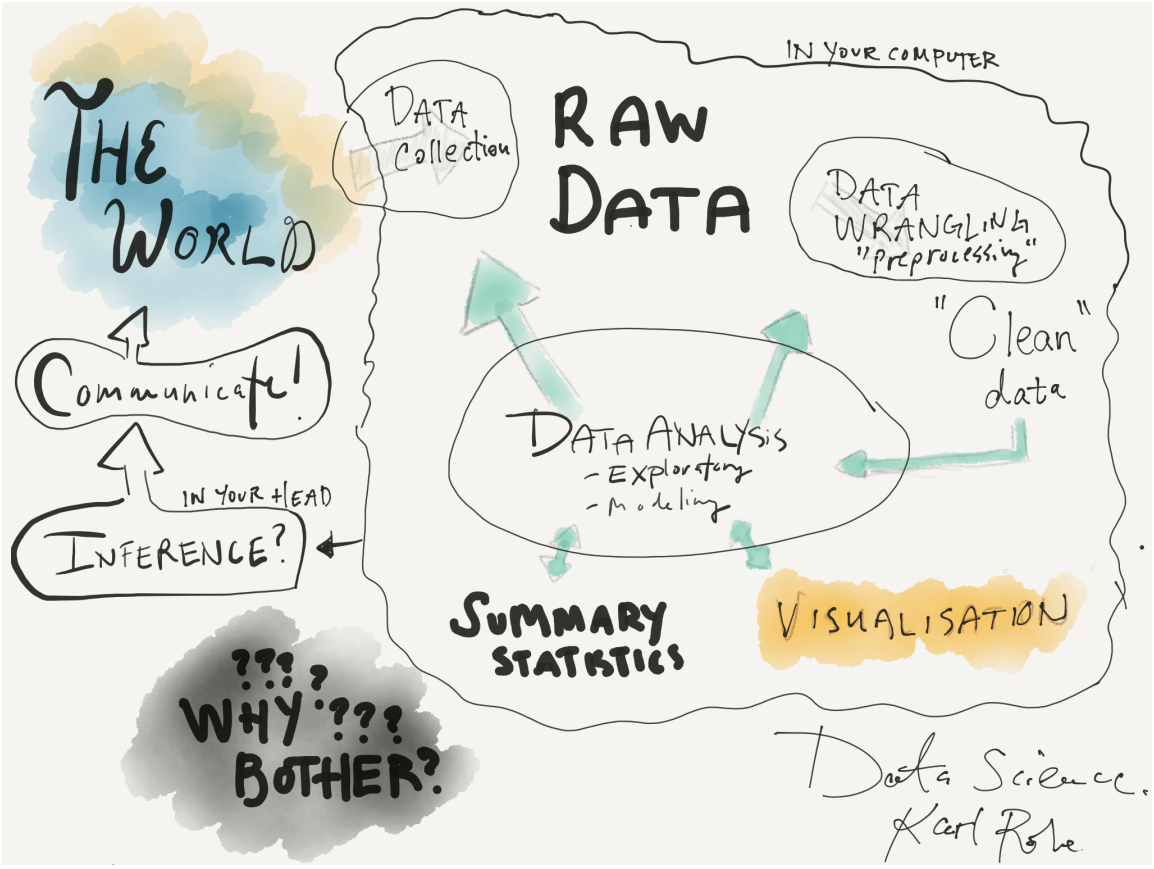
Notes on High-dimensional Statistics

Haojie Liang *

Department of Statistics, Zhejiang University



Last updated on 14:34 September 4, 2021

*E-mail: hjliang@zju.edu.cn

Contents

1 Principal component analysis

这一部分我参考了**高惠璇 2005 应用多元统计分析**的第七章和 Making sense of principal component analysis, eigenvectors & eigenvalues. 后者举了酒的例子，我觉得挺不错的.

1.1 Calculate PCA

1.1.1 Eigen-decomposition and SVD

一般有两种方法计算主成分：特征分解和奇异值分解（singular value decomposition, SVD）. 这两种方法在计算结果上有差别，但是本质上是一一对应的. 下面的讨论我主要参考了Relationship between SVD and PCA. How to use SVD to perform PCA?, amobea 的回答十分清楚，并且给出了丰富的链接，有更多的 post 阐述进一步的细节. 这里我们先列出我习惯的记号，同时列出主要的概念和计算.

Let the data matrix $\boldsymbol{X}$ be of $n \times p$ size, where n is the number of samples (sample size) and p is the number of variables. Assume that $\boldsymbol{X}$ is centered, i.e. column means have been subtracted and are now zero-mean.

(**Covariance matrix**), $p \times p$ matrix

$$C \coloneqq \boldsymbol{X}^\intercal \boldsymbol{X}/(n-1) \tag{1.1}$$

(**Eigen-decomposition**) for $\boldsymbol{C}$

$$C = VLV^\intercal \tag{1.2}$$

其中 $V = (v_1, v_2, \ldots, v_p)$, $\|v_j\|_2 = 1$

(**Eigen-values**), and λ_i is in decreasing order on the diagnal:

$$L = \operatorname{diag}\left(\lambda_1, \ldots, \lambda_p\right) \tag{1.3}$$

(**Principal direction**) Principal axes, also principal directions, eigen-vectors

$$V = (v_1, \ldots, v_p) \tag{1.4}$$

Then we have $V^\intercal V = I_p$, also for the first r eigenvectors (columns) $V^{(\cdot r)}$, we have $V^{(\cdot r)^\intercal} V^{(\cdot r)} = I_r$.

(**Principal components**) also PC scores, transformed variables: project original p variables on every eigen-vectors, project data matrix $\boldsymbol{X}$ on j-th principal axes,

$$(XV)_{\cdot j} = Xv_j \tag{1.5}$$

project i-th row of data $\boldsymbol{X}$, i.e. $(X)_{i\cdot}$ on all eigen-vectors V,

$$(XV)_{i\cdot} \tag{1.6}$$

(**Project data on Principal directions**) Since $C = VLV^\intercal$ and $V^\intercal V = I_p$, we have

$$CV = VL \tag{1.7}$$

Also, project data (covariance) $\boldsymbol{C}$ on the first r principal directons $V^{(\cdot r)}$, then

$$CV^{(\cdot r)} = VLV^\intercal \cdot V^{(\cdot r)} \tag{1.8}$$

since $VLV^\intercal = V^{(\cdot r)}L^{(r)}V^{(\cdot r)^\intercal} + V^{(\cdot -r)}L^{(-r)}V^{(\cdot -r)^\intercal}$ and $V^{(\cdot -r)^\intercal} \cdot V^{(\cdot r)} = 0$, we arrive that

$$CV^{(\cdot r)} = V^{(\cdot r)}L^{(r)} \tag{1.9}$$

(**SVD**) for data matrix $\boldsymbol{X}$

$$X = USV^\intercal \tag{1.10}$$

(**Left-singual matrix**) this is a unitary matrix, $n \times p$ matrix, $U^\intercal U = I_p$,

$$U$$

(**Singular value matrix**) $p \times p$ diagonal matrix,

$$S = \text{diag}\left(s_1, \ldots, s_p\right)$$

(**Right-singular matrix**) that is $p \times p$ eigen-vector matrix in Equation 1.4

$$V$$

(**Equivalence**) Relationship between eigen-decomposition and SVD: From the eigen-decomposition of covariance matrix, $\boldsymbol{C} = VLV^\intercal$ and the definition

$$\begin{aligned}
\boldsymbol{C} &= X^\intercal X/(n-1) \\
&= (VSU^\intercal)(USV^\intercal)/(n-1) \\
&= VSU^\intercal USV^\intercal/(n-1) \\
&= V \cdot \frac{S^2}{n-1} \cdot V^\intercal
\end{aligned}$$

which imples

$$\frac{S^2}{n-1} = L \tag{1.11}$$

that is

$$\frac{s_i^2}{n-1} = \lambda_i \tag{1.12}$$

(PC scores) in SVD form

$$XV = (USV^\intercal)V = US \tag{1.13}$$

(**the role of U**) Standard scores $\sqrt{n-1}U$

$$\begin{aligned}(US)L^{-1/2} &= (XV)L^{-1/2}\\ (US)\sqrt{n-1}S^{-1} &= (XV)\sqrt{n-1}S^{-1}\\ \sqrt{n-1}U &= \sqrt{n-1}XVS^{-1}\end{aligned}$$

(**Loadings**) $VL^{1/2}$, this is the non-scaled eigen-vectors, you can refer to Loadings vs eigenvectors in PCA: when to use one or another?

$$VL^{1/2} = VS/\sqrt{n-1} \tag{1.14}$$

(**Rank-r approximation**) denote $X^{(\cdot r)}$ as the first-r columns of X, $X^{(r)}$ as the rank-r approximation estimator,

$$X = USV^\intercal \tag{1.15}$$

$$X \leftarrow X^{(r)} = U^{(\cdot r)}S^{(\cdot r)}(V^{(\cdot r)})^\intercal \tag{1.16}$$

also

$$XV^{(\cdot r)} = U^{(\cdot r)}S^{(\cdot r)} \tag{1.17}$$

Remark: If n > p, then the rank of X is at most p. This means the last $(n-p)$ eigen-values must be 0, then

$$X = X^{(p)} = U^{(\cdot p)}S^{(\cdot p)}(V^{(\cdot p)})^\intercal \tag{1.18}$$

1.1.2 Rideg Regression

Zou et al. (2006) founded that PCA is closely connected with linear regression, exactly speaking, *Ridge Regression*. Let X be a $n \times p$ data matrix, the SVD of X be $USV^\intercal$. By Equation 1.13 the $n \times p$ matrix $US = (U_1 s_1, \ldots, U_p s_p)$ are actually the principal components of X. Now our goal is to recover eigenvector V from US (obtained from SVD) and X. This is an equivalent approach of eigen-decomposition for covariance matrix.

Theorem 1.1. *Regardless of $n > p$ or $p > n$, given data matrix, we can recover the eigen-vector/principal direction through ridge regression.*
Conditions:

1. *data matrix X*
2. *some $\lambda > 0$ (s.t the solution is unique even when $p > n$)*

Results:

1. *Let the estimate $\widehat{\beta}_{ridge}$ be the minimizer of*

$$\|U_k s_k - X\beta\|^2 + \lambda\|\beta\|^2$$

and $\widehat{v} = \widehat{\beta}_{ridge}/\|\widehat{\beta}_{ridge}\|$, then $\widehat{v}$ is the k-th eigen-vector.

Note. This is the Theorem 1 in Zou et al. (2006). 这个命题十分重要，它表明，即使 data matrix X 是非列满秩的 (especially the high-dimensional case $p > n$)，也可以通过选择合适的 tuning parameter 来恢复 eigen-vector, and the ridge penalty will not affect the estimate of eigen-vector. 因此，在样本协方差阵的 eigen-decomposition 之外，通过 ridge regression 一样可以确定原始变量的新的线性组合 (principal components). Thus we extend PCA to a more general case of ridge regression in torder to handle high-dimensional data.

1.1.3 Minimizing reconstruction error

Ridge regression method is not a perfect result. In this way, in order to recover the eigen-vector/prncipal direction, we should first perform SVD, then run the regression. Thus it is not a genuine alternative.

We can recover eigen-vector through minimizing sum of squared Euclidean distance.

这种观点有着良好的几何直观.Let X be a $n \times p$ data matrix.

Theorem 1.2.
Conditions:

1. $$\widehat{\Gamma}_k = \arg\min_{\Gamma_k} \left\| X - X\Gamma_k{}^\intercal \Gamma_k \right\|^2$$
$$\textit{subject to } \Gamma\Gamma_k{}^\intercal = I_k\,,$$

   where Γ_k *is a* $k \times p$ *matrix.*

2. k *can be* $1, \ldots, p$

Results:

1. *(Every row vector of)* $\widehat{\Gamma_k}$ *is the first* k *eigen-vectors.*

   Let Γ_k be

$$\begin{bmatrix} \eta_1{}^\intercal \\ \vdots \\ \eta_k{}^\intercal \end{bmatrix}.$$

To reveal the geometry of PCA, denote $p \times 1$ vector X_i as the i-th observation of p original variables. Let $p \times 1$ vector η_1 with $\|\eta_1\| = 1$ be a directon vector, then $\langle X_i, \eta_1 \rangle / \|\eta_1\| = X_i \cdot \frac{\eta_1}{\|\eta_1\|} = \eta_1{}^\intercal X_i$ is the η_1-corrdinate value of X_i, also the projection (or length) of X_i on η_1. The p-dimensional coordinate value[1] of $\eta_1{}^\intercal X_i$ is $\eta_1 \cdot (\eta_1{}^\intercal X_i)$

This post, Everything you did and didn't know about PCA, written by Alex Williams, is a detailed and overall review for PCA. The *An alternative optimization problem* section discuss the Euclidean reconstruction error perspective.

[1] p 个原始变量下的坐标值

1.2 Motivation and intuition

主成分分析是高维统计中一种经典的降维方法. 实际中我们会用多个变量去刻画一个（也就是“一个个”）研究对象，比如用 p 个变量，如酒精度、颜色等，去描述收集到的 100 瓶酒，每瓶酒都是一个 sample . 由于样本的随机性，p 个变量也就成了 p 维随机向量. 现实情景中 p 可能会很大，这样就不利于我们去比较 100 瓶酒的差别（比如说要给这 100 瓶酒分类）[2]. 因此，PCA 就是将多指标[3]转化为少数几个能够尽可能多地总结[4]100 个散点[5]的（新的）指标的方法，并且要求这少数几个新的变量是互不相关的.Lasso 也是一种降维的方法，确切地说，是通过变量选择直接剔除那些冗余的变量. 然而 PCA 并非如此：PCA 希望基于原始的 p 个变量构造出新的更有总结性（解释力）的变量，构造的方法是线性组合. 因此，PCA 的目标是：construct the best new characteristics that summarize the list of wines as well as only possible among all conceivable linear combinations. 这是 PCA 的直觉.

接下来用数学语言来说明这个直觉：设 X 为所感兴趣的 p 维向量（如酒精度，酒的颜色等），$\Sigma \overset{def}{=} \operatorname{cov}(X), \mu \overset{def}{=} \mathbf{E}(X)$, $a_1, \ldots, a_p$ 为 p 个线性组合.

$$
Z = \begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix} = \begin{bmatrix} a_1{}^{\intercal} X \\ \vdots \\ a_p{}^{\intercal} X \end{bmatrix} = \begin{bmatrix} a_1^{\intercal} \\ \vdots \\ a_p^{\intercal} \end{bmatrix} X \triangleq A^{\intercal} X,
$$

$$
\operatorname{Var}(Z_k) = a_k^{\intercal} \Sigma a_k,
$$

$$
\operatorname{cov}(Z) = A^{\intercal} \Sigma A.
$$

下面，我先直接列出求解主成分的 3 个目标：

1. 用方差来表示中所含有的那 100 瓶酒（也就是代表总体的随机变量 X，不是一个个变量的 p 个的全体）的信息；

2. $\|a_i\|_2 = 1$，这就是对坐标轴的量纲的限制；

3. Z_i 与 $Z_j, i \neq j$ 不含有共同的信息，即 $\operatorname{cov}(Z_i, Z_j) = 0, i \neq j$.

[2]大量的变量中还伴随着 p 个变量之间是相关的，那么 p 个分量之间的信息是有重叠的，也可以说 p 个变量略显冗余（redundant）

[3]或者说 variables, properties

[4]对“总结”的注释：或者说“summarize our list of wines (all observed objects) well”. 每个 sample 是一个向量，因此所有的研究对象，即这 100 瓶酒，就是 p 维向量空间上的一堆点. 一个主成分是去总结散点之间的差异，而不一定提炼了所有 p 个变量中的大部分信息. 究竟怎样才是“总结 100 个散点”和“总结散点之间的差异”会在之后叙述.

[5]注意，不是 p 个变量

1.3 What is "PCs summarize the information in sample"?

为了更有力地得到计算主成分的优化目标，我们先来阐述前面未解决的“总结”，也就是在统计中用什么量来表达 Z_k 含有的 100 瓶酒（也就是代表总体的随机向量）的信息.

即便在 X 的 p 个分量中，$\mathrm{Var}(() X_i)$ 也是有大有小，$\mathrm{Var}(() X_i)$ 越大表示 $X_{(t)}, t = 1,\ldots,100$ 在 X_i 上的投影（其实就是 X_i 本身，X_i 可以看成 i^{th} 坐标轴的坐标）越分散，即 X_i 反映了这 100 瓶酒较大程度的差异. 若 $\mathrm{Var}(() X_i) \approx 0$，即这 100 瓶酒在 X_i 上几乎相同，那么单从 X_i 无法分辨出 100 瓶酒之间显著的差异. 以 X_i 为酒精度为例，若 100 瓶酒包含了啤酒、红酒、黄酒和白酒等各种酒，那么酒精度从 3° 到 53° 跨越很大，通过酒精度这个量能较好分辨；若 100 瓶酒均为酒精度从 3° 到 4° 的啤酒，则酒精度对于分辨这 100 瓶酒的意义不大. 因此，some properties that strongly differ across wines 是好的，a properties that is the same for most of the wines 是坏的. 考虑一般的 $X_1,\ldots,X_p$ 的线性组合，同样要求 properties show as much variation across wines as possible. 方差最大化是理解“总结”的第一个方面.

第二个方面是从误差的角度来考虑的. 用新的坐标系去刻画样本点与原始信息之间的误差希望在某种度量下（比如 averaged squared distance）达到最小.

我们在后面会说明这两个方面其实是等价的（根据勾股定理）.

下面我给出计算总体主成分的优化目标，并将通过计算来阐述优化目标在两个方面的意义:

设 X 是一个 p 维随机向量，那么最多有 p 个主成分（实际只会使用少量几个主成分）. 从第一个主成分开始，依次是这么计算的：

$$a_i = \underset{\substack{a,\|a\|=1\\ a^\intercal \Sigma a_j=0,\\ j=1,\ldots,(i-1)}}{\arg\max} \mathrm{Var}\left(a^\intercal X\right) = \arg\max a^\intercal \Sigma a\,. \tag{1.19}$$

在代数与几何上的含义：从代数上看，主成分是原始变量的线性组合；从几何上看，是将以 $X_1,\ldots,X_p$ 为轴的坐标系旋转得到新坐标系 $Z_1,\ldots,Z_p$（或者说 $Z_1,\ldots,Z_m, m < p$）：

$$\begin{bmatrix} X_1 \\ \vdots \\ X_2 \end{bmatrix} \mapsto \begin{bmatrix} a_1^\intercal X \\ \vdots \\ a_2^\intercal X \end{bmatrix} \tag{1.20}$$

这两层含义可以在两个方面上去理解：

1. 新坐标是方差最大化的方向：

新坐标系的交点其实与原始坐标系是一样的.

$$\begin{bmatrix} X_1 - \mu_1 \\ \vdots \\ X_2 - \mu_2 \end{bmatrix} \mapsto \begin{bmatrix} a_1^\intercal (X - \mu) \\ \vdots \\ a_2^\intercal (X - \mu) \end{bmatrix} \tag{1.21}$$

在一些教材/帖子中，之所以会将新坐标系的交点移到均值，是为了更好地反映 variation of spread（因为计算方差要中心化）. 计算 a_1 时是通过样本点在 $a_1^\intercal X$ 上的投影（也就是在 $a_1^\intercal X$ 轴上的坐标）的 spread 达到最大，因此必然要用到 $a_1^\intercal X$ 的均值 $a_1^\intercal \mu$. 为此，一开始就对数据中心化，得到 $X - \mu$，然后计算主成分 $a_1^\intercal (X - \mu)$，这样更能理解为何交点在 μ 处. 为了说明 $a_1^\intercal X$ 与 $a_2^\intercal X$ 在图上垂直，且 $a_1^\intercal X$ 与 $a_2^\intercal X$ 不相关，还需要解释正交与不相关的联系，见subsection 1.4

$$\lim_{n\to\infty} \sum_{i=1}^{n} a_1^\intercal (x_i - \mu) \cdot a_2^\intercal (x_i - \mu) \tag{1.22}$$

$$= \lim \sum a_1^\intercal (x_i - \mu)(x_i - \mu)^\intercal a_2 \tag{1.23}$$

$$= a_1^\intercal \Sigma a_2 = 0 \tag{1.24}$$

我们说 $a_1^\intercal X$ 与 $a_2^\intercal X$ 垂直，并非因为 $a_1^\intercal a_2 = 0$，并且也没有 $\sum_{i=1}^{n} (a_1^\intercal x_i)(a_2^\intercal x_i) \to 0$，而是 $\operatorname{cov}\left(a_1^\intercal X, a_2^\intercal X\right) = 0$. 同时也很容易理解为什么要 $\|a\|_2 = 1$，因为以 $a^\intercal X$ 为坐标，就要让量纲从 (X_1, X_2) 到 $(a_1^\intercal X, a_2^\intercal X)$ 没有变化. 也可以这么看第一个方面：要求 the variation of values along this (new) line should be maximal，这一点在图上的几何意义是要求 the spread (variance) of the red dots 需最大化. 故我们不关心原点，处处计算与均值的距离. 新坐标是样本方差达到最大的方向，至于方向之间的所看到的垂直并非一定是不相关，而是用不相关来定义垂直.

2. 新坐标中误差**平方**距离的最小化：
   我们用新坐标 $\begin{bmatrix} a_1^\intercal X \\ a_2^\intercal X \end{bmatrix}$ 代替 $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ 时会在每一维度上产生误差. 在图上看，误差不是 X 到 $a_1^\intercal X$ 的距离，而是要将新坐标系的交点移到 μ，然后是 X 到 $a_1^\intercal (X - \mu)$ 的距离（或者一开始对所有数据中心化）. 这样做的原因是：均值 μ 是数据集的中心，我们关心 spread 偏离中心的距离是否达到最小.

   也可以这么看第二个方面：关心 reconstruction error (the average squared distance from the center of the wine) 是否达到最小.

