E(X|\tilde{Z}^{(l)}))(X - \sum_{l=1}^{K}E(X|\tilde{Z}^{(l)}))^T) \equiv \Phi, \quad n \to \infty.$$

对于项$\frac{1}{n}\sum_{i=1}^{n}\check{X}_i\sum_{l=1}^{K}\tilde{g}_{li}$, 可以证明它是$o_p(n^{-1/2})$阶的.
实际上,

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_i\sum_{l=1}^{K}\tilde{g}_{li} = \frac{1}{n}\sum_{i=1}^{n}(\xi_i + V_i + \sum_{l=1}^{K}(\eta(\tilde{Z}_i^{(l)}) - \hat{X}_{li}))\sum_{l=1}^{K}(g_{li} - \sum_{k=1}^{K}\hat{g}_{kli}),$$

其中

$$V = X - E(X|Z), \quad \eta(\tilde{Z}^{(l)}) = E(X|\tilde{Z}^{(l)}), \quad \xi = E(X|Z) - \sum_{l=1}^{K}\eta(\tilde{Z}^{(l)}).$$

通过计算,能够得到如下结果

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i\sum_{l=1}^{K}(g_{li} - \sum_{k=1}^{K}\hat{g}_{kli}) = o_p(n^{-1/2}), \quad \frac{1}{n}\sum_{i=1}^{n}V_i\sum_{l=1}^{K}(g_{li} - \sum_{k=1}^{K}\hat{g}_{kli}) = o_p(n^{-1/2}),$$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{K}(\eta(\tilde{Z}_i^{(l)}) - \hat{X}_{li}))\sum_{l=1}^{K}(g_{li} - \sum_{k=1}^{K}\hat{g}_{kli}) = o_p(n^{-1/2}),$$

参考Fan和Li(2003)中引理B.1.(i),引理B.4,引理B.1.(ii),引理B.5和定理2.1中(II)的证明.
因此, $\frac{1}{n}\sum_{i=1}^{n}\check{X}_i\sum_{l=1}^{K}\tilde{g}_{li} = o_p(n^{-1/2})$.

对于剩余项, 有$n^{-\frac{1}{2}}\sum_{i=1}^{n}\check{X}_i(\zeta_{Ki} - \sum_{l=1}^{K}\hat{\zeta}_{li}) = n^{-\frac{1}{2}}\sum_{i=1}^{n}(V_i + \xi_i)(\zeta_{Ki} - \sum_{l=1}^{K}\hat{\zeta}_{li}) + o_p(1) =$
$n^{-\frac{1}{2}}\sum_{i=1}^{n}((V_i + \xi_i)\zeta_{Ki} - \xi_i\sum_{l=1}^{K}\hat{\zeta}_{li}) + o_p(1)$, 其中
$n^{-\frac{1}{2}}\sum_{i=1}^{n}(\sum_{l=1}^{K}(\eta(\tilde{Z}_i^{(l)}) - \hat{X}_{li}))(\zeta_{Ki} - \sum_{l=1}^{K}\hat{\zeta}_{li}) = o_p(1), \quad n^{-\frac{1}{2}}\sum_{i=1}^{n}V_i\sum_{l=1}^{K}\hat{\zeta}_{li} = o_p(1),$
参考Fan和Li(2003)中引理B.5,引理B.3.(ii)和定理2.1中(II)的证明.

由于$\tilde{Z}^{(l)}, l = 1, \cdots, K$是独立的, $Z = (Z^{(1)}, \cdots, Z^{(q)})^T$其中$Z^{(l)} = \sum_{j=1}^{q}F_{lj}(\tilde{Z}^{(j)}), l = 1, \cdots, q$, 根据van der Vaart (2000)中的*Hájek*投影引理,$E(X|Z) = \sum_{l=1}^{q}E(X|\tilde{Z}^{(l)}) -$

$(q-1)EX$. 因为$E(\tilde{Z}^{(l)}|X)=0, K<l\leq q$, 所以$E(X|\tilde{Z}^{(l)})=EX, K<l\leq q$, $E(X|Z)=\sum_{l=1}^{K}E(X|\tilde{Z}^{(l)})+(1-K)EX$. 根据Fan和Li(2003)中引理B.3.(ii),

$$n^{-\frac{1}{2}}\sum_{i=1}^{n}\xi_i\sum_{l=1}^{K}\hat{\zeta}_{li}\leq n^{-\frac{1}{2}}(1-K)EX\sum_{l=1}^{K}c\|\hat{\zeta}_l\|=o_p(1).$$

这样，总的剩余项等于$n^{-1/2}\sum_{i=1}^{n}(V_i+\xi_i)\zeta_{Ki}+o_p(1)$.

因此, (3.5.1)可化成

$$\sqrt{n}(\hat{\beta}_A-\beta)=\Phi^{-1}n^{-1/2}\sum_{i=1}^{n}(V_i+\xi_i)\zeta_{Ki}+o_p(1), \tag{3.5.2}$$

根据中心极限定理,便可得到下面的渐近分布$n^{1/2}(\hat{\beta}_A-\beta)\longrightarrow_d N(0,\Phi^{-1}E(\zeta_K^2DD^T)\Phi^{-1})$, 其中$D=X-\sum_{l=1}^{K}E(X|\tilde{Z}^{(l)})$. □

第四章　稀疏部分线性回归模型的两阶段重新建模和估计

§4.1　引言

考虑如下的部分线性回归模型

$$Y = \beta^T X + \gamma^T Z + f(U) + \varepsilon, \tag{4.1.1}$$

其中, Y是1-维响应变量, X和Z分别是p-维和q-维协变量, p是有限的而且满足$p \ll q$, 对应的参数为β和γ, $f(\cdot)$是满足$Ef(U) = 0$的未知函数, ε是随机误差项. 模型(4.1.1)包含了全部有关的变量, 从而是条件无偏的, 即$E(\varepsilon|X, Z, U) = 0$. 本章中我们称之为全模型. 为简化起见, 假设$U$是1-维变量. 实际上, $f(\cdot)$可推广到U是多维变量时的可加结构. 假设$(Y_i, X_i, Z_i, U_i), i = 1, \cdots, n$是来自全模型的$(Y, X, Z, U)$的独立同分布观测值.

假设γ为稀疏的, 我们对参数β感兴趣, 本章的主要目标是估计参数β. 主要的思路是基于以下子模型进行重新建模

$$Y = \beta^T X + f(U) + \eta. \tag{4.1.2}$$

根据已知条件, 如果$\gamma^T E(Z|X, U) \neq 0$, 则$E(\eta|X = x, U = u)$是一个非零函数, 它依赖于两部分, 其一是$E(Z|X, U)$, 它描述了变量$X$和变量$Z$之间的相关性, 其二是参数$\gamma$, 它由剔除部分中参数的稀疏性决定. 这样, 基于子模型(4.1.2)的参数估计可能是不相合的.

下面将利用这两部分信息. 特别地, 在第一阶段, 为了对子模型进行纠偏, 利用变量X和变量Z之间的相关性信息, 按照第三章的方法我们将首先构造一个多步调整模型. 由于只有Z中与X相关的分量才可能对β的估计产生影响, 所以在每一步中将对子模型增加一个一维非参项, 最终将得到一个全局无偏的工作模型. 然而, 当Z中与X相关的分量个数很大时,工作模型中非参调整项的个数也将会很大. 在实际应用中这样的模型是不合适的. 因此, 在第二阶段, 利用半参SCAD变量选择方法进一步对上述调整模型进行简化. 对于半参变系数部分线性模型, 在稀疏条件下, 该方法能够同时选择重要的参数项和非参项, 详见文献Zhao和Xue(2009). 相关的文献见Fan和Li(2001), Wang,Chen和Li(2007), Wang,Li和Huang(2008), 等等. 经过两阶段重新建模, 最终得到的模型是条件无偏的而且是充分约简的. 基于这个模型, 估计和模型预测的效果能够明显得到改进.

本章的后续部分组织如下. 在§4.2节, 首先基于子模型提出了一个全局无偏的调整模型, 然后通过半参SCAD变量选择方法对调整模型进一步进行化简. 基于该简化模型, 导出了感兴趣参数β的一个新估计, 还得到了这个新估计的收敛速度和渐近正态性以及非参项估计的相合性. 第§4.3节给出了正态分布和非正态分布下的一些模拟结

果. 第§4.4节对于本章提出的建模方法和估计方法做了一些备注和说明. 第§4.5节列举了一些正则条件, 并提供了一些理论结果的证明过程.

§4.2 感兴趣参数的新估计

本章中假定p是有限的, $p \ll q$, $E(Z)=0$, $E(\varepsilon|X,Z,U)=0$, $Var(\varepsilon|X,Z,U)=\sigma^2$. 记$W=(X^T,Z^T)^T$, 假定矩阵$B=E((W-E(W|U))(W-E(W|U)^T)$正定, 从而全模型(4.1.1)是可识别的. 而且,协变量X和U以及参数β提前给定. 因此，子模型(4.1.2)是固定模型.

§4.2.1 基于相关性的多步调整

在本小节中我们首先通过多步调整方法将子模型调整为条件无偏的.

如果Z为正态分布变量, 则对Z的总体应用主成分分析(PCA)方法. 设Σ_Z为Z的协方差阵, 则存在一个$q\times q$正交矩阵Q, 使得$Q\Sigma_Z Q^T=\Lambda$, Λ是对角阵$\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_q)$, 其中$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_q\geq 0$是$\Sigma_Z$的特征值. 记$Q^T=(\tau_1,\tau_2,\cdots,\tau_q)$, $\tilde{Z}^{(j)}=\tau_j^T Z$.

如果Z为中心化的非正态分布变量, 则可以利用独立成分分析(ICA)方法. 假设Z由一组独立变量$\tilde{Z}^{(i)}$的非线性组合生成, 即$Z=F(\tilde{Z})$, 其中$F(\cdot)$是一个$R^q\to R^q$的未知的非线性映射, $\tilde{Z}$是一个未知的由独立分量组成的随机向量. 对非线性映射F或独立成分$\tilde{Z}^{(i)}$施加一定的条件, 独立成分$\tilde{Z}^{(i)}$能够被适当估计. 关于估计独立成分的一些主要统计原则和算法, 见综述性文献Simas Filho和Seixas(2007),Acharya和Panda(2008). 为简化起见, 本章中假定$Z=(Z^{(1)},\cdots,Z^{(q)})^T$, $Z^{(l)}=\Sigma_{j=1}^q F_{lj}(\tilde{Z}^{(j)}), l=1,\cdots,q$, 其中$F_{lj}(\cdot)$是标量函数.

在上述两种情况下, $\tilde{Z}^{(j)}$之间均互相独立.. 设K_0是集合$M_0=\{j: E(\tilde{Z}^{(j)}|X,U)\neq 0, 1\leq j\leq q\}$的大小. 不失一般性, 记$M_0=\{1,\cdots,K_0\}$.

构造如下的调整模型

$$Y=\beta^T X+\sum_{j=1}^{K_0} g_j(\tilde{Z}^{(j)})+f(U)+\zeta_{K_0}, \tag{4.2.1}$$

其中, $g_j(\tilde{Z}^{(j)})=E(Y-\beta^T X-f(U)|\tilde{Z}^{(j)})=\gamma^T E(Z|\tilde{Z}^{(j)}), j=1,\cdots,K_0$, $\zeta_{K_0}=Y-\beta^T X-g_1(\tilde{Z}^{(1)})-\cdots-g_{K_0}(\tilde{Z}^{(K_0)})-f(U)$. 模型(4.2.1)基于协变量$Z$的总体构造, 依赖于$X$, U和Z的分布. 容易看到模型(4.2.1)是条件无偏的, 即$E(\zeta_{K_0}|X,U,\tilde{Z}^{(j)},1\leq j\leq K_0)=0$.

调整模型(4.2.1)是一个特殊的可加部分线性模型, 其中, $\beta^T X$是参数部分, $f(U)$和$g_j(\tilde{Z}^{(j)}), j=1,\cdots,K_0$是单指标非参部分, ζ_{K_0}是随机误差项. 与子模型(4.1.2)相比, 非参部分$g_j(\tilde{Z}^{(j)}), j=1,\cdots,K_0$可看作是对子模型误差项$\eta$的纠偏项. 对于中心化的$Z$,

$E(g_j(\tilde{Z}^{(j)}))=0, j=1,\cdots,K_0$, 所以非参部分$g_1(\tilde{Z}^{(1)}),\cdots,g_{K_0}(\tilde{Z}^{(K_0)})$能够被适当估计. 实际上, 中心化的$Z$可被放宽至满足条件$\gamma^T E(Z)=0$.

当Z为中心化的正态分布时, 非参部分$g_j(\tilde{Z}^{(j)})=\alpha_j\tau_j^T Z=\alpha_j\tilde{Z}^{(j)}, j=1,\cdots,K_0$. 所以多步调整模型(4.2.1)实际上是一个部分线性模型

$$Y=\beta^T X+\alpha^T\tilde{Z}_{K_0}+f(U)+\zeta_{K_0} \tag{4.2.2}$$

其中, $\alpha=(\alpha_1,\cdots,\alpha_{K_0})^T$, $\tilde{Z}_{K_0}=(\tilde{Z}^{(1)},\cdots,\tilde{Z}^{(K_0)})^T$. 如果$Z$中的变量不是联合正态分布的, 非参部分$g_j$可能是高度非线性的, 这类似于边际回归的结果; 见文献Fan等(2011).

§4.2.2 模型化简

当全模型中的多数特征具有相关性时, 则K_0将非常大甚至接近于q. 在这种情况下, 多步调整模型将不适合在实际中使用. 为了进一步简化模型(4.2.1), 我们采用Wang,Chen和Li(2007)提出的组SCAD回归方法、Zhao和Xue(2009)提出的半参SCAD变量选择方法.

令$\mathcal{M}_*=\{1\le j\le K_0: E(g_j(\tilde{Z}^{(j)}))^2>0\}$, $s=|\mathcal{M}_*|$, 假设模型(4.2.1)是稀疏的, 即s是小的. 定义半参惩罚最小二乘如下

$$F(\beta,g(\cdot),f(\cdot))=\sum_{i=1}^{n}\{Y_i-\beta^T X_i-\sum_{j=1}^{K_0}g_j(\tilde{Z}_i^{(j)})-f(U_i)\}^2+n\sum_{j=1}^{K_0}p_{\lambda_j}(\|g_j(\tilde{Z}^{(j)})\|), \tag{4.2.3}$$

其中, $\|g_j(\tilde{Z}^{(j)})\|=(E(g_j(\tilde{Z}^{(j)}))^2)^{1/2}$, λ_j是调整参数, $p_\lambda(\cdot)$是SCAD惩罚函数, 定义如下

$$p'_\lambda(w)=\lambda\{I(w\le\lambda)+\frac{(a\lambda-w)_+}{(a-1)\lambda}I(w>\lambda)\},$$

满足$a>2$, $w>0$, $p_\lambda(0)=0$. 在(4.2.3)中, $g(\cdot)$代表$\{g_j(\tilde{Z}^{(j)}), j=1,\ldots,K_0\}$. 由于$g_j$和$f$是非参函数, 所以不能直接对(4.2.3)进行最小化. 这里将对$f(\cdot)$和$g(\cdot)$用基函数进行近似. 对$1\le j\le K_0$, 令$\{\Psi_{jk}, k=1,\cdots,L\}$满足如下条件的正交基函数

$$E(\Psi_{jk}\Psi_{jl})\equiv\int_{supp}\Psi_{jk}(\tilde{Z}^{(j)})\Psi_{jl}(\tilde{Z}^{(j)})r_j(\tilde{Z}^{(j)})dZ=\delta_{kl}=\begin{cases}0, & k\ne l;\\ 1, & k=l,\end{cases}$$

其中, $r_j(\tilde{Z}^{(j)})$是$\tilde{Z}^{(j)}$的密度函数. 类似地, 令$\{\Psi_{0k}, k=1,\cdots,L\}$满足上述条件的基函数, 条件中对应改为$U$的支撑集和密度函数. 记$\mathbf{\Psi}_j(\tilde{Z}^{(j)})=(\Psi_{j1}(\tilde{Z}^{(j)}),\cdots,\Psi_{jL}(\tilde{Z}^{(j)}))^T$, $\mathbf{\Psi}_0(U)=(\Psi_{01}(U),\cdots,\Psi_{0L}(U))^T$. 这样, $g_j(\tilde{Z}^{(j)})$和$f(U)$可被如下近似表示

$$g_j(\tilde{Z}^{(j)})\approx\theta_j^T\mathbf{\Psi}_j(\tilde{Z}^{(j)}), f(U)\approx\nu^T\mathbf{\Psi}_0(U).$$

记$\|\theta_j\|_2=(\theta_j^T\theta_j)^{1/2}$, 由于$E(\mathbf{\Psi}_j(\tilde{Z}^{(j)})\mathbf{\Psi}_j^T(\tilde{Z}^{(j)}))=\mathbf{I}_L$单位矩阵, 可以得到

$$F(\beta,\theta,\nu)=\sum_{i=1}^{n}\{Y_i-\beta^T X_i-\theta^T\mathbf{\Psi_i}-\nu^T\mathbf{\Psi_{0i}}\}^2+n\sum_{j=1}^{K_0}p_{\lambda_j}(\|\theta_j\|_2), \tag{4.2.4}$$

其中, $\theta = (\theta_1^T, \cdots, \theta_{K_0}^T)^T$, $\mathbf{\Psi_i} \equiv \mathbf{\Psi}(\tilde{Z}_i) = Vec(\mathbf{\Psi}_1(\tilde{Z}_i^{(1)}), \cdots, \mathbf{\Psi}_{K_0}(\tilde{Z}_i^{(K_0)}))$, $\mathbf{\Psi_{0i}} \equiv \mathbf{\Psi}_0(U_i)$.

