北京師範大學

本科生毕业论文（设计）

毕业论文（设计）题目：

部分线性混杂变量模型的双去偏 Lasso 估计

部　　院　　系：　统计学院

专　　　　业：　统计学

学　　　　号：　**201911011123**

学 生 姓 名：　陈致远

指 导 教 师：　李高荣

指导教师职称：　教授

指导教师单位：　北京师范大学统计学院

年　　月　　日

北京师范大学本科毕业论文（设计）诚信承诺书

本人郑重声明：所呈交的毕业论文（设计），是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

本人签名：　　日期：　　年　　月　　日

北京师范大学本科毕业论文（设计）使用授权书

本人完全了解北京师范大学有关收集、保留和使用毕业论文（设计）的规定，即：本科生毕业论文（设计）工作的知识产权单位属北京师范大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许毕业论文（设计）被查阅和借阅；学校可以公布毕业论文（设计）的全部或部分内容，可以采用影印、缩印或扫描等复制手段保存、汇编毕业论文（设计）。保密的毕业论文（设计）在解密后遵守此规定。

本论文（是、否）保密论文。

保密论文在　　年　　月解密后适用本授权书。

本人签名：　　年　　月　　日

导师签名：　　年　　月　　日

部分线性混杂变量模型的双去偏 Lasso 估计

摘　要

在计量经济学和基因组学中，统计推断的建立往往被混杂变量干扰，造成伪相关而导致因果关系失真。双去偏 Lasso 方法则旨在有效地估计高维情形下的隐混杂变量模型，它能同时去除混杂变量和正则化造成的偏差，从而得到无偏估计。

部分线性模型是一类半参数模型，将线性回归部分和非线性函数同时纳入其中。以部分线性模型为框架可以方便地同时引入非线性趋势和混杂变量，符合现实数据分析中面对的大规模非参数或半参数估计中受混杂影响的问题。

本文在受混杂变量影响的部分线性模型的设定下，利用最小二乘估计和局部线性估计的方法论将模型变形为适用于双去偏 Lasso 方法的形式，再按照双去偏 Lasso 的算法求解。本文的数值模拟部分在低维和高维情形下分别进行了实验，研究了新方法受样本量、维数和协变量相关性变化的影响，并得到了不错的结果。

关键词： 混杂变量　双去偏 Lasso　部分线性模型　局部线性估计

Doubly Debiased Lasso for Partially Linear Model under Confounding

ABSTRACT

In econometrics and genomics, the establishment of statistical inferences is often distorted by confounding variables that cause pseudo-correlation and lead to distortion of causality. The doubly debiased lasso method, on the other hand, aims to efficiently estimate implicitly confounding variable models in the high-dimensional case by simultaneously removing the bias caused by confounding variables and regularization, resulting in unbiased estimates.

Partially linear model is a kind of semiparametric model, including both linear regression and nonlinear functions in itself. Using a partially linear model as a framework allows for the convenient introduction of both nonlinear trends and confounding variables, which is consistent with the problem of large-scale nonparametric or semiparametric estimation under confounding in realistic data analysis.

In this paper, in the setting of a partially linear model under confounding, the model is deformed into a form applicable to the doubly debiased lasso using the methodologies of least squares estimation and local linear estimation, and then solved according to the algorithm of doubly debiased lasso. The performance of the new method in the low and high dimensional cases and the influence by changes in sample size, dimensionality and covariate correlations are experimented in the numerical simulation section, and good results are obtained.

KEY WORDS: confounding doubly debiased lasso partially linear models local linear estimation

目 录

主要符号对照表

$X_{(i)} \in \mathbb{R}^p$	矩阵 X 的第 i 行，但为列向量，对其他矩阵同理
$X_j \in \mathbb{R}^n$	矩阵 X 的第 j 列
$X_{-j} \in \mathbb{R}^{n\times(p-1)}$	矩阵 X 除去第 j 列后的子矩阵
$Y \in \mathbb{R}^n$	响应变量向量
$X \in \mathbb{R}^{n\times p}$	协变量矩阵
$H \in \mathbb{R}^{n\times q}$	混杂变量矩阵
$E \in \mathbb{R}^{n\times p}$	协变量矩阵中不被混杂的部分
$\beta \in \mathbb{R}^p$	回归系数向量
$\phi \in \mathbb{R}^q$, $\Psi \in \mathbb{R}^{q\times p}$	混杂变量的系数向量和矩阵
$e \in \mathbb{R}^n$, $\epsilon \in \mathbb{R}^n$	随机扰动项
$b \in \mathbb{R}^p$	混杂变量造成的误差
s	β 中非零分量的个数
Ω_E	E 的精度矩阵
Σ_E	E 的协方差矩阵
S	压缩矩阵
ρ, ρ_j	压缩矩阵的修剪调节参数
Q, $P^{(j)}$	谱变换矩阵
$\hat{\beta}^{init}$	β 的初始估计
λ, λ_j	Lasso 的惩罚调节参数
Z_j, $\hat{\gamma}$	投影方向与投影系数
$g(\cdot)$	非线性函数
T_i	一维协变量
$\lambda_j(M)$	矩阵 M 的第 j 大的奇异值
$\widetilde{Y}$, $\widetilde{X}$, $\widetilde{H}$	加权求和后的新响应变量向量、新协变量矩阵和新混杂变量矩阵
$a_n \lesssim b_n$	对所有 n，存在 $C > 0$ 使得 $a_n \le Cb_n$

1　绪论

1.1　混杂变量与双去偏 Lasso 方法

混杂变量 (confounding，confounder 或 confounding variable)，又称混杂、干扰因子、混淆变量等，它是指那些能同时影响自变量和因变量的变量[1-2]，如图1所示。在实际数据分析和建模中，很多混杂变量通常是未被观测到的，所以有时也称为隐混杂变量 (hidden confounding model)。混杂变量的存在可能造成伪相关，从而导致因果关系或相关关系失真，特别是在大规模观察性研究中，很多协变量都会被混杂影响[3]。

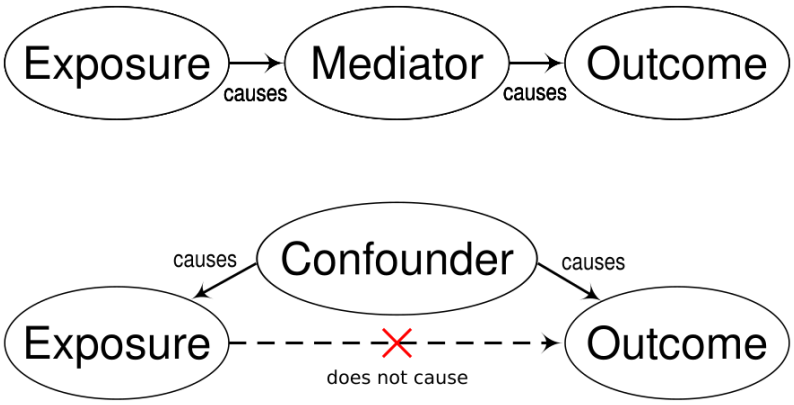

图 1　混杂变量和中间变量的区别

本文所关注的双去偏 Lasso 方法 (Doubly Debiased Lasso，DDL)，是由 Guo，Ćevid 和 Bühlmann[4] 提出的一个旨在去除混杂变量影响的模型。在高维线性模型的设定下，DDL 可以同时纠正由于高维参数估计引起的偏差和混杂变量引起的偏差。Guo，Ćevid 和 Bühlmann[4] 建立了详细的方法论、算法和程序，并在理论上证明了渐近正态性和高斯—马尔可夫意义上的有效性。

这种泛化方法的意义是显而易见的。例如，在计量经济学的语境下，隐混杂变量带来的问题被归于内生变量的框架下，通常使用工具变量解决这一类问题。然而，构造一个巧妙的工具变量需要大量相关领域的知识；同时，在高维情形下很难使用工具变量，因为需要构造和大于等于内生变量数量的工具变量，即所谓的“秩条件”[5]。而 DDL 恰好能同时解决这两个问题。另外，生物统计领域也有大量高维数据被混杂影响，DDL 在 GTEx 数据库中的基因组上良好的实际数据分析表现说明了这一点。

1.2 非线性情形与部分线性模型

尽管 DDL 方法在高维线性设定下处理混杂变量的表现非常优异，也不能回避线性回归在现实预测中明显不足的问题。很多情况下需要估计受混杂影响的非线性模型，并且一些学者已经研究了相关的方法。Albert[6] 研究了如何用中间变量帮助分析带混杂的非线性模型；Hahn，Murray 和 Carvalho[7] 提出的贝叶斯因果森林模型 (Bayesian causal forest model) 也是一个专注解决强烈混杂影响的非线性模型。可见，在非线性的假设下，模型同样会受到混杂变量的影响，或者说，拟合现实数据需要能够有效估计带混杂的非线性模型。这样，将 DDL 这一可以高效解决混杂影响的方法推广到非线性假设下就是有意义的。

在众多非线性设定中，部分线性模型 (partially linear models) 是一类半参数模型，因为它同时含有参数和非参数部分。Engle，Granger，Rice 和 Weiss[8] 最早在研究天气变化和电力需求时研究了部分线性模型，他们用线性模型拟合电价和收入，并用非线性模型拟合每天的温度，这样，非线性和线性的成分能够被同时纳入模型以达到更好的预测效果。许多国内外学者都对部分线性模型的理论和方法论展开了深入研究。例如，Shi 和 Lau[9] 研究了用经验似然方法估计部分线性模型；Liang，Liu，Li 和 Tsai[10] 研究了部分线性单指标模型的 profile 最小二乘估计 (profile least-squares estimator)，同时利用 SCAD 惩罚来选择变量；Li 和 Xue[11] 研究了部分线性误差模型的经验对数比似然统计量的性质；Härdle，Liang 和 Gao[12] 在专著中总结了部分线性模型应用在各种统计问题上的方法论，包括最小二乘回归、渐近有效估计、自助采样、删失数据分析、线性测量误差模型、非线性测量模型、非线性和非参数时间序列模型。