我们可以用 2 维的例子来阐述两个方面的等价性：$a_1^\intercal (X - \mu)$ 本身即为 $X - \mu$ 在 $a_1^\intercal X$ 上的投影到 $a_1^\intercal \mu$ 的距离，X 与 μ 的距离即为欧氏距离 $\operatorname{Var}(() X_1) + \operatorname{Var}(() X_2)$，

这是不依赖于 a_1 的常数，而在 $a_1^\intercal(X-\mu)$ 上的投影的长度为 $\mathrm{Var}(()a_1^\intercal X) = a_1^\intercal \Sigma a_1$. 因此由勾股定理知 $\max \mathrm{Var}(()a_1^\intercal)$ 等价于最小化距离平方和.

1.4 Orthogonality and uncorrelation

我觉得(https://stats.stackexchange.com/users/17023/a-donda) (n.d.) 将这两个概念阐述得非常清楚，这里面还嵌着恰到好处、表达到位的其他 links.

不相关是十分清楚的，先单独阐述 orthogonality (adj, orthogonal)，这可以译为正交和垂直. 这本是 geometry 与 linear algebra 中的概念，在几何中可以反映为垂直，在线性代数中则对应于内积空间.

在引出正交之前，必须先定义 inner product. 注意内积是可以有多种不同的定义方式的. 常用的向量内积是对应分量得乘积和. 当考虑 2 个随机变量 X,Y 的内积时，其定义可以直接沿袭或稍微拓展向量内积的定义方式，有以下 3 种：

1. 一般地，可以用 $\mathrm{cov}(X,Y)$ 作为 X 与 Y 的内积；
2. 若有 X 与 Y 的 n 次 realizations，那么也就有 a sequence of numbers，则 X 与 Y 的内积可定义为 $\sum_{i=1}^n x_i y_i = 0$；
3. 也有用 $\mathbf{E}(XY)$ 作为 X 与 Y 的内积.

后两种定义方式本质上是一样的，是向量内积的直接继承；而第一种方式的正交，也就意味着不相关了. 因此，需要注意上下文中随机变量内积的定义形式.

在 Regression 中往往有"orthogonal design" 的说法，这其实指的是各 independent variables 两两之间乘积的期望为 0，即 $\mathbf{E}(()XY) = 0$. 记 design matrix 为

$$X = \left[\ell_1, \ldots, \ell_p\right], \tag{1.25}$$

那么各个列向量表示各个 independent variables，故 $\mathbf{E}\left(X_i X_j\right) = \ell_i^\intercal \ell_j = 0, i \neq j$. 实际中往往还有每个变量的归一化：$\ell_i^\intercal \ell_i = 1$，从而有 $X^\intercal X = I_p$. 正交设计中的 orthogonality 没有太多统计含义，只是为了推导结论的便捷.

1.5 Why does PCA involve eigen theory?

首先从 spectral theorem 来看，Λ can be diagnalized by choosing a new orthogonal coordinate system, given by its eigenvectors. Λ 的对角线元素即为 eigenvalues.

$$\Lambda = A^\intercal \Sigma A,\ Z = \mathrm{cov}(A^\intercal X). \tag{1.26}$$

第二个方面是，the variance of any projection will be given by a weighted average of the eigenvalues. Any projection 也就是 $X_1, X_2, \ldots, X_p$ 的新的线性组合，或者说 $Z_1, Z_2, \ldots, Z_p$ 的某种线性组合，那么方差即为

$$Z = A^\intercal X,\ X = AZ,, \tag{1.27}$$

$$若Y = LX, \tag{1.28}$$

$$那么Y = LAZ \tag{1.29}$$

$$= \begin{bmatrix} \ell_1^\intercal \\ \ell_2^\intercal \end{bmatrix} \begin{bmatrix} a_1, a_2 \end{bmatrix} Z \tag{1.30}$$

$$= \begin{bmatrix} \ell_1^\intercal a_1 & \ell_1^\intercal a_2 \\ \ell_2^\intercal a_1 & \ell_2^\intercal a_2 \end{bmatrix} Z \tag{1.31}$$

其中我认为最后一个式子的线性变换必然可以证明也是一个正交阵. 因此，最大的方差即为将权重全部放在最大特征根上，也就是只采用第一特征向量.

1.6 Properties of PCs

先再次交代一下记号：$Z = A^\intercal X = \begin{bmatrix} a_1^\intercal \\ \vdots \\ a_p^\intercal \end{bmatrix} X, \Lambda \overset{def}{=} \mathrm{cov}(Z) = \mathrm{diag}\,(()\, \lambda_1, \ldots, \lambda_p), \Sigma \overset{def}{=} \mathrm{cov}(X) = (\sigma_{ij})_{p\times p}$ 为 Σ 的 p 个特征根

性质 1：$\Lambda = \mathrm{diag}\,(()\, \lambda_1, \ldots, \lambda_p)$, 即主成分的方差为相对应的 eigenvalue，且主成分之间是互不相关的.

Proof: 由 spectral theorem，$\Sigma = \sum_{i=1}^{p} \lambda_i a_i a_i^\intercal$ 这就是为了分解出方差能达到最大的各个方向.

$$\mathrm{Var}\,(()\, a_i^\intercal X) = a_i^\intercal \Sigma a_i \tag{1.32}$$

$$= \sum_{j=1}^{p} a_i^\intercal (\lambda_j a_j a_j^\intercal) a_i \tag{1.33}$$

$$= \lambda_i \tag{1.34}$$

性质 2：$\sum_{i=1}^{p} \sigma_{ii} = \sum_{k=1}^{p} \lambda_l$，也就是说 PCA 的本质为将原始变量 X 的分量的方差再分配，分解为主成分的方差的和 Proof:

$$\because Z = A^\intercal X \tag{1.35}$$

$$\therefore X = AZ \tag{1.36}$$

$$\therefore \Sigma = A\Lambda A^\intercal \tag{1.37}$$

$$\mathrm{tr}\,(\Sigma) = \mathrm{tr}\,(() A\Lambda A^\intercal) = \mathrm{tr}\,(() A^\intercal A\Lambda) = \sum_{k=1}^{p} \lambda_k\,. \tag{1.38}$$

性质 3：主成分 Z_k 与原始变量 X_i 的相关系数为

$$\rho(Z_k, X_i) = \frac{\mathrm{cov}\left(a_k^\intercal X, e_i^\intercal X\right)}{\sqrt{\lambda_k \sigma_{ii}}} \tag{1.39}$$

$$= \frac{a_k^\intercal \Sigma e_i}{\sqrt{\lambda_k \sigma_{ii}}} \tag{1.40}$$

$$= \sqrt{\frac{\lambda_k}{\sigma_{ii}}} a_{ik}\,, \tag{1.41}$$

这也称为因子负荷量. 第 k 列才意味着第 k 个主成分

性质 4：$\sum_{k=1}^{p} \rho^2(Z_k, X_i) = \sum_{k=1}^{p} \frac{\lambda_k}{\sigma_{ii}} a_{ik}^2 = 1$. Proof: 先交代记号：$A = \begin{bmatrix} \ell_1^\intercal \\ \vdots \\ \ell_p^\intercal \end{bmatrix}$，其中 $\ell_i^\intercal = (a_{i1}, \ldots, a_{ip})$.

对这一性质有如下的解释：首先可知各个主成分 Z_k 是原始变量 $X_1, \ldots, X_p$ 的线性组合，即 $Z = A^\intercal X$，那么反过来，X_i 也是 $Z_1, \ldots, Z_p$ 的线性组合，即 $X = AZ = \begin{bmatrix} \ell_1^\intercal \\ \vdots \\ \ell_p^\intercal \end{bmatrix} Z$，而 $Z_1, \ldots, Z_p$ 是互不相关的，从而

$$\mathrm{Var}\,(() X_i) = \sum_{k=1}^{p} \mathrm{Var}\,(() a_{ik} Z_k) \tag{1.42}$$

本质含义：将每个原始变量的信息分配给各个主成分.

性质 5：$\sum_{i=1}^{p} \sigma_{ii} \rho^2(Z_k, X_i) = \sum_{i=1}^{p} \lambda_k a_{ik}^2 = \lambda_k = \mathrm{Var}\,(() Z_k)$ Proof: 对这一性质有如下的解释：Z_k 可以表示成 $X_1, \ldots, X_p$ 的线性组合，但是 $X_1, \ldots, X_p$ 之间是相关的，从而没有像性质 4 那样的计算形式（我猜可能可以从回归分析的全相关系数对性质 4 和性质 5 给出统一的解释），只不过性质 4 中的 independent variables 是互不相关的，而性质 5 中是相关的.

1.6.1 PCs for standardized variables

对各原始变量标准化以消除量纲.

性质 2：$p = \sum_{i=1}^{p} \sigma_{ii} = \sum_{k=1}^{p} \lambda_k$

性质 3(相关系数)：$\rho(Z_k, X_i) = \sqrt{\frac{\lambda_k}{\sigma_{ii}}} a_{ik} = \sqrt{\lambda_k} a_{ik}$

性质 4：$\sum_{k=1}^{p} \rho^2(Z_k, X_i) = \sum_{k=1}^{p} \frac{\lambda_k a_{ik}^2}{\sigma_{ii}} = \sum_{i=1}^{p} \lambda_k a_{ik}^2 = 1$

性质 5：$\sum_{i=1}^{p} \sigma_{ii} \rho^2(Z_k, X_i) = \sum_{i=1}^{p} \rho^2(Z_k, X_i) = \lambda_k$

1.7 Sparse PCA

下面将主要来阐释Jolliffe et al. (2003) 是如何将 PCA 问题和线性回归联系起来，然后加上 lasso penalty. 先交代一些记号：

- $\boldsymbol{X}$: d×n data matrix
- $\widehat{\Gamma} = \begin{bmatrix} \widehat{\eta}_1^\intercal \\ \vdots \\ \widehat{\eta}_d^\intercal \end{bmatrix}$: d×d eigenvector matrix. 由正交分解定理知，这是一个正交阵；
- $\widehat{\Gamma}_k = \begin{bmatrix} \widehat{\eta}_1^\intercal \\ \vdots \\ \widehat{\eta}_k^\intercal \end{bmatrix}$: k×d, first k eigenvector matrix;
- $\widehat{W} = \widehat{\Gamma}\boldsymbol{X}$: full PC scores matrix, every column-vector records the pc scores of corresponding sample in $\boldsymbol{X}$;
- $\widehat{W}^{(k)} = \widehat{\Gamma}_k \boldsymbol{X}$: j-th row vector records the j-th PC scores of (all) samples.

This is the second perspective based on the data matrix $\boldsymbol{X}$ reconstruction.

$$\boldsymbol{X} \leftarrow \widehat{\boldsymbol{X}} = \widehat{\Gamma}_k^\intercal \widehat{\Gamma}_k \boldsymbol{X}$$

1.8 Other discussion

$a_1^\intercal X$ 在原始坐标系中的坐标是 $a_1 = (a_{11}, a_{12})$，$a_2^\intercal X$ 即为 $a_2 = (a_{21}, a_{22})$，这也就是为何要对 a_1 与 a_2 的进行 ℓ_2-norm 限制的原因：这时 $a_1^\intercal X$ 才是 X 在 $a_1^\intercal X$ 上的坐标，这是可以通过一下计算来验证的：

首先在原始坐标系中，$X = x_1\vec{e}_1 + x_2\vec{e}_2$, 设 $\vec{e}_3$ 与 $\vec{e}_4$ 为新的坐标系，那么 $\vec{e}_3 = (a_{11}, a_{12})$, $\vec{e}_4 = (a_{21}, a_{22})$. 只需要验证 $a_1^\intercal X\vec{e}_3 + a_2^\intercal X\vec{e}_4 = x_1\vec{e}_1 + x_2\vec{e}_2$.

$$a_1^\intercal X \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} + a_2^\intercal X \begin{bmatrix} a_{21} \\ a_{22} \end{bmatrix} \tag{1.43}$$

$$= (a_{11}x_1 + a_{12}x_2) \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} + (a_{21}x_1 + a_{22}x_2) \begin{bmatrix} a_{21} \\ a_{22} \end{bmatrix} \tag{1.44}$$

$$= \begin{bmatrix} (a_{11}^2 + a_{21}^2)x_1 + (a_{12}a_{11} + a_{22}a_{21}x_2) \\ (a_{11}a_{12} + a_{21}a_{22})x_1 + (a_{12}^2 + a_{22}^2)x_2 \end{bmatrix} \tag{1.45}$$

$$= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{1.46}$$

这里最后一个等式利用了 A 是一个正交阵.

2 Factor model

高惠璇 (2005)

2.1 Relationship with PCA

因子模型和 PCA 有相同点与不同点，它们之间的联系可以概括为

1. 都是对高维数据进行**降维**的手段，但是具体方法不同. 尽管两种方法直观上都是希望用一组少量的变量来近似 p 维原始变量，但是，PCA 通过线性组合将 p 维原始变量转化为少数几个主成分，而因子模型则假设存在一组未知的公因子可以线性地近似原始变量[6]；

2. 都是对原始数据的**协方差阵的近似**，但是具体方法不同. 如前所述，PCA 可以看成时秩 k 逼近，也就是用谱分解的前 k 项去近似原始变量的协方差阵. 但是，因子模型[7]是用因子载荷矩阵和特殊因子的对角阵来近似协方差阵. 若采用主成分法去估计因子模型（其实就是估计因子载荷矩阵），则 PCA 可以看成是因子模型的一个特例. 但是，显然，因子模型更有一般性；

3. 因子模型是一种线性**统计模型**，而 PCA 只是一种转换数据的方法（而不是统计模型，因此它们不同），但同时 PCA 又可以从线性模型的角度去理解（它们又是相似的）. 因子模型的有效性依赖于模型假设是否成立，而 PCA 可以对一切原始变量之间存在相关性的数据使用. 然而，正如之前所述，可以从最小化欧氏距离平方和的角度看 PCA，并且还有岭回归方法还原特征向量，我们可以将 PCA 放到线性回归模型的背景下来看，这时它和因子模型又是相似的.

在数据分析中，可以遵循如下准则去判断用 PCA 还是因子模型[8]:

- Run factor analysis if you assume or wish to test a theoretical model of latent factors causing observed variables;

- Run principal component analysis If you want to simply reduce your correlated observed variables to a smaller set of important independent composite variables.

[6]公因子模型的模型假设是需要检验的

[7]公因子模型

[8]Refer to"What are the differences between Factor Analysis and Principal Component Analysis?", and the answer of Jeromy Anglim

2.2 Introduction

接下来我们从一个实际的例子引出最一般的因子模型. 首先是总体形式：设 $X = (X_1, \ldots, X_p)^\intercal$ 是某学生 p 门学科的成绩的（中心化后的）p 维随机向量. 假设 p 个分量 $X_1, \ldots, X_p$ 受到几个不可观测的共同的因素的影响，例如学生的数理能力、表达能力等. 令 $F = (F_1, \ldots, F_k)^\intercal$ 是这 k 个未知的共同因素 (往往要求 k 较小，否则没有降维的效果). 为了用 F 的线性组合去近似原始变量 X，让 $\epsilon = (\epsilon_1, \ldots, \epsilon_p)^\intercal$ 为各个科目的特殊因子（互不相关的噪声），不妨建立如下线性模型

$$X_j = a_j^\intercal F + \epsilon_j, \quad j = 1, \ldots, p \tag{2.1}$$

or

$$X = \begin{bmatrix} a_1^\intercal \\ \vdots \\ a_p^\intercal \end{bmatrix} F + \epsilon =: AF + \epsilon\,. \tag{2.2}$$

其中 a_j 是 $k \times 1$ 的因子载荷向量，那么 A 是 $p \times k$ 的因子载荷矩阵. 这就是总体形式下的因子模型. 再来看经验形式：设 $X = (x_1, \ldots, X_n)^\intercal$ 为 $n \times p$ 的信息矩阵，其中 ith 行向量 $x_i^\intercal$ 为 ith 学生的 p 门学科的成绩，那么

$$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix} = \begin{bmatrix} a_1^\intercal \\ \vdots \\ a_p^\intercal \end{bmatrix} F_i + \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ip} \end{bmatrix} =: AF_i + \epsilon_i\,, \tag{2.3}$$

or

$$X = \begin{bmatrix} x_1^\intercal \\ \vdots \\ x_n^\intercal \end{bmatrix} = \begin{bmatrix} F_1^\intercal \\ \vdots \\ F_n^\intercal \end{bmatrix} (a_1, \ldots, a_p) + \begin{bmatrix} \epsilon_1^\intercal \\ \vdots \\ \epsilon_n^\intercal \end{bmatrix} =: FA^\intercal + \epsilon\,, \tag{2.4}$$

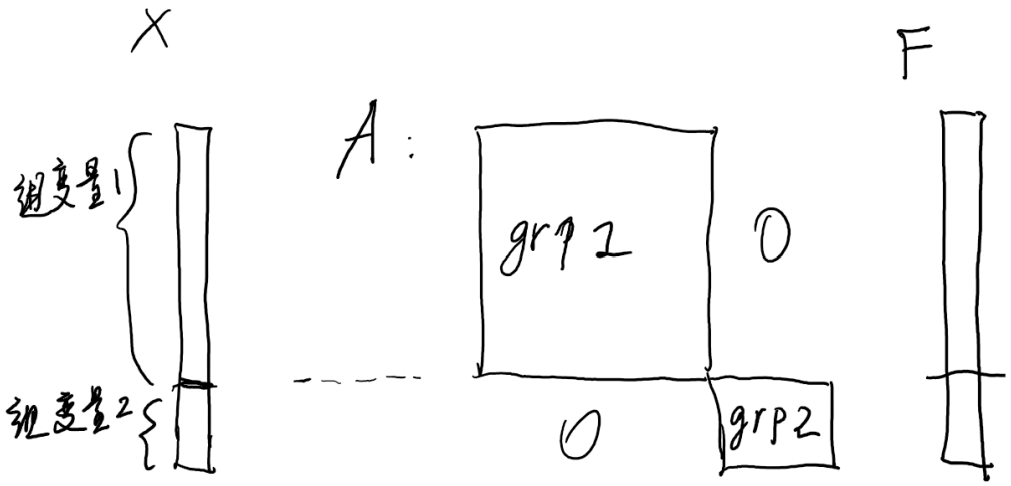
其中需假设样本 $(X_i)_{i=1}^n$ 是同质的（有着共同的 A，即相同的模型）. 试想，从中国和美国的学校，或者某地的普高和职高，中抽取学生，依赖关系 A 应该是有组结构的差异，其中中国/美国，和普高/职高就是分组. 从经验形式的因子模型，可以更清楚地看到，因子模型将一个 p 维的 X 与 k 维的公因子联系起来，为了分析 X 只需要研究 A 与 F 即可.

可以说，因子分析将多个原始变量综合为少数几个因子，用这少数几个因子去近似/再现原始变量间的相关关系. 那么研究原始变量就转变为研究因子和因子载荷，我们也可以通过因子去近似或者预测原始变量[1]. 因子分析可以分成两类：R 型和 Q 型[2]，分别研究变量和样本，相应的：

1. 降维：对 p 降维，对 n 降维[3]；

2. 分类：对 p 个变量进行分类，对 n 个样品进行分类[4].

Figure 2.1: 利用因子载荷阵为变量分类



2.3 Orthogonal Factor Model

正交因子模型是一个最常见的因子模型. 我们先讨论总体形式的正交因子模型. 设 $X \sim (\mu, \Sigma)$, 并要求如下的模型假设：

1. $\mathrm{cov}(F) = I_k$[5] ,

2. $\mathbf{E}(\epsilon) = 0$,$\mathrm{cov}(\epsilon) = \mathrm{diag}(\cdots) =: \Lambda$,

3. ϵ 与 F 不相关

$A = (a_1, \ldots, a_p)^\intercal$ 是 $p \times k$ 的因子载荷矩阵，那么

$$X - \mu = AF + \varepsilon, \tag{2.5}$$

其中，A 是未知的的，故估计模型时重要的一步就是估计 A. 上述 3 条模型假设和因子模型的思想息息相关：可以直接地看到如何用因子结构来近似原始数据的相关性. 因子模型用 $k + p$ 个变量, $F_1, \ldots, F_k, \epsilon_1, \ldots, \epsilon_p$，来线性逼近原始变量. 这也是和线性回归模型的一个差异，因为线性模型是用可观测的 k 个相关的解释变量和 p 个

白噪声来近似响应变量. 回到因子模型，“近似”可以理解为用因子载荷与对角阵去近似原始变量的协方差阵：

$$\Sigma = AA^\intercal + \Lambda, \tag{2.6}$$

从 Equation 2.6 我们可以看到，这与 PCA 的秩 k 逼近是不一样的，但只相差 Λ，两者也是很类似的.

下面我们还要来谈 3 个问题：A 的统计含义，原始变量的共同度（同时会得到因子模型的拟合优度），还有 k 个因子的解释能力的强弱. 这两个问题对于我们后面讨论经验形式下的因子模型时，模型估计和模型应用是有直接的帮助的.

A 有良好的统计含义：它表示 X 与 F 的协方差的矩阵，因为，在 3 条模型假设下

$$\begin{aligned}
\operatorname{cov}(X, F) &= \mathbf{E}\left(\right) X - \mu)F^\intercal \\
&= \mathbf{E}\left(\right) AF + \epsilon)F^\intercal \\
&= A + \mathbf{E}\left(\epsilon\right) F^\intercal \\
&= A\,.
\end{aligned}$$

若先对 X 进行标准化，则表示相关系数的矩阵. 这一良好的统计含义也是可以很容易地理解的：Equation 2.5表明每个原始变量 X_j 是互不相关的变量, $F_1, \ldots, F_k, \epsilon_j$, 的线性组合. 这个线性组合也直接指向了第 2 个问题：什么是 p 个原始变量的公共部分. 从方差的角度考虑，根据线性组合，可以写出原始变量的方差的分解：

$$\operatorname{Var}\left(\right) X_j) = \operatorname{Var}\left(\right) \sum_{\ell=1}^{k} a_{j\ell} F_\ell) + \operatorname{Var}\left(\right) \epsilon_j) \tag{2.7}$$

$$= \sum_{\ell=1}^{k} a_{j\ell}^2 + \sigma_j^2 \tag{2.8}$$

右式的第一个部分是公因子贡献的方差，体现的是 p 个原始变量之间的相关关系，因此称 $h_j = \sum_{\ell=1}^{k} a_{j\ell}^2$ 为变量 X_j 的共同度 (communality). 相应的，σ_j^2 体现的是没有被解释的方差，称为剩余方差. 剩余方差所占的比例越小，则因子模型的拟合优度越高.