分别用$\hat{\beta}$, $\hat{\theta} = (\hat{\theta}_1^T, \cdots, \hat{\theta}_{K_0}^T)^T$ 和$\hat{\nu}$记基于如上惩罚函数(4.2.4)的最小二乘估计, 即

$$(\hat{\beta}, \hat{\theta}, \hat{\nu}) = \arg\min_{\beta \in R^p, \theta_j \in R^L, \nu \in R^L} F(\beta, \theta, \nu).$$

令$\hat{g}_j \equiv \hat{g}_j(\tilde{Z}^{(j)}) = \hat{\theta}_j^T \mathbf{\Psi}_j(\tilde{Z}^{(j)})$, $\hat{f} \equiv \hat{f}(U) = \hat{\nu}^T \mathbf{\Psi}_0(U)$, 则$\hat{g}_j$是$g_j(\tilde{Z}^{(j)})$的一个估计, $\hat{f}$是$f(U)$的一个估计.

令$\hat{\mathcal{M}}_n = \{1 \le j \le K_0 : \hat{\theta}_j \ne 0\}$, $K_n = |\hat{\mathcal{M}}_n|$. 为简化起见, 假设$\mathcal{M}_* = \{1, 2, \cdots, s\}$, $\hat{\mathcal{M}}_n = \{1, 2, \cdots, K_n\}$. 从而得到以下化简的工作模型

$$Y = \beta^T X + \sum_{j=1}^{K_n} g_j(\tilde{Z}^{(j)}) + f(U) + \zeta_{K_n}, \tag{4.2.5}$$

其中, $g_j(\tilde{Z}^{(j)}) = E(\gamma^T Z | \tilde{Z}^{(j)}), j = 1, \cdots, K_n$, $\zeta_{K_n} = Y - \beta^T X - g_1(\tilde{Z}^{(1)}) - \cdots - g_{K_n}(\tilde{Z}^{(K_n)}) - f(U)$. 在稀疏性假设下, 模型(4.2.5)包含了所有重要的非参函数, 而且充分利用了协变量之间的相关性信息和对应于协变量Z的模型稀疏性.

如果Z是中心化的、协方差阵为单位矩阵的正态分布变量, 即$\Sigma_Z = I_q$, 则有$\tau_j = e_j, j = 1, \cdots, q$, 其中$e_j$表示第$j$个分量为1、其余分量为0的单位方向向量. 这时$\alpha$是稀疏的, 而且$\alpha_j = \gamma^T \tau_j = \gamma_j$. 因此模型(4.2.2)是稀疏的. 对于线性模型, 在这种特殊情况下, 可以直接采用Fan和Li(2001)中的SCAD惩罚方法同时去选择$\tilde{Z}_{K_0}$中的变量、估计参数α和β. 选出的协变量以及对应的参数分别用$\tilde{Z}_{K_n}$和α_{K_n}表示, 相应的参数估计分别用$\hat{\alpha}_{K_n}$和$\hat{\beta}$表示. 最终, 我们可以采用如下的化简模型

$$Y = \beta^T X + \alpha_{K_n}^T \tilde{Z}_{K_n} + \zeta_{K_n} \tag{4.2.6}$$

进行模型预测. 在稀疏性条件下, 这个化简模型的维数远小于多步调整模型(4.2.2)和全模型(4.1.1)的维数.

§4.2.3 点估计的渐近性质

令β_0, θ_0, ν_0, $g_{j0}(\cdot)$, $f_0(\cdot)$ 分别是模型(4.2.1)中β, θ, ν, $g_j(\cdot)$, $f(\cdot)$的真实值. 不失一般性, 假设$g_{j0}(\tilde{Z}^{(j)}) = 0, j = s+1, \cdots, K_0$, 而$g_{j0}(\tilde{Z}^{(j)}), j = 1, \cdots, s$是非零部分.

假设$g_j(\tilde{Z}^{(j)}) = \sum_{k=1}^{\infty} \theta_{jk} \Psi_{jk}(\tilde{Z}^{(j)})$, $j = 1, \cdots, K_0$, $f(U) = \sum_{k=1}^{\infty} \nu_k \Psi_{0k}(U)$, 其中$\theta_j$和$\nu$属于Sobolev椭球$S(r, M) = \{\theta : \sum_{k=1}^{\infty} \theta_k^2 k^{2r} \le M, M > 0, r > 0\}$.

下面定理给出了惩罚SCAD估计的相合性结果.

定理4.1 假设第§4.5.1节中的正则性条件C4.1-C4.5成立, 基函数的截断阶数$L = O_p(n^{1/(2r+1)})$. 则有,

(i) $\|\hat{\beta}-\beta_0\| = O_p(n^{-r/(2r+1)}+a_n)$,

(ii) $\|\hat{g}_j(\cdot)-g_{j0}(\cdot)\| = O_p(n^{-r/(2r+1)}+a_n), j=1,\cdots,K_0$,

(iii) $\|\hat{f}(\cdot)-f_0(\cdot)\| = O_p(n^{-r/(2r+1)}+a_n)$,

其中, $a_n = \max_j\{|p'_{\lambda_j}(\|\theta_{j0}\|_2)| : \theta_{j0}\neq 0\}$.

特别地, 对于线性模型, 当Z是正态分布时, 调整模型(4.2.1)实际上是一个线性模型(4.2.6). 通过正交基函数, 例如幂级数, 有$r=\infty$, 则$\|\hat{\beta}-\beta_0\| = O_p(n^{-1/2}+a_n)$, 这意味着新估计$\hat{\beta}$具有与Fan和Li(2001)中的SCAD估计相同的收敛速度.

定理4.2 假设第§4.5.1节中的正则性条件C4.1-C4.6成立, 基函数的截断阶数$L = O_p(n^{1/(2r+1)})$. 令$\lambda_{max} = \max_j\{\lambda_j\}$, $\lambda_{min} = \min_j\{\lambda_j\}$. 如果$n\to\infty$时, $\lambda_{max}\to 0$, $n^{r/(2r+1)}\lambda_{min}\to\infty$, 则有, 以趋于1的概率, $\hat{g}_j(\cdot)\equiv 0$, $j=s+1,\cdots,K_0$.

注4.1 按Fan和Li(2001)中的注1, 当$n\to\infty$时若$\lambda_{max}\to 0$, 则$a_n\to 0$. 因此, 应用定理4.1和定理4.2, 通过适当地选择调整参数, 这个变量选择方法是相合的, 而且就像系数为0的子集已知一样, 非参部分的估计能够达到最优收敛速率; 见文献Stone(1982).

令$\theta^* = (\theta_1^T,\cdots,\theta_s^T)^T$是$\theta$中的非零部分, 相应的协变量记为$\mathbf{\Psi}_i^*, i=1,\cdots,n$. 并且, 令

$$\Sigma = \frac{1}{\sigma^2_{K_0}}\{E(XX^T)-E(X\mathbf{\Psi}^{*T})E^{-1}(\mathbf{\Psi}^*\mathbf{\Psi}^{*T})E(\mathbf{\Psi}^*X^T)-E(X\breve{\Psi}_0^T)E^{-1}(\breve{\Psi}_0\breve{\Psi}_0^T)E(\breve{\Psi}_0X^T)\},$$

其中, 对于等方差情形$\sigma^2_{K_0} = Var(\zeta_{K0i})$, $\breve{\Psi}_0 = \mathbf{\Psi}_0 - E(\mathbf{\Psi}_0\mathbf{\Psi}^{*T})E^{-1}(\mathbf{\Psi}^*\mathbf{\Psi}^{*T})\mathbf{\Psi}^*$.

定理4.3 假设第§4.5.1节中的正则性条件C4.1-C4.6成立, 基函数的截断阶数$L = O_p(n^{1/(2r+1)})$. 如果Σ可逆, 则有

$$\sqrt{n}(\hat{\beta}-\beta_0)\longrightarrow_d N(0,\Sigma^{-1}),$$

其中"$\longrightarrow_d$"表示依分布收敛.

注4.2 定理4.1和定理4.3表明新的惩罚估计具有oracle性质. 进一步, 感兴趣参数β的估计具有与真实子模型相同的渐近分布.

§4.2.4 调整模型的实现

在调整模型(4.2.2)中, 使用了$\tau_j, j=1,\cdots,K_0$. 当总体分布未知时, 需要通过估计来进行近似. 当Z服从正态分布时, 如果其协方差阵Σ_Z的特征值$r_j, j=1,\cdots,q$互不相同, 则$\sqrt{n}(u_j-\tau_j)$渐近于正态分布$N(\mathbf{0},V_j)$, 其中$V_j = \sum_{l\neq j}^q \frac{r_jr_l}{(r_j-r_l)^2}\tau_l\tau_l^T$, u_j是$S = \frac{1}{n-1}\sum_{i=1}^n(Z_i-\bar{Z})(Z_i-\bar{Z})^T$ 的第j个特征向量, $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$; 见文献Anderson(2003). 当分布总体的维数很大而且与样本容量相当时, 如果协方差矩阵是稀疏的, 则可以采用文献Rütimann和Bühlmann(2009)或者Cai和Liu(2011)中的方法去估计协方差矩阵. 这样可以用u_j去近似τ_j. 当模型(4.2.2)中的τ_j用这些相合估计来替代时, 能够发现近似误差可被忽略而不改变方法的渐近性质.

调整模型中的非参项$g_i(\tilde{Z}^{(i)})$依赖于一维变量$\tilde{Z}^{(i)}, i = 1, \cdots, K_0$. 因此, 需要首先确定调整步数$K_0$. 在实际的实现过程中, 我们计算出$\tilde{Z}^{(l)}$ $(l = 1, \cdots, q)$与X和U的所有q个复相关系数. 给定一个小量$\delta > 0$, 然后选择集合$R = \{\tilde{Z}^{(l)} : |\mathrm{mcorr}(\tilde{Z}^{(l)}, (X, U))| \geq \delta, l = 1, \cdots, q\}$, 其中$\mathrm{mcorr}(u, V)$表示变量$u$与$V$之间的复相关系数. 根据文献Anderson (2003), 可以采用复相关系数的样本形式进行近似. 取$K_0 = |R|$即可.

为了实现两阶段重新建模过程, 有一些调整参数需要选取. 文献Fan和Li(2001)表明在很多情况下采用$a = 3.7$的SCAD惩罚表现很好. 因此, 在本章中我们采用这个建议. 另外, 还需选取基函数的截断阶数L和惩罚函数的调整参数λ_j. 类似于Zou(2006)中的适应性Lasso方法, 建议取$\lambda_j = \lambda/\|\hat{\theta}_j^{(0)}\|_2$, 其中$\hat{\theta}_j^{(0)}$是$\theta_j$的初始估计, 它可以通过对模型(4.2.4)中的第一部分应用最小二乘方法得到. 这样就只有L和λ两个参数需要同时选取, 可以采用leave-one-out CV或GCV方法进行选取; 详见文献Zhao和Xue(2009).

§4.3 数值模拟

本节将通过数值模拟表明新方法的性能.

§4.3.1 正态分布协变量下的线性模型

全模型(4.1.1)的维数和子模型(4.1.2) 的维数分别取为100和5. 设定

$$\beta = (0.5, 3.5, 2.5, 1.5, 4.0)^T, \gamma = (\gamma_1, \gamma_2, \mathbf{0}_{55}^T)^T,$$

其中,

$\gamma_2 \sim Unif[-0.5, 0.5]^{30}$, 为$[-0.5, 0.5]^{30}$上的一个30-维均匀分布;

γ_1 按如下两种方式设定:

情形(I): $\gamma_1 \sim Unif[0.5, 1.0]^{10}$;

情形(II): $\gamma_1 = (1.0, 1.0, 1.0, 1.5, 1.5, 1.5, 2.0, 2.0, 2.0, 2.0)$.

假设$(X^T, Z^T)^T \sim N\left((\mathbf{1}_5^T, \mathbf{0}_{40}^T, \mathbf{1}_{55}^T)^T, \Sigma\Sigma^T\right)$, 其中

$$\Sigma = (\sigma_{ij}), \qquad \sigma_{ij} = \sigma_{ji} = \begin{cases} 1.0, & j = i, \ \ i = 1, \cdots, p+q; \\ c, & j = i+p, \ \ i = 1, 3, \cdots, q; \\ 0, & \text{otherwise}, \end{cases}$$

参数c取0.5或0.8. 误差项ε服从正态分布$N(0, 0.3^2)$.

这里, 记子模型(4.1.2)为模型(I), 多步调整线性模型(4.2.2)为模型(II), 两阶段调整模型(4.2.6)为模型(III), 全模型(4.1.1)为模型(IV). 将对基于模型(III)的新两阶段估计$\hat{\beta}_{TS}$的均方误差(mean square errors, 简记为MSEs) 与基于子模型(I)的估计$\hat{\beta}_S$、基于模型(II)的SCAD估计$\hat{\beta}_{SCAD}$和多步调整估计$\hat{\beta}_M$、基于模型(IV)的最小二乘估计$\hat{\beta}_F$的

均方误差进行比较. 还比较了对于上述模型的基于相应估计的均方预测误差(mean square prediction errors, 简记为MSPEs).

样本数据通过全模型(4.1.1)模拟生成, 样本容量$n = 100$, 重复估计次数$m = 1000$. 我们采用基于样本的主成分分析方法来近似代替τ_j. SCAD惩罚函数中的参数a选为3.7, 参数λ通过leave-one-out CV方法选取.

表4.1对于采用不同方法得到的参数估计的均方误差和模型的预测误差进行了比较. 通过该表, 可以发现如下: 1) $\hat{\beta}_F$具有最大的均方误差, $\hat{\beta}_S$次之, 总体而言新估计$\hat{\beta}_{TS}$具有最小的均方误差. 2) 当$c = 0.5$, 估计$\hat{\beta}_{SCAD}$的均方误差小于估计$\hat{\beta}_M$; 当$c = 0.8$, 估计$\hat{\beta}_{SCAD}$的均方误差大于估计$\hat{\beta}_M$. 这些表明, 如果协变量之间的相关性很强, 估计$\hat{\beta}_{SCAD}$的均方误差大于$\hat{\beta}_M$, 多步调整方法是必要的, 所以基于两阶段模型的估计和预测结果得到明显改善. 3) 在情形(I)和情形(II)中, 模拟结果具有相似的表现. 4) 五个模型的均方预测误差与五个参数估计的均方误差具有类似的表现趋势. 两阶段调整模型的均方预测误差在以上模型中是最小的.

总之, 表4.1表明两阶段调整模型(4.2.6) 优于子模型、SCAD惩罚模型和多步调整模型, 明显优于全模型.

§4.3.2 非正态分布协变量下的部分线性模型

全模型(4.1.1)中线性部分的维数和子模型(4.1.2)中线性部分的维数分别取为50和5. 设定$\beta = (0.5, 3.5, 2.5, 1.5, 4.0)^T$, $\gamma = (\gamma_1, \gamma_2, \mathbf{0}_{25}^T)^T$, $f(u) = u * \sin(2u)$, 其中

$\gamma_1 = (0.5, 0.1, 0.8, 0.2, 0.5, 0.2, 0.6, 0.5, 0.1, 0.9)$,

$\gamma_2 \sim Unif[-0.3, 0.3]^{10}$, 即区间$[-0.3, 0.3]$上的一个10维均匀分布.