以部分线性模型为框架来引入非线性成分主要考虑了以下优势：

(1) 综合了线性回归的易解释性和非参数回归的稳健性的优点；

(2) 应用广泛，在纵向数据分析、EV(errors in-variables，EV) 回归、截面数据分析和生存分析中被大量使用；

(3) 形式灵活，一般形式为线性回归和非线性函数的求和，没有交叉项，比较容易转化成可以应用 DDL 解决的形式；

(4) 研究成熟，有很多有效的估计方法和推广部分线性模型的案例可以参考。

1.3 主要工作与结构安排

考虑用 DDL 估计受混杂变量影响的部分线性模型，其中部分混杂变量模型定义为

$$Y_i = X_{(i)}^\top \beta + H_{(i)}^\top \phi + g(T_i) + \epsilon_i, \quad i = 1, ..., n,$$

其中 Y_i 为响应变量，$X_{(i)} = (X_{i,1}, ..., X_{i,p})^\top$ 为 p 维协变量，$\beta = (\beta_1, ..., \beta_p)^\top$ 为 p 维回归系数向量，$H_{(i)} = (H_{i,1}, ..., H_{i,q})^\top$ 为 q 维混杂变量，$\phi = (\phi_1, ..., \phi_q)^\top$ 为 q 维混杂系数向量，$g(\cdot)$ 为未

知的光滑函数，T_i 为一元协变量，ϵ_i 为模型随机扰动项，满足 $\mathbb{E}(\epsilon_i) = 0$ 和 $\mathrm{Var}(\epsilon_i) = \sigma_\epsilon^2 < \infty$。

此模型可以被看作向部分线性模型中加入了混杂项 $H_i^\top \phi$，也可以被看作向隐混杂变量模型中加入了非线性函数 $g(T_i)$，或者是向经典的线性模型中同时加入了这两项。所以，不妨首先假设线性系数 β 和 ϕ 已知，则可利用核密度估计给出非线性函数的伪估计 $\widetilde{g}(t;\beta,\phi)$[11]。再借鉴部分线性模型的最小二乘估计和局部线性估计中用加权求和项表示 $\widetilde{g}(t;\beta,\phi)$ 的形式[13-14]

$$\widetilde{g}(t;\beta,\phi) = \sum_{i=1}^{n} W_i(t)\left(Y_i - X_{(i)}^\top \beta - H_{(i)}^\top \phi\right),$$

其中 $W_i(t)$ 为核权重。将伪估计代回原模型，则可以把原模型数据中的 X, Y, H 和加权求和的同类项合并，最终得到新的混杂线性模型[11-12]

$$\widetilde{Y}_i = \widetilde{X}_{(i)}^\top \beta + \widetilde{H}_{(i)}^\top \phi + \widetilde{\epsilon}_i, \quad i = 1, ..., n,$$

其中 $\widetilde{Y}_i = Y_i - \sum_{j=1}^n W_j(T_i)\, Y_j$ 为新响应变量，$\widetilde{X}_{(i)} = X_{(i)} - \sum_{j=1}^n W_j(T_i)\, X_{(j)}$ 为新协变量，$\widetilde{H}_{(i)} = H_{(i)} - \sum_{j=1}^n W_j(T_i)\, H_j$, 为新混杂变量，$\widetilde{\epsilon}_i = g(T_i) - \widetilde{g}(T_i;\beta,\phi) + \epsilon_i$ 为新随机扰动项。新的混杂线性模型可以直接使用 DDL 求解。

文章的其余部分安排如下：第2章介绍 DDL 的方法论以及估计量的理论性质；第3章介绍部分线性模型及其最小二乘估计和局部线性估计的方法论；第4章推导了用 DDL 估计部分线性混杂变量模型的方法论，并给出了估计算法；第5章利用数值模拟检验了新方法的有效性，并将结论总结在第6章。

2 双去偏 Lasso 方法

2.1 隐混杂变量模型

DDL 方法所面对的模型，即隐混杂变量模型的定义如下

$$Y_i = X_{(i)}^\top \beta + H_{(i)}^\top \phi + e_i, \quad X_{(i)}^\top = H_{(i)}^\top \Psi + E_{(i)}^\top, \quad i = 1, ..., n, \qquad \text{式（2-1）}$$

其中 Y_i 为响应变量，$X_{(i)} = (X_{i,1}, ..., X_{i,p})^\top$ 为 p 维协变量，$\beta = (\beta_1, ..., \beta_p)^\top$ 为 p 维回归系数向量，$H_{(i)} = (H_{i,1}, ..., H_{i,q})^\top$ 为 q 维混杂变量，$E_{(i)} = (E_{i,1}, ..., E_{i,p})^\top$ 为 p 维的协变量中未被混杂的部分。$\phi = (\phi_1, ..., \phi_q)^\top$ 为刻画 Y_i 和 $H_{(i)}$ 之间线性回归关系的 q 维混杂系数向量，Ψ 为刻画 $X_{(i)}$ 和 $H_{(i)}$ 之间线性回归关系的 $q \times p$ 维混杂系数矩阵。e_i 为模型随机扰动项，满足 $\mathbb{E}(e_i) = 0$ 和 $\mathrm{Var}(e_i) = \sigma_e^2 < \infty$。

为了推导的方便，隐混杂变量模型（2-1）可以被改写为线性模型

$$\begin{aligned} &Y_i = X_{(i)}^\top (\beta + b) + \epsilon_i, \quad X_{(i)}^\top = H_{(i)}^\top \Psi + E_{(i)}^\top, \\ &\epsilon_i = e_i + H_{(i)}^\top \phi - X_{(i)}^\top b, \quad b = \Sigma_X^{-1} \Psi^\top \phi, \quad i = 1, ..., n. \end{aligned} \qquad \text{式（2-2）}$$

令 $Y \in \mathbb{R}^n$ 和 $X \in \mathbb{R}^{n \times p}$ 分别代表响应变量和观测到的协变量；$H \in \mathbb{R}^{n \times q}$ 代表未被观测到的混杂变量，$E \in \mathbb{R}^{q \times p}$ 是不含混杂变量的成分，系数矩阵与向量 $\Psi \in \mathbb{R}^{q \times p}$，$\phi \in \mathbb{R}^q$。模型（2-1）和（2-2）也可改写为如下矩阵形式

$$Y = X(\beta + b) + \epsilon, \quad X = H\Psi + E. \qquad \text{式（2-3）}$$

模型（2-3）中只含有响应变量 Y 和协变量 X，但在估计参数 β 时会出现一个误差 b，这个误差就是由混杂变量 H 带来的。DDL 所面对的模型一般考虑 β 是稀疏的，其中非零分量 s 的数目很小。

2.1.1 稠密混杂

DDL 模型的一个重要假设是稠密混杂 (dense confounding)，即隐混杂变量 $H_{(i)}$ 和协变量 $X_{(i)}$ 中的很多个分量相关[4]。假设1通过限制系数矩阵 Ψ 来描述了这种关系。

假设 1 稠密混杂假设

系数矩阵 $\Psi \in \mathbb{R}^{q \times (p-1)}$ 的第 q 大的奇异值满足

$$\lambda_q(\Psi_{-j}) \gg l(n, p, g) := \max\left(M\sqrt{\frac{qp}{n}}(\log p)^{3/4},\ \sqrt{Mq} \cdot p^{1/4}(\log p)^{3/8},\ \sqrt{qn \cdot \log p} \right), \qquad \text{式（2-4）}$$

其中 M 是 $X_{(i)}$ 的亚高斯范数 (sub-Gaussian norm)。进一步可以得到

$$\max\left(\|\Psi(\Omega_E)_j\|_2,\ \|\Psi_j\|_2,\ \|\phi\|_2\right) \lesssim \sqrt{q}(\log p)^c, \qquad \text{式（2-5）}$$

其中 c 是一个常数，且 $0 < c \leq 1/4$，$\Omega_E = \Sigma_E^{-1}$ 是 E 的精度矩阵。

假设1对于推导有关 β_j 和 b 的理论性质非常重要，详见2.3节；同时，其中用到的亚高斯范数如定义1所示。

定义 1　亚高斯范数

$$\|X\|_{\psi_2} := \inf\left\{t \geq 0 :\ \mathbb{E}\left[\psi_2\left(\frac{|X|}{t}\right)\right] \leq 1\right\}, \qquad \text{式（2-6）}$$

其中 $X \in \mathbb{R}$, $\psi_2(x) = e^{x^2} - 1$。

2.1.2　混杂变量与飙升的奇异值

判断数据集是否被混杂变量影响的一个方法是考虑协变量矩阵 X 的奇异值大小，飙升的奇异值通常意味着混杂变量的存在。Guo，Ćevid 和 Bühlmann[4] 给出的模拟图像显示带混杂变量的矩阵有一些奇异值明显大于不带混杂变量的矩阵的奇异值范围，如图2所示。自然也可以想到，压缩数据的奇异值也许能帮助减少混杂变量的影响。