最后我们来讨论 k 个因子中哪个因子最能解释 p 个原始变量. 我们引入因子的贡献度

$$\text{Contribution}(F_\ell) = \sum_{j=1}^{p} a_{j\ell}^2,$$

显然这是依赖于 A 的. 一般在讨论因子的贡献度时，需要先对原始变量进行标准化.

最后我们还要指出正交因子模型的一个特点：原始变量的量纲发生变化，但是因子矩阵 F 可以不. Consider that

$$\mathrm{cov}(X) = \mathrm{cov}(AF) + \mathrm{cov}(\epsilon), \tag{2.9}$$

让 C 为表示量纲的对角阵, then

$$\begin{aligned} \mathrm{cov}(CX) &= C\Sigma C && (2.10)\\ &= C(AA^\intercal + \Lambda)C^\intercal && (2.11)\\ &= CAA^\intercal C^\intercal + C\Lambda C^\intercal && (2.12) \end{aligned}$$

we arrive the new factor model:

$$CX = CAF + C\epsilon\,. \tag{2.13}$$

从 Equation 2.13 我们可以看出，新的因子模型中因子矩阵依然是 F, 而因子载荷矩阵和特殊因子的方差对角阵关于量纲做了线性变换.

估计因子模型的一般思路：先估计 $\widehat{\Sigma}$, 然后去算 A 和 D

2.4 Geometric representation

利用 $(F_1, \ldots, F_k, \epsilon_1, \ldots, \epsilon_p)$ 张成的空间去近似 X. 分量 X_j 的坐标是 $(a_{j1}, \ldots, a_{jk}, 0, \ldots, \epsilon_j, \ldots, 0)$. X_j 与 F_ℓ 的夹角余弦

3 High-dimensional Factor Models

3.1 Factor Models of Large Dimensions in Panel Data

Bai (2003) 中首次给出了面板数据高维因子模型 PCA 估计的渐近分布，其中相合性的证明首次出现在Bai and Ng (2002) 的 theorem1. 我接下来主要来讨论Bai (2003) 这篇文章中用到的证明方法，主要是差项的思路.

thm 1 (1) 的证明步骤：

1. decomposition identity
2. lemma A.1
3. ...

并且，lemma A.1 的证明也会用到 decomposition identity，因此，我们先给出这个等式的计算. 在计算这个等式之前，我们还要先去算 $\widetilde{F}$ 与 $F^\star$ 之间相差的可逆变换（可逆矩阵）. 之所以相差这一个矩阵是因为因子模型的不唯一性问题：即因子乘上一个可逆阵而载荷乘上一个该矩阵的逆，则因子模型不变. 因此在证明相合性时要考虑到 $\widetilde{F}$ 与 $F^\star$ 之间相差的可逆变换. 直观上可以想到，该可逆变换依赖于 $\widetilde{F}$ 和 $F^\star$.

3.2 Rotation Matrix of Factors

Bai (2003) 采用 PCA 因子模型估计, 出发点是对 $T \times T$ 的样本协方差阵做谱分解：

$$\frac{1}{NT} XX^\intercal = \sum_{i=1}^{T} \lambda_i \eta_i \eta_i', \tag{3.1}$$

其中，$\frac{1}{NT}$ 这个常数是因为面板数据共包含 NT 条观测数据，有截面和时间两个维度.$\frac{1}{NT}$ 不影响估计 eigen-vec η_i, 而只影响 eigenvalue λ_i 的量纲. 之后Bai (2003) 将证明 $V_{NT} = \text{diag}([)\,\lambda_1, \lambda_2, \ldots, \lambda_r]$ 的相合性. 而因子的估计 $\widetilde{F}$ 就来自于上述谱分解的前 r 个特征向量（乘上 $\sqrt{T}$）:

$$\widetilde{F} = \sqrt{T}(\eta_1, \ldots, \eta_r), \tag{3.2}$$

利用 $\widetilde{F}$，对原始数据的样本协方差阵 $XX^\intercal$ 有一个秩 r 逼近（因子的协方差结构估计原始数据的协方差结构）

$$\frac{1}{NT} XX^\intercal \leftarrow \left(\frac{\widetilde{F}}{\sqrt{T}}\right) V_{NT} \left(\frac{\widetilde{F}^\intercal}{\sqrt{T}}\right). \tag{3.3}$$

以上还是一个约等于的关系，为了得到等式，考虑将左边**投影**[6]到正交子空间 $\widetilde{F}$ 上，那么进一步，通过两边右乘一个 $\widetilde{F}$，利用 $\widetilde{F}^\intercal\widetilde{F}/T = I_r$ 得

$$\frac{1}{NT}XX^\intercal\widetilde{F} = \widetilde{F}V_{NT} \tag{3.4}$$

然后，用 true DGP $X = F\Lambda^{\star\intercal} + \epsilon$ 替换，得到

$$\frac{1}{NT}\Big(F\Lambda^{\star\intercal} + \epsilon\Big)\Big(F\Lambda^{\star\intercal} + \epsilon\Big)^\intercal\widetilde{F} = \widetilde{F}V_{NT} \tag{3.5}$$

that is

$$\frac{1}{NT}F\Lambda^{\star\intercal}\Lambda^\star F^\intercal + I_1 + I_2 + I3 = \widetilde{F}V_{NT} \tag{3.6}$$

由于统计建模中会假设真实的因子模型中的误差项 e 对估计的影响很小，that is $I_1 + I_2 + I_3 \to \mathrm{o}_p(1)$. Thus

$$\frac{1}{NT}F^\star\Lambda^{\star\intercal}\Lambda^\star F^{\star\intercal}\widetilde{F}V_{NT}^{-1} \leftarrow \widetilde{F} \tag{3.7}$$

从而

$$\begin{aligned} H &= \frac{1}{NT}\Lambda^{\star\intercal}\Lambda^\star F^{\star\intercal}\widetilde{F}V_{NT}^{-1} \\ &= (\Lambda^{\star\intercal}\Lambda^\star/N)(F^{\star\intercal}\widetilde{F}/T)V_{NT}^{-1} \end{aligned}$$

然后我们需要从 $\widetilde{F} - F^\star H$ 得到 t 时刻的恒等式. 我们希望能从 $\widetilde{F} - F^\star H$ 过渡到 $\widetilde{F}_t - H'F_t^\star$，因此目标是把矩阵整理成方阵乘上列向量阵的形式. 首先我们有

$$\widetilde{F} = \frac{1}{NT}XX^\intercal\widetilde{F}V_{NT}^{-1} \tag{3.8}$$

故

$$\begin{aligned} \widetilde{F} - F^\star H &= \frac{1}{NT}XX^\intercal\widetilde{F}V_{NT}^{-1} - \frac{1}{NT}F^\star\Lambda^{\star\intercal}\Lambda^\star F^{\star\intercal}\widetilde{F}V_{NT}^{-1} \\ &= \frac{1}{NT}(XX^\intercal - F^\star\Lambda^{\star\intercal}\Lambda^\star F^{\star\intercal})\widetilde{F}V_{NT}^{-1} \end{aligned}$$

注意，中间括号中的是因子协方差结构和样本协方差的差距，这个差距应该是比较小的（之后也证明了这三项是随机无穷小的）. 因为

$$X = F^\star\Lambda^{\star\intercal} + e\,, \tag{3.9}$$

那么

$$\begin{aligned} XX^\intercal &= (F^\star\Lambda^{\star\intercal} + e)(e' + \Lambda^\star F^{\star\intercal}) \\ &= F^\star\Lambda^{\star\intercal}\Lambda^\star F^{\star\intercal} + F^\star\Lambda^{\star\intercal}e' + e\Lambda^\star F^{\star\intercal} + ee' \end{aligned}$$

继续写，得到

$$= \frac{1}{NT}(F^\star\Lambda^{\star\intercal}e' + e\Lambda^\star F^{\star\intercal} + ee')\widetilde{F}V_{NT}^{-1} \tag{3.10}$$

注意，上面括号中的是方阵，但是前两项不一定是对称矩阵，所以不能合并. 写到这里我们还没有整理出一个列向量的形式[9]，还要继续写

$$= \frac{1}{NT}(F^\star\Lambda^{\star\intercal}(e'\widetilde{F}) + e\Lambda^\star(F^{\star\intercal}\widetilde{F}) + e(e'\widetilde{F}))V_{NT}^{-1} \tag{3.11}$$

$$= \frac{1}{NT}(F^\star\Lambda^{\star\intercal}(\sum_{s=1}^{T} e_s\widetilde{F}_s') + e\Lambda^\star(\sum_{s=1}^{T} F_s^\star\widetilde{F}_s') + e(\sum_{s=1}^{T} e_s\widetilde{F}_s'))V_{NT}^{-1} \tag{3.12}$$

那么

$$\begin{aligned}\widetilde{F}' - H'F^{\star\intercal} &= (\widetilde{F}_1 - H'F_1^0, \ldots, \widetilde{F}_T - H'F_T^\star) \\ &= \frac{1}{NT}V_{NT}^{-1}(\sum_{s=1}^{T}\widetilde{F}_s e_s'\Lambda^\star F^{\star\intercal} + \sum_{s=1}^{T}\widetilde{F}_s F_s^{\star\prime}\Lambda^{\star\intercal}e' + \sum_{s=1}^{T}\widetilde{F}_s e_s'e'),\end{aligned}$$

注意到 $F^{\star\intercal} = (F_1^0, \ldots, F_T^\star)$ 那么

$$\widetilde{F}_t - H'F_t^\star = \frac{1}{NT}V_{NT}^{-1}(\sum_{s=1}^{T}\widetilde{F}_s(e_s'\Lambda^\star F_t^\star) + \sum_{s=1}^{T}\widetilde{F}_s F_s^{\star\prime}\Lambda^{\star\intercal}e_t + \sum_{s=1}^{T}\widetilde{F}_s e_s'e_t) \tag{3.13}$$

$$= \frac{1}{NT}V_{NT}^{-1}(\sum_{s=1}^{T}\widetilde{F}_s F_t^{\star\prime}\Lambda^{\star\intercal}e_s + \sum_{s=1}^{T}\widetilde{F}_s F_s^{\star\prime}\Lambda^{\star\intercal}e_t + \sum_{s=1}^{T}\widetilde{F}_s e_s'e_t) \tag{3.14}$$

$$= V_{NT}^{-1}(\mathrm{I}_4 + \mathrm{I}_3 + (\mathrm{I}_1 + \mathrm{I}_2)) \tag{3.15}$$

其中，可以单独的证明 $V_{NT} = \mathcal{O}_p(1)$. 另外，还有 I_4 中要把 $\widetilde{F}_s F_t^{\star\prime}$ 放在一起，这在后面的 prop 证明了依概率收敛；$\mathrm{I}_1 + \mathrm{I}_2$ 需要扣掉真值 $\gamma_N(s,t) = \frac{1}{N}e_s'e_t = \sum_{i=1}^{N} e_{is}e_{it}$.

Lemma. *(A.1, consistency of factors) Under assumptions A-D,*

$$\delta_{NT}^2\left(\frac{1}{T}\sum_{t=1}^{T}\left\|\widetilde{F}_t - H'F_t^\star\right\|^2\right) = \mathcal{O}_p(1)$$

然后我们来证明 lemma A.1. 我们要证

$$\frac{1}{T}\sum_{t=1}^{T}\left\|\widetilde{F}_t - H'F_t^\star\right\|^2 = \mathcal{O}_p(1)(\frac{1}{\delta_{NT}^2}) = \mathcal{O}_p(1)(\frac{1}{N}) + \mathcal{O}_p(1)(\frac{1}{T}) \tag{3.16}$$

[9]这一点我后面会补上

等价于去证

$$\sum_{t=1}^{T}\left\|\widetilde{F}_t - H'F_t^{\star}\right\|^2 = \mathcal{O}_p(1)(\frac{T}{N}) + \mathcal{O}_p(1)(1)$$

根据等式 (3.15)，$\widetilde{F}_t - H'F_t^{\star}$ 可以分解成 4 项，视每一项为内积中的一项，再运用 Cauchy-Scharwtz 不等式[10]

$$\begin{aligned}\left\|\widetilde{F}_t - H'F_t^{\star}\right\|^2 &= \left\|\sum_{k=1}^{4} 1 \times \mathrm{I}_k\right\|^2 \leq \|\mathbf{1}_4\|^2 \cdot \|(\mathrm{I}_1,\mathrm{I}_2,\mathrm{I}_3,\mathrm{I}_4)\|_F^2 \\ &= 4\cdot(\|\mathrm{I}_1\|^2 + \|\mathrm{I}_2\|^2 + \|\mathrm{I}_3\|^2 + \|\mathrm{I}_4\|^2) \\ &=: 4(a_t + b_t + c_t + d_t)\end{aligned}$$

其中

$$\begin{aligned}a_t &= \left\|\frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s\gamma_N(s,t)\right\|^2, \quad b_t = \left\|\frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s\zeta_{st}\right\|^2 \\ c_t &= \left\|\frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s\eta_{st}\right\|^2, \quad d_t = \left\|\frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s\xi_{st}\right\|^2\end{aligned}$$

用 Cauchy-Schwartz 不等式来证明 Bai and Ng (2002) theorem 1. 现在先来证：

$$\left\|\sum_{s=1}^{T}\widetilde{F}_s\gamma_N(s,t)\right\|^2 \leq \left(\sum_{s=1}^{T}\left\|\widetilde{F}_s\right\|^2\right)\left(\sum_{s=1}^{T}\gamma_N^2(s,t)\right) \tag{3.17}$$

将 $\sum_{s=1}^{T}\widetilde{F}_s\gamma_N(s,t)$ 看成内积，那么就是

$$\widetilde{F}' = (\widetilde{F}_1,\ldots,\widetilde{F}_T)$$

and

$$\gamma = (\gamma_N(1,t),\ldots,\gamma_N(T,t))$$

即矩阵和向量之间的内积产生一个向量，得到

$$\left\|\left\langle\widetilde{F}',\gamma\right\rangle\right\|_F \leq \left\|\widetilde{F}\right\|_F\|\gamma\|_F$$

[10]关于 Cauchy-Schwartz inequality 参见14.1

左边内积的结果 $\left\langle \widetilde{F}', \gamma \right\rangle$ 是一个向量，ℓ_2-norm 与 F norm 是一样的，因此得到

$$\|\langle \cdot, \cdot \rangle\| \leq \|\cdot\| \|\cdot\|$$

即

$$\left\| \sum_{s=1}^{T} \widetilde{F}_s \gamma_N(s,t) \right\|^2 \leq \left(\sum_{s=1}^{T} \left\| \widetilde{F}_s \right\|^2 \right) \left(\sum_{s=1}^{T} \gamma_N^2(s,t) \right)$$

其中，$\sum_{s=1}^{T} \gamma_N^2(s,t)$ 反映了时间维度上的异质性和序列相关性. 我们要去证明 $a_t = \mathcal{O}_p(1)(\frac{1}{N})\text{or}\mathcal{O}_p(1)(\frac{1}{T})$，也等价于要去证明

$$\begin{aligned} \sum_{t=1}^{T} a_t &= \mathcal{O}_p(1)(\frac{T}{N}) + \mathcal{O}_p(1) \\ &= \mathrm{o}_p(1) + \mathcal{O}_p(1) \end{aligned}$$

其他三项也是这样. 我们来证 a_t,

$$\begin{aligned} a_t = \left\| \sum_{s=1}^{T} (\frac{\widetilde{F}_s}{T}) \gamma_N(s,t) \right\|^2 &\leq \left(\sum_{s=1}^{T} \frac{1}{T^2} \left\| \widetilde{F}_s \right\|^2 \right) \left(\sum_{s=1}^{T} \gamma_N^2(s,t) \right) \\ &= (\frac{1}{T} \sum_{s=1}^{T} \left\| \widetilde{F}_s \right\|^2)(\frac{1}{T} \sum_{s=1}^{T} \gamma_N^2(s,t)) \end{aligned}$$

同时，注意有 $\frac{1}{T} \sum_{s=1}^{T} \left\| \widetilde{F}_s \right\|^2 = \mathcal{O}_p(1)$，因为

$$\begin{aligned} \widetilde{F}' &= (\widetilde{F}_1, \ldots, \widetilde{F}_T) \\ \frac{1}{T} \sum_{s=1}^{T} \left\| \widetilde{F}_s \right\|^2 &= \frac{1}{T} \left\| \widetilde{F} \right\|_F^2 = \frac{1}{T} \operatorname{tr}(()\, \widetilde{F}' \widetilde{F}) \\ &= \frac{1}{T} \operatorname{tr}(()\, T \cdot T_r) \\ &= r = \mathcal{O}_p(1) \end{aligned}$$

这里我有一个问题：我们的限制性条件 $\widetilde{F}'\widetilde{F}/T = I_r$ 使得这里面都出现了 $\mathcal{O}(1)$ 而非 $\mathcal{O}_p(1)$ ，这样合适嘛？

$$\begin{aligned} \therefore \sum a_t &= \sum_{t=1}^{T} (\frac{1}{T} \mathcal{O}_p(1))(\frac{1}{T} \sum_{s=1}^{T} \gamma_N^2(s,t)) \\ &= \mathcal{O}_p(1) \cdot \sum_{t=1}^{T} \frac{1}{T} \sum_{s=1}^{T} \gamma_N^2(s,t) \end{aligned}$$

之前的引理可以证明 $\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\gamma_N^2(s,t)=\mathcal{O}(1)$, 证明如下：先利用如下 3 个条件

$$\gamma_N(s,t)=(\gamma_N(s,s)\cdot\gamma_N(t,t))^{1/2}\cdot\rho(s,t) \tag{3.18}$$

以及时间维度上地异质性一致有界

$$|\gamma_N(s,s)|\leq M,\quad |\rho(s,t)|\leq 1 \tag{3.19}$$

时间序列相关性的一致有界

$$\frac{1}{T}\sum\sum|\gamma_N(s,t)|\leq M \tag{3.20}$$

我们可以得到

$$\begin{aligned}\frac{1}{T}\sum\sum\gamma_N^2(s,t)&=\frac{1}{T}\sum\sum\gamma_N(s,s)\gamma_N(t,t)\rho^2(s,t) &(3.21)\\ &\leq\frac{M}{T}\sum\sum(\gamma_N(s,s)\gamma_N(t,t))^{1/2}|\rho(s,t)| &(3.22)\\ &=\frac{M}{T}\sum\sum|\gamma(s,t)|\leq M^2 &(3.23)\end{aligned}$$

then we arrive that

$$\frac{1}{T}\sum_{t=1}^{T}a_t=\frac{1}{T}\mathcal{O}_p(1)\mathcal{O}(1)=\mathcal{O}_p(1)(\frac{1}{T})$$

我们再来证明:

$$\frac{1}{T}\sum_{t=1}^{T}b_t=\frac{1}{T}\sum_{t=1}^{T}\left\|\frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s\zeta_{st}\right\|^2=\mathrm{o}_p(1)$$

与之前证明相类似

$$\begin{aligned}\sum_{t=1}^{T}\left\|\frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s\zeta_{st}\right\|^2&=\frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T}\sum_{u=1}^{T}\widetilde{F}_s'\widetilde{F}_u\zeta_{st}\zeta_{ut}\\ &=\frac{1}{T^2}\sum_{s,u}\left(\left(\widetilde{F}_s'\widetilde{F}_u\right)\sum_t\zeta_{st}\zeta_{ut}\right)\\ &\leq\frac{1}{T^2}\left(\sum_{s,u}\left(\widetilde{F}_s'\widetilde{F}_u\right)^2\right)^{1/2}\left(\sum_{s,u}\left(\sum_t\zeta_{st}\zeta_{ut}\right)^2\right)^{1/2}\end{aligned}$$

注意上面的 $\widetilde{F}_s'\widetilde{F}_u$ 和 $\zeta_{st}\zeta_{ut}$ 都是数字. 然后，利用 F-norm 的次可乘性得到

$$\leq\frac{1}{T^2}\left(\sum\left\|\widetilde{F}_k\right\|^2\right)\left(\sum_s\sum_u\left(\sum_{t=1}^{T}\zeta_{st}\zeta_{ut}\right)^2\right)^{1/2}$$

对 $\left(\sum_{t=1}^{T}\zeta_{st}\zeta_{ut}\right)^2$ 反复利用 Cauchy-Schwartz 不等式，可以得到

$$\left(\sum_{t=1}^{T}\zeta_{st}\zeta_{ut}\right)^2 \leq T^2 \max_s \mathbf{E}\left(\zeta\right)_{st}^4$$

同时，利用 Markov 不等式和 Assumption，得到

$$\mathbf{E}\left(|\zeta_{st}|^4\right) = \frac{1}{N^2}\mathbf{E}\left(\left|\frac{1}{N}\sum_{i=1}^{N}(e_{it}e_{is}-\gamma_N(s,t))\right|^4\right) \leq \frac{M}{N^2}$$

then

从而根据 Markov 不等式得到

$$\left(\sum_s\sum_u\left(\sum_{t=1}^{T}\zeta_{st}\zeta_{ut}\right)^2\right)^{1/2} = \left(\mathcal{O}_p(1)\left(\frac{T^4}{N^2}\right)\right)^{1/2} = \mathcal{O}_p(1)\left(\frac{T^2}{N}\right) \tag{3.24}$$

then

$$\sum_{t=1}^{T} b_t \leq \mathcal{O}_p(1)\mathcal{O}_p(1)\left(\frac{T}{N}\right)$$

类似地我们也可以证明其他 2 项，完毕.