假设协变量的分布按如下两种方式:

情形(I): $(X^T, Z^T, U)^T \sim t_5\left(\mathbf{0}_{51}^T, \Sigma\right)$, 即一个自由度$df = 5$的51维t分布, 其中

$$\Sigma = (\sigma_{ij}), \qquad \sigma_{ij} = \sigma_{ji} = \begin{cases} 1.0, & j = i, \quad i = 1, \cdots, p+q+1; \\ 0.9, & j = i+p, \quad i = 1, 3, \cdots, q+1; \\ 0, & \text{otherwise.} \end{cases}$$

情形(II): $X = \frac{1}{1+c}(W_1 + cV), Z = (Z_1^T, Z_2^T, Z_3^T, Z_4^T)^T$, $Z_1 = \frac{1}{1+c}(W_2 + cV), Z_2 = W_3, Z_3 = \frac{1}{1+c}(W_4 + cV), Z_4 = W_5$, $U = W_5^{(1)}$, 其中$W_1, W_2, W_3, W_4 \sim Unif[-1.0, 1.0]^5$, $W_5 \sim Unif[-1.0, 1.0]^{30}$, $V \sim Unif[-1.0, 1.0]^5$, 即$[-1.0, 1.0]$上的均匀分布, 常数$c = 0.1$. 所有W_1, W_2, W_3, W_4, W_5和V是相互独立的.

误差项ε服从正态分布$N(0, 0.3^2)$.

记子模型(4.1.2)为模型(I)', 多步调整部分线性模型(4.2.1)为模型(II)', 两阶段模型(4.2.5)为(III)', 全模型(4.1.1)为(IV)'. 将对基于模型(III)'的新两阶段估计$\hat{\beta}_{TS}$的均方误差(mean square errors, 简记为MSEs) 与基于子模型(I)'的估计$\hat{\beta}_S$、基于模型(II)'的多

表 4.1: 兴趣参数估计的均方误差(MSEs)和模型的均方预测误差(MSPEs).

No.	item	$\hat{\beta}_S$	$\hat{\beta}_{SCAD}$	$\hat{\beta}_M$	$\hat{\beta}_{TS}$	$\hat{\beta}_F$
		0.3079	0.0457	0.0660	0.0571	1.6105×10^3
		0.1763	0.0206	0.0346	0.0176	1.0940×10^3
情形(I)	MSEs	0.1396	0.0481	0.0631	0.0461	4.2049×10^3
c=0.5		0.1870	0.0196	0.0349	0.0186	5.0183×10^3
		0.1131	0.0517	0.0609	0.0430	6.2615×10^3
	MSPEs	3.4780	1.1896	1.6512	1.0679	3.0499×10^2
		0.1568	0.6191	0.0934	0.0826	1.2494×10^3
		0.6239	0.1060	0.0090	0.0083	1.0456×10^2
情形(I)	MSEs	0.8829	0.8173	0.0895	0.1039	2.6368×10^2
c=0.8		0.5882	0.0919	0.0107	0.0100	7.6452×10^1
		1.0799	0.9829	0.0961	0.0929	1.1610×10^3
	MSPEs	4.7930	2.6700	0.8354	0.7771	1.3223×10^2
		0.4272	0.0660	0.0849	0.0557	4.3002×10^2
		0.6371	0.0318	0.0499	0.0295	3.7893×10^3
情形(II)	MSEs	0.4560	0.0715	0.0927	0.0588	1.2784×10^3
c=0.5		0.5926	0.0306	0.0491	0.0287	6.7354×10^3
		0.9052	0.0734	0.0874	0.0583	2.5047×10^2
	MSPEs	6.8634	1.5096	2.0780	1.2077	5.0464×10^3
		0.6764	0.4263	0.1212	0.0960	1.3904×10^3
		0.9721	0.1060	0.0107	0.0102	4.0743×10^2
情形(II)	MSEs	0.6242	0.4756	0.1146	0.1003	1.0498×10^3
c=0.8		1.0282	0.0954	0.0112	0.0098	5.6031×10^2
		1.3420	0.5474	0.1341	0.1124	9.9632×10^2
	MSPEs	7.9928	2.1165	0.9514	0.8469	2.3110×10^2

步调整估计$\hat{\beta}_M$、基于模型(IV)的最小二乘估计$\hat{\beta}_F$的均方误差进行比较. 还比较了$f(\cdot)$的非参估计的均方平均误差(mean average square errors, 简记为MASEs) 和对于上述模型的基于相应估计的均方预测误差(mean square prediction errors, 简记为MSPEs).

样本数据通过全模型(4.1.1)模拟生成, 样本容量$n = 100$, 重复估计次数$m = 500$. 我们采用基于样本的独立成分分析方法来近似, 详细参见文献Hyvärinen和Oja(1997). SCAD惩罚函数中的参数a选为3.7, 阶数L和参数λ通过GCV方法选取. 基函数采用标

准的傅立叶(Fourier)正交基.

表4.2列举了感兴趣参数β的不同点估计的均方误差、非参函数$f(\cdot)$的估计的均方平均误差以及相应的不同模型的均方预测误差. 通过该表, 可以发现如下结果: 1) $\hat{\beta}_F$具有最大的均方误差, 其均方误差远大于其它估计的均方误差, 新的两步调整估计$\hat{\beta}_{TS}$具有最小的均方误差. 2) 非参函数$f(\cdot)$的估计的均方平均误差具有与参数β估计的均方误差相类似的趋势, 但是非参估计的差别不是很明显. 3) 类似于参数点估计的均方误差, 在四个模型中两阶段调整模型的均方预测误差是最小的. 4) 在情形(II)中, 由于协变量的相关结构, 其模拟结果比情形(I)更好一些.

总之, 表4.2表明两阶段调整模型(4.2.5)的表现优于子模型, 明显优于全模型和多步调整模型.

表 4.2: 兴趣参数估计的均方误差(MSEs), 非参函数估计的均方平均误差(MASEs) 和模型的均方预测误差(MSPEs).

No.	item	$\hat{\beta}_S$	$\hat{\beta}_M$	$\hat{\beta}_{TS}$	$\hat{\beta}_F$
		0.4352	5.0403	0.3267	2.9753×10^{1}
		0.6859	1.2820×10^{1}	0.3328	1.4593×10^{1}
case(I)	MSEs	1.1152	8.1542	0.3723	1.4391×10^{1}
		1.8489	7.2055	1.3194	2.4036×10^{1}
		3.3079	1.6144×10^{1}	1.9989	4.8575×10^{1}
	MASEs	3.0887	5.9814	3.0175	3.0633
	MSPEs	4.6047	7.0331×10^{1}	3.5536	3.9648
		0.0377	0.6144	0.0191	-[1]
		0.0449	1.0876	0.0305	-
case(II)	MSEs	0.0332	3.7510	0.0246	-
		0.0396	0.4324	0.0238	-
		0.0512	1.1995	0.0335	-
	MASEs	0.4722	0.5220	0.4126	0.4380
	MSPEs	0.9221	9.3068	0.8053	-

§4.4 附注

本章的主要目标是相合地估计感兴趣的参数β. 在估计参数β时, 估计的偏主要取决于与其相关的变量,其它变量可能会对估计的方差产生影响. 因为变量选择主要依赖

[1]'-' 表示算法异常终止而且没有得到返回值.

于参数的稀疏性,当直接对部分线性模型采用变量选择方法时, 一些系数非零但与X不相关的变量可能会被选入模型. 这样可能会影响参数β估计的有效性和稳定性. 因此基于预先给定的子模型, 本章提出了一个新的两阶段重建模方法. 在新的重建模过程中, 充分利用了协变量(X, Z, U)之间的相关性和回归结构的稀疏性. 从而最终的选模型被充分化简, 同时还是条件无偏的. 基于这个简化模型, 参数估计和模型预测效果得到了明显改善. 一般而言, 经过第一阶段后得到的调整模型是一个可加部分线性模型. 所以, 本章提出的重建模方法可被应用于部分线性模型, 线性模型只是一个特例.

从重建模过程可以看出, 该方法可被直接应用于可加部分线性模型, 其中非参项$f(U)$具有分量可加的形式. 至于非参项为多维非参函数的一般部分线性模型, 需要借助于多维的非参估计方法. 如果协变量U是高维的, 可能会面临"维数祸根"问题.

在模型简化的过程中, 采用了正交级数估计方法. 这只是出于技术方便的考虑, 因为半参惩罚最小二乘(4.2.3) 可以被直接转为参数型惩罚最小二乘(4.2.4), 进而得到方法的理论结果. 尽管其它的非参数估计方法, 例如核方法和样条方法, 可以没有本质性困难地使用, 但是它们难以直接达到这个目的. 与核方法相比, 对于正交级数方法而言, 在一定条件下难以获得非参项$f(U)$估计的渐近正态性结果.

§4.5 附录

§4.5.1 正则条件(C4.1)-(C4.6)

(C4.1) $(\tilde{Z}, U)$具有有限的、非退化的紧支撑集, 记为$\mathrm{supp}(\tilde{Z}, U)$.

(C4.2) 存在常数L_1和L_2, $\tilde{Z}^{(j)}$的密度函数$r_j(t)$和U的密度函数$r_0(t)$在其支撑集上满足$0 < L_1 \le r_j(t) \le L_2 < \infty$ $(0 \le j \le K_0)$, 而且密度函数是连续可微的.

(C4.3) $G(\tilde{Z}, U) = E(XX^T|\tilde{Z}, U)$和$E(\zeta_{K_0}^2|\tilde{Z}, U)$是连续的. 对于给定的$z$和$u$, $G(z, u)$是正定的, 而且它的特征值是有界的.

(C4.4) $\sup_{(z,u)\in \mathrm{supp}(\tilde{Z},U)} E(\|X\|^3|\tilde{Z} = z, U = u) < \infty$, $Ef(U) = 0$, 而且$f(\cdot)$的前两阶导数是一阶Lipschitz连续的.

(C4.5) $b_n = \max_j\{p''_{\lambda_j}(\|\theta_{j0}\|_2) : \theta_{j0} \ne 0\} \to 0$, 当$n \to \infty$时.

(C4.6) $\liminf_{n\to\infty} \liminf_{\|\theta_{j0}\|_2 \to 0} \lambda_j^{-1} p'_{\lambda_j}(\|\theta_{j0}\|_2) > 0$, 对于$j = s+1, \cdots, K_0$.

条件C4.1-C4.3是关于协变量的一些常规性约束, 条件C4.4是关于回归结构的一些约束, 见文献Härdle(2000). 条件C4.5-C4.6是关于惩罚函数的一些假设, 类似于文献Fan和Li(2001), Wang,Li和Huang(2008).

§4.5.2 定理4.1的证明

证明. 令$\delta = n^{-r/(2r+1)} + a_n$, $\beta = \beta_0 + \delta T_1$, $\theta = \theta_0 + \delta T_2$, $\nu = \nu_0 + \delta T_3$, $T = (T_1^T, T_2^T, T_3^T)^T$. 首先, 将证明$\forall \epsilon > 0, \exists C > 0$, $P\{\inf_{\|T\|=C} F(\beta, \theta, \nu) > F(\beta_0, \theta_0, \nu_0)\} \ge 1 - \epsilon$.

记$D(\beta,\theta,\nu)=L^{-1}\{F(\beta,\theta,\nu)-F(\beta_0,\theta_0.\nu_0)\}$, 有

$$\begin{aligned}
D(\beta,\theta,\nu) &= \frac{1}{L}\sum_{i=1}^{n}[(T_1^TX_i+T_2^T\Psi(\tilde{Z}_i)+T_3^T\Psi_0(U_i))(-2\delta Y_i)\\
&\quad +2\delta(\beta_0^TX_i+\theta_0^T\Psi(\tilde{Z}_i)+\nu_0^T\Psi_0(U_i))(T_1^TX_i+T_2^T\Psi(\tilde{Z}_i)+T_3^T\Psi_0(U_i))\\
&\quad +\delta^2(T_1^TX_i+T_2^T\Psi(\tilde{Z}_i)+T_3^T\Psi_0(U_i))^2]+\frac{n}{L}\sum_{j=1}^{K_0}(p_{\lambda_j}(\|\theta_j\|_2)-p_{\lambda_j}(\|\theta_{j0}\|_2))\\
&\geq -\frac{2\delta}{L}\sum_{i=1}^{n}(\zeta_{K_0i}+R(\tilde{Z}_i,U_i))(T_1^TX_i+T_2^T\Psi(\tilde{Z}_i)+T_3^T\Psi_0(U_i))\\
&\quad +\frac{\delta^2}{L}\sum_{i=1}^{n}(T_1^TX_i+T_2^T\Psi(\tilde{Z}_i)+T_3^T\Psi_0(U_i))^2+\frac{n}{L}\sum_{j=1}^{s}(p_{\lambda_j}(\|\theta_j\|_2)-p_{\lambda_j}(\|\theta_{j0}\|_2))\\
&\equiv I_1+I_2+I_3,
\end{aligned}$$

其中, $R(\tilde{Z}_i,U_i)=\sum_{j=1}^{K_0}R_j(\tilde{Z}_i)+R_0(U_i)$, $R_j(\tilde{Z}_i)=g_j(\tilde{Z}_i^{(j)})-\theta_j^T\Psi_j(\tilde{Z}_i^{(j)}), j=1,\cdots,K_0$, $R_0(U_i)=f(U_i)-\nu^T\Psi_0(U_i)$.

根据条件C4.1和C4.2, $g_j(\tilde{Z}^{(j)})$的最大均方偏为

$$\|R_j(\tilde{Z}^{(j)})\|^2=\sum_{k=L+1}^{\infty}\theta_{jk}^2\leq\sum_{k=L+1}^{\infty}\theta_{jk}^2(\frac{k}{L})^{2r}\leq ML^{-2r},$$

因此$\|R_j(\tilde{Z}^{(j)})\|=O(L^{-r})$. 类似地, 有$\|R_0(U)\|=O(L^{-r})$. 这样,

$$\sum_{i=1}^{n}R(\tilde{Z}_i,U_i)(T_1^TX_i+T_2^T\Psi(\tilde{Z}_i)+T_3^T\Psi_0(U_i))=O_p(nK_0L^{-r}\|T\|).$$

注意到$E(\zeta_{K_0}|X,\tilde{Z},U)=0$, 根据Zhao和Xue(2009), 有

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\zeta_{K_0i}(T_1^TX_i+T_2^T\Psi(\tilde{Z}_i)+T_3^T\Psi_0(U_i))=O_p(\|T\|).$$

所以

$$I_1=-\frac{2\delta}{L}[O_p(nK_0L^{-r}\|T\|)+O_p(\sqrt{n}\|T\|)]=O_p(1+n^{\frac{r}{2r+1}}a_n)\|T\|.$$

类似地, 有

$$0<I_2=O_p(nL^{-1}\delta^2\|T\|^2)=O_p(1+2n^{\frac{r}{2r+1}}a_n+n^{\frac{2r}{2r+1}}a_n^2)\|T\|^2.$$

通过适当地选取一个充分大的数C, 在$\|T\|=C$上, 相比于I_1, I_2一致地占主导地位.

采用Taylor展开,

$$\begin{aligned}
I_3 &= \sum_{j=1}^{s}\{\frac{n}{L}p'_{\lambda_j}(\|\theta_{j0}\|_2)(\|\theta_{j0}\|_2)'\delta T_{2j}\\
&\quad +\frac{n}{2L}\{p''_{\lambda_j}(\|\theta_{j0}\|_2)[(\|\theta_{j0}\|_2)'\delta T_{2j}]^2+p'_{\lambda_j}(\|\theta_{j0}\|_2)[\delta^2T_{2j}^T(\|\theta_{j0}\|_2)''T_{2j}]\}(1+o(1))\}\\
&\equiv I_{31}+I_{32}.
\end{aligned}$$

满足

$$|I_{31}| \le \frac{n}{L}\delta a_n l_1 \sum_{j=1}^{s} \|T_{2j}\| \le \sqrt{s}\frac{n}{L}\delta a_n l_1 \|T\| = O_p(n^{\frac{r}{2r+1}} a_n + n^{\frac{2r}{2r+1}} a_n^2)\|T\|,$$

$$I_{32} \le \frac{n}{2L}\delta^2 \{b_n l_2 \sum_{j=1}^{s} \|T_{2j}\|^2 + a_n l_3 \sum_{j=1}^{s} \|T_{2j}\|^2\} \le \frac{n}{2L}\delta^2 (b_n l_2 + a_n l_3)\|T\|^2,$$

其中, l_1, l_2和l_3是一些整数. 可以发现, 在$\|T\| = C$上, 相比于I_{31}, I_2也是一致地占主导地位. 在条件C4.4下, 有

$$0 < I_{32} \le O_p(1 + 2n^{\frac{r}{2r+1}} a_n + n^{\frac{2r}{2r+1}} a_n^2)\|T\|^2.$$

因此, 通过适当地选取一个充分大的C, $P\{\inf_{\|T\|=C} F(\beta, \theta, \nu) > F(\beta_0, \theta_0, \nu_0)\} \ge 1 - \epsilon$, 这意味着, 至少有$1-\epsilon$的概率, 在球$\{\beta_0 + \delta T_1 : \|T_1\| \le C\}$内$F(\beta, \theta, \nu)$存在一个局部最小点. 记这个局部最小点为$\hat{\beta}$, 则有

$$\|\hat{\beta} - \beta_0\| = O_p(\delta) = O_p(n^{-r/(2r+1)} + a_n).$$

同理, 在球$\{\theta_0 + \delta T_2 : \|T_2\| \le C\}$内, $F(\beta, \theta, \nu)$存在一个局部最小点, 而且这个局部最小点$\hat{\theta}$满足

$$\|\hat{\theta} - \theta_0\| = O_p(n^{-r/(2r+1)} + a_n).$$

对于非参项$g(\cdot)$, 注意到

$$\begin{aligned} \|\hat{g}_j - g_{j0}\|^2 &= E\{\hat{g}_j(\tilde{Z}^{(j)}) - g_{j0}(\tilde{Z}^{(j)})\}^2 \\ &= E\{\mathbf{\Psi}_j(\tilde{Z}^{(j)})\hat{\theta}_j - \mathbf{\Psi}_j(\tilde{Z}^{(j)})\theta_{j0} + R_{j0}(\tilde{Z}^{(j)})\}^2 \\ &\le 2E\{\mathbf{\Psi}_j(\tilde{Z}^{(j)})\hat{\theta}_j - \mathbf{\Psi}_j(\tilde{Z}^{(j)})\theta_{j0}\}^2 + 2E\{R_{j0}(\tilde{Z}^{(j)})\}^2 \\ &= 2(\hat{\theta}_j - \theta_{j0})^T(\hat{\theta}_j - \theta_{j0}) + 2E\{R_{j0}(\tilde{Z}^{(j)})\}^2, \end{aligned}$$

由于$\|R_j(\tilde{Z}^{(j)})\| = O(L^{-r})$, 所以

$$E\{R_{j0}(\tilde{Z}^{(j)})\}^2 = O_p(n^{-2r/(2r+1)}).$$

这样, 可以得到

$$\|\hat{g}_j - g_{j0}\| = O_p(n^{-r/(2r+1)} + a_n).$$

类似地, 存在一个局部最小点$\hat{\nu}$满足$\|\hat{\nu} - \nu_0\| = O_p(n^{-r/(2r+1)} + a_n)$. 从而$\|\hat{f} - f_0\| = O_p(n^{-r/(2r+1)} + a_n)$.