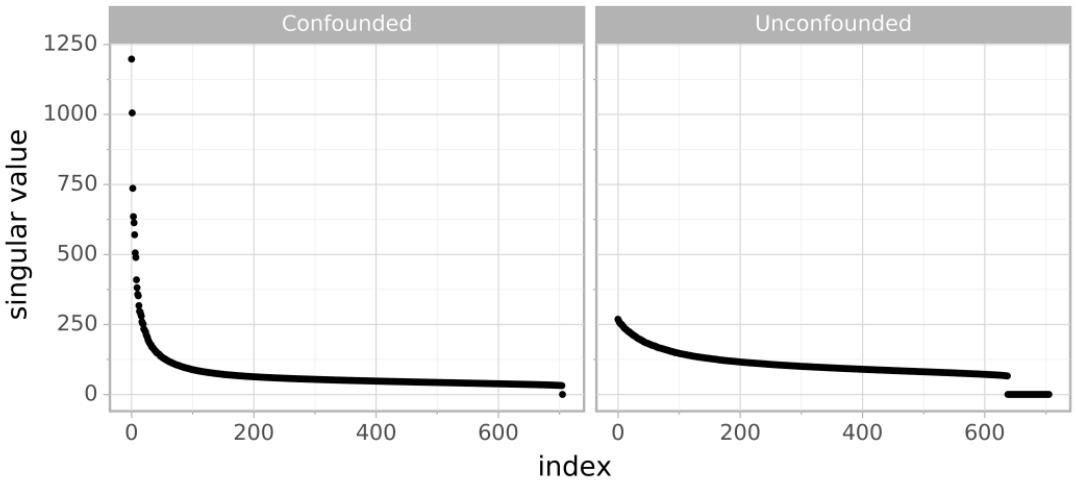

图 2　左：GTEx 数据库中的某基因表达矩阵的奇异值；右：该数据去除 65 个混杂变量之后的奇异值[4]。

2.2　两步估计

为了估计系数 β，DDL 方法使用了两步估计 (two-step estimator)。对于每一个 β_j，第一步通过谱变换后的惩罚似然估计得到去混杂的初始估计 $\hat{\beta}^{init}$；第二步先依赖初始估计 $\hat{\beta}^{init}$

得到 β_{-j}，再计算出无偏估计 $\hat{\beta}_j$。

2.2.1 谱变换

在 DDL 中，去除混杂影响的核心手段是对原数据做一个谱变换。例如，2.2.3节中求解过程需要的一个谱变换矩阵 $P^{(j)}$ 作用在 X_{-j} 上后要使 $P^{(j)}X_{-j}$ 的过大奇异值被压缩回正常范围，从而达到去除混杂偏差的目的。

构造谱变换矩阵的方法是修剪变换 (trim transform)[15]，这里以 $P^{(j)}$ 为例介绍构造过程。首先对 X_{-j} 作奇异值分解 (SVD)

$$X_{-j} = U\left(X_{-j}\right)\Lambda\left(X_{-j}\right)V\left(X_{-j}\right)^{\top}, \qquad \text{式（2-7）}$$

不妨定义 $P^{(j)}$ 为

$$P^{(j)} = U\left(X_{-j}\right)S\left(X_{-j}\right)U\left(X_{-j}\right)^{\top}, \qquad \text{式（2-8）}$$

则 $P^{(j)}X_{-j}$，即施加了谱变换后的 X_{-j} 为

$$P^{(j)}X_{-j} = U\left(S\Lambda\right)V^{\top}. \qquad \text{式（2-9）}$$

所以谱变换将原来的奇异值矩阵 Λ 变为了 $S\Lambda$，只要合理定义 S 中的元素就能压缩 $P^{(j)}X_{-j}$ 的奇异值。修剪变换限制每个奇异值不大于阈值 τ，即对 $1 \le l \le m$，定义压缩矩阵 S 如下

$$S_{l,l} = \begin{cases} \frac{\tau}{\Lambda_{l,l}} & \text{if } \Lambda_{l,l} > \tau, \\ 1 & \text{otherwise.} \end{cases} \qquad \text{式（2-10）}$$

更一般地，可以用任意百分位数 $\rho_j \in (0,1)$ 来压缩上 $(100\rho_j)\%$ 的奇异值。这样定义的 ρ_j - 修剪矩阵 $P^{(j)} \in \mathbb{R}^{n\times n}$ 为

$$P^{(j)} = U\left(X_{-j}\right)S\left(X_{-j}\right)U\left(X_{-j}\right)^{\top} \text{ with } S_{l,l}\left(X_{-j}\right) = \begin{cases} \frac{\Lambda_{\rho_j m,\rho_j m}}{\Lambda_{l,l}} & \text{if } l \le |\rho_j m|, \\ 1 & \text{otherwise.} \end{cases} \qquad \text{式（2-11）}$$

对于任意的百分位数 ρ_j，$S_{l,l}$ 可以将前 $\rho_j m$ 个最大的奇异值全部压缩而后面较小的奇异值保持不变。压缩比例 ρ_j 越大，对混杂的修正越强；但对原始数据的扭曲也越大，可能造成更大的误差。百分位数阈值的一个好选择为 $\rho = \rho_j = 1/2$，这样可以把过大的奇异值压缩至所有奇异值的中位数。

2.2.2 求解初始估计 $\hat{\beta}^{init}$

两步估计的第一步是求解初始估计 $\hat{\beta}^{init}$，这里需要使用谱变换矩阵 Q。Q 的构造方式和 $P^{(j)}$ 基本一致，只是 $P^{(j)}$ 的目标是压缩 X_{-j} 的奇异值，而 Q 的目标是压缩 X 的奇异值。对于任意的百分位数 $\rho \in (0,1)$，定义 ρ - 修剪矩阵 Q 为

$$Q = U(X)S(X)U(X)^{\top} \text{ with } S_{l,l}(X) = \begin{cases} \frac{\Lambda_{\rho m,\rho m}}{\Lambda_{l,l}} & \text{if } l \le |\rho m|, \\ 1 & \text{otherwise.} \end{cases} \quad \text{式（2-12）}$$

为了得到初始估计 $\hat{\beta}^{init}$，还需要对变换后的数据 QX 和 QY 做 Lasso 估计

$$\hat{\beta}^{init} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|Q(y - X\beta)\|_2^2 + \lambda \sum_{j=1}^{p} \frac{\|QX_j\|_2}{\sqrt{n}}|\beta_j| \right\}, \quad \text{式（2-13）}$$

其中 $\lambda = A\sigma_e\sqrt{\log p/n}$ 为调节参数，$A > \sqrt{2}$。λ 的值用 10 折交叉验证方法可以确定。

求解初始估计 $\hat{\beta}^{init}$ 的过程中，谱变换 Q 通过削弱混杂的影响来帮助估计 β，Lasso 估计则作为应对高维稀疏情形的一个良好的正则化方法选择[15]。

2.2.3 双去偏求解 $\hat{\beta}_j$

DDL 每次只估计 β 的一个固定的分量 β_j。两步估计的第二步通过拆分关心的分量 β_j 所对应的第 j 列 $X_j \in \mathbb{R}^n$ 和不关心的分量 β_{-j} 所对应的子矩阵 $X_{-j} \in \mathbb{R}^{n\times(p-1)}$，施之以谱变换去除混杂的影响，再将偏差项舍去来达到去除偏差的目的。

从线性模型（2-3）开始推导

$$\begin{aligned} Y &= X(\beta + b) + \epsilon, \\ Y &= X_j\left(\beta_j + b_j\right) + X_{-j}\left(\beta_{-j} + b_{-j}\right) + \epsilon \quad \text{for} \quad \text{j} \in 1, ..., \text{p}, \\ Y - X_{-j}\hat{\beta}_{-j}^{init} &= X_j\left(\beta_j + b_j\right) + X_{-j}\left(\beta_{-j} - \hat{\beta}_{-j}^{init}\right) + X_{-j}b_{-j} + \epsilon. \end{aligned} \quad \text{式（2-14）}$$

从式（2-14）中可以看到两个来源的偏差：$X_{-j}\left(\beta_{-j} - \hat{\beta}_{-j}^{init}\right)$ 源于初始估计 $\hat{\beta}^{init}$ 的 Lasso 有偏性，$X_{-j}b_{-j}$ 源于代表偏差影响的扰动向量 b。注意到两个偏差项都依赖于 X_{-j}，将谱变换矩阵 $P^{(j)}$ 作用于上式得

$$P^{(j)}\left(Y - X_{-j}\hat{\beta}_{-j}^{init}\right) = P^{(j)}X_j\left(\beta_j + b_j\right) + P^{(j)}X_{-j}\left(\beta_{-j} - \hat{\beta}_{-j}^{init}\right) + P^{(j)}X_{-j}b_{-j} + P^{(j)}\epsilon. \quad \text{式（2-15）}$$

又因为估计目标 β_j 和 $P^{(j)}X_j$ 绑定，而两个偏差项都与 $P^{(j)}X_{-j}$ 绑定。则构造投影方向向量 $P^{(j)}Z_j \in \mathbb{R}^n$ 作为 $P^{(j)}X_j$ 对 $P^{(j)}X_{-j}$ 回归的残差

$$P^{(j)}Z_j = P^{(j)}X_j - P^{(j)}X_{-j}\gamma, \quad \text{式（2-16）}$$

其中系数 $\hat{\gamma}$ 通过 Lasso 求解

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}} \left\{ \frac{1}{2n} \| P^{(j)} X_j - P^{(j)} X_{-j} \gamma \|_2^2 + \lambda_j \sum_{l \neq j} \frac{\| P^{(j)} X_l \|_2}{\sqrt{n}} |\gamma_l| \right\}, \quad \text{式（2-17）}$$