在得到相合性后，为了证明渐近正态性，我们必须去更加精细地确定这四项各自的收敛速度，因此我们要出证明如下引理

Lemma 1. *(A.2:) Under Assumptions A-F, we have*

(a) $T^{-1}\sum_{s=1}^{T}\widetilde{F}_s\gamma_N(s,t) = \mathcal{O}_p(1)(\frac{1}{\sqrt{T}\delta_{NT}})$

(b) $T^{-1}\sum_{s=1}^{T}\widetilde{F}_s\zeta_{st} = \mathcal{O}_p(1)(\frac{1}{\sqrt{N}\delta_{NT}})$

(c) $T^{-1}\sum_{s=1}^{T}\widetilde{F}_s\eta_{st} = \mathcal{O}_p(1)(\frac{1}{\sqrt{N}})$

(d) $T^{-1}\sum_{s=1}^{T}\widetilde{F}_s\xi_{st} = \mathcal{O}_p(1)(\frac{1}{\sqrt{N}\delta_{NT}})$

We want to prove

$$\frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s\gamma_N(s,t) = \mathcal{O}_p(1)\left(\frac{1}{\sqrt{N}\delta_{NT}}\right) = \mathcal{O}_p(1)\left(\frac{1}{T}\right) + \mathcal{O}_p(1)\left(\frac{1}{\sqrt{NT}}\right)$$

一个基本的套路是先扣掉真值，拆成两项：

$$\begin{aligned}\frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s\gamma_N(s,t) &= \frac{1}{T}\sum_{s=1}^{T}(\widetilde{F}_s - H'F_s^0 + H'F_s^0)\gamma_N(s,t)\\ &= \frac{1}{T}\sum_{s=1}^{T}(\widetilde{F}_s - H'F_s^0)\gamma_N(s,t) + H'\frac{1}{T}\sum_{s=1}^{T}F_s^\star)\gamma_N(s,t)\end{aligned}$$

then, for the second term, we want to use Markov's inequality. Thus, we have to show that the expectaton of the absolute second is bounded.

$$\begin{aligned}\left\|\sum_{s=1}^{T}F_s^\star)\gamma_N(s,t)\right\| &\leq \sum_{s=1}^{T}\left\|F_s^\star\gamma_N(s,t)\right\|\\ &= \sum_{s=1}^{T}|\gamma_N(s,t)|\cdot\left\|F_s^\star\right\|\end{aligned}$$

其中，先是利用了 norm 的次可加性，再利用了齐次性. 然后，希望利用 Markov 不等式证明依概率有界. Then

$$\mathbf{E}\left(\left\|\sum_{s=1}^{T}F_s^\star\gamma_N(s,t)\right\|\right) \leq \left(\sum_{s=1}^{T}|\gamma_N(s,t)|\right)\max_s\mathbf{E}\left(\left\|F_s^\star\right\|\right) \leq M$$

For the first term,

$$\begin{aligned}\left\|\frac{1}{T}\sum_{s=1}^{T}(\widetilde{F}_s - H'F_s^0)\gamma_N(s,t)\right\| &\leq \frac{1}{\sqrt{T}}\left(\frac{1}{T}\sum_{s=1}^{T}\left\|\widetilde{F}_s - H'F_s^0\right\|^2\right)^{1/2}\left(\sum_{s=1}^{T}\gamma_N^2(s,t)\right)^{1/2}\\ &= \mathcal{O}_p(1)\left(\frac{1}{\sqrt{T}}\frac{1}{\delta_{NT}}\right)\end{aligned}$$

since

$$\sum_{s=1}^{T}|\gamma_N(s,t)| \leq M,$$
$$|\gamma_N(s,s)| \leq M$$

然后我们来证明Prop. 3.1，这是Bai (2003) 中的 prop1

Proposition 3.1. *(Consistency) Under Asuumptions A-D and G,*

$$\frac{\widetilde{F}'F^\star}{T} \xrightarrow{P} Q$$

The matrix Q is invertible and is given by $Q = V^{1/2}\Upsilon'\Sigma_\Lambda^{1/2}$ where $V = \text{diag}\,(()\, v_1, v_2, \ldots, v_r)$ are the eigenvalues of $\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}$, and Υ is the corresponding eigenvector matrix such that $\Upsilon'\Upsilon = I_r$.

因为

$$\frac{1}{NT}XX^{\intercal} \approx \frac{\widetilde{F}}{\sqrt{T}}V_{NT}\frac{\widetilde{F}'}{\sqrt{T}} \tag{3.25}$$

故有

$$\frac{1}{NT}XX^{\intercal}\widetilde{F} = \widetilde{F}V_{NT} \tag{3.26}$$

两边同时左乘 $\frac{1}{T}(\Lambda^{\star\intercal}\Lambda^{\star}/N)^{1/2}F^{\star\intercal}$ 得

$$\left(\frac{\Lambda^{\star\intercal}\Lambda^{\star}}{N}\right)^{1/2}\left(\frac{1}{T}F^{\star\intercal}\right)\left(\frac{XX^{\intercal}}{NT}\right)\widetilde{F} = \left(\frac{\Lambda^{\star\intercal}\Lambda^{\star}}{N}\right)^{1/2}\left(\frac{F^{\star\intercal}\widetilde{F}}{T}\right)V_{NT} \tag{3.27}$$

利用

$$XX^{\intercal} = F^{\star}\Lambda^{\star\intercal}\Lambda^{\star}F^{\star\intercal} + (\mathrm{I}+\mathrm{II}+\mathrm{III}) \tag{3.28}$$

and

$$d_{NT} \coloneqq \frac{1}{T}F^{\star\intercal}(\mathrm{I}+\mathrm{II}+\mathrm{III})\widetilde{F} = \mathrm{o}_p(1) \tag{3.29}$$

and

$$\left(\frac{\Lambda^{\star\intercal}\Lambda^{\star}}{N}\right)^{1/2} = \mathcal{O}_p(1) \tag{3.30}$$

continue

$$\left(\frac{\Lambda^{\star\intercal}\Lambda^{\star}}{N}\right)^{1/2}\left(\frac{F^{\star\intercal}F^{\star}}{T}\right)\left(\frac{\Lambda^{\star\intercal}\Lambda^{\star}}{N}\right)\left(\frac{F^{\star\intercal}\widetilde{F}}{T}\right) + d_{NT} = \left(\frac{\Lambda^{\star\intercal}\Lambda^{\star}}{N}\right)^{1/2}\left(\frac{F^{\star\intercal}\widetilde{F}}{T}\right)V_{NT}$$

and let

$$B_{NT} \coloneqq \left(\frac{\Lambda^{\star\intercal}\Lambda^{\star}}{N}\right)^{1/2}\left(\frac{F^{\star\intercal}F^{\star}}{T}\right)\left(\frac{\Lambda^{\star\intercal}\Lambda^{\star}}{N}\right)^{1/2}$$

and

$$R_{NT} \coloneqq \left(\frac{\Lambda^{\star\intercal}\Lambda^{\star}}{N}\right)^{1/2}\left(\frac{F^{\star\intercal}\widetilde{F}}{T}\right)$$

then

$$B_{NT}R_{NT} + d_{NT} = R_{NT}V_{NT}(\mathrm{or}V_{NT}R_{NT}) \tag{3.31}$$

and we also have

$$B_{NT} \xrightarrow{P} \Sigma_\Lambda^{1/2} \Sigma_F \Sigma_\Lambda^{1/2} \tag{3.32}$$

因为已经假设 $r \times r$ 矩阵 $\Sigma_\Lambda \Sigma_F$ 有 r 个不同的特征根，这意味着 B_{NT} 也会有 r 个不同的特征根.

同时还有一个纯粹技术性的处理：由于 R_{NT} 的每一列还没有单位化，因此还不能说是 B_{NT} 的特征向量，那么我们还要做一下单位化：让

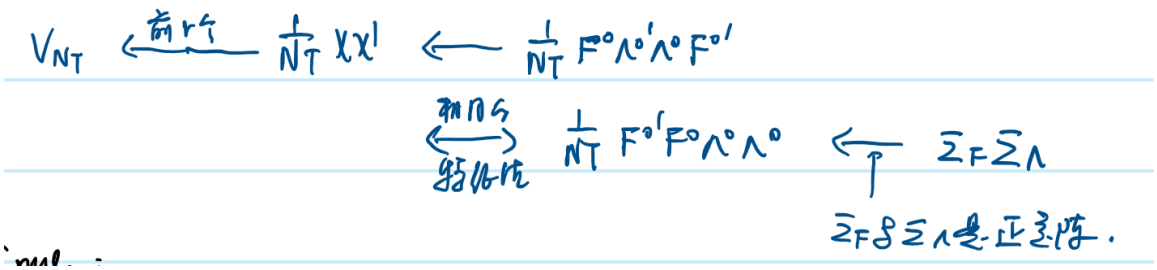
$$V_{NT}^* = \operatorname{diag}(() R_{NT}' R_{NT})$$
$$\Upsilon_{NT} = R_{NT} {V_{NT}^*}^{-1/2}$$

那么

$$B_{NT}\Upsilon_{NT} + d_{NT} = \Upsilon_{NT} V_{NT}$$

即，对左右两边的 R_{NT} 进行单位化，同时，我们还注意到 V_{NT} 是 B_{NT} 的特征根，这是直观的，可以参见Figure 3.1

Figure 3.1: Relationship among eigenvalue and sample matrices



Continue：由于 $\Sigma_\Lambda \Sigma_F$ 有 r 个不相等[7]的（正的）特征根，故由特征值的连续性，B_{NT} 的特征值 V_{NT}（严格来说还不是 V_{NT} 因为还相差 d_{NT}）也有 r 个不相等的对角元，从而 Υ_{NT} 的每一列在相差正负号下是唯一确定的.

由于 R_{NT} 的每个列向量都由 $\widetilde{F}$ 的对应列向量唯一确定，故 R_{NT}, Υ_{NT} 的每一列的符号也是由 $\widetilde{F}$ 唯一确定的，这样就得到：确定了 $\widetilde{F}$，则 Υ_{NT} 唯一.

由特征向量的扰动理论，存在 $\Sigma_\Lambda^{1/2} \Sigma_F \Sigma_\Lambda^{1/2}$ 的唯一的特征向量矩阵 Υ 使得

$$R_{NT} = \left(\frac{\Lambda^{\star\mathsf{T}} \Lambda^\star}{N}\right)^{1/2} \left(\frac{F^{\star\mathsf{T}} \widetilde{F}}{T}\right)$$
$$\Upsilon_{NT} {V_{NT}^*}^{1/2} = \left(\frac{\Lambda^{\star\mathsf{T}} \Lambda^\star}{N}\right)^{1/2} \left(\frac{F^{\star\mathsf{T}} \widetilde{F}}{T}\right)$$
$$\Upsilon_{NT} \xrightarrow{P} \Upsilon$$
$${V_{NT}^*}^{1/2} \xrightarrow{P} V^{1/2}$$

应该说 V_{NT}^{*} 是十分接近 V_{NT} 的，因为

$$\begin{aligned} R_{NT}' R_{NT} &= \frac{\widetilde{F}' F^{\star}}{T} \left(\frac{\Lambda^{\star\intercal} \Lambda^{\star}}{N} \right) \frac{F^{\star\intercal} \widetilde{F}}{T} \\ &\approx \frac{1}{T} \left(\frac{X X^{\intercal}}{NT} \right) \leftrightarrow V_{NT} \end{aligned}$$

continue

$$\frac{F^{\star\intercal} \widetilde{F}}{T} \xrightarrow{P} \Sigma_{\Lambda}^{-1/2} \Upsilon V^{1/2} \tag{3.33}$$

3.2.1 Thm 1

首先将 $\widetilde{F}_t - H' F_t^{\star}$ 根据等式拆成 4 项，应用 Lemma A.2 得到

$$\begin{aligned} \widetilde{F}_t - H' F_t^{\star} &= \mathcal{O}_p(1) \left(\mathcal{O}_p(1)(\frac{1}{\sqrt{T}\delta_{NT}}) + \mathcal{O}_p(1)(\frac{1}{\sqrt{N}\delta_{NT}}) + \mathcal{O}_p(1)(\frac{1}{\sqrt{N}}) + \mathcal{O}_p(1)(\frac{1}{\sqrt{N}\delta_{NT}}) \right) \\ &= \mathcal{O}_p(1)(\frac{1}{\sqrt{T}\delta_{NT}}) + \mathcal{O}_p(1)(\frac{1}{\sqrt{N}\delta_{NT}}) + \mathcal{O}_p(1)(\frac{1}{\sqrt{N}}) + \mathcal{O}_p(1)(\frac{1}{\sqrt{N}\delta_{NT}}) \\ &=: \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3 + \mathrm{I}_4 \end{aligned}$$

本文更关心 $N \to \infty$ 的速度比 $T \to \infty$ 的速度更快的情形，因此可以先肯定的是 I_2 与 I_4 是 I_1 的高阶无穷小，而 I_1 与 I_3 需要分情况讨论. 因为

$$\begin{aligned} \delta_{NT} &= \min\{\sqrt{N}, \sqrt{T}\} \\ \frac{1}{T} &= \frac{1}{\sqrt{N}} \end{aligned}$$

1. 当 $T < N < T^2$ 时，I_1 是 I_3 的高阶无穷小，因此 I_3 是主项；

2. 当 $N \geq T^2$ 时，则 I_1 是主项

先讨论 $\frac{\sqrt{N}}{T} \to 0$ ，即 $N < T^2$，那么有

$$\begin{aligned} \widetilde{F}_t - H' F_t^{\star} &= V_{NT}^{-1} \cdot \frac{1}{T} \sum_{s=1}^{T} \widetilde{F}_s F_s^{\star\prime} \Lambda^{\star\intercal} e_t \cdot \frac{1}{N} \\ &= V_{NT}^{-1} \cdot \frac{1}{\sqrt{N}} \left(\frac{1}{T} \frac{1}{T} \sum_{s=1}^{T} \widetilde{F}_s F_s^{\star\prime} \right) \left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \lambda_i^0 e_{it} \right) \end{aligned}$$

and

$$V_{NT} \xrightarrow{P} V, V_{NT} = \mathcal{O}_p(1)$$
$$\frac{1}{T}\frac{1}{T}\sum_{s=1}^{T}\widetilde{F}_s F_s^{\star\prime} \xrightarrow{P} V^{1/2}\Upsilon'\Sigma_{\Lambda}^{-1/2}$$
$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\lambda_i^0 e_{it} \xrightarrow{d} N(0,\Gamma_t)$$

那么

$$\sqrt{N}\left(\widetilde{F}_t - H'F_t^{\star}\right) \xrightarrow{d} N(0, V^{-1/2}\Upsilon'\Sigma^{-1/2}\Gamma_t\Sigma^{-1/2}\Upsilon V^{-1/2})$$

如果 $N \to \infty$ 快一些，即 $\liminf \frac{\sqrt{N}}{T} \geq \tau > 0$，那么

$$\widetilde{F}_t - H'F_t^{\star} = \mathcal{O}_p(1)(\frac{1}{\sqrt{T}\sqrt{T}}) + \mathcal{O}_p(1)(\frac{T}{\sqrt{N}})$$
$$\because \limsup \frac{T}{\sqrt{N}} \leq \frac{1}{\tau}$$
$$\therefore \left(\widetilde{F}_t - H'F_t^{\star}\right)\cdot T = \mathcal{O}_p(1) + \mathcal{O}_p(1)(\frac{T}{\sqrt{N}}) = \mathcal{O}_p(1)$$

故收敛速度是 T

关于收敛速度的讨论:

- case 1 中 $\frac{\sqrt{N}}{T} \to 0$ 并且 $\min\{N,T\} \to \infty$ 包含了 $T < N < T^2$ 和 $N \leq T$ 的情形；由于高维因子模型关注 $N > T$，因此这时我们只去讨论 $T < N < T^2$，比如说 $N = T^{3/2}$ ，这时 $\delta_{NT} = \sqrt{T}$

- case 2 中 $\liminf \frac{\sqrt{N}}{T} \geq \tau > 0$ 时，意味着 $N \geq T^2$ 这时 $N \to \infty$ 的速度比 case 1 更快，比如 $N = T^2$

3.3 Linear Regression with Factor Structure Error

The model is

$$Y_i = X_i\beta + F\lambda_i + \epsilon_i, \quad i = 1,\ldots,N \tag{3.34}$$

subject to $F^{\intercal}F/T = I_r$($\frac{r(r+1)}{2}$ 个约束条件) and $\Lambda^{\intercal}\Lambda = \text{diag}([]\cdot]$($\frac{r(r-1)}{2}$ 个约束条件). $X_i = \left(X_{i1},\ldots,X_{iT}\right)^{\intercal}$ is a $T \times p$ matrix.

Note. 由于因子模型在相差一个 $r \times r$ 的可逆阵[11]之下是等价的，而 $r \times r$ 的可逆阵意味着 r^2 个自由元素，因此需要 r^2 个线性约束条件. 这方面可以参考Connor and Korajczyk (1986) and Stock and Watson (2002)[8].

我还有个问题：在 OLS 中，是否有如下的等价关系？

1. $X^\intercal X$ full rank
2. $P_X = X(X^\intercal X)^{-1}X^\intercal$ full rank
3. $M_X = I - P_X$ full rank

9

Then, we come to the estimation process.

3.3.1 Estimation

Firstly, estimate the regression coefficients, $\widehat{\beta}$. 首先，先将 β 看成已知的，求出 $\widehat{\Lambda}$:

$$Y_i - X_i\beta = F\lambda_i + \epsilon_i, \quad i = 1, \ldots, N \tag{3.35}$$

$$\widehat{\lambda_i} = \left(F^\intercal F\right)^{-1} F^\intercal (Y_i - X_i\beta) \tag{3.36}$$

其中，对于每个 i，以上是一个有 T 条样本, r 个协变量的时间序列回归；对于每个 t，让 $i = 1, \ldots, N$ 是截面回归.Then replace the λ_i above with the least-square estimator, we have

$$Y_i - X_i\beta = F\left(F^\intercal F\right)^{-1} F^\intercal (Y_i - X_i\beta) + \epsilon_i \tag{3.37}$$

$$= P_F(Y_i - X_i\beta) + \epsilon_i \tag{3.38}$$

since $M_F = I_T - P_F$, we can rewrite the above as

$$M_F Y_i = M_F X_i\beta + \epsilon_i, \quad i = 1, \ldots, N \tag{3.39}$$

这可以看成是在因子空间的补空间上的线性回归. Note that this is a vector response linear regression rather than univariate response case for each i, then we need to

[11]在 Jushan Bai 的文章中，经常记为 H

differentiate the sum square loss function with respect to β

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \|M_F Y_i - M_F X_i \beta\|^2 \tag{3.40}$$

$$= \sum_{i=1}^{N} \left(Y_i^\intercal M_F Y_i - 2Y_i^\intercal M_F X_i \beta + \beta^\intercal X_i^\intercal M_F X_i \beta\right) \tag{3.41}$$

$$\frac{\partial \cdot}{\partial \beta} = \sum_{i=1}^{N} \left(2X_i^\intercal M_F X_i \beta - 2X_i^\intercal M_F Y_i\right) = 0 \tag{3.42}$$

that is

$$\left(\sum_{i=1}^{N} X_i^\intercal M_F X_i\right)\beta = \sum_{i=1}^{N} X_i^\intercal M_F Y_i \tag{3.43}$$

we arrive that

$$\widehat{\beta} = \left(\sum_{i=1}^{N} X_i^\intercal M_F X_i\right)^{-1} \sum_{i=1}^{N} X_i^\intercal M_F Y_i \tag{3.44}$$

Note. 1. $\widehat{\beta}$ 来自 $N \times T$ 条数据，因此由截面和时间序列回归共同决定. 若 F 是已知的，那么 $\widehat{\beta}$ 存在当且仅当 $\sum_{i=1}^{N} X_i^\intercal M_F X_i$ 是满秩的. 但是 F 实际上是不可观测的，因此需要更强的条件；

2. 因为 true slope 是

$$\beta^\star = \left(\sum_{i=1}^{N} X_i^\intercal M_F X_i\right)^{-1} \sum_{i=1}^{N} X_i^\intercal M_F X_i \beta^\star \tag{3.45}$$

then we get

$$\widehat{\beta} - \beta^\star = \left(\sum_{i=1}^{N} X_i^\intercal M_F X_i\right)^{-1} \sum_{i=1}^{N} X_i^\intercal M_F \left(F^\star \lambda_i^0 + \epsilon_i\right) \tag{3.46}$$

主要讨论因子误差的加权和的相合性[10].