□

§4.5.3 定理4.2的证明

证明. 根据$p'_\lambda(w)$的形式, 当$\lambda_{max} \to 0$时, 对于大的n有$a_n = 0$. 根据定理4.1, 只需要证明: 对于任何满足$\|\beta - \beta_0\| = O_p(n^{-r/(2r+1)})$的$\beta$, 满足$\|\theta_j - \theta_{j0}\| = O_p(n^{-r/(2r+1)})$的$\theta_j$, $j = 1, \cdots, s$, 满足$\|\nu - \nu_0\| = O_p(n^{-r/(2r+1)})$的$\nu$, 对于给定的小量$\iota_n = Cn^{-r/(2r+1)}$, 当$n \to \infty$时, 以趋于1的概率, 有

$$\frac{\partial F(\beta,\theta,\nu)}{\partial \theta_j} > 0, \text{对于} \quad 0 < \theta_j < \iota_n, j = s+1, \cdots, K_0$$

$$\frac{\partial F(\beta,\theta,\nu)}{\partial \theta_j} < 0, \text{对于} -\iota_n < \theta_j < 0, j = s+1, \cdots, K_0.$$

因此$F(\beta,\theta,\nu)$在$\theta_j = 0, j = s+1, \cdots, K_0$处达到最小值.

实际上,

$$\begin{aligned}
\tfrac{\partial F(\beta,\theta,\nu)}{\partial \theta_j} &= -2\sum_{i=1}^{n} \boldsymbol{\Psi}_j(\tilde{Z}_i^{(j)})(Y_i - \beta^T X_i - \theta^T \boldsymbol{\Psi}(\tilde{Z}_i) - \nu^T \boldsymbol{\Psi}_0(U_i)) + np'_{\lambda_j}(\|\theta_j\|_2)(\|\theta_j\|_2)' \\
&= -2\sum_{i=1}^{n} \boldsymbol{\Psi}_j(\tilde{Z}_i^{(j)})(\zeta_{K0i} + R(\tilde{Z}_i, U_i)) - 2\sum_{i=1}^{n} \boldsymbol{\Psi}_j(\tilde{Z}_i^{(j)}) X_i^T(\beta_0 - \beta) \\
&\quad -2\sum_{i=1}^{n} \boldsymbol{\Psi}_j(\tilde{Z}_i^{(j)}) \boldsymbol{\Psi}^T(\tilde{Z}_i)(\theta_0 - \theta) - 2\sum_{i=1}^{n} \boldsymbol{\Psi}_j(\tilde{Z}_i^{(j)}) \boldsymbol{\Psi}_0^T(U_i)(\nu_0 - \nu) \\
&\quad +np'_{\lambda_j}(\|\theta_j\|_2)(\|\theta_j\|_2)' \\
&= n\lambda_j\{O_p(\lambda_j^{-1} n^{-r/(2r+1)}) + \lambda_j^{-1} p'_\lambda(\|\theta_j\|_2)\frac{\theta_j}{\|\theta_j\|_2}\}.
\end{aligned}$$

根据条件$\liminf_{n\to\infty} \liminf_{\|\theta_{j0}\|_2 \to 0} \lambda_j^{-1} p'_{\lambda_j}(\|\theta_{j0}\|_2) = C > 0$和$\lambda_j n^{\frac{r}{2r+1}} > \lambda_{min} n^{\frac{r}{2r+1}} \to \infty$, 有$\frac{\partial F(\beta,\theta,\nu)}{\partial \theta_j} = O_p(n\lambda_j \frac{\theta_j}{\|\theta_j\|_2})$. 因此, 导数的符号由$\theta_j$决定.

因此, 以趋于1的概率, 有$\hat{\theta}_j = 0, j = s+1, \cdots, K_0$. 从而在$\sup_Z \|\boldsymbol{\Psi}_j(\tilde{Z}^{(j)})\| = O(1)$条件下,有$\hat{g}_j(\tilde{Z}^{(j)}) = \hat{\theta}_j^T \boldsymbol{\Psi}_j(\tilde{Z}^{(j)}) \equiv 0, j = s+1, \cdots, K_0$.

□

§4.5.4 定理4.3的证明

证明. 根据定理4.1和4.2, 当$n \to \infty$时, 以趋于1的概率, 有$F(\beta,\theta,\nu)$在$\hat{\beta}$、$(\theta^{*T}, 0)^T$和$\hat{\nu}$处达到局部最小值. 令$F_{1n}(\beta,\theta,\nu) = \partial F(\beta,\theta,\nu)/\partial\beta$, $F_{2n}(\beta,\theta,\nu) = \partial F(\beta,\theta,\nu)/\partial\theta^*$, $F_{3n}(\beta,\theta,\nu) = \partial F(\beta,\theta,\nu)/\partial\nu$, 则有

$$\frac{1}{n} F_{1n}(\hat{\beta}, (\hat{\theta}^{*T}, 0)^T, \hat{\nu}) = \frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}^T X_i - \hat{\theta}^{*T} \boldsymbol{\Psi}_i^* - \hat{\nu}^T \boldsymbol{\Psi}_{0i}) = 0, \qquad (4.5.1)$$

$$\frac{1}{n}F_{2n}(\hat{\beta},(\hat{\theta}^{*T},0)^T,\hat{\nu})=\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Psi}_i^*(Y_i-\hat{\beta}^TX_i-\hat{\theta}^{*T}\mathbf{\Psi}_i^*-\hat{\nu}^T\mathbf{\Psi}_{0i})+\sum_{j=1}^{s}p'_{\lambda_j}(\|\hat{\theta}_j\|_2)\frac{\hat{\theta}_j}{\|\hat{\theta}_j\|_2}=0, \tag{4.5.2}$$

$$\frac{1}{n}F_{3n}(\hat{\beta},(\hat{\theta}^{*T},0)^T,\hat{\nu})=\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Psi}_{0i}(Y_i-\hat{\beta}^TX_i-\hat{\theta}^{*T}\mathbf{\Psi}_i^*-\hat{\nu}^T\mathbf{\Psi}_{0i})=0. \tag{4.5.3}$$

从式(4.5.1), 可以得到

$$\frac{1}{n}\sum_{i=1}^{n}X_i((\beta_0-\hat{\beta})^TX_i+(\theta_0^*-\hat{\theta}^*)^T\mathbf{\Psi}_i^*+(\nu_0-\hat{\nu})^T\mathbf{\Psi}_{0i}+R^*(\tilde{Z}_i,U_i)+\zeta_{K_0i})=0 \tag{4.5.4}$$

其中$R^*(\tilde{Z}_i,U_i)=\sum_{j=1}^{s}R_j^*(\tilde{Z}_i)+R_0(U_i)$. 应用Taylor展开, 有

$$p'_{\lambda_j}(\|\hat{\theta}_j\|_2)=p'_{\lambda_j}(\|\hat{\theta}_{j0}\|_2)+\{p''_{\lambda_j}(\|\hat{\theta}_{j0}\|_2)\frac{\hat{\theta}_j}{\|\hat{\theta}_j\|_2}+o_p(1)\}(\hat{\theta}^*-\theta_0^*).$$

进一步, 条件C4.6意味着$p''_{\lambda_j}(\|\hat{\theta}_{j0}\|_2)=o_p(1)$, 注意到$p'_{\lambda_j}(\|\hat{\theta}_{j0}\|_2)=0$当$\lambda_{max}\to 0$时, 则有$p'_{\lambda_j}(\|\hat{\theta}_j\|_2)=o_p(\hat{\theta}^*-\theta_0^*)$. 因此, 从式(4.5.2)能够得到

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Psi}_i^*((\beta_0-\hat{\beta})^TX_i+(\theta_0^*-\hat{\theta}^*)^T\mathbf{\Psi}_i^*+(\nu_0-\hat{\nu})^T\mathbf{\Psi}_{0i}+R^*(\tilde{Z}_i,U_i)+\zeta_{K_0i})+o_p(\theta_0^*-\hat{\theta}^*)=0. \tag{4.5.5}$$

令$\Phi_n=n^{-1}\sum_{i=1}^{n}\mathbf{\Psi}_i^*\mathbf{\Psi}_i^{*T}$, $\Gamma_n=n^{-1}\sum_{i=1}^{n}\mathbf{\Psi}_i^*X_i^T$, $\Pi_n=n^{-1}\sum_{i=1}^{n}\mathbf{\Psi}_i^*\mathbf{\Psi}_{0i}^T$, 则有

$$\hat{\theta}^*-\theta_0^*=[\Phi_n+o_p(1)]^{-1}\{\Gamma_n(\beta_0-\hat{\beta})+\Pi_n(\nu_0-\hat{\nu})\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Psi}_i^*(R^*(\tilde{Z}_i)+\zeta_{K_0i})\}. \tag{4.5.6}$$

将式(4.5.6)代入式(4.5.4), 有

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\ & X_i\{(\mathbf{\Psi}_{0i}-\Pi_n^T\Phi_n^{-1}\mathbf{\Psi}_i^*)^T(\hat{\beta}-\beta_0)+(X_i-\Gamma_n^T\Phi_n^{-1}\mathbf{\Psi}_i^*)^T(\hat{\nu}-\nu_0)\}+o_p(\hat{\beta}-\beta_0)\\ &+o_p(\hat{\nu}-\nu_0)\ =\ \frac{1}{n}\sum_{i=1}^{n}X_i\{\zeta_{K_0i}+R^*(\tilde{Z}_i,U_i)-\mathbf{\Psi}_i^{*T}(\Phi_n^{-1}+o_p(1))\Lambda_n\},\end{aligned} \tag{4.5.7}$$

其中$\Lambda_n=n^{-1}\sum_{i=1}^{n}\mathbf{\Psi}_i^*(R^*(\tilde{Z}_i,U_i)+\zeta_{K_0i})$.

从式(4.5.3), 可以得到

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{\Psi}_{0i}((\beta_0-\hat{\beta})^TX_i+(\theta_0^*-\hat{\theta}^*)^T\mathbf{\Psi}_i^*+(\nu_0-\hat{\nu})^T\mathbf{\Psi}_{0i}+R^*(\tilde{Z}_i,U_i)+\zeta_{K_0i})=0 \tag{4.5.8}$$

将式(4.5.6)代入式(4.5.8), 有

$$\frac{1}{n}\sum_{i=1}^{n} \Psi_{0i}\{(X_i-\Gamma_n^T\Phi_n^{-1}\Psi_i^*)^T(\hat{\beta}-\beta_0)+(\Psi_{0i}-\Pi_n^T\Phi_n^{-1}\Psi_i^*)^T(\hat{\nu}-\nu_0)\}+o_p(\hat{\beta}-\beta_0)$$

$$+o_p(\hat{\nu}-\nu_0) = \frac{1}{n}\sum_{i=1}^{n}\Psi_{0i}\{\zeta_{K0i}+R^*(\tilde{Z}_i,U_i)-\Psi_i^{*T}(\Phi_n^{-1}+o_p(1))\Lambda_n\}, \tag{4.5.9}$$

注意到

$$n^{-1}\sum_{i=1}^{n}\Pi_n^T\Phi_n^{-1}\Psi_i^*\{\Psi_{0i}^T-\Psi_i^{*T}\Phi_n^{-1}\Pi_n\}=0$$

和

$$n^{-1}\sum_{i=1}^{n}\Pi_n^T\Phi_n^{-1}\Psi_i^*\{\zeta_{K0i}+R^*(\tilde{Z}_i,U_i)-\Psi_i^{*T}\Phi_n^{-1}\Lambda_n\}=0,$$

式(4.5.9)可以被改写为

$$\frac{1}{n}\sum_{i=1}^{n}\Psi_{0i}\check{X}_i^T(\hat{\beta}-\beta_0)+o_p(\hat{\beta}-\beta_0)+\frac{1}{n}\sum_{i=1}^{n}\check{\Psi}_{0i}\check{\Psi}_{0i}^T(\hat{\nu}-\nu_0)+o_p(\hat{\nu}-\nu_0)$$

$$=\frac{1}{n}\sum_{i=1}^{n}\check{\Psi}_{0i}\{\zeta_{K0i}+R^*(\tilde{Z}_i,U_i)-\Psi_i^{*T}(\Phi_n^{-1}+o_p(1))\Lambda_n\},$$

其中$\check{X}_i=X_i-\Gamma_n^T\Phi_n^{-1}\Psi_i^*$, $\check{\Psi}_{0i}=\Psi_{0i}-\Pi_n^T\Phi_n^{-1}\Psi_i^*$. 令$\Xi_n=n^{-1}\sum_{i=1}^{n}\check{\Psi}_{0i}\check{\Psi}_{0i}^T$, 则有

$$\hat{\nu}-\nu_0=\Xi_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\Psi_{0i}\check{X}_i^T(\beta_0-\hat{\beta})+\Xi_n^{-1}\frac{1}{n}\sum_{i=1}^{n}\check{\Psi}_{0i}(\zeta_{K0i}+R^*(\tilde{Z}_i,U_i)-\Psi_i^{*T}\Phi_n^{-1}\Lambda_n))+o_p(\hat{\beta}-\beta_0). \tag{4.5.10}$$

将式(4.5.10)代入式(4.5.7), 同时注意到

$$n^{-1}\sum_{i=1}^{n}\Gamma_n^T\Phi_n^{-1}\Psi_i^*\{X_i-\Psi_i^{*T}\Phi_n^{-1}\Gamma_n\}=0,$$

$$n^{-1}\sum_{i=1}^{n}\Gamma_n^T\Phi_n^{-1}\Psi_i^*\{\zeta_{K0i}+R^*(\tilde{Z}_i)-\Psi_i^{*T}\Phi_n^{-1}\Lambda_n\}=0,$$

$$n^{-1}\sum_{i=1}^{n}\check{\Psi}_{0i}X_i^T=n^{-1}\sum_{i=1}^{n}\Psi_{0i}\check{X}_i^T.$$

容易证明

$$\begin{aligned}
&(\check{\Phi}_n-\Upsilon_n^T\Xi_n^{-1}\Upsilon_n+o_p(1))\sqrt{n}(\hat{\beta}-\beta_0)\\
=\;&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\check{X}_i-\Upsilon_n^T\Xi_n^{-1}\check{\Psi}_{0i})(\zeta_{K0i}+R^*(\tilde{Z}_i,U_i)-\Psi_i^{*T}[\Phi_n^{-1}+o_p(1)]\Lambda_n)\\
=\;&\tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n}\check{X}_i\zeta_{K0i}-\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\check{X}_i\Psi_i^{*T}[\Phi_n^{-1}+o_p(1)]\Lambda_n+\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n}\check{X}_iR^*(\tilde{Z}_i,U_i)\\
\equiv\;&I_1+I_2+I_3,
\end{aligned} \tag{4.5.11}$$

其中$\check{\Phi}_n = n^{-1}\sum_{i=1}^n \check{X}_i\check{X}_i^T$, $\Upsilon_n = n^{-1}\sum_{i=1}^n \check{\mathbf{\Psi}}_{0i}X_i^T$, $\tilde{X}_i = \check{X}_i - \Upsilon_n^T\Xi_n^{-1}\check{\mathbf{\Psi}}_{0i}$.