其中 $\lambda_j = A\sigma_j \sqrt{\log p / n}$ 为调节参数，$A > \sqrt{2}$。λ_j 的值同样可以用 10 折交叉验证确定。

在式（2-15）的等式两边左乘上 $\left(P^{(j)} Z_j\right)^\top$，则可以使 $\left(P^{(j)} Z_j\right)^\top P^{(j)} X_j$ 被投影到 $\mathbb{R}$ 上

$$\begin{aligned} \left(P^{(j)} Z_j\right)^\top P^{(j)} \left(Y - X_{-j} \hat{\beta}_{-j}^{init}\right) = & \left(P^{(j)} Z_j\right)^\top P^{(j)} X_j \left(\beta_j + b_j\right) + \left(P^{(j)} Z_j\right)^\top \\ & \times \left[P^{(j)} X_{-j} \left(\beta_{-j} - \hat{\beta}_{-j}^{init}\right) + P^{(j)} X_{-j} b_{-j} + P^{(j)} \epsilon \right], \end{aligned} \quad \text{式（2-18）}$$

于是，可以直接用四则运算分离关心的真实值 β_j

$$\begin{aligned} \beta_j + b_j = & \frac{\left(P^{(j)} Z_j\right)^\top P^{(j)} \left(Y - X_{-j} \hat{\beta}_{-j}^{init}\right)}{\left(P^{(j)} Z_j\right)^\top P^{(j)} X_j} - \frac{\left(P^{(j)} Z_j\right)^\top P^{(j)} X_{-j} \left(\beta_{-j} - \hat{\beta}_{-j}^{init}\right)}{\left(P^{(j)} Z_j\right)^\top P^{(j)} X_j} \\ & - \frac{\left(P^{(j)} Z_j\right)^\top P^{(j)} X_{-j} b_{-j}}{\left(P^{(j)} Z_j\right)^\top P^{(j)} X_j} - \frac{\left(P^{(j)} Z_j\right)^\top P^{(j)} \epsilon}{\left(P^{(j)} Z_j\right)^\top P^{(j)} X_j}. \end{aligned} \quad \text{式（2-19）}$$

根据上文提出的两个偏差来源，令 β 的估计 $\hat{\beta}_j$ 为

$$\hat{\beta}_j = \frac{\left(P^{(j)} Z_j\right)^\top P^{(j)} \left(Y - X_{-j} \hat{\beta}_{-j}^{init}\right)}{\left(P^{(j)} Z_j\right)^\top P^{(j)} X_j}, \quad \text{式（2-20）}$$

估计 $\hat{\beta}_j$ 和真实值 β 之差为

$$\begin{aligned} \beta_j - \hat{\beta}_j = & \frac{\left(P^{(j)} Z_j\right)^\top P^{(j)} \epsilon}{\left(P^{(j)} Z_j\right)^\top P^{(j)} X_j} + \frac{\left(P^{(j)} Z_j\right)^\top P^{(j)} X_{-j} \left(\beta_{-j} - \hat{\beta}_{-j}^{init}\right)}{\left(P^{(j)} Z_j\right)^\top P^{(j)} X_j} \\ & + \frac{\left(P^{(j)} Z_j\right)^\top P^{(j)} X_{-j} b_{-j}}{\left(P^{(j)} Z_j\right)^\top P^{(j)} X_j} + b_j. \end{aligned} \quad \text{式（2-21）}$$

观察式（2-21）右边可以发现：

（1）第一项为方差；

（2）第二项中，由于式（2-17）中 $\hat{\gamma}$ 的构造，残差 $P^{(j)} Z_j$ 和投影 $P^{(j)} X_{-j}$ 是正交的；式（2-13）中 Lasso 估计使得 $\hat{\beta}^{init}$ 和 β_j 在 l_1 范数下非常接近。所以 $\left(P^{(j)} Z_j\right)^\top P^{(j)} X_{-j} \left(\beta_{-j} - \hat{\beta}_{-j}^{init}\right)$ 的值很小；

（3）第三项中，因为 $P^{(j)}$ 缩减了 X_{-j} 的奇异值，Guo，Ćevid 和 Bühlmann[4] 证明了 $\left\| \frac{1}{\sqrt{n}} P^{(j)} X_{-j} b_{-j} \right\| = O_p\left(\frac{1}{\min(n,p)}\right)$；

（4）第四项中，在稠密混杂，即假设1的限制下，b 是稠密的且当 p 很大时 $\|b\|_2$ 很小[4]。由引理1可得 $|b_j| \leq \frac{q\sqrt{\log p}}{1+\lambda_q^2(\Psi)}$；

所以，后三项误差相对于方差都是可以忽略的，即估计 $\hat{\beta}_j$ 和真实值 β 的差距仅为估计的方差，意味着它就是无偏的。

2.2.4 方差估计

构造 β_j 的置信区间还需要随机扰动项方差 $\sigma_e^2 = \mathbb{E}(e_i^2)$ 的相合估计。构造相合估计的思路基于对线性模型施加谱变换，来先求另一个随机扰动项方差 σ_ϵ^2 的估计 $\hat{\sigma}_\epsilon^2$

$$\begin{aligned} Y &= X(\beta + b) + \epsilon, \\ Y - X\hat{\beta}^{init} &= X\left(\beta + b - \hat{\beta}^{init}\right) + \epsilon, \\ QY - QX\hat{\beta}^{init} &= Q\epsilon + QX\left(\beta - \hat{\beta}^{init}\right) + QXb. \end{aligned} \qquad \text{式（2-22）}$$

观察式（2-22）右边可以发现：

（1）$\hat{\beta}^{init}$ 估计 β 有好的相合性；

（2）类似地，谱变换 Q 缩减了 X 的奇异值，同样 Guo，Ćevid 和 Bühlmann[4] 证明了 $\left\|\frac{1}{\sqrt{n}}QXb\right\| = O_p\left(\frac{1}{\min(n,p)}\right)$。

所以，后两项的影响可以忽略，$\|Q\epsilon\|_2^2 \ / \ \mathrm{Tr}(Q^2)$ 即为 σ_ϵ^2 的相合估计。令 $\hat{\sigma}_e^2$ 为

$$\hat{\sigma}_e^2 = \hat{\sigma}_\epsilon^2 = \frac{1}{\mathrm{Tr}(Q^2)} \|Q\epsilon\|_2^2 = \frac{1}{\mathrm{Tr}(Q^2)} \left\|Qy - QX\hat{\beta}^{init}\right\|_2^2, \qquad \text{式（2-23）}$$

由引理1可得 σ_e^2 和 σ_ϵ^2 的差距很小，则可以证明式（2-23）中的 $\hat{\sigma}_e^2$ 为 σ_e^2 的相合估计。

2.2.5 置信区间

有了2.2.4节中 σ_e 的相合估计 $\hat{\sigma}_e$，以及2.3节中 $\hat{\beta}_j$ 的渐近正态分布，可以继续估计 β_j 的标准差为

$$\hat{\mathrm{sd}}(\beta_j) = \sqrt{\frac{\hat{\sigma}_e^2 \cdot Z_j^\top \left(P^{(j)}\right)^4 Z_j}{\left[Z_j^\top (P^{(j)})^2 X_j\right]^2}},$$

于是可以构造 β_j 的渐近置信度为 $1-\alpha$ 的置信区间

$$\mathrm{CI}\ (\beta_j) = \left(\hat{\beta}_j - \hat{\mathrm{sd}}(\beta_j)\, z_{1-\alpha/2},\ \hat{\beta}_j + \hat{\mathrm{sd}}(\beta_j)\, z_{1-\alpha/2}\right), \qquad \text{式（2-24）}$$

其中 $z_{1-\alpha/2}$ 为标准正态分布的 $1-\alpha/2$ 分位数。

2.3 理论结果

Guo，Ćevid 和 Bühlmann[4] 完成了有关 DDL 方法的详细而完整的理论性质证明。这里摘取部分以便于理解2.2节中的方法论。

首先需要补充关于 E 的精度矩阵 (precision matrix) 的假设。

假设 2 精度矩阵 $\Omega_E = \left[\mathbb{E}(E_{(i)}, E_{(i)}^\top)\right]^{-1}$ 满足 $c_0 \leq \lambda_{\min}(\Omega_E) \leq \lambda_{\max}(\Omega_E) \leq C_0$ 和 $\|(\Omega_E)_j^\top\|_0 \leq s$，其中 $C_0 > 0$ 和 $c_0 > 0$ 是正常数，s 表示稀疏程度，且可以随着 n 和 p 增长。

假设1和假设2都是高维统计中常见的假设。基于这两个假设，Guo，Ćevid 和 Bühlmann[4] 推导了关于混杂造成的误差 b 和误差项方差 σ_ϵ^2 和 σ_e^2 的引理。

引理 1 若假设1和假设2成立，则

$$|b_j| \lesssim \frac{q(\log p)^{1/2}}{1+\lambda_q^2(\Psi)}, \quad \|b\|_2 \lesssim \frac{\sqrt{q}(\log p)^{1/4}}{\lambda_q(\Psi)}, \qquad \text{式（2-25）}$$

以及

$$\left|\sigma_\epsilon^2 - \sigma_e^2\right| = \left|\phi^\top\left(\mathrm{I}_q^\top - \Psi\Sigma_X^{-1}\Psi^\top\right)\phi\right| \lesssim \frac{q(\log p)^{1/2}}{1+\lambda_q^2(\Psi)}. \qquad \text{式（2-26）}$$