在得到回归部分的估计后，再来估计因子部分. Let $W_i = Y_i - X_i\beta$, then

$$W_i = F\lambda_i + \epsilon_i, \quad i = 1, \ldots, N \tag{3.47}$$

称此为 pure factor model. 我们用 PCA 去恢复因子模型，这也等价于最小化 reconstruction error:

$$\min_{F,\Lambda} \sum_{i=1}^{N} \|W_i - F\lambda_i\|^2 = \sum_{i=1}^{N} (W_i - F\lambda_i)^\intercal (W_i - F\lambda_i) \tag{3.48}$$

$$= \sum_{i=1}^{N} W_i^\intercal W_i - 2\sum_{i=1}^{N} W_i^\intercal F\lambda_i + \sum_{i=1}^{N} \lambda_i^\intercal F^\intercal F\lambda_i \tag{3.49}$$

since $\Lambda = (\lambda_1, \ldots, \lambda_N)^\intercal$ ($N \times r$ matrix) and $F\Lambda^\intercal = (F\lambda_1, \ldots, F\lambda_N)$ and $W = (W_1, \ldots, W_N)$, then

$$\min_{F,\Lambda} \sum_{i=1}^{N} \|W_i F \lambda_i\|^2 = \operatorname{tr}\left(W^\intercal W\right) - 2\operatorname{tr}\left(W^\intercal F \Lambda^\intercal\right) + \operatorname{tr}\left(\Lambda F^\intercal F \Lambda^\intercal\right) \tag{3.50}$$

尽管 F 和 Λ 都是未知的，但是可以先视 F 是已知的，从第 i 个时间序列回归中用 LSE 求出 λ_i 的估计：from $W_i = F\lambda_i + \epsilon_i$, we get

$$\widehat{\lambda_i} = \left(F^\intercal F\right)^{-1} F^\intercal W_i = F^\intercal W_i / T \tag{3.51}$$

then

$$\begin{aligned} \Lambda &= (\lambda_1, \ldots, \lambda_N)^\intercal && (3.52)\\ &= \left(F^\intercal W_1, \ldots, F^\intercal W_N\right)^\intercal / T && (3.53)\\ &= \begin{bmatrix} W_1{}^\intercal F \\ \vdots \\ W_N{}^\intercal F \end{bmatrix} / T = W^\intercal F / T && (3.54) \end{aligned}$$

带入消掉 Λ 得到

$$\min_{F,\Lambda} \sum_{i=1}^{N} \|W_i F \lambda_i\|^2 = \operatorname{tr}\left(WW^\intercal\right) - \operatorname{tr}\left(W^\intercal F F^\intercal W\right) / T \tag{3.55}$$

then we find that minimizing reconstruction error is equal to maximize $\operatorname{tr}\left(W^\intercal F F^\intercal W\right)$, or

$$\max_{F} \operatorname{tr}\left(F^\intercal W W^\intercal F\right) \tag{3.56}$$

then $F = (\ell_1, \ldots, \ell_r)\sqrt{T}$, where $\ell_1, \ldots, \ell_r$ are the first r normalized eigen-vec of $T \times T$ $WW^\intercal$.

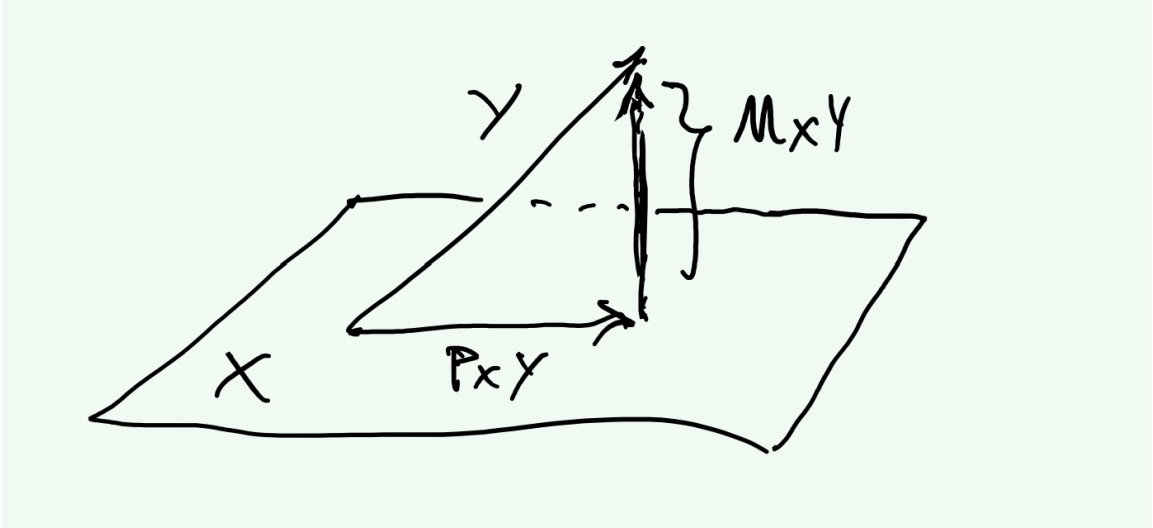
Note. 以上是从 $WW^\intercal$ 中得到的前 r 个 eigen-vec. 由于 eigen-vecs 的长度是单位化的，因而也可以从 $\frac{1}{NT}WW^\intercal$ 中求得 $\ell_1, \ldots, \ell_r$. 这里 $\frac{1}{NT}$ 体现了对 NT 条数据的平均化.

Note. 我们可以将此归纳为

$$\frac{1}{NT} W W^\intercal \widehat{F} = \widehat{F} V_{NT} \tag{3.57}$$

where V_{NT} is the diagonal matrix consisting of first r eigen-vals. We can get $\widehat{\beta}$ and $\left(\widehat{F}, \widehat{\Lambda}\right)$ through iterations.

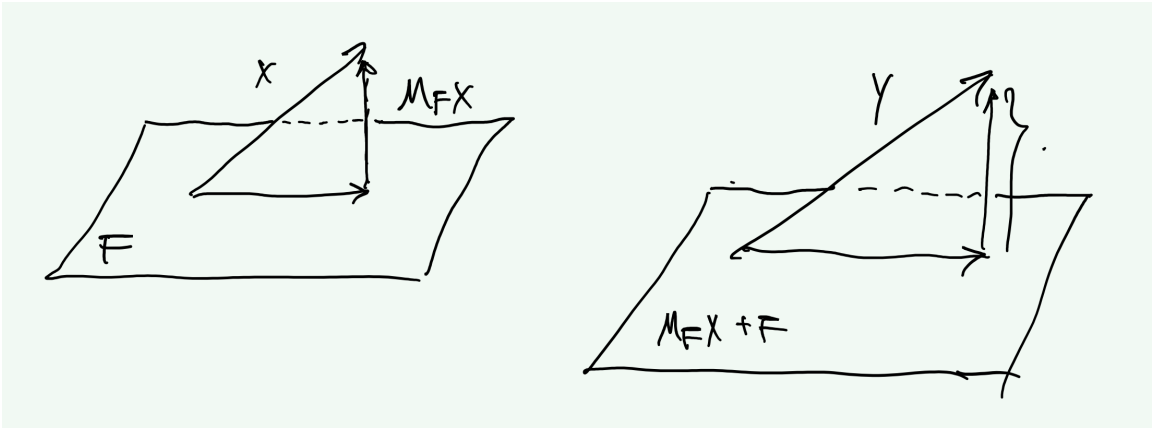
Figure 3.2: Linear regression



3.3.2 Consistency of Slope $\widehat{\beta}$

In linear regression, we can represent LSE as Figure 3.2 Then, in panel data regression with factor error, we will project Y on X and F, and it's possible that X is correlated with F. Thus we project X on F first and then, use $M_F X$ and F as predictors to interpret Y. Figure 3.3 represents it.

Figure 3.3: Linear regression with factor error



Note. 本文没有再引入 $\Lambda^\star$, 因为默认 $F^\star$ 的 LSE 得到的即为 true factor loading[11]

在证明 consistency 之前，还需要注意到一个事实：$\frac{1}{T}\left\|X_i^\intercal F^\star\right\| = \mathcal{O}_p(1)$, 它限制了 X_i 与 F 在时间上的相关性. 我们可以证明这个事实：

Proof: by sub-multiplicative property of F-norm,

$$\left\|X_i^\intercal F^\star\right\|/T \leq \|X_i\|\left\|F^\star\right\|/T \tag{3.58}$$

by Cauchy-Schwartz inequality(in expectation form),

$$\mathbf{E}\left(\|X_i\|\left\|F^\star\right\|\right)/T = \left[\mathbf{E}\left(\|X_i\|^2\right)/T\right]^{1/2}\left[\mathbf{E}\left(\|Y_i\|^2\right)/T\right]^{1/2} \tag{3.59}$$

$$= \left[\mathbf{E}\left(\frac{1}{T}\sum_{t=1}^{T}\|X_{it}\|^2\right)\right]^{1/2}\left[\mathbf{E}\left(\frac{1}{T}\sum_{t=1}^{T}\left\|F_t^0\right\|^2\right)\right]^{1/2} \tag{3.60}$$

$$= \left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{E}\left(\|X_{it}\|^2\right)\right]^{1/2}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{E}\left(\left\|F_t^0\right\|^2\right)\right]^{1/2} \tag{3.61}$$

also,since

$$\mathbf{E}\left(\|X_{it}\|^4\right) \leq M \text{ and } \mathbf{E}\left(\|F_t\|^4\right) \leq M \tag{3.62}$$

we have

$$\mathbf{E}\left(\|X_{it}\|^2 \cdot 1\right) \leq \left(\mathbf{E}\left(\|X_{it}\|^4\right)\right)^{1/2} \leq M^{1/2} \tag{3.63}$$

and

$$\mathbf{E}\left(\|F_t\|^2 \cdot 1\right) \leq \left(\mathbf{E}\left(\|F_t\|^4\right)\right)^{1/2} \leq M^{1/2} \tag{3.64}$$

we arrive that

$$\mathbf{E}\left(\|X_i\|\left\|F^\star\right\|\right)/T \leq M^{1/2} \tag{3.65}$$

thus

$$\frac{1}{T}\left\|X_i^\intercal F^\star\right\| = \mathcal{O}_p(1) \tag{3.66}$$

□

Note. 1. 随机变量的高阶矩有界，则低阶矩也一定有界 (by Cauchy-Schwartz inequalityor Hölder's inequality)

2. 允许 X_i 与 $F_j, j = 1, \ldots, r$ 有较高的相关性，甚至是线性相关.

由之前的分析知, $\left(\widehat{\beta}, \widehat{F}\right)$ 来自于如下的目标函数 $S_{NT}(\beta, F)$:

$$S_{NT}(\beta, F) = \frac{1}{NT}\sum_{i=1}^{N}(Y_i - X_i\beta)^\intercal M_F(Y_i - X_i\beta) - \frac{1}{N}\sum_{i=1}^{N}\epsilon_i^\intercal M_{F^\star}\epsilon_i \tag{3.67}$$

$$= \frac{1}{NT}\sum_{i=1}^{N}\left(F^\star\lambda_i + \epsilon_i - X_i\beta\right)^\intercal M_F\left(F^\star\lambda_i + \epsilon_i - X_i\beta\right) - \frac{1}{N}\sum_{i=1}^{N}\epsilon_i^\intercal M_{F^\star}\epsilon_i \tag{3.68}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\Big(\lambda_i^\intercal F^{\star\intercal} M_F F^0\lambda_i - 2\lambda_i^\intercal F^{\star\intercal} M_F X_i\beta + \beta^\intercal X_i^\intercal M_F X_i\beta \tag{3.69}$$

$$+\epsilon_i^\intercal M_F\epsilon_i + 2\lambda_i^\intercal F^{\star\intercal} M_F\epsilon_i - 2\beta^\intercal X_i^\intercal M_F\epsilon_i\Big) - \frac{1}{N}\sum_{i=1}^{N}\epsilon_i^\intercal M_{F^\star}\epsilon_i \tag{3.70}$$

$$=: \widetilde{S}_{NT}(\beta, F) + I_1 + 2I_2 - 2I_3 \tag{3.71}$$

and

$$\widetilde{S}_{NT} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i^\intercal F^{\star\intercal} M_F F^0\lambda_i - 2\left(\frac{1}{N}\sum_{i=1}^{N}\lambda_i^\intercal F^{\star\intercal} M_F X_i\right)\beta + \beta^\intercal\left(\frac{1}{N}\sum_{i=1}^{N}X_i^\intercal M_F X_i\right)\beta \tag{3.72}$$

$$I_1 = \frac{1}{NT}\sum_{i=1}^{N}\epsilon_i^\intercal(P_{F^\star} - P_F)\epsilon_i$$

$$I_2 = \frac{1}{NT}\sum_{i=1}^{N}\lambda_i^\intercal F^{\star\intercal} M_F\epsilon_i$$

$$I_3 = \beta^\intercal\left(\frac{1}{NT}\sum_{i=1}^{N}X_i^\intercal M_F\epsilon_i\right)$$

Also, notice that $F^\star\lambda_i$ is a $T \times 1$ vector, we can rewrite the first term in the trace form, which is the inner product of $F^{\star\intercal} M_F F^0/T$ and $\Lambda^\intercal\Lambda/N$

$$\Lambda = \begin{pmatrix}\lambda_1 & \dots & \lambda_N\end{pmatrix}^\intercal \quad \text{a } N \times r \text{ matrix} \tag{3.73}$$

$$F\Lambda^\intercal = \begin{pmatrix}F\lambda_1 & \dots & F\lambda_N\end{pmatrix} \tag{3.74}$$

thus

$$\frac{1}{N}\sum_{i=1}^{N}\lambda_i^\intercal F^{\star\intercal} M_F F^0\lambda_i = \frac{1}{NT}\operatorname{tr}\left([\Lambda F^{\star\intercal} M_F{}^\intercal][M_F F^0\Lambda^\intercal]\right) \tag{3.75}$$

$$= \operatorname{tr}\left(\frac{F^{\star\intercal} M_F F^0}{T}\frac{\Lambda^\intercal\Lambda}{N}\right) \tag{3.76}$$

3.3.3 Main Proof

Newey and McFadden (1994) 给出了证明 $\widehat{\beta} = \arg\min f_N(\beta)$ 依概率收敛到 $\beta^\star$ 的一般思路，其中 $f_N(\beta)$ 是在 N 条数据下模型的损失函数. Assume that for some M s.t $\|\beta\| \leq M$. 若能证明 $f_N(\beta) \xrightarrow{P} f(\beta)$, as $N \to \infty$, 其中 f 是不依赖于样本量 N 的函数. 若 $\beta^\star$ 是 $f(\cdot)$ 唯一的最小值点，那么有 $\beta \xrightarrow{P} \beta$. 然而，在我们的目标函数 $S_{NT}(\beta, F)$ 中，F 是一个 $T \times r$ matrix, and $\|F\|^2 = \operatorname{tr}\left(F^\intercal F\right) = T \operatorname{tr}\left(F^\intercal F/T\right) = T \cdot r$, 从而当样本量 T 增加时，F 不是一个有界的待定参数. Bai (1994) 改进了 Newey and McFadden 的做法，使得允许 (dimension of) parameter space increases with the sample size. Bai (2009) 沿用了这一做法：去证明存在 $\widetilde{S}_{NT}$ 使得 $S_{NT} - \widetilde{S}_{NT} = \mathrm{o}_p(1)$，并且 $\widetilde{S}_{NT}$ 在 $\beta^\star$ 与 $F^\star$ 处取得最小值. 由于 $\|\beta\| \leq M$，从而 $\widehat{\beta} \xrightarrow{P} \beta^\star$，而 $\|F\| = \sqrt{T \cdot r}$，因此还要继续要论在其他意义下 $\widehat{F}$ 的相合性. 本文证明了在如下 3 种意义下 $\widehat{F}$ 与 $F^\star$ 是渐近等价的：

1. $\left\|P_{\widehat{F}} - P_{F^\star}\right\|_F \xrightarrow{P} 0$. This is the part (ii) in prop 1. This result claims that the space spanned by $\widehat{F}$ and $F^\star$ are asymptotically the same;
2. componentwise consistency
3. $\frac{1}{T}\left\|\widehat{F} - F^\star H\right\|^2 = \mathcal{O}_p(\|\widehat{\beta} - \beta^\star\|^2) + \mathcal{O}_p(\frac{1}{\min[N,T]})$. This result shows the average norm consistency. This is the part (ii) in prop A.1.

Note. 因为 $P_{F^\star} = P_{F^0 H}, \forall r \times r$ invertible matrix H，从而实际上上述 3 个结果证明了 $\widehat{F}$ 与 $F^\star$ 是渐近等价的，其中 F 可以进一步表示成 $\widehat{F}$ 与 F_0 的函数，$H = \left(\frac{\Lambda\Lambda^\intercal}{N}\right)\left(\frac{F^{\star\intercal}\widehat{F}}{T}\right)V_{NT}^{-1}$，细节参见subsection 3.2.

Proposition 3.2. *consistency*

由之前的计算得

$$\left\|S_{NT}(\beta, F) - \widetilde{S}_{NT}(\beta, F)\right\| \leq \sum_{k=1}^{3} \|I_k\|, \tag{3.77}$$

由 Lemma 1 知，$\sum_{k=1}^{3} \|I_k\| = \mathrm{o}_p(1)$，因此

$$S_{NT}(\beta, F) = \widetilde{S}_{NT}(\beta, F) + \mathrm{o}_p(1) \tag{3.78}$$

然后，我们要证 $\widetilde{S}_{NT}$ 只在 $\beta^\star$，$F^0 H$ 处取得最小值. 由于之前扣掉真值，因此要证

$$\widetilde{S}_{NT}(\beta^\star, F^0 H) = 0 \tag{3.79}$$

and

$$\widetilde{S}_{NT}(\beta, FH) > 0, \quad \forall (\beta, F) \neq (\beta^\star, F^0 H)$$

为了方便，不妨令 $\beta^\star = 0$. 由于 $P_F F = F(F^\intercal F)^{-1} F^\intercal F = F$，从而 $M_F F = 0$，从容易证明第一条. 接下来往证第二条.

我们希望将 $\widetilde{S}_{NT}(\beta, F)$ 改写为二次型的和的形式，再然后只要假设二次型矩阵是一个正定阵就证明了第二条. 我们逐个考虑 $\widetilde{S}_{NT}(\beta, F)$ 中的 3 项. 先考虑第一项，之前已经得到

$$\frac{1}{N}\sum_{i=1}^{N} \lambda_i^\intercal F^{\star\intercal} M_F F^0 \lambda_i = \operatorname{tr}\left(\frac{F^{\star\intercal} M_F F^0}{T}\frac{\Lambda^\intercal \Lambda}{N}\right) \tag{3.80}$$

那么，利用 $\operatorname{tr}(\cdot)$ 与 $\operatorname{vec}(\cdot)$[12]的性质, 将 $\frac{\Lambda^\intercal \Lambda}{N}$ 作为 $r \times r$ 的二次型矩阵，得到

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \lambda_i^\intercal F^{\star\intercal} M_F F^0 \lambda_i &= \frac{1}{T}\operatorname{tr}\left(M_F F^0 \frac{\Lambda^\intercal \Lambda}{N} F^{\star\intercal} M_F{}^\intercal\right) && (3.81)\\
&= \frac{1}{T}\sum_{t=1}^{T} \widetilde{F}_t^\intercal \frac{\Lambda^\intercal \Lambda}{N} \widetilde{F}_t && (3.82)\\
&= \frac{1}{T}\operatorname{vec}\left(F^{\star\intercal} M_F\right)^\intercal \left(\frac{\Lambda^\intercal \Lambda}{NT} \otimes I_T\right) \operatorname{vec}\left(F^{\star\intercal} M_F\right) && (3.83)\\
&=: \eta^\intercal B \eta && (3.84)
\end{aligned}$$

其中，$\operatorname{vec}\left(F^{\star\intercal} M_F\right)$ 将 $r \times T$ 的矩阵转换为 $r \cdot T \times 1$ 的列向量, 而 I_T/T 代替了 $\sum_{t=1}^{T}$

对于交叉项 $2\left(\frac{1}{N}\sum_{i=1}^{N} \lambda_i^\intercal F^{\star\intercal} M_F X_i\right)\beta$, 它可以改写成下面这个式子

$$\operatorname{vec}\left(F^{\star\intercal} M_F\right)^\intercal \left(\frac{1}{NT}\sum_{i=1}^{N} \lambda_i \otimes M_F X_i\right)\beta =: 2\eta^\intercal C\beta \tag{3.85}$$

但是，为什么可以写成这个式子呢？[12]令 $A = \frac{1}{N}\sum_{i=1}^{N} X_i^\intercal M_F X_i$, 则 Equation 3.72可以写成

$$\widetilde{S}_{NT} = \eta^\intercal B\eta - 2\eta^\intercal C\beta + \beta^\intercal A\beta$$

我们可以根据它的一维版本将上面这个式子写成两个二次型的和：

$$\begin{aligned}
ax^2 + by^2 + 2cxy &= a\left(x^2 + \frac{2c}{a}xy\right) + by^2\\
&= a\left(x + \frac{c}{a}y\right) - a\cdot\frac{c^2}{a^2}y^2 + by^2\\
&= a\left(x + \frac{c}{a}y\right)^2 + \left(b - \frac{c^2}{a}\right)y^2
\end{aligned}$$

[12]wikipedia for "vectorization(mathematics)"

or

$$\begin{aligned} ax^2 + by^2 + 2cxy &= b\left(y^2 + \frac{2c}{b}xy\right) + ax^2 \\ &= b\left(y + \frac{c}{b}x\right)^2 + ax^2 - \frac{c^2}{b}x^2 \\ &= \left(a - \frac{c^2}{b}x\right)^2 + b\left(y + \frac{c}{b}x\right)^2 \\ &= x\left(a - c \cdot \frac{1}{b} \cdot c\right)x + \left(y + x \cdot c \cdot \frac{1}{b}\right)b\left(y + x \cdot c \cdot \frac{1}{b}\right) \end{aligned}$$

thus, we rewrite the $\widetilde{S}_{NT}$ as

$$\begin{aligned} \widetilde{S}_{NT} &= \beta^\intercal A\beta + \eta^\intercal B\eta + 2\eta^\intercal C\beta \\ &= \beta^\intercal \left(A - C^\intercal B^{-1}C\right)\beta + \left(\eta^\intercal + \beta^\intercal C^\intercal B^{-1}\right)B\left(\eta + B^{-1}C\beta\right) \end{aligned}$$

但是我不明白为什么有 $A - C^\intercal B^{-1}C = D(F)$. 若先承认这件事情，那么假设 $\inf_F D(F) > 0$，从而上面两个二次型矩阵均为正定阵，那么 $\widetilde{S}_{NT} \geq 0$. 并且，第一项等于 0 当且仅当 $\beta = 0$; 当第一项等于 0 时，第二项等于 $\eta^\intercal B\eta$，从而（在第一项等于 0 的条件下）第二项等于 0 当且仅当 $\eta = 0$. 因此我们证明了 $\widetilde{S}_{NT}(\beta, FH) > 0, \quad \forall (\beta, F) \neq (\beta^\star, F^\star H)$.