利用中心极限定理(Central Limit Theorem), 能够得到

$$I_1 \longrightarrow_d N(0, \sigma_{K_0}^2\Sigma_0),$$

其中, "$\longrightarrow_d$"表示依分布收敛,

$$\Sigma_0 = E(XX^T) - E(X\mathbf{\Psi}^{*T})E^{-1}(\mathbf{\Psi}^*\mathbf{\Psi}^{*T})E(\mathbf{\Psi}^*X^T) - E(X\check{\Psi}_0^T)E^{-1}(\check{\Psi}_0\check{\Psi}_0^T)E(\check{\Psi}_0X^T).$$

另外, 注意到$\sum_{i=1}^n \check{X}_i\mathbf{\Psi}_i^{*T} = 0$, $\sum_{i=1}^n \check{\Psi}_{0i}\mathbf{\Psi}_i^{*T} = 0$, 则有$I_2 = 0$. 进一步, 得到

$$\begin{aligned} I_3 &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\{X_i - E(\Gamma_n^T)E^{-1}(\Phi_n)\mathbf{\Psi}_i^*\}R^*(\tilde{Z}_i, U_i) \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^n\{E(\Gamma_n^T)E^{-1}(\Phi_n) - \Gamma_n^T\Phi_n^{-1}\}\mathbf{\Psi}_i^*R^*(\tilde{Z}_i, U_i) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \Upsilon_n^T\Xi_n^{-1}\check{\Psi}_{0i}R^*(\tilde{Z}_i, U_i) \\ &\equiv I_{31} + I_{32} + I_{33}. \end{aligned}$$

由于$E\{[X_i - E(\Gamma_n^T)E^{-1}(\Phi_n)\mathbf{\Psi}_i^*]\mathbf{\Psi}_i^{*T}\} = 0$, 根据Zhao和Xue(2009), 则有

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n(X_i - E(\Gamma_n^T)E^{-1}(\Phi_n)\mathbf{\Psi}_i^*)\mathbf{\Psi}_i^{*T} = O_p(1).$$

结合$\|\mathbf{\Psi}_j(\tilde{Z}^{(j)})\| = O(1)$和$\|R(\tilde{Z}, U)\| = o(1)$, 得到$I_{31} = o_p(1)$. 类似地, $I_{32} = o_p(1)$. 注意到$\frac{1}{\sqrt{n}}\sum_{i=1}^n \Upsilon_n^T\Xi^{-1}\mathbf{\Psi}_{0i}^*\mathbf{\Psi}_i^{*T} = 0$, 同理可得到$I_{33} = o_p(1)$. 因此, $I_3 = o_p(1)$.

根据大数定律(the law of large numbers), 有$\frac{1}{n}\sum_{i=1}^n \tilde{X}_i\tilde{X}_i^T \longrightarrow_p \Sigma_0$, 其中, "$\longrightarrow_p$"表示依概率收敛. 利用Slutsky定理, 则得到$\sqrt{n}(\hat{\beta} - \beta_0) \longrightarrow_d N(0, \sigma_{K_0}^2\Sigma_0^{-1})$.

□

第五章　高维线性回归模型中基于松弛投影的参数估计

§5.1　引言

本章中将考虑如下的高维线性回归模型

$$Y = \beta^T X + \gamma^T Z + \varepsilon, \tag{5.1.1}$$

其中, Y是1-维响应变量, p是固定有限的, q可趋于无穷; X和Z分别是服从正态分布的p-维和q-维协变量; β是感兴趣参数, ε是正态分布的随机误差项. 假设Z中仅有一小部分分量对响应变量Y有影响, 即$S_0 = \{1 \le j \le q : \gamma_j \neq 0\}$, 令$s_0 = |S_0|$表示集合内的元素个数, 有$s_0 \ll q$. 模型(5.1.1)包含了全部有关的变量, 从而是条件无偏的, 本章中我们称之为全模型. 令$(Y_i, X_i, Z_i), i = 1, \cdots, n$是来自全模型的$(Y, X, Z)$的独立同分布观测值, 记$n$维响应向量$\mathbb{Y} = (Y_1, \cdots, Y_n)^T$, $n \times p$设计矩阵$\mathbb{X} = (\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(p)}) = (X_1, \cdots, X_n)^T$, $n \times q$设计矩阵$\mathbb{Z} = (\mathbf{Z}^{(1)}, \cdots, \mathbf{Z}^{(q)}) = (Z_1, \cdots, Z_n)^T$. 本章的主要目标是估计参数$\beta$.

在高维数据中, 当协变量之间高度相关(highly correlated)时, 如果采用边际相关(marginal correlation)作为度量各协变量对响应变量的贡献, 其结果将变得不可信. 在Cho和Fryzlewicz (2012)的文章中, 提出了一种度量每一个协变量分量对响应变量的贡献的新方法. 该方法首先基于协变量的样本相关阵C, 对于任一个指标j, 筛选出与变量$X^{(j)}$相关性不可忽略(non-negligible)的指标集C_j. 记$\tilde{\mathbf{X}}_j$为由$\{\mathbf{X}^{(k)}, k \in C_j\}$作为列构成的矩阵, 对应于$X^{(j)}$的倾斜变量(tilted variable, 记作X_j^*)为$\mathbf{X}^{(j)}$到$\tilde{\mathbf{X}}_j$的正交补空间的投影, 对应于$X^{(j)}$和Y的关于调整因子s_j的倾斜相关(tilted correlation) 定义为$c_j^* = s_j^{-1} X_j^{*T} \mathbb{Y}$. 当$C_j = \emptyset$时, 对应于$X^{(j)}$和$Y$的倾斜相关也是边际相关, 此时若$X^{(j)}$达到最大的边际相关, 则$X^{(j)}$很可能与$Y$相关; 当$C_j \neq \emptyset$时, $X^{(j)}$与Y之间的高度边际相关可能是由$X^{(j)}$ 与$X^{(k)}, k \in C_j$ 之间的相关性导致, 这时, 通过计算变量$X^{(k)}, k \in C_j \cup \{j\}$对应的倾斜相关, 选取倾斜相关绝对值达到最大的变量作为可能与Y相关的变量. 但是, 这个方法依赖于几个条件, 特别是其中的条件3、条件5和条件6. 直观地, 条件3要求对于任一变量筛选出的与其相关的变量不能太多; 条件5要求$X^{(j)}$在其对应的倾斜变量X_j^*上的投影长度不能太小, 更不能为零; 条件6要求对任一满足$S_0 \subseteq C_j$的指标j, 响应变量$\mathbb{Y}$中参数部分$\mathbf{X}_{S_0}^T \beta_{S_0}$ 到$\tilde{\mathbf{X}}_j$的正交补空间的投影长度不能太小. 显然, 当协变量相关矩阵的非对角元上非零元素很多或分量之间存在线性相关时, 上述条件中的部分条件将不能满足.

Zhang和Zhang(2012)从变量投影的角度来看待传统意义上的线性模型最小二乘估计. 对于固定设计的线性模型, 将最小二乘估计改写为$\hat{\beta}_j = \mathbb{Y}^T X_j^{\perp} / \|X_j^{\perp}\|_2^2$, 其中$X_j^{\perp}$是$\mathbf{X}^{(j)}$到列空间$\mathbf{X}^{(-j)} = \{\mathbf{X}^{(k)}, k \neq j\}$的正交补空间的投影. 对于高维数据$\dim(X) > n$,

在一般情况下$\mathrm{rank}(\mathbf{X}^{(-j)}) = n$. 因此$X_j^{\perp} = 0$, 从而估计$\hat{\beta}_j$是没有定义的. 为了弥补这个缺点同时保留线性估计形式, Zhang和Zhang(2012)通过一个松弛投影(relaxed projection, 记为z_j)去替代$X_j^{\perp}$的作用. 实际上, 当$X_j^{\perp} = 0$时, $\mathbf{X}^{(j)}$与$\mathbf{X}^{(-j)}$之间线性相关, 即$\mathbf{X}^{(j)} = \mathbf{X}^{(-j)}\alpha_{-j}$. 更一般地, 当$\|X_j^{\perp}\|_2$很小时, 可假设$\mathbf{X}^{(j)}$与$\mathbf{X}^{(-j)}$之间满足线性关系, 即

$$\mathbf{X}^{(j)} = \mathbf{X}^{(-j)}\alpha_{-j} + e^{(j)}, \tag{5.1.2}$$

其中$e^{(j)}$与$\mathbf{X}^{(-j)}$独立. 取z_j为上述残余向量$e^{(j)}$的估计即可. 从而问题归结为线性模型(5.1.2)中参数α_{-j}的估计. 设$\hat{\alpha}_{-j}$为α_{-j}的一个相合估计, 可取$z_j = \mathbf{X}^{(j)} - \mathbf{X}^{(-j)}\hat{\alpha}_{-j}$.

本章的主要思路是基于以下子模型,

$$Y = \beta^T X + \eta, \tag{5.1.3}$$

将Cho和Fryzlewicz(2012)、Zhang和Zhang(2012)中的方法有机结合, 从而进行重新建模. 如果$\gamma^T E(Z|X) \neq 0$, 则$E(\eta|X = x)$是一个非零函数. 这样, 单纯基于子模型(5.1.3)的感兴趣参数的估计可能是不相合的. 所以, 首先对该子模型, 采用Cho和 Fryzlewicz (2012)的方法, 将协变量Z中与X相关的分量扩充到模型中, 得到一个调整模型. 然后, 计算X对应的倾斜变量. 如果倾斜变量的长度不太小, 可直接基于倾斜变量和调整模型进行参数β的估计; 如果存在长度很小的倾斜变量, 则需要按照Zhang和Zhang (2012)的方法对投影进行放松. 最后, 基于放松后的松弛投影和调整模型即可对参数β进行估计, 包括点估计和置信区间估计.

本章后续部分组织如下. 在§5.2节, 基于子模型提出了一个利用相关性的调整模型, 然后导出了一个基于倾斜变量的的参数估计, 并给出了其相合性结果. 第§5.3节提出了一个基于松弛投影的改进估计, 并给出了这个新估计的分布函数的渐近正态性和区间估计的覆盖率结果. 第§5.4节给出了一些模拟结果. 第§5.5节提供了一些理论结果的证明过程.

§5.2 基于投影的倾斜变量估计方法

假定p是有限的, $p \ll q$, $EZ = 0$, $\varepsilon \sim N(0, \sigma^2/n)$. 而且, 协变量$X$和参数$\beta$提前给定. 本章将构造出参数$\beta$的一个相合估计.

对于n维向量$\mathbf{u} \in \mathbf{R}^n$, 记$\mathrm{supp}(\mathbf{u}) = \{j : u_j \neq 0\}$, 其$l_0$,$l_r(0 < r < \infty)$和$l_\infty$范数分别定义为$\|\mathbf{u}\|_0 = |\mathrm{supp}(\mathbf{u})| = \sharp\{j : u_j \neq 0\}$, $\|\mathbf{u}\|_r = (\sum_{j=1}^n |u_j|^r)^{1/r}$, $\|\mathbf{u}\|_\infty = \max_{1\leq j\leq n} |u_j|$. 假设$\mathbb{X}, \mathbb{Z}$的每一个列向量的$l_2$范数均为1. 记$\mathcal{C}$为指标集$j = \{1, \cdots, q\}$的一个子集, 则$\mathbf{Z}_{\mathcal{C}}$表示设计矩阵$\mathbb{Z}$的一个$n \times |\mathcal{C}|$子矩阵, 其列由$\mathbf{Z}^{(j)}, j \in \mathcal{C}$组成.

取区间$(0, 1)$上一个小的数π_n做为阀值, 记$c_k = \|\mathbb{X}^T\mathbf{Z}^{(k)}\|_\infty$. 令子集$\mathcal{C}_X = \{1 \leq k \leq q : c_k > \pi_n\}$, 则$\{Z^{(k)}, k \in \mathcal{C}_X\}$为$Z$中与$X$相关的分量, 记作$Z_{\mathcal{C}_X}$(或简记为$Z_X$). 构造如

下的调整模型

$$Y = \beta^T X + \sum_{k\in\mathcal{C}_X} \gamma_k Z^{(k)} + \zeta \equiv \beta^T X + \gamma_X^T Z_{\mathcal{C}_X} + \zeta, \tag{5.2.1}$$

其中$\zeta = \varepsilon + \sum_{k\in\mathcal{J}\setminus\mathcal{C}_X} \gamma_k Z^{(k)}$.

记$\tilde{\mathbf{Z}}_X$表示设计矩阵$\mathbb{Z}$的一个$n \times |\mathcal{C}_X|$子矩阵, 其列由$Z_{\mathcal{C}_X}$ (即$Z^{(j)}, j \in \mathcal{C}_X$) 组成; Π_{Z_X}为到由$\mathbf{Z}^{(j)}, j \in \mathcal{C}_X$生成的空间的投影, 即$\Pi_{Z_X} \equiv \tilde{\mathbf{Z}}_X(\tilde{\mathbf{Z}}_X^T\tilde{\mathbf{Z}}_X)^{-1}\tilde{\mathbf{Z}}_X^T$. 协变量$\mathbb{X}$对协变量$\mathbb{Z}$的倾斜变量定义为$\mathbf{U}_0 \equiv (\mathbf{I}_n - \Pi_{Z_X})\mathbb{X}$. 从而, 倾斜变量$\mathbf{U}_0$与$\tilde{\mathbf{Z}}_X$的相关性降低为0.

基于样本观测点$(Y_i, X_i, Z_i), i = 1, \cdots, n$, 利用倾斜变量$\mathbf{U}_0$ 对模型(5.2.1)两边同乘, 可得到$\mathbf{U}_0^T\mathbb{Y} = \mathbf{U}_0^T\mathbb{X}\beta + \sum_{k\in\mathcal{C}_X} \mathbf{U}_0^T\mathbf{Z}^{(k)}\gamma_k + \mathbf{U}_0^T\zeta$. 根据倾斜变量$\mathbf{U}_0$的性质, 从而近似地有$\mathbf{U}_0^T\mathbf{Y} \approx \mathbf{U}_0^T\mathbf{X}\beta$. 因此, 当$p \times p$矩阵$\mathbf{U}_0^T\mathbb{X}$可逆时, 即得到参数$\beta$ 的估计如下:

$$\hat{\beta}_T = (\mathbf{U}_0^T\mathbb{X})^{-1}\mathbf{U}_0^T\mathbb{Y}. \tag{5.2.2}$$

为了讨论估计$\hat{\beta}_T$的理论性质, 对模型及变量做如下假设:

假设5.1. 参数γ中非零分量的个数s_0满足$s_0 = O(n^\delta)$,其中$\delta \in [0, 1/2)$.

假设5.2. 协变量Z的维数q满足$\log(q) = O(n^\theta)$, 其中$\theta \in [0, 1-2\varrho), \varrho \in (\delta, 1/2)$.

假设5.3. 阀值π_n选取为$\pi_n = C_1 n^{-\varrho}$, 其中常数$C_1 > 0$. 存在常数$C_2 > 0$ 使得$\mathcal{C}_X$ 中的元素个数满足$|\mathcal{C}_X| \le C_2 n^\xi$, 其中$\xi \in [0, 2(\varrho - \delta))$.

假设5.4. 存在常数$M \in (0, \infty)$和$\mu \in [0, \varrho - \delta - \xi/2)$, 兴趣参数$\beta$满足$\max_{1\le j\le p} |\beta_j| < M$和$n^\mu \min_{1\le j\le p} |\beta_j| \to \infty$, 参数$\gamma$的非零分量满足$\max_{j\in S_0} |\gamma_j| < M$.

假设5.5. 矩阵$\mathbf{I}_p - \mathbb{X}^T\Pi_{Z_X}\mathbb{X}$的特征值记为$v_1, v_2, \cdots, v_p$, 存在常数$a \in (0, 1)$, 使得$\min_{1\le j\le p}\{v_j\} > a$.

在假设5.1和假设5.2中, 允许参数γ的稀疏性s_0和维数q随着样本容量n增长. 假设5.3限定了阀值的阶数, 并要求协变量Z中与X相关的分量不能太多. 当参数β中分量为0或非零分量距离0太近时则难以识别, 因此假设5.4对于参数距离0的下界幅度做了要求, 但仍然允许随着n的增大绝对值最小的参数可以趋于0; 同时对参数的绝对值上界进行了限制, 目的是为了确保随着n的增大参数绝对值的最大值与最小值之比不要增长得太快. 假设5.5是为了排除变量之间的强共线性情形, 使得估计(5.2.2)有定义. 关于这些假设的详细解释以及满足假设的例子详见文献Cho和Fryzlewicz(2012).