引理1说明了误差项 ϵ_i 和 e_i 的方差非常相近。

理论结果中最重要的部分是估计 $\hat{\beta}_j$（2-20）的性质。当高维情形下限制 $n, p \to \infty$，$c^* = \lim p/n \in (0, \infty]$ 时，Guo，Ćevid 和 Bühlmann[4] 推导了 $\hat{\beta}_j$ 的渐近正态分布。

定理 1 DDL 估计 $\hat{\beta}_j$ 满足

$$\frac{1}{V}\left(\hat{\beta}_j - \beta_j\right) \xrightarrow{d} N(0,1), \qquad \text{式（2-27）}$$

其中

$$V = \frac{\sigma_e^2 Z_j^\top (P^{(j)})^4 Z_j}{\left[Z_j^\top (P^{(j)})^2 X_j\right]^2} \quad \text{and} \quad V^{-1}\frac{\sigma_e^2 \mathrm{Tr}\left[(P^{(j)})^4\right]}{\sigma_j^2 \mathrm{Tr}^2\left[(P^{(j)})^2\right]} \xrightarrow{p} 1. \qquad \text{式（2-28）}$$

3　部分线性模型

部分线性模型的形式被定义为

$$Y_i = X_{(i)}^{\top}\beta + g(T_i) + \epsilon_i, \quad i = 1, ..., n, \qquad \text{式（3-1）}$$

其中 Y_i 为响应变量，$X_{(i)} = (X_{i,1}, ..., X_{i,p})^{\top}$ 为 p 维协变量，$\beta = (\beta_1, ..., \beta_p)^{\top}$ 为 p 维线性回归系数向量，$g(\cdot)$ 为未知的光滑函数，T_i 为一元协变量，ϵ_i 为模型随机扰动项，满足 $\mathbb{E}(\epsilon_i) = 0$ 和 $\mathrm{Var}(\epsilon_i) = \sigma_{\epsilon}^2 < \infty$。

接下来介绍两种常用的估计部分线性模型的方法，其中的思路将用于第4章中的求解。

3.1　最小二乘估计

如果对于对模型（3-1）非参数部分的某些假设成立，那么最小二乘估计适用于部分线性模型[12]。不妨假设 β 已知，可以定义光滑函数 $g(\cdot)$ 的一个估计为

$$\hat{g}(t;\beta) = \sum_{j=1}^{n} \omega_j(t) \left(Y_j - X_{(j)}^{\top}\beta\right), \qquad \text{式（3-2）}$$

其中 $\omega_j(t) = \omega_j(t; T_1, ..., T_n)$ 为依赖 t 和 $T_1, ..., T_n$ 的权重函数。用 $\hat{g}(T_i)$ 替换模型（3-1）中的 $g(T_i)$ 得

$$Y_i = X_{(i)}^{\top}\beta - \sum_{j=1}^{n} \omega_j(T_i) \left(Y_j - X_{(j)}^{\top}\beta\right) + \epsilon_i, \quad i = 1, ..., n, \qquad \text{式（3-3）}$$

则可以将模型（3-1）变形为线性模型

$$\begin{aligned} Y_i - \sum_{j=1}^{n} \omega_j(T_i)\, Y_j &= X_{(i)}^{\top}\beta - \sum_{j=1}^{n} \omega_j(T_i)\, X_{(j)}^{\top}\beta + \epsilon_i, \\ \widetilde{Y}_i &= \widetilde{X}_{(i)}^{\top}\beta + \epsilon_i, \quad i = 1, ..., n, \end{aligned} \qquad \text{式（3-4）}$$

其中 $\widetilde{Y}_i = Y_i - \sum_{j=1}^{n} \omega_j(T_i)\, Y_j$，$\widetilde{X}_{(i)} = X_{(i)} - \sum_{j=1}^{n} \omega_j(T_i)\, X_{(j)}$。模型（3-4）也可以写成矩阵形式

$$\widetilde{Y} = \widetilde{X}\beta + \epsilon, \qquad \text{式（3-5）}$$

其中 $\widetilde{X}^{\top} = \left(\widetilde{X}_{(1)}, ..., \widetilde{X}_{(n)}\right)$，$\widetilde{Y}^{\top} = \left(\widetilde{Y}_1, ..., \widetilde{Y}_n\right)$。对模型（3-5）用最小二乘法估计模型（3-5）即可得到 β 的估计

$$\widetilde{\beta}_{LS} = \left(\widetilde{X}^{\top}\widetilde{X}\right)^{-1} \widetilde{X}^{\top}\widetilde{Y}. \qquad \text{式（3-6）}$$

再代回 $\hat{g}(t;\beta)$ 即得到非参数估计

$$\widetilde{g}(t)=\sum_{j=1}^{n}\omega_j(t)\left(Y_j-X_{(j)}^{\top}\widetilde{\beta}_{LS}\right). \qquad \text{式（3-7）}$$

3.2 局部线性估计

局部线性估计 (local linear estimation) 是一种常用的非参数拟合方法，也可以用于部分线性模型的估计[13-14]。局部线性估计有很多优良的性质，作为局部多项式估计的特殊情形，它的理论更加完善；而相比 N-W 核密度估计，局部线性估计的渐近偏差更小，且在自变量的边界区间拟合得更好[16]。

假设 β 已知，可以将模型（3-1）转化为非参数回归模型

$$Y_i-X_{(i)}^{\top}\beta=g(T_i)+\epsilon_i,\quad i=1,...,n. \qquad \text{式（3-8）}$$

当 T_i 在 t 的一个小邻域内时，可以用线性函数近似 $g(T_i)$

$$g(T_i)\approx g(t)+g'(t)(T_i-t)=:a+b(T_i-t),\quad i=1,...,n. \qquad \text{式（3-9）}$$

于是 $g(T_i)$ 的估计转化为对 a 和 b 的估计，由于 b 可以由 a 求导得到，所以主要是对 a 的估计。假设 β 已知时，a 和 b 的估计转化为优化问题

$$\min_{a,b}\sum_{i=1}^{n}\left[Y_i-X_{(i)}^{\top}\beta-a-b(T_i-t)\right]^2\cdot K_h(T_i-t), \qquad \text{式（3-10）}$$

其中 $K_h(\cdot)=K(\cdot/h)/h$ 是核函数，h 为窗宽。令 $\hat{a}$ 和 $\hat{b}$ 为最小化的解，经计算可得

$$\hat{a}=\frac{\sum_{j=1}^{n}w_j(t)\ \left(Y_j-X_{(j)}^{\top}\beta\right)}{\sum_{j=1}^{n}w_j(t)}, \qquad \text{式（3-11）}$$

其中

$$\begin{aligned} w_j(t)&=K_h(T_j-t)\cdot\left[S_{n,2}(t)-(T_j-t)S_{n,1}(t)\right],\\ S_{n,l}(t)&=\frac{1}{n}\sum_{j=1}^{n}K_h(T_j-t)\cdot(T_j-t)^l,\quad t=1,2. \end{aligned} \qquad \text{式（3-12）}$$

即可得到 $g(t)$ 的伪估计 $\hat{g}(t,\beta)$ 为

$$\hat{g}(t,\beta)=\sum_{j=1}^{n}W_{nj}(t)\ \left(Y_j-X_{(j)}^{\top}\beta\right), \qquad \text{式（3-13）}$$

其中

$$W_{nj}(t)=\frac{w_j(t)}{\sum_{j=1}^{n}w_j(t)}. \qquad \text{式（3-14）}$$

此时 β 未知，还无法得到 $g(t)$ 的估计。Li 和 Xue[11] 在继续求得了 β 的极大经验似然估计 (maximum empirical likelihood estimator) $\widetilde{\beta}$ 后，将其带入式（3-13）后得到 $g(t)$ 的估计 $\widetilde{g}_n(t) = \sum_{j=1}^{n} W_{nj}(t) \left(Y_j - X_{(j)}^{\top}\widetilde{\beta}\right)$。但此部分与本文主体方法论的思路无关，故在此不再赘述。

4 用双去偏 Lasso 估计部分线性混杂变量模型

在部分线性模型中加入隐混杂得到部分线性混杂变量模型

$$\begin{aligned} Y_i &= X_{(i)}^\top \beta + H_{(i)}^\top \phi + g(T_i) + \epsilon_i, \\ X_{(i)}^\top &= H_{(i)}^\top \Psi + E_{(i)}^\top, \quad i = 1, ..., n, \end{aligned} \tag{式（4-1）}$$

其中 Y_i 为响应变量，$X_{(i)} = (X_{i,1}, ..., X_{i,p})^\top$ 为 p 维协变量，$\beta = (\beta_1, ..., \beta_p)^\top$ 为 p 维回归系数向量，$H_{(i)} = (H_{i,1}, ..., H_{i,q})^\top$ 为 q 维混杂变量，$E_{(i)} = (E_{i,1}, ..., E_{i,p})^\top$ 为 p 维的协变量中未被混杂的部分。$\phi = (\phi_1, ..., \phi_q)^\top$ 为刻画 Y_i 和 $H_{(i)}$ 之间线性回归关系的 q 维混杂系数向量，Ψ 为刻画 $X_{(i)}$ 和 $H_{(i)}$ 之间线性回归关系的 $q \times p$ 维混杂系数矩阵。$g(\cdot)$ 为未知的光滑函数，T_i 为一元协变量。ϵ_i 为模型随机扰动项，满足 $\mathbb{E}(\epsilon_i) = 0$ 和 $\mathrm{Var}(\epsilon_i) = \sigma_\epsilon^2 < \infty$。