至此，我们证明了 $\widetilde{S}_{NT}$ 只在 $\beta^\star$，$F^0 H$ 处取得最小值. 但是，由于 $\boldsymbol{\beta} \leq M$ and $\|F\| = \sqrt{T \cdot r}$, 从而只能证明 $\widehat{\beta} \xrightarrow{P} \beta^\star$. Next we show that $P_{\widehat{F}} \xrightarrow{P} P_{F^\star}$. Since

$$0 \leq S_{NT}(\beta, F) = \widetilde{S}_{NT}(\beta, F) + \mathrm{o}_p(1) \tag{3.86}$$

$$\widetilde{S}_{NT}(\beta^\star, F^\star) = 0 \tag{3.87}$$

$$S_{NT}(\widehat{\beta}, \widehat{F}) \leq S_{NT}(\beta^\star, F^\star) \tag{3.88}$$

then we have

$$0 \geq S_{NT}(\widehat{\beta}, \widehat{F}) = \widetilde{S}_{NT}(\widehat{\beta}, \widehat{F}) + \mathrm{o}_p(1) \tag{3.89}$$

thus

$$\widetilde{S}_{NT}(\widehat{\beta}, \widehat{F}) = \mathrm{o}_p(1) \tag{3.90}$$

since we have shown that $\widehat{\beta} \xrightarrow{P} \beta^\star = 0$, this implies that $\eta^\intercal C\beta \xrightarrow{P} 0$ and $\beta^\intercal A\beta \xrightarrow{P} 0$. Then we must have $\eta^\intercal B\eta = \mathrm{o}_p(1)$, that is

$$\mathrm{tr}\left(\frac{F^{\star\intercal} M_{\widehat{F}} F^\star}{T}\frac{\Lambda^\intercal \Lambda}{N}\right) = \mathrm{o}_p(1) \tag{3.91}$$

由于 $\Lambda^\intercal\Lambda/N \to \Sigma_\Lambda > 0$, 从而当 N 充分大时必有 $\Lambda^\intercal\Lambda/N > 0$, 那么有 $\frac{F^{\star\intercal} M_{\widehat{F}} F^\star}{T} = \mathrm{o}_p(1)$[13], 也就是

$$\frac{F_0^\intercal \left(I_T - P_{\widehat{F}} F^\star\right)}{T} = \mathrm{o}_p(1) \tag{3.92}$$

or we can write is as

$$\frac{F^{\star\intercal} F^\star}{T} = \frac{F^{\star\intercal} P_{\widehat{F}} F^\star}{T} + \mathrm{o}_p(1) \tag{3.93}$$

$$= \frac{F^{\star\intercal}\left(\widehat{F}\widehat{F}^\intercal\right)F^\star}{T^2} + \mathrm{o}_p(1) \tag{3.94}$$

$$= \frac{F^{\star\intercal}\widehat{F}}{T} \cdot \frac{F^{\star\intercal}\widehat{F}^\intercal}{T} + \mathrm{o}_p(1) \tag{3.95}$$

从谱分解的角度可知，$AA^\intercal$ 可逆则意味着方阵 A 也是可逆的，据此证明了 $\frac{F^{\star\intercal}\widehat{F}}{T}$ 是一个可逆阵.

3.3.4 Technical Lemmas

Lemma 2. *Three terms containing noise converge to 0 in probability.*
Conditions:

1. $\mathbf{E}\left(\|X_{it}\|^4\right) \le M$
2. $\underset{F\in\mathcal{F}}{D(F)} > 0$
3. $\mathbf{E}\left(\|F_t\|^4\right) \le M$
4. $\frac{1}{T}\sum_{t=1}^T F_t F_t^\intercal \to \Sigma_F > 0$
5. $\mathbf{E}\left(\|\lambda_i\|^4\right) \le M$
6. $\Lambda^\intercal\Lambda/N = \frac{1}{N}\sum_{i=1}^N \lambda_i\lambda_i^\intercal \to \Sigma_\Lambda > 0$
7. $\left|\sigma_{ij,ts}\right| \le \tau_{ts}$
8. $\mathbf{E}\left(|\epsilon_{it}|^8\right) \le M$
9. $\frac{1}{T}\sum_{ts}\tau_{ts} \le M$, 控制时间相依性, *and we can rewrite it as*

$$\frac{1}{\sqrt{T}}\sum_t \frac{1}{\sqrt{T}}\sum_s \tau_{ts} \le M \tag{3.96}$$

但是，对于这个条件我有一个问题[14]

10. $\mathbf{E}\left(\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N}(\epsilon_{it}\epsilon_{is}-\mathbf{E}(\epsilon_{it}\epsilon_{is}))\right|^4\right)\leq M$, 控制截面相关性

Results:

1. $\frac{1}{N}\sum_{i=1}^{N}X_i^\intercal\epsilon_i=\mathrm{o}_p(1)$

2. $\sup_F\frac{1}{N}\sum_{i=1}^{N}X_i^\intercal P_F\epsilon_i=\mathrm{o}_p(1)$

Proof: To prove 1, we write that

$$\frac{1}{NT}\left\|\sum_{i=1}^{N}X_i^\intercal\epsilon_i\right\|=\frac{1}{NT}\left\|\sum_{i=1}^{N}\sum_{t=1}^{T}X_{it}\epsilon_{it}\right\| \tag{3.97}$$

$$\leq\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\|X_{it}\epsilon_{it}\| \tag{3.98}$$

$$=\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathrm{tr}\left(X_{it}\epsilon_{it}\epsilon_{it}{X_{it}}^{\intercal\intercal}\right) \tag{3.99}$$

$$=\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left({X_{it}}^\intercal X_{it}\left(\epsilon_{it}^2-\mathbf{E}\left(\epsilon_{it}^2\right)+expt\epsilon_{it}^2\right)\right) \tag{3.100}$$

$$=\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left({X_{it}}^\intercal X_{it}\left(\epsilon_{it}^2-\mathbf{E}\left(\epsilon_{it}^2\right)\right)\right)+\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left({X_{it}}^\intercal X_{it}\mathbf{E}\left(\epsilon_{it}^2\right)\right) \tag{3.101}$$

接下去的证明是否可以用 Cauchy-Schwartz 不等式?

To prove 2, we want to show that $\frac{1}{N}\sum_{i=1}^{N}X_i^\intercal P_F\epsilon_i=\mathcal{O}_p(1)\mathrm{o}(1)$. Using $P_F=FF^\intercal/T$, we write it as

$$\frac{1}{N}\sum_{i=1}^{N}X_iP_F\epsilon_i=\frac{1}{N}\sum_{i=1}^{N}X_iFF^\intercal\epsilon_i/T \tag{3.102}$$

$$=\frac{1}{N}\sum_{i=1}^{N}\left(\frac{X_iF}{T}\right)\left(\frac{1}{T}\sum_{t=1}^{T}F_t\epsilon_{it}\right) \tag{3.103}$$

then

$$\frac{1}{NT}\left\|\sum_{i=1}^{N} X_i P_F \epsilon_i\right\| \leq \frac{1}{NT}\sum_{i=1}^{N}\|X_i P_F \epsilon_i\| \quad \text{subadditivity} \tag{3.104}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\|\left(\frac{X_i F}{T}\right)\left(\frac{1}{T}\sum_{t=1}^{T} F_t \epsilon_{it}\right)\right\| \tag{3.105}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left\|\left(\frac{X_i F}{T}\right)\right\|\left\|\left(\frac{1}{T}\sum_{t=1}^{T} F_t \epsilon_{it}\right)\right\| \quad \text{submultiplicative property} \tag{3.106}$$

$$\leq \frac{1}{N}\left(\sum_{i=1}^{N}\left\|\frac{X_i F}{T}\right\|^2\right)^{1/2}\left(\sum_{i=1}^{N}\left\|\frac{1}{T}\sum_{t=1}^{T} F_t \epsilon_{it}\right\|^2\right)^{1/2} \quad \text{Cauchy-Schwartz inequality} \tag{3.107}$$

$$= \left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{X_i F}{T}\right\|^2\right)^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T}\sum_{t=1}^{T} F_t \epsilon_{it}\right\|^2\right)^{1/2} \tag{3.108}$$

I have question about the last second step[15]. 我们去证明以上两块分别是 $\mathcal{O}_p(1)$ 和 $\mathrm{o}_p(1)$. Since

$$\left\|\frac{X_i F}{T}\right\| \leq \frac{1}{T}\|X_i\|\|F\| \tag{3.109}$$

and using $FF^\intercal/T = I_r$, we have

$$\|F\| = \left(\mathrm{tr}\left(FF^\intercal\right)\right)^{1/2} = \sqrt{T}\left(\mathrm{tr}\left(FF^\intercal/T\right)\right)^{1/2} = \sqrt{T}\cdot\sqrt{r} \tag{3.110}$$

thus

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{X_i F}{T}\right\|^2 \leq \frac{r}{N}\sum_{i=1}^{N}\|X_i\|^2/T \tag{3.111}$$

$$= \frac{r}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T}\|X_{it}\|^2 \quad \text{using } \mathbf{E}\left(\|X_{it}\|^2\right) \leq M \text{ and Markov inequality} \tag{3.112}$$

$$= \mathcal{O}_p(1) \tag{3.113}$$

Then we need to show that

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T}\sum_{t=1}^{T} F_t \epsilon_{it}\right\|^2 = \mathrm{o}_p(1) = \mathcal{O}_p(1)\mathrm{o}(1) \tag{3.114}$$

At first we express the above in trace form

$$\sum_{i=1}^{N}\left\|\frac{1}{T}\sum_{t=1}^{T}F_t\epsilon_{it}\right\|^2 = \mathrm{tr}\left(\left\{\frac{1}{T}\sum_{t=1}^{T}F_t\epsilon_{it}\right\}\left\{\frac{1}{T}\sum_{s=1}^{T}\sum_{i=1}^{N}\epsilon_{is}F_s^\intercal\right\}\right) \tag{3.115}$$

$$= \mathrm{tr}\left(\frac{1}{T^2}\sum_{st}F_tF_s^\intercal\sum_{i=1}^{N}\epsilon_{is}\epsilon_{it}\right) \tag{3.116}$$

$$= \frac{1}{T^2}\sum_{st}F_s^\intercal F_t\left\{\sum_{i=1}^{N}\epsilon_{is}\epsilon_{it}\right\} \tag{3.117}$$

$$= \frac{1}{T^2}\sum_{st}F_s^\intercal F_t\left\{\sum_{i=1}^{N}\epsilon_{is}\epsilon_{it}-\mathbf{E}\left(\epsilon_{is}\epsilon_{it}\right)\right\}+\frac{1}{T^2}\sum_{st}F_s^\intercal F_t\left\{\sum_{i=1}^{N}\mathbf{E}\left(\epsilon_{is}\epsilon_{it}\right)\right\} \tag{3.118}$$

$$\leq \frac{1}{T^2}\left|\sum_{st}F_s^\intercal F_t\left\{\sum_{i=1}^{N}\epsilon_{is}\epsilon_{it}-\mathbf{E}\left(\epsilon_{is}\epsilon_{it}\right)\right\}\right|+\frac{1}{T^2}\left|\sum_{st}F_s^\intercal F_t\left[\sum_{i=1}^{N}\mathbf{E}\left(\epsilon_{is}\epsilon_{it}\right)\right]\right| \tag{3.119}$$

$$=: \frac{1}{T^2}I_1+\frac{1}{T^2}I_2 \tag{3.120}$$

其中，倒数第二个式子，先扣掉真值，然后再加上绝对值放大. Thus we arrive that

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{T}\sum_{t=1}^{T}F_t\epsilon_{it}\right\|^2 = \frac{1}{NT^2}I_1+\frac{1}{NT^2}I_2 \tag{3.121}$$

For I_1, using Cauchy-Schwartz inequality, we have

$$I_1 \leq \left(\sum_{st}(F_s^\intercal F_t)^2\right)^{1/2}\left(\sum_{st}\left\{\sum_{i=1}^{N}\epsilon_{is}\epsilon_{it}-\mathbf{E}\left(\epsilon_{is}\epsilon_{it}\right)\right\}^2\right)^{1/2} \tag{3.122}$$

$$\leq \sum_{st}\|F_s\|^2\|F_t\|^2(\cdot)^{1/2} \quad \text{using sub-multiplicative property} \tag{3.123}$$

$$= \left(\sum_{t}\|F_t\|^2\right)(\cdot)^{1/2} \tag{3.124}$$

$$= \|F\|(\cdot)^{1/2} \tag{3.125}$$

$$= Tr\left(\sum_{st}\left\{\sum_{i=1}^{N}\epsilon_{is}\epsilon_{it}-\mathbf{E}\left(\epsilon_{is}\epsilon_{it}\right)\right\}^2\right)^{1/2} \tag{3.126}$$

we have that

$$\frac{1}{\sqrt{N}T^2}I_1 \leq r\left(\frac{1}{T^2}\sum_{st}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\epsilon_{is}\epsilon_{it}-\mathbf{E}\left(\epsilon_{is}\epsilon_{it}\right)\right\}^2\right)^{1/2} \tag{3.127}$$

the term in $\{\cdot\}$ is equal to $\mathcal{O}_p(1)$ because of condition 10. Thus the first part of the above identity is equal to $\mathcal{O}(N^{-1/2})\mathcal{O}_p(1)$

Then we need to show that $\frac{1}{NT^2}I_2 = \mathrm{o}_p(1)$.

$$\frac{1}{NT^2}I_2 = \frac{1}{NT^2}\left|\sum_{st} F_s^\top F_t \left[\sum_{i=1}^{N} \mathbf{E}\left(\epsilon_{is}\epsilon_{it}\right)\right]\right| \tag{3.128}$$

$$\leq \frac{1}{NT^2}\left(\sum_{st}\left(F_s^\top F_t\right)^2\right)^{1/2}\left(\sum_{st}\left[\sum_{i=1}^{N} \mathbf{E}\left(\epsilon_{is}\epsilon_{it}\right)\right]^2\right)^{1/2} \tag{3.129}$$

$$= \left(\frac{1}{T^2}\sum_{st}\left(F_s^\top F_t\right)^2\right)^{1/2}\frac{1}{\sqrt{T}}\left(\frac{1}{T}\sum_{st}\left[\frac{1}{N}\sum_{i=1}^{N} \mathbf{E}\left(\epsilon_{is}\epsilon_{it}\right)\right]^2\right)^{1/2} \tag{3.130}$$

$$= r \cdot \frac{1}{\sqrt{T}}\left(\frac{1}{T}\sum_{st}\tau_{ts}^2\right)^{1/2} \tag{3.131}$$

where the last equality follows from $[\cdot]^2 \leq \tau_{ts}^2$, that is

$$\left|\frac{1}{N}\sum_{i=1}^{N} \mathbf{E}\left(\epsilon_{is}\epsilon_{it}\right)\right| \leq \frac{1}{N}\sum_{i=1}^{N}\left|\mathbf{E}\left(\epsilon_{is}\epsilon_{it}\right)\right| \tag{3.132}$$

$$\leq \frac{1}{N}\sum_{i=1}^{N}\tau_{ts}(=\tau_{ts}) \tag{3.133}$$

where the last inequality follows from condition 7. Since

$$\left|\mathbf{E}\left(\epsilon_{it}\epsilon_{js}\right)\right| \leq \left(\mathbf{E}\left(\epsilon_{it}^2\right)\right)^{1/2}\left(\mathbf{E}\left(\epsilon_{js}^2\right)\right)^{1/2} \leq M \tag{3.134}$$

thus

$$\tau_{ts} \leq M \tag{3.135}$$

then

$$\frac{1}{T}\sum_{st}\tau_{ts}^2 \leq \frac{M}{T}\sum_{st}\tau_{ts} \tag{3.136}$$

$$= \mathcal{O}_p(1) \quad \text{condition 9} \tag{3.137}$$

therefore

$$\frac{1}{NT^2}I_2 \leq r \cdot \frac{1}{\sqrt{T}}\mathcal{O}_p(1) \tag{3.138}$$

□

Lemma 3.
Conditions:

1. $\frac{1}{NT^2}\sum_{tsuv}\sum_{ij}\left|\operatorname{cov}\left(\epsilon_{it}\epsilon_{is},\epsilon_{ju}\epsilon_{jv}\right)\right| \leq M$
2. ϵ_{it} *is independent of* X_{js}, λ_i, *and* F_s
3. $\mathbf{E}\left(\|X_{it}\|^4\right) \leq M$
4. $\mathbf{E}\left(\|F_t\|^4\right) \leq M$

Results:

1. $\mathbf{E}\left(\frac{1}{NT^2}\left\|\sum_i\sum_{ts}F_sF_t^{\intercal}(\epsilon_{it}\epsilon_{is}-\mathbf{E}\left(\epsilon_{it}\epsilon_{is}\right))\right\|^2\right) \leq M$
2. $\mathbf{E}\left(\frac{1}{NT^2}\left\|\sum_{k=1}^N\sum_{ts}X_{it}(\epsilon_{kt}\epsilon_{ks}-\mathbf{E}\left(\epsilon_{kt}\epsilon_{ks}\right))F_{hs}\right\|^2\right) \leq M$

Proof: To prove result 1,

$$\frac{1}{NT^2}\left\|\sum_i\sum_{ts}F_sF_t^{\intercal}(\epsilon_{it}\epsilon_{is}-\mathbf{E}\left(\epsilon_{it}\epsilon_{is}\right))\right\|^2 \tag{3.139}$$

$$= \frac{1}{NT^2}\operatorname{tr}\left(\left[\sum_i\sum_{ts}F_sF_t^{\intercal}(\epsilon_{it}\epsilon_{is}-\mathbf{E}\left(\epsilon_{it}\epsilon_{is}\right))\right]\left[\sum_j\sum_{uv}\left(\epsilon_{ju}\epsilon_{jv}-\mathbf{E}\left(\epsilon_{ju}\epsilon_{jv}\right)\right)F_vF_u^{\intercal}\right]\right) \tag{3.140}$$

$$= \frac{1}{NT^2}\operatorname{tr}\left(\left[\sum_{ts}F_sF_t^{\intercal}\left(\sum_i\epsilon_{it}\epsilon_{is}-\mathbf{E}\left(\epsilon_{it}\epsilon_{is}\right)\right)\right]\left[\sum_{uv}F_vF_u^{\intercal}\left(\sum_j\epsilon_{ju}\epsilon_{jv}-\mathbf{E}\left(\epsilon_{ju}\epsilon_{jv}\right)\right)\right]\right) \tag{3.141}$$

$$= \frac{1}{NT^2}\operatorname{tr}\left(\sum_{tsuv}F_sF_t^{\intercal}F_vF_u^{\intercal}\left[\left(\sum_i\epsilon_{it}\epsilon_{is}-\mathbf{E}\left(\epsilon_{it}\epsilon_{is}\right)\right)\left(\sum_j\epsilon_{ju}\epsilon_{jv}-\mathbf{E}\left(\epsilon_{ju}\epsilon_{jv}\right)\right)\right]\right) \tag{3.142}$$

$$= \frac{1}{NT^2}\sum_{tsuv}F_u^{\intercal}F_sF_t^{\intercal}F_v\left[\left(\sum_i\epsilon_{it}\epsilon_{is}-\mathbf{E}\left(\epsilon_{it}\epsilon_{is}\right)\right)\left(\sum_j\epsilon_{ju}\epsilon_{jv}-\mathbf{E}\left(\epsilon_{ju}\epsilon_{jv}\right)\right)\right] \tag{3.143}$$

thus, since condition 2

$$\mathbf{E}\left(\frac{1}{NT^2}\left\|\sum_i\sum_{ts}F_sF_t^{\intercal}(\epsilon_{it}\epsilon_{is}-\mathbf{E}\left(\epsilon_{it}\epsilon_{is}\right))\right\|^2\right) \tag{3.144}$$

$$=\frac{1}{NT^2}\sum_{tsuv}\mathbf{E}\left(F_u{}^{\intercal}F_sF_t{}^{\intercal}F_v\right)\mathbf{E}\left(\left[\left(\sum_i\epsilon_{it}\epsilon_{is}-\mathbf{E}\left(\epsilon_{it}\epsilon_{is}\right)\right)\left(\sum_j\epsilon_{ju}\epsilon_{jv}-\mathbf{E}\left(\epsilon_{ju}\epsilon_{jv}\right)\right)\right]\right) \tag{3.145}$$

$$\leq\frac{1}{NT^2}\sum_{tsuv}\mathbf{E}\left(\|F_s\|^4\right)\sum_{ij}\left|\mathrm{cov}\left(\epsilon_{it}\epsilon_{is},\epsilon_{ju}\epsilon_{jv}\right)\right|\leq M^4\cdot M \tag{3.146}$$