条件5.1. 存在常数$C_3 > 0$, 对于任意的$k \in S_0 \setminus \mathcal{C}_X$, 有$|(\Pi_{Z_X}\mathbb{X})^T\mathbf{Z}^{(k)}| \le C_3 n^{-\varrho}$.

定理5.1 在假设5.1 - 假设5.5下, 如果条件5.1成立, 则以趋于1的概率, 有下述结论成立:

(a) $\hat{\beta}_T \to \beta$, 其中β为参数的真值.

(b) 对$1 \le j \le p$, 有$n^\mu \hat{\beta}_{Tj} \to \infty$.

条件5.1表明, 当协变量$\mathbb{X}$投影到由$\tilde{\mathbf{Z}}_X$的列向量扩展生成的空间时, 任何$\mathbf{Z}^{(k)}, k \in S_0$与$\mathbb{X}$不接近时(在$k \notin \mathcal{C}_X$意义下), 则也不会太接近于其投影$\Pi_{Z_X}\mathbb{X}$.

令子集$\mathcal{C}_l = \{1 \le k \le q : |\mathbf{Z}^{(k)T}\mathbf{Z}^{(l)}| > \pi_n\}$, 记$\mathcal{K} = S_0 \cup \mathcal{C}_X \cup (\cup_{j\in S_0}\mathcal{C}_j)$ 表示j 的一个子集, 使得$Z^{(k)}, k \in \mathcal{K}$对真模型$Y = \beta^T X + \Sigma_{l\in S_0}\gamma_l Z^{(l)} + \varepsilon$显著, 即$k \in S_0$, 或者与真模型的至少一个变量高度相关, 即$k \in \mathcal{C}_X \cup (\cup_{j\in S_0}\mathcal{C}_j)$.

条件5.2. 对任一$k \in \mathcal{K} \setminus \mathcal{C}_X$, 则$\mathcal{C}_k \cap \mathcal{C}_X = \emptyset$.

定理5.2 在假设5.1 - 假设5.5下, 如果条件5.2成立, 则以趋于1的概率, 有下述结论成立:

(a) $\hat{\beta}_T \to \beta$, 其中β为参数的真值.

(b) 对$1 \le j \le p$, 有$n^{\mu}\hat{\beta}_{Tj} \to \infty$.

条件5.2表明, 对任一$\mathbf{Z}^{(k)}, k \in \mathcal{K}$, 如果$\mathbf{Z}^{(k)}$不与$\mathbb{X}$ 高度相关, 则不存在$\mathbf{Z}^{(l)}, l \ne k$, 使得其与$\mathbf{Z}^{(k)}$和$\mathbb{X}$的样本相关系数同时大于阀值$\pi_n$.

基于式(5.2.2)可以对感兴趣参数β进行估计, 通过定理5.1 和定理5.2 可以了解估计$\hat{\beta}_T$ 的性质. 关于参数γ中非零分量的估计, 与文献Cho和Fryzlewicz(2012)中的方法相同, 此处不再赘述.

§5.3 基于松弛投影的改进方法

在§5.2节的估计方法中, 要求矩阵$\mathbf{U}_0^T\mathbb{X} = \mathbb{X}^T(\mathbf{I}_n - \Pi_{Z_X})\mathbb{X}$ 可逆, 实际对应为假设5.5成立. 但是, 当协变量X与Z之间的相关性很强时该矩阵将可能退化, 从而该方法将不可用. 具体计算时, 可采用矩阵$\mathbf{U}_0^T\mathbb{X}$的条件数进行判定矩阵是否退化. 因此, 下面考虑对投影进行放松. 为了叙述方便, 本节中首先考虑协变量X和Z是固定设计的情形, 假定随机误差项$\varepsilon \sim N(0, \sigma^2)$.

更一般地, 取$\mathbf{U} = \mathbb{X} - \mathbf{Z}_{\mathcal{C}_X}\hat{\theta}$, 估计量$\hat{\theta}$满足:

$$\hat{\theta} = \arg\min_{B\in\mathbb{R}^{d\times p}}\Big\{\frac{\mathrm{tr}((\mathbb{X} - \mathbf{Z}_{\mathcal{C}_X}B)^T(\mathbb{X} - \mathbf{Z}_{\mathcal{C}_X}B))}{2n} + \lambda\|B\|_1\Big\}, \tag{5.3.1}$$

其中, $d = |\mathcal{C}_X|$, $\mathrm{tr}(V)$表示矩阵V的迹. 为方便起见, 上式按LASSO方法进行估计, 取参数$\lambda = \sqrt{(2/n)\log(p * d)}$.

利用松弛投影后的倾斜变量$\mathbf{U}$, 对模型(5.2.1)两边同乘, 可得到$\mathbf{U}^T\mathbb{Y} = \mathbf{U}^T\mathbb{X}\beta + \sum_{k\in\mathcal{C}_X}\mathbf{U}^T\mathbf{Z}^{(k)}\gamma_k + \mathbf{U}^T\zeta$. 根据倾斜变量$\mathbf{U}$的性质, 从而近似地有$\mathbf{U}^T\mathbb{Y} \approx \mathbf{U}^T\mathbb{X}\beta$. 根据$\hat{\theta}$的取法, 可知$p \times p$矩阵$\mathbf{U}^T\mathbb{X}$可逆. 因此, 得到参数$\beta$的一个线性估计如下:

$$\hat{\beta}_L = (\mathbf{U}^T\mathbb{X})^{-1}\mathbf{U}^T\mathbb{Y}. \tag{5.3.2}$$

由于采用了松弛投影, 从而需要对估计$\hat{\beta}_L$进行纠偏. 假设$(\hat{\beta}^{(init)}, \hat{\gamma}^{(init)})$为模型(5.1.1)

中参数(β,γ)的一个初值, 满足

$$\{\hat{\beta}^{(init)},\hat{\gamma}^{(init)},\hat{\sigma}\}=\arg\min_{b\in\mathbb{R}^p,r\in\mathbb{R}^q,\sigma\in\mathbb{R}^1_+}\Big\{\frac{\|\mathbb{Y}-\mathbb{X}b-\mathbb{Z}r\|_2^2}{2n\sigma}+\frac{\sigma}{2}+\lambda_0\|r\|_1\Big\},\quad(5.3.3)$$

其中, 取参数$\lambda_0=\sqrt{(2/n)\log(q)}$. 进而, 构造参数$\beta$的一个新的纠偏估计如下:

$$\hat{\beta}_U=\hat{\beta}_L-(\mathbf{U}^T\mathbb{X})^{-1}\mathbf{U}^T\mathbb{Z}\hat{\gamma}^{(init)}.\quad(5.3.4)$$

则新估计$\hat{\beta}_U$的误差为

$$\hat{\beta}_U-\beta=(\mathbf{U}^T\mathbb{X})^{-1}\mathbf{U}^T(\varepsilon+\mathbb{Z}(\gamma-\hat{\gamma}^{(init)})).\quad(5.3.5)$$

其中, 误差项的方差正比于

$$V=(V_{jk})p\times p=(\mathbf{U}^T\mathbb{X})^{-1}(\mathbf{U}^T\mathbf{U})(\mathbb{X}^T\mathbf{U})^{-1}=\sigma^{-2}Var((\mathbf{U}^T\mathbb{X})^{-1}(\mathbf{U}^T\varepsilon)).\quad(5.3.6)$$

对于稀疏向量a, 例如$\|a\|_0=2$对两个回归系数进行比较, 一个近似的$(1-\alpha)100\%$ 置信区间为

$$|a^T\hat{\beta}_U-a^T\beta|\le\hat{\sigma}\Phi^{-1}(1-\alpha/2)(a^TVa)^{1/2},\quad(5.3.7)$$

其中, $\hat{\sigma}$为(5.3.3)中σ的估计, Φ 是标准正态分布函数.

令$\lambda_{univ}=\sqrt{(2/n)\log(q)}$. 假设参数$\gamma$在下述更一般意义下是稀疏的

$$\Sigma_{j=1}^q\min\{|\gamma_j|/(\sigma\lambda_{univ}),1\}\le s.\quad(5.3.8)$$

令$S=\{1\le j\le q:|\gamma_j|>\sigma\lambda_{univ}\}$, 记$|S|$表示集合$S$内的元素个数, 则有$|S|\le s\le|S_0|=s_0$.

以下讨论这个新的纠偏估计$\hat{\beta}_U$的理论性质. 假定$s\log(q)\ll n^{1/2}$, $\sigma^*=\|\varepsilon\|_2/\sqrt{n}$, 初始估计$\hat{\gamma}^{(init)}$满足

$$P\{\|\hat{\gamma}^{(init)}-\gamma\|_1\ge C_1s\sigma^*\sqrt{(2/n)\log(q/\delta)}\}\le\delta,\quad(5.3.9)$$

其中, C_1为固定常数, $\alpha_0/q^2\le\delta\le1$, $\alpha_0\in(0,1)$是提前给定的一个常数; 估计$\hat{\sigma}$满足

$$P\{|\hat{\sigma}/\sigma^*-1|\ge C_2s(2/n)\log(q/\delta)\}\le\delta,\quad(5.3.10)$$

其中, C_2为固定常数, $\alpha_0/q^2\le\delta\le1$.

令$r\ge1$, 定义CF因子(Compatibility Factor)如下

$$CF(r,S)=\inf\{(|S|\cdot u^T\mathbb{Z}^T(I_n-\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X})\mathbb{Z}u)^{1/2}/(n^{1/2}\|u_S\|_1):u\ne0,\|u_{S^c}\|_1\le r\|u_S\|_1\}.$$

定理5.3 假设$\{A,r,c_0\}$为固定常数, 满足$r>1$, 而且$A>(r+1)/(r-1)$. 令$\hat{\gamma}^{(init)}$和$\hat{\sigma}$是通过(5.3.3)式得到的初值, 其中$\lambda_0=A\sqrt{(2/n)\log(q/\delta)}$. 假设$CF^2(r,S)\ge c_0$和(5.3.8)

成立, 而且存在一个常数$\mu_* > 0$时使得$(s \vee 1)(2/n)\log(q/\delta) \leq \mu_*$. 则有, 对充分大的$n$、式(5.3.9)和式(5.3.10)成立, 常数C1和C2依赖于$\{A, r, c_0\}$. 记$\triangle = (\mathbf{U}^T\mathbf{U})^{1/2}$, $\tau = \triangle^{-1}\mathbf{U}^T\mathbb{X}$, $\xi = \max_{j\in\jmath}|\triangle^{-1}\mathbf{U}^T\mathbf{Z}^{(j)}|$. 令$\max(\delta_n', \delta_n'') \to 0_+$, 常数$\xi^* > 0$满足$\|\xi\|_\infty \leq \xi^*$, 则如下结论成立:

(a)如果$\xi^* C_1 s\sqrt{(2/n)\log(q/\delta)} \leq \delta_n'$, 则有

$$P\{|\tau(\hat{\beta}_U - \beta) - \triangle^{-1}\mathbf{U}^T\varepsilon| > \sigma^*\delta_n'\mathbf{1}\} \leq \delta. \tag{5.3.11}$$

(b)如果$C_2 s(2/n)\log(q/\delta) \leq \delta_n''$, 则$\forall \mathbf{t} \in \mathbb{R}^p$满足$\mathbf{t} \geq (1+\delta_n')/(1-\delta_n'')\mathbf{1}$, 有

$$P\{|\tau(\hat{\beta}_U - \beta)| \geq \hat{\sigma}\mathbf{t}\} \leq 2\mathbf{\Phi}_{n,p}(-(1-\delta_n'')\mathbf{t} + \delta_n'\mathbf{1}) + 2\delta, \tag{5.3.12}$$

其中, $\mathbf{\Phi}_{n,p}(\cdot)$是自由度为$n$的$p$维$t$分布函数.

(c)对任一固定的数$m \leq p$, 有

$$\lim_{n\to\infty}\inf_{a\in\mathfrak{A}(p,m)} P\{|a^T\hat{\beta}_U - a^T\beta| \leq \hat{\sigma}\Phi^{-1}(1-\alpha/2)(a^TVa)^{1/2}\} = 1-\alpha, \tag{5.3.13}$$

其中, $\Phi(\cdot)$为一元正态随机变量$N(0,1)$的分布函数, 向量族$\mathfrak{A}(p,m) = \{a : \|a\|_0 \leq m, \max_{1\leq j\leq p}|a_j|\xi_j \leq \xi^*\}$.

注5.1 定理5.3给出了估计$\hat{\beta}_U$的相关概率结果和渐近结果. 在实际使用过程中可按照Zhang和Zhang(2012)中的迭代方法寻找合适的参数λ以及ξ和τ, 以使估计$\hat{\beta}_U$具有小的均方误差. 特别地, 对于分量$j \in \jmath \setminus C_X$应着重考虑.

定理5.3的主要条件是$s(\log(q))/n \to 0$时$CF^2(r,S) \geq c_0$. 对于一类正态分布的随机设计矩阵, 这个条件是满足的, 详见以下定理.

定理5.4 假设协变量X和Z服从正态分布$N(0,\Sigma)$, 其中协方差阵$\Sigma \triangleq \begin{pmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{pmatrix}$. 记$\Sigma_{Z|X} \triangleq \Sigma_{ZZ} - \Sigma_{ZX}\Sigma_{XX}^{-1}\Sigma_{XZ}$, 满足对角元$\mathrm{diag}(\Sigma_{Z|X}) = I_{q\times q}$, 其特征值$\mathrm{eigenvalues}(\Sigma_{Z|X}) \subseteq [c_*, c^*]$, 其中$0 < c_* < c^* < \infty$是固定的, 而且$\Sigma_{Z|X}^{-1}$的行向量属于向量族$\mathfrak{B}(s, \lambda_{univ}) = \{\mathbf{b} \in \mathbb{R}^q : \Sigma_{j=1}^q \min(|b_j|/\lambda_{univ}, 1) \leq s\}$. 令$\hat{\gamma}^{(init)}$和$\hat{\sigma}$是通过(5.3.3)式得到的初值, 其中$\lambda_0 = A\sqrt{(2/n)\log(q/\delta)}$, $A > 1$. 当$s(\log(q))/n \to 0$时, 则以趋于1的概率有$CF^2(r,S) \geq c_0 > 0$. 从而, 以趋于1的概率定理5.3的(a),(b),(c)结论成立.

§5.4 数值模拟

本节将通过数值模拟来反映各方法的估计效果. 我们对新估计$\hat{\beta}_U$、基于松弛投影的线性估计$\hat{\beta}_L$、第4章中两步估计$\hat{\beta}_{TS}$、基于SCAD的估计$\hat{\beta}_{SCAD}$、基于子模型的估计$\hat{\beta}_S$等进行比较. 将比较相应点估计的效果, 另外对于新估计$\hat{\beta}_U$的区间估计效果还进行了展示.

我们将参考Zhang和Zhang(2012)进行数值模拟用例的设计. 在全模型(5.1.1)中, 取样本容量$n = 100$, 算法重复次数$m = 100$, 维数$p = 5$和$q = 300$, 惩罚参数$\lambda_{univ} = \sqrt{(2/n)\log(q)}$, 模型参数$\beta$和$\gamma$分别取为

$$\beta = (0.5, -3.5, 2.5, -1.5, 4.0)^T;$$

$$\gamma = (\gamma_j)_{j=1}^q, \gamma_j = \begin{cases} 3\lambda_{univ}, & j = 150, 180, 210, \cdots, 300; \\ 3\lambda_{univ}/(\mathrm{mod}(j, 100) + 1)^\alpha, & \text{otherwise.} \end{cases}$$

对于协变量X和Z, 首先按联合正态分布$N(0, \Sigma)$生成容量为n的样本矩阵$\mathbb{W}_{n\times(p+q)}$, 其中$\Sigma = (\rho^{|j-k|})_{j,k}$. 记集合$\imath = \{50, 100, 150, 200, 250\}$, 取由$\mathbb{W}$的第$\imath$ 列组成样本阵$\mathbb{X}$, 其余列组成样本阵$\mathbb{Z}$. 误差项ε服从正态分布$N(0, 0.3^2)$, $\mathbb{Y}$ 由模型(5.1.1) 生成. 本节将考虑(A),(B),(C),(D)四种情形, 分别对应$(\alpha, \rho) = (2, 1/5), (1, 1/5), (2, 4/5), (1, 4/5)$. 在算法的具体实现过程中, 我们参考了Sun和Zhang(2012), Yuan,Ekici和Lu等(2007)中的算法程序.