模型（4-1）也可以写成矩阵形式

$$\begin{aligned} Y &= X\beta + H\phi + g(T) + \epsilon, \\ X &= H\Psi + E, \end{aligned} \tag{式（4-2）}$$

其中 $T = (T_1, ..., T_n)^\top$，其余变量与模型（2-3）中所解释一致。

将模型（4-1）的主模型分离线性和非线性部分后得到

$$Y_i - X_{(i)}^\top \beta + H_{(i)}^\top \phi = g(T_i) + \epsilon_i \quad i = 1, ..., n. \tag{式（4-3）}$$

不妨假设 β 和 ϕ 是固定的，则首先对 $g(T_i)$ 进行估计。参考3.2节局部线性估计的结论，可得类似式（3-13）的伪估计 $\widetilde{g}(t; \beta, \phi)$

$$\begin{aligned} \widetilde{g}(t; \beta, \phi) &= \arg\min_{a,b} \left[Y_i - X_{(i)}^\top \beta - H_{(i)}^\top \phi - a - b(T_i - t) \right]^2 \cdot K_h(T_i - t) \\ &= \sum_{i=1}^n W_i(t) \left(Y_i - X_{(i)}^\top \beta - H_{(i)}^\top \phi \right), \quad i = 1, ..., n. \\ W_i(t) &= \frac{1}{n} \cdot \frac{K_h(T_i - t) \cdot [S_{n,2}(t) - (T_i - t) S_{n,1}]}{S_{n,0}(t) S_{n,2}(t) - S_{n,1}^2}, \quad i = 1, ..., n. \\ S_{n,e}(t) &= \frac{1}{n} \sum_{i=1}^n K_h(T_i - t) (T_i - t)^e, \quad e = 0, 1, 2. \end{aligned} \tag{式（4-4）}$$

由于伪估计依赖于 β 和 ϕ，将 $\widetilde{g}(t; \beta, \phi)$ 代回模型（4-1）得

$$\begin{aligned} Y_i &= X_{(i)}^\top \beta + H_{(i)}^\top \phi + \widetilde{g}(T_i; \beta, \phi) + g(T_i) - \widetilde{g}(T_i; \beta, \phi) + \epsilon_i \\ &= X_{(i)}^\top \beta + H_{(i)}^\top \phi + \sum_{j=1}^n W_j(T_i) \left(Y_j - X_{(j)}^\top \beta - H_{(j)}^\top \phi \right) + g(T_i) - \widetilde{g}(T_i; \beta, \phi) + \epsilon_i. \end{aligned} \tag{式（4-5）}$$

类似3.1节最小二乘估计中式（3-4）的做法，将 $Y_i, X_{(i)}, H_{(i)}$ 与它们的加权求和项分别合并为新响应变量 $\widetilde{Y}_i$，新协变量 $\widetilde{X}_{(i)}$ 和新混杂变量 $\widetilde{H}_{(i)}$，并定义新随机扰动项为 $\widetilde{\epsilon}_i$

$$\begin{aligned}
&\widetilde{Y}_i = Y_i - \sum_{j=1}^{n} W_j(T_i)\, Y_j, \quad \widetilde{X}_{(i)} = X_{(i)} - \sum_{j=1}^{n} W_j(T_i)\, X_{(j)}, \\
&\widetilde{H}_{(i)} = H_{(i)} - \sum_{j=1}^{n} W_j(T_i)\, H_j, \quad \widetilde{\epsilon}_i = g(T_i) - \widetilde{g}(T_i;\beta,\phi) + \epsilon_i, \quad i = 1, ..., n,
\end{aligned} \tag{式（4-6）}$$

则式（4-5）被改写为新的混杂线性模型

$$\widetilde{Y}_i = \widetilde{X}_{(i)}^{\top}\beta + \widetilde{H}_{(i)}^{\top}\phi + \widetilde{\epsilon}_i, \quad i = 1, ..., n, \tag{式（4-7）}$$

也可以写成以下矩阵形式

$$\widetilde{Y} = \widetilde{X}\beta + \widetilde{H}\phi + \widetilde{\epsilon}. \tag{式（4-8）}$$

其中 $\widetilde{X}^{\top} = \left(\widetilde{X}_{(1)}, ..., \widetilde{X}_{(n)}\right)$，$\widetilde{Y}^{\top} = \left(\widetilde{Y}_1, ..., \widetilde{Y}_n\right)$，$\widetilde{H}^{\top} = \left(\widetilde{H}_{(1)}, ..., \widetilde{H}_{(n)}\right)$。

从模型（4-8）的形式不难看出，对于新响应变量 $\widetilde{Y}$，新协变量 $\widetilde{X}$ 和新混杂变量 $\widetilde{H}$，用 DDL 估计 β 即可。所以，估计带隐混杂的部分线性模型的求解过程如算法1所示。

算法 1 部分线性混杂变量模型的双去偏 Lasso 估计

输入： 数据 $X \in \mathbb{R}^{n\times p}, Y \in \mathbb{R}^n$ 和 $\{T_i,\ 1 \le i \le n\}$; 指标 j; 调节参数 $\rho, \rho_j \in (0,1)$ 和 $\lambda > 0, \lambda_j > 0$

输出： 点估计 $\hat{\beta}_j$, 随机扰动项方差估计 $\hat{\sigma}_e^2$ 和置信区间 $\mathrm{CI}\left(\beta_j\right)$

1: $\widetilde{g}(t;\beta,\phi) \leftarrow \mathrm{LLE}(X, Y, T_i)$ ▷ 求伪估计，见式（4-4）

2: $\left(\widetilde{X}, \widetilde{Y}\right) \leftarrow \mathrm{Unite}(X, Y, \widetilde{g})$ ▷ 合并同类得新项，见式（4-6）

3: $Q \leftarrow \mathrm{TrimTransform}\left(\widetilde{X}, \rho\right)$ ▷ 构造 ρ - 修剪矩阵，见式（2-11）

4: $\hat{\beta}^{init} \leftarrow \mathrm{Lasso}\left(Q\widetilde{X}, Q\widetilde{Y}, \lambda\right)$ ▷ 计算初始估计，见式（2-13）

5: $P^{(j)} \leftarrow \mathrm{TrimTransform}\left(\widetilde{X}_{-j}, \rho_j\right)$ ▷ 构造 ρ_j - 修剪矩阵，见式（2-12）

6: $\hat{\gamma} \leftarrow \mathrm{Lasso}\left(P^{(j)}\widetilde{X}_{-j}, P^{(j)}\widetilde{X}_j, \lambda_j\right)$ ▷ 计算投影的回归系数，见式（2-17）

7: $P^{(j)}Z_j \leftarrow P^{(j)}\widetilde{X}_j - P^{(j)}\widetilde{X}_{-j}\hat{\gamma}$ ▷ 令投影方向向量为残差，见式（2-16）

8: $\hat{\beta}_j \leftarrow \mathrm{DDL}\left(\hat{\beta}^{init}, P^{(j)}\widetilde{X}_j, P^{(j)}\widetilde{X}_{-j}, P^{(j)}Z_j\right)$ ▷ 计算 DDL 估计，见式（2-20）

9: $\hat{\sigma}_e^2 \leftarrow \mathrm{NoiseLevel}\left(\widetilde{X}, \widetilde{Y}, Q, \hat{\beta}^{init}\right)$ ▷ 计算随机扰动项方差估计，见式（2-23）

10: $\mathrm{CI}(\beta_j) \leftarrow \mathrm{CI}\left(\hat{\beta}_j, P^{(j)}\widetilde{X}_j, P^{(j)}Z_j, \hat{\sigma}_e^2, \alpha\right)$ ▷ 计算 $1-\alpha$ 置信区间，见式（2-24）

5 数值模拟

这一章将给出用 DDL 估计部分线性混杂变量模型 (以下简称“新方法”) 的数值模拟结果，过程按照算法1执行。全部程序使用 R 语言完成，主要依赖求解部分线性模型的包 **PLRModels** 和 DDL 模型的包 **DDL**。

5.1 参数设定与数据生成

对于本文所关心的部分线性混杂变量模型，即模型（4-1）

$$Y_i = X_{(i)}^\top \beta + H_i^\top \phi + g(T_i) + \epsilon_i,$$
$$X_{(i)}^\top = H_{(i)}^\top \Psi + E_{(i)}^\top, \quad i = 1, ..., n,$$

在所有模拟中设置参数 $q = 3$，$s = 5$，$\beta = (1, 1, 1, 1, 1, 0, ..., 0)^\top$，$g(t) = 2 + 2t\cos(4\pi t)$。为了研究该估计受数据维数变化的影响，模拟过程中将改变样本量 n 和维数 p 的数值并记录结果：在低维情形下，当样本量 $n = 200$ 和 400 时，让维数 p 从 25 增加到 150；在高维情形下，当样本量 $n = 200$ 时，令维数 $p = 200$ 和 250。另外，为了研究估计受数据协方差的影响，令 $(\Sigma_E)_{i,j} = (\Sigma_H)_{i,j} = \kappa^{|i-j|}$，其中 $\kappa \in \{0.3, 0.7\}$，Σ_H 为 $q \times q$ 维矩阵，Σ_E 为 $p \times p$ 维矩阵。