Then, to prove result 2,

$$\left\|\sum_{k=1}^N\sum_{ts}X_{it}(\epsilon_{kt}\epsilon_{ks}-\mathbf{E}\left(\epsilon_{kt}\epsilon_{ks}\right))F_{hs}\right\|^2 \tag{3.147}$$

$$=\mathrm{tr}\left(\left[\sum_{k=1}^N\sum_{ts}X_{it}(\epsilon_{kt}\epsilon_{ks}-\mathbf{E}\left(\epsilon_{kt}\epsilon_{ks}\right))F_{hs}\right]\left[\sum_{j=1}^N\sum_{uv}F_{hv}\left(\epsilon_{ju}\epsilon_{jv}-\mathbf{E}\left(\epsilon_{ju}\epsilon_{jv}\right)\right)X_{iu}{}^{\intercal}\right]\right) \tag{3.148}$$

$$=\mathrm{tr}\left(\sum_{tsuv}X_{it}X_{iu}{}^{\intercal}F_{hs}F_{hv}\sum_{kj}(\epsilon_{kt}\epsilon_{ks}-\mathbf{E}\left(\epsilon_{kt}\epsilon_{ks}\right))\left(\epsilon_{ju}\epsilon_{jv}-\mathbf{E}\left(\epsilon_{ju}\epsilon_{jv}\right)\right)\right) \tag{3.149}$$

$$=\sum_{tsuv}X_{iu}{}^{\intercal}X_{it}F_{hs}F_{hv}\sum_{kj}(\epsilon_{kt}\epsilon_{ks}-\mathbf{E}\left(\epsilon_{kt}\epsilon_{ks}\right))\left(\epsilon_{ju}\epsilon_{jv}-\mathbf{E}\left(\epsilon_{ju}\epsilon_{jv}\right)\right) \tag{3.150}$$

then, use condition 2, we have

$$\mathbf{E}\left(\left\|\sum_{k=1}^N\sum_{ts}X_{it}(\epsilon_{kt}\epsilon_{ks}-\mathbf{E}\left(\epsilon_{kt}\epsilon_{ks}\right))F_{hs}\right\|^2\right) \tag{3.151}$$

$$=\sum_{tsuv}\mathbf{E}\left(X_{iu}{}^{\intercal}X_{it}F_{hs}F_{hv}\right)\sum_{kj}\mathrm{cov}\left(\epsilon_{kt}\epsilon_{ks},\epsilon_{ju}\epsilon_{jv}\right) \tag{3.152}$$

$$\leq\sum_{tsuv}\left|\mathbf{E}\left(X_{iu}{}^{\intercal}X_{it}F_{hs}F_{hv}\right)\right|\sum_{kj}\left|\mathrm{cov}\left(\epsilon_{kt}\epsilon_{ks},\epsilon_{ju}\epsilon_{jv}\right)\right| \tag{3.153}$$

then we only need to show that $\left|\mathbf{E}\left(X_{iu}{}^{\intercal} X_{it} F_{hs} F_{hv}\right)\right| \leq M$. Using Cauchy-Schwartz inequality,

$$\left|\mathbf{E}\left(X_{iu}{}^{\intercal} X_{it} F_{hs} F_{hv}\right)\right| \leq M \leq \left[\mathbf{E}\left(X_{iu}{}^{\intercal} X_{it}\right)^2\right]^{1/2} \left(\mathbf{E}\left(F_{hs} F_{hv}\right)^2\right)^{1/2} \tag{3.154}$$

$$\leq \left(\mathbf{E}\left(\|X_{iu}\|^2 \|X_{it}\|^2\right)\right)^{1/2} \left[\mathbf{E}\left(\|F_{hs}\|^2 \|F_{hv}\|^2\right)\right]^{1/2} \tag{3.155}$$

$$\leq \left\{\left[\mathbf{E}\left(\|X_{iu}\|^4\right)\right]^{1/2} \left[\mathbf{E}\left(\|X_{it}\|^4\right)\right]^{1/2}\right\}^{1/2} \left\{\left[\mathbf{E}\left(\|F_{hs}\|^4\right)\right]^{1/2} \left[\mathbf{E}\left(\|F_{hv}\|^4\right)\right]^{1/2}\right\}^{1/2} \tag{3.156}$$

$$\leq M \tag{3.157}$$

where the last $\leq$ follows from conditions 3 & 4 □

4 Further Topics: Grouped Factor Models

4.1 Introduction

4.2 Fixed number of groups

当我们采用 LSE 去估计 factor loadings 时，目标函数可以改写成

$$\sum_{i=1}^{N} \left\| y_i - X_i\beta - F_{g_i}\lambda_{g_i i} \right\|^2 = \sum_{i=1}^{N} (y_i - X_i\beta)^\intercal M_{F_{g_i}} (y_i - X_i\beta) \tag{4.1}$$

为了证明 joint estimator $(\widehat{\beta}, \widehat{F}, \widehat{G})$ 是 $(\beta^\star, F^\star, G^0)$ 的相合估计，要去考虑 $\sum_{i=1}^{N} \left\| y_i - X_i\beta - F_{g_i}\lambda_{g_i i} \right\|^2$ 与 $\sum_{i=1}^{N} \left\| y_i - X_i\beta^\star - F^0_{g_i}\Lambda^\star_{g_i i} \right\|^2$ 之间的差距. 由于 true DGP 为

$$y_i = X_i\beta^\star + F^0_{g_i}\Lambda^\star_{g_i i} + \epsilon_i\,, \tag{4.2}$$

因此将 true DGP 代入得

$$\begin{aligned}
&\sum_{i=1}^{N} \left\| y_i - X_i\beta^\star - F^0_{g_i}\Lambda^\star_{g_i i} \right\|^2 \\
=& \sum_{i=1}^{N} (y_i - X_i\beta^\star)^\intercal M_{F^0_{g_i}} (y_i - X_i\beta^\star) \\
=& \sum_{i=1}^{N} \epsilon_i{}^\intercal M_{F^0_{g_i{}^\star}} \epsilon_i
\end{aligned}$$

其中利用到 $M_{F^0_{g_i{}^\star}} F^0_{g_i} = 0$. 由于 $M_{F_{g_i}} F^0_{g_i}$ 是真实的因子得分在估计得到的因子空间 F_{g_i} 的补空间上的投影，一般不等于 0. 因此将 true DGP 代入 U_{NT} 时会多出交叉项，即为

$$\begin{aligned}
&(y_i - X_i\beta)^\intercal M_{F_{g_i}} (y_i - X_i\beta) \\
=& (X_i(\beta^\star - \beta) + F^0_{g_i}\Lambda^\star_{g_i i} + \epsilon_i)^\intercal M_{F_{g_i}} (X_i(\beta^\star - \beta) + F^0_{g_i}\Lambda^\star_{g_i i} + \epsilon_i) \\
=& (\beta^\star - \beta)^\intercal X_i{}^\intercal M_{F_{g_i}} X_i(\beta^\star - \beta) + 2(\beta^\star - \beta)^\intercal X_i{}^\intercal M_{F_{g_i}} F^0_{g_i}\Lambda^\star_{g_i i} \\
&\qquad + 2(\beta^\star - \beta)^\intercal X_i{}^\intercal M_{F_{g_i}} \epsilon_i \\
&\qquad + \lambda^{0\intercal}_{g_i} F^{0\intercal}_{g_i} M_{F_{g_i}} F^0_{g_i}\Lambda^\star_{g_i i} \\
&\qquad + 2\lambda^{0\intercal}_{g_i} F^{0\intercal}_{g_i} M_{F_{g_i}} \epsilon_i \\
&\qquad + \epsilon_i{}^\intercal M_{F_{g_i}} \epsilon_i
\end{aligned}$$

不妨让 $\beta^\star = 0$，那么

$$\begin{aligned} U_{NT} = \beta^\intercal \left(\frac{1}{NT} X_i^\intercal M_{F_{g_i}} X_i\right)\beta &+ \frac{1}{N}\sum_{i=1}^{N} \lambda_{g_i}^{0\intercal} F_{g_i}^{0\intercal} M_{F_{g_i}} F_{g_i}^0 \Lambda_{g_i i}^\star \\ &+ 2\beta^\intercal \left(\frac{1}{N}\sum_{i=1}^{N} X_i^\intercal M_{F_{g_i}} F_{g_i}^0 \Lambda_{g_i i}^\star\right) \\ &+ 2\beta^\intercal \left(\frac{1}{N}\sum_{i=1}^{N} X_i^\intercal M_{F_{g_i}} \epsilon_i\right) \\ &+ 2\frac{1}{N}\sum_{i=1}^{N} \lambda_{g_i}^{0\intercal} F_{g_i}^{0\intercal} M_{F_{g_i}} \epsilon_i \\ &+ \frac{1}{N}\sum_{i=1}^{N} \frac{1}{NT} \epsilon_i^\intercal \left(P_{F_{g_i}^0} - P_{F_{g_i}}\right)\epsilon_i \end{aligned}$$

Note. 这里，有 3 项含有 $M_{F_{g_i}} F_{g_i}^0$ ，若 F_{g_i} 与 $F_{g_i\star}$ 相差较大，该如何处理？

Note. 不妨令 $D_j = \frac{1}{NT}\sum_{i:g_i=j} X_i^\intercal M_{F_j} X_i$, 因此 D_j 是从 $\{i : g_i = j\}$ 这些样本中恢复的因子空间；相比于 $\frac{1}{N}\sum_{i=1}^{N} X_i^\intercal M_F X_i$ ，$\{D_j\}_{j=1}^{S}$ 的优势在哪里？

在Bai (2009) 中，我们证明了 $\left\|X_i^\intercal F\right\|/T \leq .$ 在Ando and Bai (2016) 中，依然可以证明，由于次可乘性

$$\left\|X_i^\intercal F_j\right\|/T \leq \|X_i\|\|F_j\|/T$$

以及

$$\left\|F_j\right\|^2/T = \operatorname{tr}\left(F_j^\intercal F_j/T\right) = r_j$$

故有

$$\left\|X_i^\intercal F_j\right\|/T \leq T^{1/2}\sqrt{r_j}\|X_i\|$$

由 Assumption D_2 知，$X_{it}, t = 1, \ldots, T$ 满足

$$\max_{1\leq i\leq N} T^{-1}\|X_i\|^2 = \mathcal{O}_p\left(N^{\alpha/2}\right)$$

若假设 $\exists r \text{ s.t } r_j \leq r, j = 1, \ldots, S,$ 即每一组的异质性是一致有界的，那么

$$\left\|X_i^\intercal F_j\right\|/T = \mathcal{O}_p\left(N^{\alpha/2}\right)$$

Lemma A_1 证明了含有 ϵ_i 的三项均是随即无穷小的，且趋于无穷小的速度与Bai (2009) 相同.

先去证明第一条：

$$\frac{1}{NT}\sum_{i=1}^{N}\left\|X_i^\intercal M_{F_j}\epsilon_i\right\| = \mathcal{O}_p\left(T^{1/4}\right) + \mathcal{O}_p\left(N^{1/4}\right)$$

考虑 $\left\|X_i^\intercal M_{F_j}\epsilon_i\right\|/T$ 的收敛速度. 因为

$$\begin{aligned}\frac{1}{T}\left\|X_i^\intercal M_{F_j}\epsilon_i\right\| &= \frac{1}{T}\left\|X_i^\intercal\epsilon_i - X_i^\intercal P_{F_j}\epsilon_i\right\| \\ &\overset{(i)}{\leq} \frac{1}{T}\left\|X_i^\intercal\epsilon_i\right\| + \frac{1}{T}\left\|X_i^\intercal P_{F_j}\epsilon_i\right\|\end{aligned}$$

其中，$\left\|X_i^\intercal\epsilon_i\right\|/T = \mathcal{O}_p\left(\frac{1}{\sqrt{NT}}\right)$ 相当于没有因子结构的线性回归. 我不清楚，step (i) 的放大是否体现了分组的优势？分组的好坏会不会影响这一步？

利用 $P_{F_j} = F_jF_j^\intercal/T$，则第二项可以写成

$$\frac{1}{T}\left\|X_i^\intercal P_{F_j}\epsilon_i\right\| = \left\|X_i^\intercal F_j/T\right\|\left\|F_j^\intercal\epsilon_i/T\right\|$$

由 Cauchy-Schwartz inequality 得，

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\left\|X_i^\intercal P_{F_j}\epsilon_i\right\| &\leq \left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{N}\sum_{i=1}^{N}X_i^\intercal F_j/T\right\|^2\right)^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N}\left\|F_j^\intercal\epsilon_i/T\right\|\right)^{1/2} \\ &= \mathcal{O}_p\left(N^{\alpha/2}\right)\cdot\left(\frac{1}{N}\sum_{i=1}^{N}\left\|F_j^\intercal\epsilon_i/T\right\|\right)^{1/2}\end{aligned}$$

接下来考虑 $\left\|F_j^\intercal\epsilon_i/T\right\|^2$ 的收敛速度，不妨将 F_j 写成 F_{g_i}，

$$\begin{aligned}\left\|F_j^\intercal\epsilon_i/T\right\|^2 &= \left\|\sum_{j=1}^{S}\frac{1}{T}\mathbb{1}_{(g_i=j)}F_j^\intercal\epsilon_i\right\|^2 \\ &\overset{(i)}{\leq} \sum_{j=1}^{S}\left\|F_j^\intercal\epsilon_i/T\right\|^2\end{aligned}$$

the setp (i) follows form sub-additivity. Then, since S is a fixed constant, we have

$$\text{above} \leq S\cdot\sup_F\left\|\frac{1}{T}F^\intercal\epsilon_i\right\|^2$$

因此，Ando and Bai (2016) 考虑的只是 fixed number of groups，用次可加性还是有效的，但是假设 S N, 那么便不太好了. 之前我们假设了 S 个因子空间的异质性有

共同的上界 r，那么对于 $j-$th group data $\{(X_i, y_i\} : g_i = j)$，$\exists$ 阶数为 r 的因子空间 $\widetilde{F}_j$ s.t

$$\left\|F_j{}^{\intercal}\epsilon_i/T\right\|^2 \leq \left\|\widetilde{F}_j^{\intercal}\epsilon_i/T\right\|^2$$

而 Bai (2009) 证明了

$$\sup_F \left\|F^{\intercal}\epsilon_i/T\right\|^2 = \mathcal{O}_p\left(N^{-1/2}\right) + \mathcal{O}_p\left(T^{-1/2}\right)$$

因此，

$$\left\|F_{g_i}{}^{\intercal}\epsilon_i/T\right\|^2 = \mathcal{O}_p\left(N^{-1/2}\right) + \mathcal{O}_p\left(T^{-1/2}\right)$$

那么

$$\frac{1}{N}\sum_{i=1}^{N}\left\|X_i^{\intercal}P_{F_{g_i}}\epsilon_i\right\| = \mathcal{O}_p\left(N^{\alpha/2}\right)\cdot\left[\mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)\right]$$

由于必须要 $\alpha/2 - 1/4 < 0$, 因此这里要求 $\alpha < 1/2$

在证明了 Lemma A1 后，就证明了

$$U_{NT} = \widetilde{U}_{NT} + \mathcal{O}_p\left(\frac{1}{N^{1/4}}\right) + \mathcal{O}_p\left(\frac{1}{T^{1/4}}\right)$$

为了证明 $\widetilde{U}_{NT}$ 在 $(\beta^\star, G^0, F^\star)$ 处取得最小值，则把 $\widetilde{U}_{NT}$ 写成二次型的和的形式，之前已经证明了

$$\begin{aligned}\widetilde{U}_{NT} = \beta^{\intercal}\left(\frac{1}{NT}X_i^{\intercal}M_{F_{g_i}}X_i\right)\beta + \frac{1}{N}\sum_{i=1}^{N}\lambda_{g_i}^{0\,\intercal}F_{g_i}^{0\intercal}M_{F_{g_i}}F_{g_i}^0\Lambda_{g_ii}^\star \\ + 2\beta^{\intercal}\left(\frac{1}{N}\sum_{i=1}^{N}X_i^{\intercal}M_{F_{g_i}}F_{g_i}^0\Lambda_{g_ii}^\star\right)\end{aligned}$$

根据 $g_i = j, j = 1,\ldots,S, g_i^\star = k, k = 1,\ldots,S$ ，将以上 3 项均为分若干组

$$\frac{1}{NT}\sum_{i=1}^{N}X_i^{\intercal}M_{F_{g_i}}X_i = \sum_{j=1}^{S}\left(\frac{1}{NT}\sum_{i:g_i=j}X_i^{\intercal}M_{F_j}X_i\right)$$

那么，类似于 Bai (2009) 中，记 $A = \frac{1}{NT}\sum_{i=1}^{N}X_i^{\intercal}M_FX_i$, 这里记 $D_j = \frac{1}{NT}\sum_{i:g_i=j}X_i^{\intercal}M_{F_j}X_i$，那么第一项等于 $\beta^{\intercal}\left(\sum_{j=1}^{S}D_j\right)\beta = \sum_{j=1}^{S}\beta^{\intercal}D_j\beta$ 这里没有体现出分组正确率对模型估计的影响. 接下来考虑后面的 2 项. 根据 g_i 与 $g_i^\star$ 进行分组

$$\frac{1}{N}\sum_{i=1}^{N}\lambda_{g_i}^{0\,\intercal}F_{g_i}^{0\intercal}M_{F_{g_i}}F_{g_i}^0\Lambda_{g_ii}^\star = \sum_{j=1}^{S}\sum_{k=1}^{S}\left(\frac{1}{NT}\sum_{\substack{i:\,g_i=j\\ g_i^\star=k}}\lambda_k^{0\intercal}F_k^{0\intercal}M_{F_j}F_k^0\Lambda_{ki}^\star\right)$$

在 Bai (2009) 中，将因子结构二次型这一项写成了

$$\frac{1}{NT}\sum_{i=1}^{N}\lambda_i^{\intercal}F^{\star\intercal}M_F F^0\lambda_i = \operatorname{vec}\left(F^{\star\intercal}M_F\right)^{\intercal}\left(\frac{\Lambda^{\intercal}\Lambda}{NT}\otimes I_r\right)\operatorname{vec}\left(F^{\star\intercal}M_F\right)$$

其中 $\Lambda = (\lambda_1,\ldots,\lambda_N)^{\intercal}$ 是 $N\times r$ 矩阵，是因子空间 F 下的坐标。而 $\operatorname{vec}()$ 取出了 $r\times T$ 的矩阵 $F^{\star\intercal}M_F$ 的每个列向量，是每个时刻下 r 个方向上的因子得分。从而上式可以继续写成

$$\begin{aligned}&= \operatorname{vec}\left(F_k^{0\intercal}M_{F_j}\right)^{\intercal}\left(\frac{\Lambda_{jk}{}^{\intercal}\Lambda_{jk}}{NT}\otimes I_T\right)\operatorname{vec}\left(F_k^{0\intercal}M_{F_j}\right)\\ &=: \zeta_{jk}{}^{\intercal}\left(\frac{1}{T}E_{jk}\right)\zeta_{jk}\end{aligned}$$

其中 $\Lambda_{jk}{}^{\intercal}\Lambda_{jk} = \sum_{\substack{i:\ g_i=j\\ g_i^{\star}=k}}\lambda_{ki}^0\cdot\lambda_{ki}^{0\intercal} = \sum_{\substack{i:\ g_i=j\\ g_i^{\star}=k}}\lambda_{ki}^0\otimes\lambda_{ki}^0$ 是 $r_k\times r_k$ 矩阵。修正文中记号 E_{jk} 的定义，改为 $E_{jk} = \frac{1}{NT}\left(\Lambda_{jk}{}^{\intercal}\Lambda_{jk}\otimes I_T\right)$ 若 F_j 是 F_k^0 的好的估计，那么 $M_{F_j}F_k^0\approx 0$.

从而

$$\begin{aligned}\frac{1}{NT}\sum_{i=1}^{N}\lambda_{g_i}^{0\intercal}F_{g_i}^{0\intercal}M_{F_{g_i}}F_{g_i}\lambda_{g_i} &= \sum_{j=1}^{S}\sum_{k=1}^{S}\zeta_{jk}{}^{\intercal}E_{jk}\zeta_{jk}{}^{\intercal}\\ &= \sum_{j=1}^{S}\zeta_j{}^{\intercal}E_j\zeta\end{aligned}$$

其中 $\zeta_j{}^{\intercal}\left[\zeta_{j1}{}^{\intercal},\ldots,\zeta_{jS}{}^{\intercal}\right]$, $E_j = \operatorname{diag}\left(E_{j1},\ldots,E_{jS}\right)$

类似地考虑第 3 项

$$\frac{1}{NT}\sum_{i=1}^{N}X_i^{\intercal}M_{F_{g_i}}F_{g_i^{\star}}^0\lambda_{g_i^{\star}i}^0 = \sum_{j=1}^{S}\sum_{k=1}^{S}\left(\frac{1}{NT}\sum_{\substack{i:\ g_i=j\\ g_i^{\star}=k}}X_i^{\intercal}M_{F_j}F_k^0\lambda_{ki}^0\right)$$

由 Bai (2009) 知

$$\frac{1}{NT}\sum_{i=1}^{N}\lambda_i^{\intercal}F^{\star\intercal}M_F X_i = \operatorname{vec}\left(F^{\star\intercal}M_F\right)^{\intercal}\left(\frac{1}{NT}\sum_{i=1}^{N}\lambda_i\otimes M_F X_i\right)$$

Table 4.1: Group the units by true and estimated membership

true est	1	2	...	S
1				
2				
⋮				
S				

从而

$$
\begin{aligned}
&= \sum_{j=1}^{S}\sum_{k=1}^{S}\left(\frac{1}{NT}\sum_{\substack{i:\ g_i=j\\ g_i^\star=k}}\lambda_{ki}^0\otimes M_{F_j}X_i\right)\operatorname{vec}\left(F_k^{0\intercal}M_{F_j}\right)\\
&= \sum_{j=1}^{S}\sum_{k=1}^{S}L_{jk}{}^{\intercal}\zeta_{jk}\\
&= \sum_{j=1}^{S}L_j{}^{\intercal}\zeta_j
\end{aligned}
$$

其中 $L_j{}^{\intercal}=(L_{j1}{}^{\intercal},\ldots,L_{jS}{}^{\intercal})$ 因此 $\widetilde{U}_{NT}$ 可以根据 $g_i=j,j=1,\ldots,S$ 改写成如下形式

$$
\widetilde{U}_{NT}=\sum_{j=1}^{S}\left(\beta^{\intercal}D_j\beta+\zeta_j{}^{\intercal}E_j\zeta_j+2\beta^{\intercal}L_j{}^{\intercal}\zeta_j\right)
$$

根据 Bai (2009) 的写法，我们又可以改写成

$$
\begin{aligned}
\widetilde{U}_{NT} &= \sum_{j=1}^{S}\left[\beta^{\intercal}(D_j-L_j{}^{\intercal}E_j^{-1}L_j)\beta+(\zeta_j{}^{\intercal}+\beta^{\intercal}L_j{}^{\intercal}E_j^{-1})E_j(\zeta_j+E_j^{-1}L_j\beta)\right]\\
&= \beta^{\intercal}\left[\sum_{j=1}^{S}(D_j-L_j{}^{\intercal}E_j^{-1}L_j)\right]\beta+\sum_{j=1}^{S}\left[(\zeta_j{}^{\intercal}+\beta^{\intercal}L_j{}^{\intercal}E_j^{-1})E_j(\zeta_j+E_j^{-1}L_j\beta)\right]
\end{aligned}
$$

因此，假设 $\left[\sum_{j=1}^{S}(D_j-L_j{}^{\intercal}E_j^{-1}L_j)\right]$ 与 $E_j,j=1,\ldots,S$ 是正定阵，则当且仅当 $\beta^\star,G^0,F^\star$ 时 $\widetilde{U}_{NT}$ 取得最小值 0