表5.1列出了四种情形下采用各估计方法对感兴趣参数β 的绝对值最小分量β_1的点估计结果, 表5.2列出了对参数β的绝对值最大分量β_5 的点估计结果, 包括偏差(bias), 标准差(std)和均方误差(MSE). 总体而言, 新估计$\hat{\beta}_U$各项指标的表现明显优于其它几种估计, 说明在这种用例环境下新方法确实是有效果的; 表5.1的数据与表5.2的数据没有明显的差别; 线性估计$\hat{\beta}_L$、两步估计$\hat{\beta}_{TS}$、SCAD估计$\hat{\beta}_{SCAD}$ 之间没有明显差别, 其在部分指标上的表现不同程度地优于基于子模型的估计$\hat{\beta}_S$. 由于α对参数γ的稀疏性影响, 情形(B)和(D)的各项指标总体上比情形(A)和(C) 的各项指标表现较差; 而ρ对协变量相关性强弱的影响, 对于估计$\hat{\beta}_{SCAD}$和$\hat{\beta}_S$有较大影响, 可以看出对应于这两个估计, 情形(C)和(D)的各项指标总体上比情形(A)和(B) 的各项指标表现较差.

取置信水平$a = 0.05$, 重复模拟10次, 每次估计100次, 从而计算各分量区间估计的覆盖率和平均覆盖率. 表5.3列出了四种情形下采用§5.3节中新方法对感兴趣参数β所有分量区间估计的平均覆盖率结果; 各分量的区间估计覆盖率结果见图5.1. 可以看出, 在情形(A)和(B)下各分量的区间估计覆盖率和平均覆盖率都接近于0.95, 情形(A)和(B)的结果没有较大差别; 在情形(C)和(D)下只有2个分量的覆盖率接近于0.95, 剩余3个分量的覆盖率约在0.75到0.85之间, 而且情形(C) 的结果优于情形(D)的结果. 这主要是由于情形(A)和(B)的协变量相关性较弱, 情形(C)和(D)的相关性较强, 同时由于α的影响在情形(D)下模型系数的稀疏性较弱导致.

表 5.1: 各情形下新方法与不同估计方法关于分量β_1的点估计结果.

No.	Items	$\hat{\beta}_U$	$\hat{\beta}_L$	$\hat{\beta}_{TS}$	$\hat{\beta}_{SCAD}$	$\hat{\beta}_S$
	bias	-0.0772	-0.1164	-0.0069	0.0349	0.0268
(A)	std	0.1740	0.3109	0.2574	0.2587	0.2588
	MSE	0.0362	0.1102	0.0663	0.0681	0.0677
	bias	0.0834	0.0487	0.0215	0.0399	0.0412
(B)	std	0.2335	0.3306	0.3809	0.3684	0.3630
	MSE	0.0614	0.1116	0.1455	0.1373	0.1334
	bias	-0.0194	-0.1083	-0.0115	0.0045	0.0036
(C)	std	0.1208	0.2548	0.2976	0.3243	0.3188
	MSE	0.0149	0.0766	0.0887	0.1052	0.1016
	bias	0.0145	-0.2074	-0.0099	0.2577	0.2254
(D)	std	0.1896	0.4910	0.3694	0.3532	0.3082
	MSE	0.0361	0.2841	0.1365	0.1911	0.1458

表 5.2: 各情形下新方法与不同估计方法关于分量β_5的点估计结果.

No.	Items	$\hat{\beta}_U$	$\hat{\beta}_L$	$\hat{\beta}_{TS}$	$\hat{\beta}_{SCAD}$	$\hat{\beta}_S$
	bias	-0.0071	0.1826	0.1619	0.1188	0.1266
(A)	std	0.1808	0.3206	0.2178	0.2577	0.2594
	MSE	0.0327	0.1362	0.0736	0.0805	0.0833
	bias	-0.0522	0.0035	0.0945	0.1300	0.1255
(B)	std	0.2167	0.3637	0.4105	0.3934	0.3923
	MSE	0.0497	0.1322	0.1774	0.1717	0.1697
	bias	0.0647	0.2037	0.3029	0.3982	0.4071
(C)	std	0.1389	0.2554	0.2963	0.3790	0.3872
	MSE	0.0234	0.1067	0.1796	0.3022	0.3157
	bias	0.1088	-0.0547	0.1522	0.4319	0.4776
(D)	std	0.1418	0.2746	0.3369	0.4175	0.4415
	MSE	0.0319	0.0784	0.1367	0.3608	0.4231

§5.5 附录

§5.5.1 定理5.1的证明

证明. 根据估计$\hat{\beta}_T$的形式, 有$\hat{\beta}_T = \beta + (\mathbf{U}_0^T\mathbb{X})^{-1}[\mathbf{U}_0^T\varepsilon + \mathbf{U}_0^T(\sum_{k\in S_0\setminus\mathcal{C}_X}\gamma_k\mathbf{Z}^{(k)})]$. 从而定理5.1的证明可分解为两步进行. 其中, 在第1步中首先控制倾斜变量$\mathbf{U}_0$与误差项ε 的

表 5.3: 各情形下新方法对于参数β所有分量区间估计的平均覆盖率结果.

	(A)	(B)	(C)	(D)
coverage	0.9404	0.9484	0.8722	0.8496

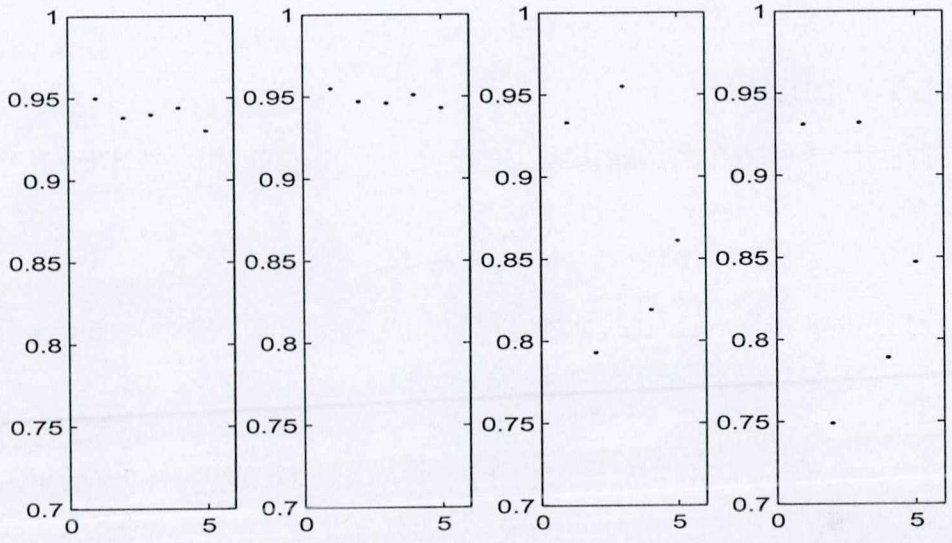

图 5.1: 新方法对于参数β各分量的区间估计覆盖率结果, 从左到右分别对应(A),(B),(C),(D)四种情形.

内积大小, 在第2步中控制第二项$I \equiv \sum_{k\in S_0\setminus\mathcal{C}_X}\gamma_k\mathbf{U}_0^T\mathbf{Z}^{(k)}$, 从而得到定理结果.

第1步: 对$\varepsilon \sim N_n(\mathbf{0}, n^{-1}\sigma^2\mathbf{I}_n)$, 以趋于1的概率, 对于单位范数的向量$\mathbf{Z}^{(1)},\cdots,\mathbf{Z}^{(q)} \in \mathbb{R}^n$, 有$\max_{1\le j\le q}|<\varepsilon,\mathbf{Z}^{(j)}>| \le \sigma\sqrt{2\log(q)/n}$. 根据假设5.2, 有$\sigma\sqrt{2\log(q)/n} \le Cn^{-\varrho}$; 根据假设5.5, 有$\|\mathbf{U}_{0j}\|_2 > \sqrt{a} > 0$. 从而, 定义$\wp = \max_j|\mathbf{U}_{0j}^T\varepsilon| < Cn^{-\varrho}$, 有$P(\wp)\to 1$.

第2步: 当$S_0\setminus\mathcal{C}_X = \emptyset$时, 显然$I=0$. 当$S_0\setminus\mathcal{C}_X \neq \emptyset$时, 对于$j\in S_0\setminus\mathcal{C}_X$, 有$\mathbf{U}_0^T\mathbf{Z}^{(j)} = \mathbb{X}^T(\mathbf{I}_n-\Pi_{Z_X})\mathbf{Z}^{(j)} = \mathbb{X}^T\mathbf{Z}^{(j)} - \mathbb{X}^T\Pi_{Z_X}\mathbf{Z}^{(j)}$, 从而根据假设5.3和条件5.1有$|\mathbf{U}_0^T\mathbf{Z}^{(j)}| \le C_1n^{-\varrho}\mathbf{1_p}+C_3n^{-\varrho}\mathbf{1_p} = C_*n^{-\varrho}\mathbf{1_p}$, 进而, 结合假设5.1, $|\sum_{j\in S_0\setminus\mathcal{C}_X}\gamma_j\mathbf{U}_0^T\mathbf{Z}^{(j)}| \le C_*n^{-\varrho+\delta}\mathbf{1_p}$. 这样, $|I| \le MC_*n^{-\varrho+\delta}\to 0$. 结合假设5.5,有$\hat{\beta}_T = \beta + O(n^{-\varrho+\delta})\mathbf{1_p}\to\beta$, 即估计$\hat{\beta}_T$是相合的. 进一步, 结合假设5.4, $n^\mu|\hat{\beta}_T| \ge n^\mu|\beta| - C'n^{\delta+\mu-\varrho}\mathbf{1_p} > n^\mu|\beta| - C'n^{-\xi/2}\mathbf{1_p}\to\infty$.

□

§5.5.2 定理5.2的证明

证明. 类似于定理5.1的证明, 只需证明I按照一定的速率趋于0. 根据条件5.2, 当$S_0\setminus\mathcal{C}_X\neq\emptyset$时, 对于$j\in S_0\setminus\mathcal{C}_X$, 有$\mathcal{C}_j\cap\mathcal{C}_X=\emptyset$. 根据假设5.3, 有$\|\Pi_{Z_X}\mathbf{Z}^{(j)}\|_2 \le C_*n^{-(\varrho-\xi/2)}$. 由于$\mathbf{U}_0^T\mathbf{Z}^{(j)} = \mathbb{X}^T(\mathbf{I}_n-\Pi_{Z_X})\mathbf{Z}^{(j)} = \mathbb{X}^T\mathbf{Z}^{(j)} - (\Pi_{Z_X}\mathbb{X})^T\Pi_{Z_X}\mathbf{Z}^{(j)}$, 而且$\mathbb{X}$和$\mathbb{Z}$的列向量范数

为1, 因此, $|\mathbf{U}_0^T\mathbf{Z}^{(j)}| \le C_1 n^{-\varrho}\mathbf{1_p} + C' n^{-\varrho+\xi/2}\mathbf{1_p}$. 结合假设5.1有, $|\sum_{j\in S_0\setminus\mathcal{C}_X}\gamma_j\mathbf{U}_0^T\mathbf{Z}^{(j)}| = O(n^{-\varrho+\delta+\xi/2})\mathbf{1_p}$. 这样, $|I| \le C_{\#} n^{-\varrho+\delta+\xi/2}\mathbf{1_p} \to 0$. 结合假设5.5,有$\hat{\beta}_T = \beta + O(n^{-\varrho+\delta+\xi/2})\mathbf{1_p} \to \beta$, 即估计$\hat{\beta}_T$是相合的. 进一步, $n^\mu|\hat{\beta}_T| \ge n^\mu|\beta| - C_{\#} n^{-\varrho+\delta+\xi/2+\mu}\mathbf{1_p} > n^\mu|\beta| - C_{\#} \to \infty$.

□

§5.5.3 定理5.3的证明

证明. 首先证明: 当(5.3.9)和(5.3.10)满足时, (5.3.11)、(5.3.12)、(5.3.13)成立. 根据式(5.3.5), 有

$$\tau(\hat{\beta}_U - \beta) - \triangle^{-1}\mathbf{U}^T\varepsilon = \triangle^{-1}\mathbf{U}^T\mathbb{Z}(\gamma - \hat{\gamma}^{(init)}).$$

由于松弛投影$\mathbf{U}$由$\mathbb{X}$和$\mathbb{Z}$决定, 根据上述可得到$|\tau(\hat{\beta}_U - \beta) - \triangle^{-1}\mathbf{U}^T\varepsilon| \le \max_{j\in J}|\triangle^{-1}\mathbf{U}^T\mathbf{Z}^{(j)}| \cdot \|\gamma - \hat{\gamma}^{(init)}\|_1 = \xi\|\gamma - \hat{\gamma}^{(init)}\|_1$. 结合式(5.3.9), 可得到式(5.3.11). 当$|\tau(\hat{\beta}_U - \beta) - \triangle^{-1}\mathbf{U}^T\varepsilon| \le \sigma^*\delta_n'\mathbf{1}$ 而且$|\hat{\sigma}/\sigma^* - 1| \le \delta_n''$时, $|\tau(\hat{\beta}_U - \beta)| \ge \hat{\sigma}\mathbf{t}$意味着$|\triangle^{-1}\mathbf{U}^T\varepsilon| \ge \hat{\sigma}\mathbf{t} - \sigma^*\delta_n'\mathbf{1} \ge \sigma^*((1-\delta_n'')\mathbf{t} - \delta_n'\mathbf{1})$. 由于$\varepsilon \sim N(0, \sigma^2\mathbf{I})$, 而且$\mathbf{U}$仅依赖于$\mathbb{X}$和$\mathbb{Z}$, 因此$\triangle^{-1}\mathbf{U}^T\varepsilon/\sigma^* \sim \sqrt{n}\varepsilon/\|\varepsilon\|_2$. 从而, 对于$x > 1$, 有$P\{|\triangle^{-1}\mathbf{U}^T\varepsilon| \ge \sigma^* x\} \le 2\mathbf{\Phi}_{n,p}(-x)$. 由于$\max(\delta_n', \delta_n'') \to 0_+$, V可看作近似的协方差阵, 因此有(5.3.13)成立.

下面证明: 在定理条件下, 当$CF^2(r,S) \ge c_0$成立时(5.3.9)和(5.3.10)成立. 记$\sigma^* = \|\varepsilon\|_2/\sqrt{n}$. 当$CF^2(r,S) \ge c_0$成立时, 由Sun和Zhang(2012)的定理2, 参考Zhang和Zhang (2012)的定理4, 类似可证明式(5.3.9)和(5.3.10) 成立.

□

§5.5.4 定理5.4的证明

证明. 由于$\Sigma_{Z|X}$的特征值一致有界, 根据Zhang和Zhang(2012)、Zhang和Huang(2008)、Ye和Zhang(2010), 存在正的常数$\{c_0, \delta_0, \delta_1, K\}$, 对于小的$(K+1)(s+1)\log(q)/n \le \delta_1$, 以不小于$1 - e^{-\delta_0 n}$的概率有$CF^2(r,S) \ge c_0 > 0$. 即定理5.3的条件成立, 从而当$n \to \infty$时以趋于1的概率定理结论成立.

□

参 考 文 献

[1] Acharya, D.P., Panda, G. (2008). A review of independent component analysis techniques and their applications. *IETE Tech. Rev.*, **25(6)**, 320-332.

[2] Adragni, K.P., Cook, R.D. (2009). Sufficient dimension reduction and prediction in regression. *Phil. Trans. R. Soc. A.*, **367**, 4385-4405.

[3] Akaike, H. (1973). Information theorey and an extension of the maximum likelihood principle. In *Proc. 2nd International Symposium on Information Theory* (V. Petrov and F. Csáki, eds.). 267-281. Akadémiai Kiadó, Budapest.

[4] Anderson, T.W. (2003). *An introduction to multivariate statistical analysis* (3rd edition). John Wiley & Sons.

[5] Barrios, E.B., Lansangan, J.R.G. (2010). Sparse Principal Component Regression. University of the Philippines Diliman (working paper).

[6] Bates, D.M., Watts, D.G. (1988). *Nonlinear regression analysis and its application.* John Wiley & Sons.

[7] Breiman, A., Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlations (with discussion). *J. Amer. Statist. Assoc.*, **80**, 580-619.

[8] Breiman, L. (1996). Heuristics of instablity and stabilization in model selection. *Ann. Statist.*, **24**, 2350-2382.

[9] Bühlmann, P., van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer-Verlag Berlin Heidelberg.

[10] Buja, A., Hastie, T.J., Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.*, **17**, 453-555.

[11] Cai, T., Liu, W.D. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.*, **106**, 672-684.

[12] Candès, E., Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35**, 2313-2351.

[13] Chamberlain. G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, **60**, 567-596.

[14] Chen, A.Y., Bickel, P.J. (2005). Consistent independent component analysis and prewhitening. *IEEE Trans. Sig. Proc.*, **53**, 3625-3632.