$E_{(i)}$、$H_{(i)}$ 和 ϵ_i 分别从多元正态分布 $N_p(\mathbf{0}, \Sigma_E)$、多元正态分布 $N_q(\mathbf{0}, \Sigma_H)$ 和标准正态分布 $N(0, 1)$ 中生成 n 个独立同分布样本。$\Psi_{(k)}$ 和 ϕ_k 分别从多元标准正态分布 $N_p(\mathbf{0}, I_p)$ 和标准正态分布 $N(0, 1)$ 中生成 q 个独立同分布样本。令 $T_i = (i - 0.5)/n,\ i = 1, ..., n$。$X_{(i)}$ 和 Y_i 由模型（4-1）中的公式计算得到。

5.2 模拟结果

本节将分别在低维和高维情形下进行单次和重复实验，并用表格展示新方法的所有模拟结果。所有重复模拟的重复次数都为 500。

5.2.1 低维情形

由于 DDL 每次对一个固定的指标 j 估计 β 的一个分量，所以展示所有分量的估计和统计量能帮助直观了解新方法的效果。在维数 $p = 25$ 的设定下进行的单次实验结果如表1所示，其中 SE 代表标准误，Pr 代表 p 值，CI 代表对应分量的置信区间。

表 1　低维单次实验估计结果

Index	β	$\hat{\beta}^{init}$	$\hat{\beta}$	SE	Pr	CI
1	1	0.837	1.056	0.084	2.168×10^{-36}	(0.892, 1.220)
2	1	0.755	0.970	0.088	5.049×10^{-28}	(0.797, 1.143)
3	1	0.824	1.030	0.077	1.161×10^{-40}	(0.879, 1.181)
4	1	0.659	0.846	0.081	3.470×10^{-25}	(0.686, 1.005)
5	1	0.672	0.846	0.074	1.641×10^{-31}	(0.720, 1.011)
6	0	0	0.132	0.082	0.871	(-0.148, 0.174)
7	0	0	0.162	0.082	0.049	(0.001, 0.323)
8	0	0	-0.064	0.066	0.339	(-0.194, 0.067)
9	0	0	-0.011	0.073	0.879	(-0.155, 0.132)
10	0	0	-0.009	0.077	0.909	(-0.160, 0.142)
11	0	0	0.066	0.081	0.414	(-0.092, 0.223)
12	0	0	0.057	0.084	0.498	(-0.108, 0.222)
13	0	0	0.109	0.065	0.099	(-0.021, 0.238)
14	0	0	-0.051	0.080	0.528	(-0.208, 0.107)
15	0	0	0.106	0.066	0.107	(-0.023, 0.236)
16	0	0	-0.004	0.070	0.954	(-0.141, 0.133)
17	0	0	-0.047	0.066	0.474	(-0.176, 0.082)
18	0	0	-0.039	0.074	0.588	(-0.184, 0.104)
19	0	0	-0.095	0.080	0.235	(-0.253, 0.062)
20	0	0	-0.043	0.070	0.543	(-0.181, 0.095)
21	0	0	-0.058	0.076	0.453	(-0.208, 0.093)
22	0	0	0.117	0.077	0.126	(-0.033, 0.267)
23	0	0	-0.044	0.077	0.568	(-0.194, 0.106)
24	0	0	0.016	0.077	0.831	(-0.134, 0.107)
25	0	0	0.007	0.080	0.925	(-0.149, 0.164)

从表1的结果中可以发现：

(1) 初始估计 $\hat{\beta}^{init}$ 在 1-5 号信号分量的估计上还是有明显的偏差，而最终估计 $\hat{\beta}$ 对 1-5 号信号分量的估计很接近真值，说明 DDL 去除偏差的效果十分显著。最终估计 $\hat{\beta}$ 对 6-25 号噪声分量的估计也比较准确，在 0 附件波动；

(2) 1-5 号信号分量的 p 值大小显著地小，说明新方法能有效识别信号。如果以 $\alpha = 0.05$ 作为阈值，则有且仅有 7 号噪声分量会被错误识别为信号。总体而言，新方法在选择变量上表现良好；

(3) 所有分量的标准误都较小，集中在 0.076 附近。信号分量的平均置信区间长度为 0.381，噪声分量的平均置信区间长度为 0.263，相对真值而言表现良好。

新方法在低维情形下单次实验中的估计表现良好，但其稳定性还需要继续在重复模拟中观察。所以，在样本量 $n = 200$ 和 400，维数 p 从 25 增加到 150，协方差矩阵 $(\Sigma_E)_{i,j} = \kappa^{|i-j|}$ 的相关系数 κ 分别为 0.3 和 0.7 的设定下进行 500 次重复模拟实验。结果如表2所示，其中 MSE= $\|\beta - \hat{\beta}\|_2$ 代表估计 $\hat{\beta}$ 的均方误差，Correct 代表正确识别信号的平均个数，Incorrect 代表将噪声错误识别为信号的平均个数。

表 2　低维重复模拟估计结果

		MSE	Correct	Incorrect	MSE	Correct	Incorrect
n	p		$\kappa = 0.3$			$\kappa = 0.7$	
200	25	0.576	5	2.45	0.907	5	2.48
	50	0.739	5	2.48	1.148	4.99	3.03
	100	1.095	5	4.76	1.703	4.99	6.22
	150	1.431	5	9.68	2.178	4.97	11.18
400	25	0.503	5	4.40	0.742	4.99	4.02
	50	0.503	5	3.51	0.813	4.99	3.80
	100	0.707	5	5.13	1.107	5	6.06
	150	0.903	5	7.69	1.431	5	9.68

从表2的结果中可以发现：

(1) 当维数 p 增大或相关系数 κ 增大时，估计的 MSE 会显著增大。当样本量 n 增大时，估计的 MSE 会显著减小。

(2) 在所有情况下，新方法都几乎能全部识别信号。

(3) 当维数 p 增大或相关系数 κ 增大时，新方法会更多地将噪声识别为信号。

(4) 样本量 n 的变化对 Incorrect 的影响没有明显的单调趋势：在维数 p 较小时，样本量 n 增大使新方法更多地将噪声识别为信号；而在维数 p 较大时，样本量 n 增大反而会减少对噪声的错误识别。这种不单调的影响是否是一般性的需要在后续实验中确定，比如增加样本量 $n = 600$ 和 800 等变化。

5.2.2　高维情形

DDL 方法不仅是为处理混杂变量，也是为应对高维情形而设计的。为了检验新方法在高维情形下的表现，首先在样本量 $n = 200$，$p = 300$ 的设定下进行单次实验，为了方便，表3中仅展示前十个分量。

表 3　高维单次实验估计结果

Index	β	$\hat{\beta}^{init}$	$\hat{\beta}$	SE	Pr	CI
1	1	0.515	0.769	0.104	1.625×10^{-13}	(0.564, 0.973)
2	1	0.564	0.856	0.098	2.876×10^{-18}	(0.663, 1.048)
3	1	0.583	0.967	0.0971	2.397×10^{-23}	(0.776, 1.158)
4	1	0.945	1.306	0.1037	2.230×10^{-36}	(1.103, 1.509)
5	1	0.658	1.024	0.102	7.358×10^{-24}	(0.825, 1.224)
6	0	0	0.224	0.110	0.041	(0.009, 0.439)
7	0	0	-0.173	0.105	0.097	(-0.379, 0.032)
8	0	0	-0.128	0.106	0.229	(-0.336, 0.080)
9	0	0	-0.021	0.099	0.837	(-0.216, 0.175)
10	0	0	0.122	0.103	0.256	(-0.080, 0.324)

可以发现新方法在高维情形下对 1-5 号信号分量的估计出现了肉眼可见的偏差。这种偏差应该是高维的而非混杂的，因为初始估计 $\hat{\beta}^{init}$ 的结果已经足够糟糕，而最终估计 $\hat{\beta}$ 仍向着好的方向进行了大幅修正，这说明新方法去除混杂变量的影响的能力并不会在高维情形下消失。

同样地，在样本量 $n = 200$，维数 $p = 200$ 和 250，$\kappa = 0.3$ 的设定下进行 500 次重复模拟实验。结果如表4所示，并与表2中的低维结果进行对比。

表 4　高维重复模拟估计结果

		MSE	Correct	Incorrect
n	p		$\kappa = 0.3$	
200	200	1.761	5	16.480
	250	1.906	5	18.240

如图3所示，高维情形下估计的 MSE 相比低维并没有急剧地升高，MSE 随维数 p 增长的速度基本和低维持平。Incorrect 在超过样本量 $n = 200$ 的临界点之后也没有急剧升高，甚至增长速度有所减缓。但是维数较大时 Incorrect 的值在绝对意义上已经很大，严重影响变量选择的效果。考虑到在表3所示的单次实验中，1-5 号信号分量和 6-10 号噪声分量的 p 值依然有很大的，至少是 10 个数量级以上的差距，或许可以考虑适当增大调节参数 λ 和 λ_j 的值来增强惩罚力度[17]。

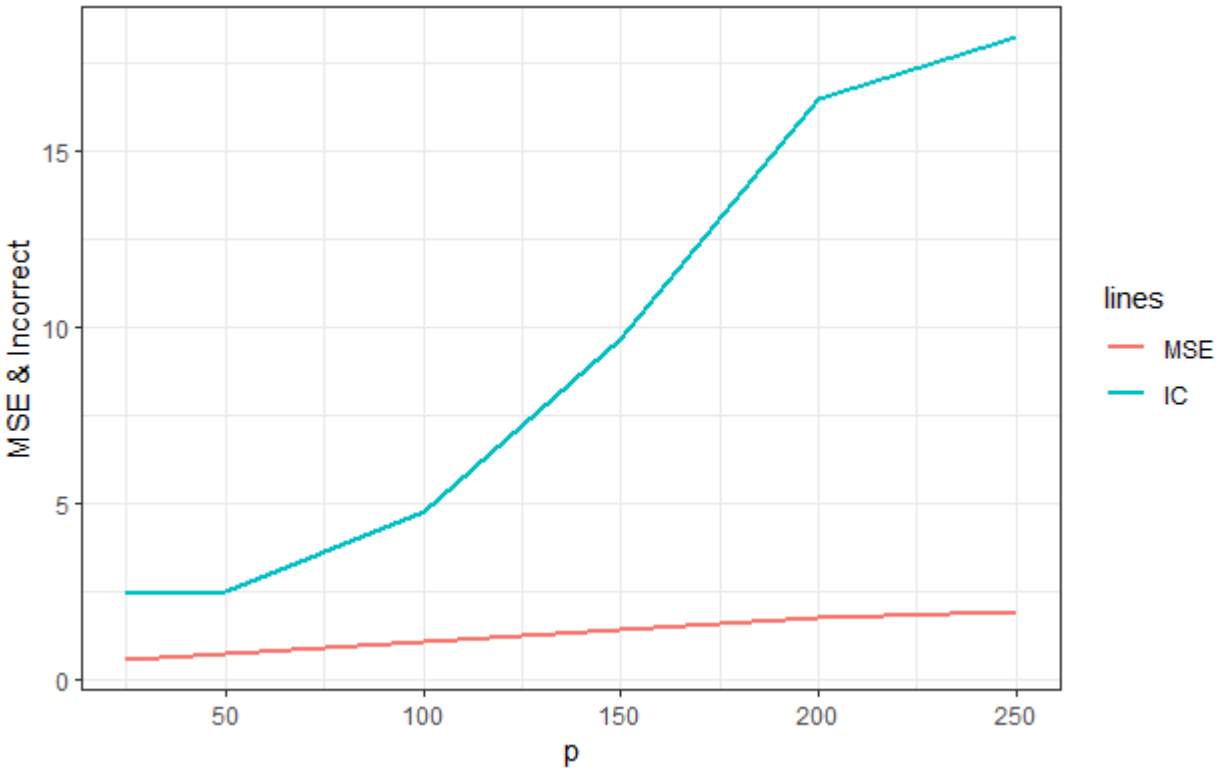

图 3　$n = 200$，从低维到高维的 MSE 与 Incorrect 变化

6 结论

本文的目的是讨论将双去偏 Lasso 方法应用于部分线性混杂变量模型的估计，主要研究了方法论和在高维、低维情形以及协变量不同相关性时的数值模拟结果。首先以拆解步骤的形式详细介绍了双去偏 Lasso 方法的估计方法论、去除偏差的逻辑和关键的理论性质证明。然后介绍了部分线性模型的两大常用估计方法：最小二乘估计和局部线性估计。利用这两种方法的思想可以将部分线性混杂变量模型变形为双去偏 Lasso 估计可以求解的形式并直接求解。

数值模拟的结果表明，在低维情形下，维数增大和协变量相关性增强时，新方法的估计效果会变差，同时会更多地将噪声变量错误识别为信号。而样本量增大时，估计效果会变好。但在所有模拟情况下新方法都能几乎选出所有的信号变量。所以，在模型稀疏的情况下，当对变量选择的准确性有较高要求时，可以对新方法选出的变量进行二次选择来去除多余噪声。

在高维情形下，新方法的估计效果不会急剧下降，去除混杂偏差的能力也不会失效。但随着维数不断增加，随之增加的高维偏差会使估计效果越来越差。这个问题的解决需要从惩罚估计 Lasso 本身入手，调节惩罚力度或尝试其他高维推断模型，如 SCAD、SIS 或 LDPE 也许是有用的[17]。

总体而言，本文研究的部分线性混杂变量模型的双去偏 Lasso 估计是有效可靠的。在实际应用中，对于具有非线性趋势且受混杂变量影响的数据，可以首先用部分线性混杂变量模型建模，再用双去偏 Lasso 估计。其在高维和稀疏情形下的优良表现也能应付更加复杂的问题，为实际工作者提供有力的支持。

参考文献

[1] VANDERWEELE T J, SHPITSER I. On the definition of a confounder[J]. Annals of Statistics, 2013, 41(1): 196-220.

[2] GREENLAND S, PEARL J, ROBINS J M. Confounding and collapsibility in causal inference[J]. Statistical Science, 1999, 14(1): 29-46.

[3] 陈强. 高级计量经济学及 Stata 应用 [M]. 北京: 高级计量经济学及 Stata 应用, 2014.

[4] GUO Z, ĆEVID D, BÜHLMANN P. Doubly debiased lasso: High-dimensional inference under hidden confounding[J]. Annals of Statistics, 2022, 50(3): 1320-1347.

[5] WOOLDRIDGE J M. Econometric Analysis of Cross Section and Panel Data[M]. [S.l.]: MIT press, 2010.

[6] ALBERT J M. Mediation analysis for nonlinear models with confounding[J]. Epidemiology (Cambridge, Mass.), 2012, 23(6): 879-888.

[7] HAHN P R, MURRAY J S, CARVALHO C M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)[J]. Bayesian Analysis, 2020, 15(3): 965-1056.

[8] ENGLE R F, GRANGER C W, RICE J, et al. Semiparametric estimates of the relation between weather and electricity sales[J]. Journal of the American Statistical Association, 1986, 81(394): 310-320.

[9] SHI J, LAU T S. Empirical likelihood for partially linear models[J]. Journal of Multivariate Analysis, 2000, 72(1): 132-148.

[10] LIANG H, LIU X, LI R, et al. Estimation and testing for partially linear single-index models[J]. Annals of Statistics, 2010, 38(6): 3811-3836.

[11] LI G, XUE L. Empirical likelihood confidence region for the parameter in a partially linear errors-in-variables model[J]. Communications in Statistics—Theory and Methods, 2008, 37(10): 1552-1564.

[12] HÄRDLE W, LIANG H, GAO J. Partially Linear Models[M]. [S.l.]: Springer Science & Business Media, 2000.

[13] FAN J. Local linear regression smoothers and their minimax efficiencies[J]. Annals of Statistics, 1993, 21(1): 196-216.

[14] HAMILTON S A, TRUONG Y K. Local linear estimation in partly linear models[J]. Journal of Multivariate Analysis, 1997, 60(1): 1-19.

[15] ĆEVID D, BÜHLMANN P, MEINSHAUSEN N. Spectral deconfounding via perturbed sparse linear models[J]. Journal of Machine Learning Research, 2020, 21(1): 9442-9482.

[16] HASTIE T, TIBSHIRANI R, FRIEDMAN J H, et al. The Elements of Statistical Learning: Data mining, Inference, and Prediction[M]. [S.l.]: Springer, 2009.

[17] ZHANG C H, ZHANG S S. Confidence intervals for low dimensional parameters in high dimensional linear models[J]. Journal of the Royal Statistical Society: Series B: Statistical Methodology, 2014, 76(1): 217-242.

致　谢

在论文完成之际，我首先要对我的导师李高荣教授表示最真挚的感谢。本文能够顺利完成和李老师热心、耐心并且高水平的指导是分不开的，无论是在论文的选题还是方法论的指导上，老师都给予了我很大的帮助。尤其是中途我因为突发决定出国深造以后，申请事务繁琐，时间紧迫而无暇关注毕业论文的推进时，老师也持续给我鼓励和现实的建议，让我能够回到自己舒服的节奏中完成这些工作。我在完成毕业论文的过程中扮演的角色就像是真正的学徒——学习真正的研究者们规范的工作是怎么样的，亲自尝试并获得成就感。除开毕业论文，李老师在我整个本科高年级阶段的指导和帮助也令我受益匪浅。老师的课程教给了我统计学的知识、锻炼了我的学术能力，对我留学申请的关心也令人感动。没能继续留在师大，在老师门下读研是令人遗憾的，但我会继续拥抱有挑战的未来，在人生的道路上前行。

感谢师大的平台和学术底蕴塑造了今天的我，以及统计学院的各位老师们对我的栽培。每每想到自己坐在教室里，被纯粹的知识所吸引的那些时刻，就会觉得这是人生中最高级的幸福。“怕什么真理无穷，进一寸有一寸的欢喜”，这种时刻也许以后不会再有了。但我依然会怀念那个沉迷于统计学本身的自己。

感谢我本科四年的所有同学、朋友、点头之交的人、萍水相逢的人，等等。我无法在这里列出所有人的名字，但他们同样塑造了我的全部。他们中有的人曾与我讨论各种问题，有的曾在竞争中激励我奋发向上，有的给了我很多有用的建议，有的帮我排解苦难领悟人生。落日下的球场、早晨八点的英东楼、半夜的澡堂，有过我孤独的身影，但更多的时候是陪伴。谢谢你们帮助了我的成长和成熟，我的答卷已经完成了，也祝诸君好。

感谢我的父母，他们殚精竭虑培养我，送我进入北师大，我是知道的；他们努力工作养育资助我，我是知道的；他们好好生活，不希望我为他们担心，但同时又为我而担心，我也是知道的。我自小家风朴素，家教严厉，虽应当基本达到了父母亲对我成材的期望，但也变成了在家不甚交谈的性格。此非我所愿，亦非我所选，但我希望我的这篇论文能让我的父母安心：你们的儿子不再是那个需要你们时刻担忧学业的小孩了，他有能力，也有自信面对未来的困难。

无冥冥之志者无昭昭之明，无惛惛之事者无赫赫之功。感谢始终探索，坚持热爱，保持善良，不忘初心的自己，祝我未来可期。

陈致远

2023 年 4 月