[15] Chen, A.Y., Bickel, P.J. (2006). Efficient independent component analysis. *Ann. Statist.*, **34**, 2825-2855.

[16] Chen, H. (1988). Convergence rates for parametric componens in a partially linear model. *Ann. Statist.*, **16**, 136-146.

[17] Cho, H., Fryzlewicz, P. (2012). High demensional variable selection via tilting. *J. Roy. Statist. Soc. B*, **74(3)**, 593-622.

[18] Chen, X., Zou, C., Cook, R.D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection, *Ann. Statist.*, **38**, 3696-3723.

[19] Claeskens, G., Carroll, R.J. (2007). An asymptotic theory for model selection inference in general semiparametric problems. *Biometrika*, **94**, 1-17.

[20] Cook, R.D., Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *J. Amer. Statistist. Assoc.*, **104**, 197-208.

[21] Cook, R.D., Ni, L.(2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statistist. Assoc.*, **100**, 410-428.

[22] Craven, P., Wahba, G.(1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics*, **31**, 377-403.

[23] Deng, G.H., Liang, H. (2010). Model averaging for semiparametric additive partial linear models. *Science China Mathematics*, **53**, 1363-1376.

[24] Dicker, L., Lin, X. (2009). A large sample analysis of the Dantzig slector and extensions. (manuscript).

[25] Draper, N. R., Smith, H. (1998). *Applied regression analysis* (wiley series in probability and statistics). Wiley-Interscience.

[26] Efromovich, S. (1999). *Nonparametric curve estimation: methods, theory, and applications.* Springer-Verlag New York.

[27] Ehsanes Saleh, A.K.Md. (2006). *Theory of preliminary test and stein-type estimation with applications.* John Wiley & Sons, Inc.

[28] Fan, J., Feng, Y., Song, R.(2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Amer. Statistist. Assoc.*, **106**, 544-557.

[29] Fan, J., Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman and Hall, London.

[30] Fan, J., Härdle, W, Mammen, E. (1998). Direct estimation of low-dimensional components in additive models. *Ann. Statist.*, **26**, 943-971.

[31] Fan, J., Jiang, J. (2005). Nonparametric inferences for additive models. *J. Amer. Statist. Assoc.*, **100**, 890-907.

[32] Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348-1360.

[33] Fan, J., Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Roy. Statist. Soc. B*, **70**, 849-911.

[34] Fan, J., Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, **20**, 101-148.

[35] Fan, J., Lv, J. (2011). Properties of non-concave penalized likelihood with NP-dimensionality. *IEEE Trans. Inf. Theory*, **57(8)**, 5467-5484.

[36] Fan, J., Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.*, **32**, 928-961.

[37] Fan, Y., Li, Q. (2003). A kernel-based method for estimating additive partially linear models. *Statistica Sinica*, **13**, 739-762.

[38] Friedman, J.H., Stuetzle, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.*, **87**, 998-1004.

[39] Fu, W.J. (1998). Penalized regressions: the Bridge versus the lasso. *J. Comp. Graph. Stat.*, **7**, 397-416.

[40] Gai, Y., Lin, L., Wang, X.(2011). Consistent inference for biased sub-model of high-dimensional partially linear model. *J. Statist. Plan. Infer.*, **141(5)**, 1888-1898.

[41] 高集体, 赵林城. (1992). 部分线性模型中的自适应估计. 中国科学, *A* 辑, **22(8)**, 791-803.

[42] Glad, I.K. (1998). Parametrically guided non-parametric regression. *Scand. J. Statist.*, **25** 649-668.

[43] Hall, A.R., Inoue, A. (2003). The large sample behavior of the generalized method of moments estimator in misspecified models. *J. Econometrics*, **114**, 361-394.

[44] Härdle, W., Hall, P. Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.*, **21**, 157-178.

[45] Härdle, W., Liang, H., Gao, J.T. (2000). *Partially linear models.* Physica Verlag.

[46] Härdle, W., Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, **21(4)**, 1926-1947.

[47] Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of Fit Tests.* Springer-Verlag, New York, Inc.

[48] Harville, D.(1976). Extension of the Gauss-Markov theorem to include the estimation of random effects.. *Ann. Statist.*, **4(2)**, 384-395.

[49] Hastie, T.J., Tibshirani, R. (1993) Varying-coefficient models. *J. Roy. Statist. Soc. B*, **55**, 757-796.

[50] Hjort, N.L., Claeskens, G. (2003). Frequentist model average estimators (with Discussion). *J. Amer. Statist. Assoc.*, **98**, 879-899.

[51] Hjort, N.L., Glad, I.K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.*, **23**, 882-904.

[52] Hjort, N.L., Jones, M.C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.*, **24**, 1619-1647.

[53] Hyvärinen, A., Karhunen, J., Oja, E. (2001). *Independent component analysis.* John Wiley & Sons.

[54] Hyvärinen, A., Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, **9(7)**, 1483-1492.

[55] James, G.M., Radchenko, P. (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika*, **96**, 323-337.

[56] Johnstone, I.M., Lu, A.Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.*, **104**, 682 - 693.

[57] Kariya, T., Kurata, H. (2004). *Generalized Least Squares.* John Wiley & Sons.

[58] Kim, Y., Choi, H., Oh, H.S. (2008). Smoothly clipped absolute deviation on high dimensions. *J. Amer. Statist. Assoc.*, **103**, 1665-1673.

[59] Kim, Y., Kwon, S. (2012). Global optimality of nonconvex penalized estimators. *Biometrika*, **99(2)**, 315-325.

[60] Kitamura, Y., Tripathi, G., Ahn, H. (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, **72**, 1667-1714.

[61] Kutner, M.H., Nachtsheim, C., Neter, J. (2004). *Applied linear regression models.* McGraw-Hill New York, NY.

[62] Leeb, H. (2009). Conditional predictive inference post model selection. *Ann. Statist.*, **37**, 2838-2876.

[63] Leeb, H., Poetscher, B.M. (2005). Model selection and inference: facts and fiction. *Econometric Theory*, **21**, 21-59.

[64] Leeb, H., Poetscher, B.M. (2008). Sparse estimators and the oracle property, or the return to the Hodges' estimator. *J. Econometrics.*, **142**, 201-211.

[65] Lewis, T., Odell, P. (1966). A generalization of the Gauss-Markov theorem. *J. Amer. Statist. Assoc.*, **61**, 1063-1066.

[66] 李根, 邹国华, 张新雨. (2012). 高维模型选择方法综述. 数理统计与管理, **31(4)**, 640-658.

[67] Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika*, **94**, 603-613.

[68] Li, L., Zhu, L., Zhu, L. (2011). Inference on the primary parameter of interest with the aid of dimension reduction estimation. *J. Roy. Statist. Soc. B*, **73(1)**, 59-80.

[69] Li, Q. (2000). Efficient estimation of additive partially linear models. *International Economic Review*, **41(4)**, 1073-1092.

[70] Li, R., Zhong, W., Zhu, L. (2012). Feature screening via distance correlation learning, *J. Amer. Statist. Assoc.*, **107:499**, 1129-1139.

[71] Lin, L., Cui, X., Zhu, L. (2008). An adaptive two-stage estimation method for additive models. *Scand. J. Statist.*, **36**, 248-269.

[72] Lin, L., Li, F. (2008). Stable and bias-corrected estimation for nonparametric regression models. *J. Nonparametr. Stat.*, **20**, 283-303.

[73] Lin, L., Zeng, Y., Zhu, L. (2008). A semiparametric estimation approach for biased sub-models of high-dimensional linear regression models. (manuscript).

[74] Lin, L., Zhu, L., Gai, Y. (2012). Quasi-instrumental variable-based inference for high-dimensional non-sparse models. (manuscript).

[75] Lv, J., Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.*, **37**, 3498-3528.

[76] Mallows, C.L. (1973). Some comments on Cp. *Technometrics*, **12**, 661-675.

[77] Miller, A.J. (2002). *Subset selection in regression* (2nd edition). Chapman and Hall / CRC Press, London and New York.

[78] Montgomery, D.C., Peck, E.A., Vining, G.G., Vining, J. (2001). *Introduction to linear regression analysis*. Wiley New York.

[79] Naito, K. (2004). Semiparametric density estimation by local L_2 fitting. *Ann. Statist.*, **32**, 1162-1191.

[80] Opsomer, J.D., Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *J. Amer. Statist. Assoc.*, **8**, 715-732.

[81] Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237-249.

[82] Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.*, **18**, 90-120.

[83] Owen, A. (1991). Empirical likelihood for linear models. *Ann. Statist.*, **19**, 1725-1747.

[84] Pfeffermann, D. (1984). On extensions of the Gauss-Markov theorem to the case of stochastic regression coefficients. *J. Roy. Statist. Soc. B*, **46(1)**, 139-148.

[85] Qin, J., Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300-325.

[86] Ruppert, D., Sheather, S.J., Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.*, **90**, 1257-1270.

[87] Rütimann, P., Bühlmann, P. (2009). High dimensional sparse covariance estimation via directed acyclic graphs. *Electronic J. Statist.*, **3**, 1133-1160.

[88] Salinelli, E. (2009). Nonlinear principal components, II: Characterization of normal distributions. *J. Multi. Anal.*, **100**, 652-660.

[89] Samarov, A., Tsybakov, A. (2004). Nonparametric independent component analysis. *Bernoulli*, **10**, 565-582.

[90] Schennach, S.M. (2007). Point estimation with exponentially tilted empirical likelihood. *Ann. Statist.*, **35**, 634-672.

[91] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

[92] Seber, G.A.F., Wild, C.J. (2003). *Nonlinear regression.* John Wiley & Sons.

[93] Sen, P.K. (1979). Asymptotic properties of maximum likelihood estimators based on conditional specification. *Ann. Statist.*, **5**, 1019-1033.

[94] Sen, P.K., Ehsanes Saleh, A.K.M. (1987). On preliminary test and shrinkage *M*-estimation in linear models. *Ann. Statist.*, **15**, 1580-1592.

[95] Severini, T.A. (1998). Some properties of inferences in misspecified linear models. *Statist. Probab. Lett.*, **40**, 149-153.

[96] Shen, D., Shen, H.P., Marronx, J.S. (2011). Consistency of sparse pca in high dimension,low sample size contexts. arXiv:1104.4289v1, [math.ST].

[97] Shen, X.T., Huang, H.C., Ye, J. (2004). Inference after model selection. *J. Amer. Statist. Assoc.*, **99**, 751-762.

[98] Shi, J., Lau, T.S. (2000). Empirical likelihood for partially linear models. *J. Multi. Anal.*, **72(1)**, 132-148.

[99] Simas Filho, E.F., Seixas, J.M. (2007). Nonlinear independent component analysis: theoretical review and applications. *Learning and Nonlinear Models*, **5(2)**, 99-120.

[100] Speckman, P. (1988). Kernel smoothing in partial linear models. *J. Roy. Statist. Soc. B*, **50**, 413-436.

[101] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussions). *J. Roy. Statist. Soc. B*, **36**, 111-147.

[102] Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10**, 1040-1053.

[103] Sun, T., Zhang, C. (2012). Scaled sparse linear regression. arXiv:1104.4595v2, [stat.ML], 21 Jun. 2012.

[104] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B*, **58**, 267-288.

[105] van der Vaart, A.W. (2000). *Asymptotic statistics.* Cambridge University Press.

[106] Wang, H., Li, G., Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD—Lasso. *J. Bus. Econom. Statist.*, **25**, 347-355.

[107] Wang, L., Chen, G., Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics.* **23**, 1486-1494.

[108] Wang, L., Li, H., Huang, J. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.*, **103**, 1556-1569.

[109] Wang, L., Li, H., Tsai, C.L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553-568.

[110] Wang, Q., Jing, B. (2003). Empirical likelihood for partially linear models. *Ann. Inst. Statist. Math.*, **55**, 585-595.

[111] Wu, S., Harris, T.J., McAuley, K.B. (2007). The use of simplified or misspecified models: linear case. *Can. J. Chem. Eng.*, **85**, 386-398.

[112] Ye, F., Zhang, C. (2010). Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Machine Learning Research*, **11**, 3519-3540.

[113] Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. B*, **68**, 49-67.

[114] Yuan, M., Ekici, A., Lu, Z., Monteiro, Y. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. Roy. Statist. Soc. B*, **69**, 329-346.

[115] Zhang, C., Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, **36(4)**, 1567 - 1594.

[116] Zhang, C., Zhang, S. (2012). Confidence intervals for low-dimensional parameters with high-dimensional data. arXiv:1110.2563v2, [stat.ME], 2 Nov. 2012.

[117] Zhang, P. (1992). Inference after variable selection in linear regression models. *Biometrika*, **79**, 741-746.

[118] Zhao, P., Xue, L. (2009). Variable selection for semiparametric varying coefficient partially linear models. *Statist. Probab. Lett.*, **79**, 2148-2157.

[119] Zhou, Z., Jiang, R., Qian, W. (2011). Variable selection for additive partially linear models with measurement error. *Metrika*, **74(2)**,185-202.

[120] Zhu, L., Xue, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *J. Roy. Statist. Soc. B*, **68**, 549-570.

[121] Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. B*, **67**, 301-320.

[122] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101**, 1418-1429.

致　谢

在本文即将完成定稿之际, 我首先要衷心感谢我的导师林路教授. 感谢林老师在这近六年的时间里在科学道路上给予我悉心的培养和教导. 特别是本文的写作, 从选题到最终定稿都离不开林老师的精心指导和审阅. 我不仅从林老师那里学到了非参和半参统计的基本知识, 而且也学到了做学问和做人踏实的态度. 林老师严谨的治学态度, 孜孜以求的科研作风, 诲人不倦的育人精神, 不仅使我在学习期间收获很大, 更将使我终身受益. 特别, 在我论文撰写过程中遇到困难时, 林老师对我不断地鼓励和引导, 耐心地讨论和修改, 让我走出了那段艰难低迷的时期. 对林老师的付出, 在此谨致以深深的敬意和感谢! 感谢师母欧阳海波老师对我学习,工作和生活上的关心.

衷心感谢彭实戈院士. 他广博而精深的知识, 敏锐的洞察力和独到的见解让人油然而生敬意. 彭老师治学严谨, 为人高尚而朴素, 使我强烈地感受到了一种崇高的科学精神和人格魅力. 他不仅带领我们在解决繁难的数学问题时感悟思考和探索的乐趣, 而且引导我们将科学理论与实际应用相联系.

衷心感谢陈增敬教授, 吴臻教授, 栾贻会教授, 魏刚教授, 赵卫东教授, 石玉峰教授, 嵇少林教授, 贾广岩教授以及金融基地的其他(她) 老师. 他们严谨踏实的治学态度, 独具特色的教授方法和对学生热情真诚的无私胸怀, 让我由衷的敬佩. 在山东大学金融基地的这段学习经历将是我受益一生的财富.

感谢王秀丽, 盖玉洁以及其他(她)同学的热心帮助和与我进行的有意义的讨论. 感谢几年来朝夕相处的师弟师妹, 有你们一路相伴, 让我度过了愉快难忘的学习生活. 愿我们的友谊之树常青!

感谢山东大学数学学院和山东省计算中心的领导和老师们对我在读博士期间的支持与帮助!

最后, 深深感谢我的儿子,家人和朋友, 他们一直在以各种不同的形式支持和帮助我, 是他们的理解,支持和鼓励让我顺利完成学业.

攻读博士学位期间完成论文情况

1. 曾云辉, 林路, 王秀丽. (2012). 多维线性回归有偏子模型的多步调整相合推断. *数学物理学报（中文版）*, **32A(6):** 1019-1031.

2. Yunhui Zeng, Xiuli Wang, Lu Lin. (2013). Remodeling and estimation for sparse partially linear regression models. *Abstract and Applied Analysis*, **vol. 2013**, Article ID 687151, 11 pages. doi:10.1155/2013/687151.

3. Lu Lin, Yunhui Zeng, Lixing Zhu. (2008). A semiparametric estimation approach for biased sub-models of high-dimensional linear regression models. (manuscript).

学位论文评阅及答辩情况表

	姓　名	专业技术职　务	是否博导（硕导）	所在单位	总体评价※
论文评阅人	匿名				A
	匿名				B
	艾明晶	教授	博导	北京大学	B

		姓　名	专业技术职　务	是否博导（硕导）	所在单位
答辩委员会成员	主席	陈松蹊	教授	博导	北京大学光华管理学院
	委员	栾贻会	教授	博导	山东大学
		魏刚	教授	博导	山东大学
		贾宁生	教授	博导	山东大学
		张建康	教授	博导	加拿大卡尔顿大学

答辩委员会对论文的总体评价※	A	答辩秘书	史敬涛	答辩日期	2013.5.20
备注					

※ 优秀为“A”；良好为“B”；合格为“C”；不合格为“D”。