为了得到估计量的粗糙的收敛速度，有

$$
\begin{aligned}
U_{NT}(\beta^\star,G^0,F^\star) &= \widetilde{U}_{NT}(\beta^\star,G^0,F^\star)+\mathcal{O}_p\left(N^{-1/4}\right)+\mathcal{O}_p\left(T^{-1/4}\right)\\
&\geq \widetilde{U}_{NT}(\widehat{\beta},\widehat{G},\widehat{F})+\mathcal{O}_p\left(N^{-1/4}\right)+\mathcal{O}_p\left(T^{-1/4}\right)
\end{aligned}
$$

thus

$$\widetilde{U}_{NT}(\widehat{\beta},\widehat{G},\widehat{F}) = \mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$$

由于 $\widetilde{U}_{NT}(\widehat{\beta},\widehat{G},\widehat{F})$ 可以写成 2 个（非负定）矩阵的二次型的和，因此这两项都是 $\mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$，即有

$$\beta^\intercal\left[\sum_{j=1}^{S}(D_j - L_j{}^\intercal E_j^{-1}L_j)\right]\beta = \mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$$

Note. 但当 $S \to \infty$ 时怎么办呢？

因此有

$$\left\|\widehat{\beta} - \beta^\star\right\|^2 = \mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$$

或者说

$$\left\|\widehat{\beta} - \beta^\star\right\| = \mathcal{O}_p\left(N^{-1/8}\right) + \mathcal{O}_p\left(T^{-1/8}\right)$$

由于 $\widetilde{U}_{NT}$ 可以写成 3 项的和，

$$\widetilde{U}_{NT} = \beta^\intercal\left(\frac{1}{NT}X_i{}^\intercal M_{F_{g_i}}X_i\right)\beta + \frac{1}{N}\sum_{i=1}^{N}\lambda_{g_i}^{0\intercal}F_{g_i}^{0\intercal}M_{F_{g_i}}F_{g_i}^0\Lambda_{g_i i}^\star + 2\beta^\intercal\left(\frac{1}{N}\sum_{i=1}^{N}X_i{}^\intercal M_{F_{g_i}}F_{g_i}^0\Lambda_{g_i i}^\star\right)$$

那么也有

$$\frac{1}{N}\sum_{i=1}^{N}\lambda_{g_i}^{0\intercal}F_{g_i}^{0\intercal}M_{F_{g_i}}F_{g_i}^0\Lambda_{g_i i}^\star = \mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$$

Note. 我对这一步抱有怀疑，我觉得只能得出$\mathrm{o}_p(1)$，而没有速度

那么等价于是

$$\sum_{j=1}^{S}\sum_{k=1}^{S}\left(\frac{1}{NT}\sum_{\substack{i:\ g_i=j\\ g_i^\star=k}}\lambda_{ki}^{0\intercal}F_k^{0\intercal}M_{F_j}F_k^0\Lambda_{ki}^\star\right) = \mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$$

由 Bai (2009) 知，上式又可以改写为

$$\sum_{j=1}^{S}\sum_{k=1}^{S}\operatorname{tr}\left(\left[\frac{1}{T}F_k^{0\intercal}M_{F_j}F_k^0\right]A_{kj}\right) = \mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$$

$A_{kj} = \frac{1}{NT}\sum_{\substack{i:\ g_i=j\\ g_i^\star=k}}\lambda_{ki}^0\otimes\lambda_{ki}^0$, $\operatorname{tr}(\cdot)$ 是平方和，故必定大于 0. 因此每一项的上界也一定是 $\mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$

Note. 当 $S \to \infty$ 时，这一步需要调整

我们假设当 $N \to \infty$ 时，$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{(g_i=j)} \to c_j > 0$，即 jth group 不会退化到没有。因此可以假设这些 $j-$th group 中的因子在 F_j^0 上有着较强的因子载荷，即假设 $\sum_{j=1}^{S} A_{kj} = \frac{1}{NT}\sum_{\substack{i:\ g_i=j \\ g_i^\star=k}} \lambda_{ki}^0 \otimes \lambda_{ki}^0 \to \Sigma_{\Lambda_k^0} > 0$，即表示 kth group 是非退化的。$\sum_{j=1}^{S} A_{kj}$ 是第 j 列的和，表示所有被分到 jth group 的资产的真实的载荷。由于

$$\begin{aligned}
\sum_{k=1}^{S} A_{kj} &= \sum_{k=1}^{S} \sum_{\substack{i:\ g_i=j \\ g_i^\star=k}} \frac{1}{N} \lambda_{ki}^0 \otimes \lambda_{ki}^0 \\
&= \sum_{k=1}^{S} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(g_i=j)} \lambda_{ki}^0 \otimes \lambda_{ki}^0 \\
&= \sum_{k=1}^{S} c_j \left(\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(g_i=j)} \lambda_{ki}^0 \otimes \lambda_{ki}^0 \right) \\
&= c_j \cdot \sum_{k=1}^{S} \Sigma_{\Lambda_k^0} > 0
\end{aligned}$$

其中, 最后的结果是 S 个正定阵的加权和.

由于 $A_{kj}, k = 1, \ldots, S$ 都是非负定矩阵，而它们和的极限是正定阵，从而 $\exists$ 某个 k s.t A_{kj} 的极限是正定阵.

现在考虑第 1 列，不妨设 $A_{11} \to A_{11}^0 > 0$. 因此，由于 $\operatorname{tr}\left(\left[F_1^{0\intercal} M_{\widehat{F}_1} F_1^0 / T\right] A_{11}\right) = \mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$，从而有[13]

$$\left\| F_1^{0\intercal} M_{\widehat{F}_1} F_1^0 / T \right\|_F = \mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$$

因为

$$\begin{aligned}
\frac{1}{T} F_1^{0\intercal} M_{\widehat{F}_1} F_1^0 &= \frac{1}{T} F_1^{0\intercal} F_1^0 - \frac{1}{T} F_1^{0\intercal} \left(\widehat{F}_1 \widehat{F}_1^\intercal / T \right) F_1^0 \\
&= I_{r_1}
\end{aligned}$$

所以

$$\left\| \left(\frac{F_1^{0\intercal} \widehat{F}_1}{T} \right) \left(\frac{\widehat{F}_1^\intercal F_1^0}{T} \right) - I_{r_1} \right\| = \mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$$

[13]为什么可以从 $\operatorname{tr}(\cdot)$ 到 $\|\cdot\|_F$ ？

Table 4.2: Group the units by true k and estimated j

k j	1	2	...	S
1				
2				
⋮				
S				

而 $\left\|P_{\widehat{F}_1} - P_{F_1^0}\right\|^2 = \mathrm{tr}\left(P_{\widehat{F}_1} + P_{F_1^0} - 2P_{\widehat{F}_1}P_{F_1^0}\right)$. 因为

$$\begin{aligned}\mathrm{tr}\left(P_F\right) &= \mathrm{tr}\left(F(F^\intercal F)^{-1}F^\intercal\right)\\ &= \mathrm{tr}\left(FF^\intercal/T\right)\\ &= \mathrm{tr}\left(F^\intercal F/T\right) = \mathrm{tr}\left(I_r\right) = r\end{aligned}$$

由于之前证明了

$$\left\|\left(\frac{F_1^{0\intercal}\widehat{F}_1}{T}\right)\left(\frac{\widehat{F}_1^\intercal F_1^0}{T}\right) - I_{r_1}\right\| = \mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$$

因此差不多也有

$$\mathrm{tr}\left(I_{r_1} - P_{\widehat{F}_1}P_{F_1^0}\right) = \mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$$

所以

$$\left\|P_{\widehat{F}_1} - P_{F_1^0}\right\|^2 = \mathcal{O}_p\left(N^{-1/4}\right) + \mathcal{O}_p\left(T^{-1/4}\right)$$

或者说 $\left\|P_{\widehat{F}_1} - P_{F_1^0}\right\| = \mathcal{O}_p\left(N^{-1/8}\right) + \mathcal{O}_p\left(T^{-1/8}\right)$, 然后可以证明 $A_{k1} \to A_{k1}^0 = 0, k \neq 1$

若 $\exists k \neq 1 \ \text{ s.t}\, A_{k1}^0 > 0$, 那么也可以证明 $\widehat{F}_1 \to F_k^0$. 由于极限是唯一的，从而 $A_{k1}^0 = 0, k \neq 1$.

然后考虑第 2 列: 由于 $A_{11}^0 > 0$, 则希望证明 A_{22}^0 或 $A_{23}^0 > 0$。反证法，若 A_{22} 与 A_{23} 都收敛到 0，那么由于每一行一定收敛到正定阵，从而 A_{22} 与 A_{23} 都收敛到正定阵，那么 $\widehat{F}_3 \to F_2^0, \widehat{F}_3 \to F_3^0$. 由于极限是唯一的，矛盾，因此不妨设 $A_{22}^0 > 0$, 那么有 $\widehat{F}_2 \to F_2^0$, 然后有 $\widehat{F}_3 \to F_3^0$. 在一般的写法中，即有一个排列组合 $\{\sigma(j\}, j = 1, \ldots, S)$

4.3 A New Grouping Method: Group Lasso

The model we consider is the same as Ando and Bai (2016)

$$y_{it} = x_{it}^\intercal \beta + f_{g_i t}\lambda_{g_i i} + \epsilon_{it} \tag{4.3}$$

$$= x_{it}^\intercal \beta + \sum_{j=1}^{S} \mathbb{1}_{(g_i=j)} f_{jt}\lambda_{ji} + \epsilon_{it} \tag{4.4}$$

We propose the following objective function based on the work of Ando and Bai (2016)

$$\frac{1}{2NT}\sum_{i=1}^{N}\|y_i - X_i\beta - \sum_{j=1}^{S} F_j\lambda_{ji}\|^2 + \tau_1\|\beta\|_1 + \sum_{i=1}^{N}\tau_2^i\sum_{j=1}^{S}\sqrt{r_j}\left\|\lambda_{ji}\right\| \tag{4.5}$$

where we need to tuning τ_2^i s.t $\lambda_{ji} = 0$ when $j \neq \widehat{g}_i$.

Ando and Bai proposed the following estimate for g_i:

$$\widehat{g}_i = \arg\min_{j\in\{1,\dots,S\}}\left(y_i - X_i\widehat{\beta}\right)^\intercal M_{\widehat{F}_j}\left(y_i - X_i\widehat{\beta}\right) \tag{4.6}$$

where $X_i = (x_{i1},\dots,x_{iT})^\intercal$, $F_j = \left(f_{j1},\dots,f_{jT}\right)^\intercal$. Since

$$M_{\widehat{F}_j} = I_T - \widehat{F}_j\left(\widehat{F}_j^\intercal\widehat{F}_j\right)^{-1}\widehat{F}_j^\intercal \tag{4.7}$$

$$= I_T - \widehat{F}_j\widehat{F}_j^\intercal/T \tag{4.8}$$

we can rewrite the grouping estimate as

$$\begin{aligned}\widehat{g}_i &= \arg\min_{j\in\{1,\dots,S\}}\left(y_i - X_i\widehat{\beta}\right)^\intercal\left(I_T - \widehat{F}_j\widehat{F}_j^\intercal/T\right)\left(y_i - X_i\widehat{\beta}\right)\\ &= \arg\min_{j\in\{1,\dots,S\}} -\left(y_i - X_i\widehat{\beta}\right)^\intercal\widehat{F}_j\widehat{F}_j^\intercal\left(y_i - X_i\widehat{\beta}\right)\\ &= \arg\max_{j\in\{1,\dots,S\}}\left(y_i - X_i\widehat{\beta}\right)^\intercal\widehat{F}_j\widehat{F}_j^\intercal\left(y_i - X_i\widehat{\beta}\right)\end{aligned}$$

My estimate for g_i[14] is

$$\widetilde{g}_i = \arg\max_{j\in\{1,\dots,S\}}\frac{\left\|\widehat{F}_j^\intercal\left(y_i - X_i\widehat{\beta}\right)\right\|}{\sqrt{r_j}} \tag{4.9}$$

$$= \arg\max_{j\in\{1,\dots,S\}}\left(y_i - X_i\widehat{\beta}\right)^\intercal\widehat{F}_j\widehat{F}_j^\intercal\left(y_i - X_i\widehat{\beta}\right)/r_j \tag{4.10}$$

[14] 2021.7.28 version

Thus, it seems that $\widehat{g}_i$ is the same as $\widetilde{g}_i$ if we don't consider $\sqrt{r_j}$. 然而，factor loading 是和 β 同时估计出来的，而 group membership 也是根据 factor loading 才确定的. 因此，上述写法割裂了 $\widetilde{g}_i$ 与 $\widehat{\beta}$ 的联系. 由于它们是 joint estimator，更合理的做法是 KKT 方程.

We have to finish the following work:

1. 对Equation 4.5关于 β，λ_{ji} 求偏（次）导，尤其是对 λ_{ji} 求次导数，重新计算 $\widetilde{g}_i$. 因为我在Ando and Bai (2016) 的基础上加上了 group lasso penalty，联立关于 β 与 λ_{ji} 求偏次导的 KKT 条件，解这个方程组得到的解或许不是Equation 4.10 中的结果；

2. 设 $\widehat{G}_i$ 是通过 group lasso 得到的对 g_i 的分组的估计的集合. 区别于 $\widehat{g}_i$ 是属于 $\{1,\ldots,S\}$ 的一个元素，$\widehat{G}_i$ 是含于 $\{1,\ldots,S\}$ 的一个真子集. 可以利用 lasso 过估计的特点，也就是说估计出的非零的协变量包含了真实非零的协变量，这可以改进 Ando and Bai 的工作. 可以参考Chan et al. (2014), 去类似地证明

$$P\left(g_i^{\star} \in \widehat{G}_i\right) = 1 - \mathrm{o}(1) \tag{4.11}$$

3. 如果我们还能证明 $\exists M > 0$ s.t

$$P\left(|\widehat{G}_i| \leq M\right) = 1 - \mathrm{o}(1), \quad i = 1,\ldots,N$$

那么 $\widehat{G}_i$ 与 S 就将起到Ando and Bai (2016) 中 S 的作用. 先尝试在默认这个命题成立的情况下，怎么改写Ando and Bai (2016) 的证明；再去考虑，$\widehat{G}_i$ 是一个怎样的估计？我们希望在迭代的一开始 $\widehat{G}_i$ 尽量地大一点，随着迭代的增加希望 $\widehat{G}_i$ 会变小;

4. 给出一个新的算法

Then we start the work of the first item.

Consider to minimize the objective function Equation 4.5, 展开、合并同类项得

$$\frac{1}{2NT}\sum_{i=1}^{N}\left[y_i^{\intercal}y_i + \beta^{\intercal}X_i^{\intercal}X_i\beta + \left(\sum_{j=1}^{S}\lambda_{ji}^{\intercal}F_j^{\intercal}\right)\left(\sum_{j=1}^{S}F_j\lambda_{ji}\right) + 2\beta^{\intercal}X_i^{\intercal}\sum_{j=1}^{S}F_j\lambda_{ji} - 2y_i^{\intercal}\sum_{j=1}^{S}F_j\lambda_{ji} - 2\beta^{\intercal}X_i^{\intercal}y_i\right] + \lambda_1\|\beta\|_1 + \sum_{i=1}^{N}\tau_2^i\sum_{j=1}^{S}\sqrt{r_j}\left\|\lambda_{ji}\right\|$$

可以把上式重新改写为如下的式子，方便对 β 求导

$$\frac{1}{2NT}\sum_{i=1}^{N} y_i^\intercal y_i + \frac{1}{2}\beta^\intercal\left(\frac{1}{NT}X_i^\intercal X_i\right)\beta + \frac{1}{2}\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{S}\lambda_{ji}^\intercal F_j^\intercal\right)\left(\sum_{j=1}^{S}F_j\lambda_{ji}\right) + \beta^\intercal\frac{1}{NT}\sum_{i=1}^{N}X_i^\intercal\left(\sum_{j=1}^{S}F_j\lambda_{ji} - y_i\right) - \frac{1}{N}\sum_{i=1}^{N}y_i^\intercal\left(\sum_{j=1}^{S}F_j\lambda_{ji}\right) + \lambda_1\|\beta\|_1$$

或者也可以改写成如下式子，方便对 λ_{ji} 求导

$$\frac{1}{2NT}\sum_{i=1}^{N} y_i^\intercal y_i + \frac{1}{2}\beta^\intercal\left(\frac{1}{NT}X_i^\intercal X_i\right)\beta + \frac{1}{2}\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{S}\lambda_{ji}^\intercal F_j^\intercal\right)\left(\sum_{j=1}^{S}F_j\lambda_{ji}\right) + \frac{1}{NT}\sum_{i=1}^{N}\left(\sum_{j=1}^{S}F_j\lambda_{ji}\right)^\intercal(X_i\beta - y_i) - \frac{1}{NT}\sum_{i=1}^{N}y_i^\intercal X_i\beta + \sum_{i=1}^{N}\lambda_2^i\sum_{j=1}^{S}\sqrt{r_j}\|\lambda_{ji}\|$$

对 β 求偏导，利用上面的第一个式子，得到

$$\left(\frac{1}{N}\sum_{i=1}^{N}X_i^\intercal X_i\right)\widehat{\beta} + \frac{1}{NT}\sum_{i=1}^{N}X_i^\intercal\left(\sum_{j=1}^{S}F_j\lambda_{ji} - y_i\right) + \lambda_1\widehat{s} = 0$$

where $\widehat{s} = (\widehat{s}_1,\ldots,\widehat{s}_p)^\intercal$ and $\widehat{s}_j$ is the sub-gradient of $|\widehat{\beta}_j|$. we can write it as the componentwise form for the above equation

$$\frac{-1}{NT}\sum_{i=1}^{N}X_i^{(k)\intercal}\left(y_i - X_i\widehat{\beta} - \sum_{j=1}^{S}F_j\lambda_{ji}\right) + \lambda_1\widehat{s}_k = 0, \quad k = 1,\ldots,p$$

where $X_i = \left(X_i^{(1)},\ldots,X_i^{(p)}\right)$.Since

$$X_i\widehat{\beta} = \sum_{\ell=1}^{p}X_i^{(\ell)}\widehat{\beta}_\ell = \sum_{\ell\neq k}^{p}X_i^{(\ell)}\widehat{\beta}_\ell + X_i^{(k)}\widehat{\beta}_k$$

and

$$|\widehat{s}_k| \le 1 \quad \text{when } \widehat{\beta}_k = 0$$

$$\widehat{s}_k = \frac{\widehat{\beta}_k}{\left|\widehat{\beta}_k\right|} \quad \text{when } \widehat{\beta}_k \neq 0$$

we have

$$\frac{-1}{NT}\sum_{i=1}^{N}X_i^{(k)\intercal}\left(y_i - \sum_{\ell\neq k}X_i^{(\ell)}\widehat{\beta}_\ell - \sum_{j=1}^{S}F_j\lambda_{ji}\right) + \left(\frac{1}{NT}\sum_{i=1}^{N}X_i^{(k)\intercal}X_i^{(k)}\right)\widehat{\beta}_k + \lambda_1\widehat{s}_k = 0, \quad k = 1,\ldots,p$$

thus, for $k = 1, \ldots, p$,

$$\begin{aligned}\widehat{\beta}_k = 0 &\Leftrightarrow |\widehat{s}_k| \le 1 \\ &\Leftrightarrow \lambda_1 \ge |\lambda_1 \widehat{s}_k| = \frac{1}{NT}\left|\sum_{i=1}^{N} X_i^{(k)\intercal}\left(y_i - \sum_{\ell \ne k} X_i^{(\ell)}\widehat{\beta}_\ell - \sum_{j=1}^{S} F_j \lambda_{ji}\right)\right|\end{aligned}$$

and

$$\begin{aligned}\widehat{\beta}_k \ne 0 &\Leftrightarrow \widehat{s}_k = \widehat{\beta}_k / |\widehat{\beta}_k| \\ &\Leftrightarrow \left(\frac{1}{N}\sum_{i=1}^{N} X_i^{(k)\intercal} X_i^{(k)} + \frac{\lambda_1}{|\widehat{\beta}_k|}\right)\widehat{\beta}_k = \frac{1}{N}\sum_{i=1}^{N} X_i^{(k)\intercal}\left(y_i - \sum_{\ell \ne k} X_i^{(\ell)}\widehat{\beta}_\ell - \sum_{j=1}^{S} F_j \lambda_{ji}\right) \\ &\Leftrightarrow \widehat{\beta}_k = \left(\frac{1}{N}\sum_{i=1}^{N} X_i^{(k)\intercal} X_i^{(k)} + \frac{\lambda_1}{|\widehat{\beta}_k|}\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} X_i^{(k)\intercal}\left(y_i - \sum_{\ell \ne k} X_i^{(\ell)}\widehat{\beta}_\ell - \sum_{j=1}^{S} F_j \lambda_{ji}\right)\end{aligned}$$

然后对 λ_{ji} 求偏导. 为了不引起下标的混淆，对 λ_{ku} 求偏导，$k = 1, \ldots, S$, $u = 1, \ldots, N$. 由于只考虑第 u 个个体，因此等价于对下式求偏导

$$\frac{1}{2NT}\left(\sum_{j=1}^{S} \lambda_{ji}{}^{\intercal} F_j{}^{\intercal}\right)\left(\sum_{j=1}^{S} F_j \lambda_{ji}\right) + \frac{1}{NT}\left(\sum_{j=1}^{S} F_j \lambda_{ji}\right)^{\intercal}(X_u\beta - y_u) + \lambda_2^u \sum_{j=1}^{S}\sqrt{r_j}\left\|\lambda_{ju}\right\|$$

求导得

$$\frac{1}{NT}F_k{}^{\intercal}\left(\sum_{j=1}^{S} F_j \widehat{\lambda}_{ju}\right) + \frac{1}{NT}F_k{}^{\intercal}(X_u\beta - y_u) + \lambda_2^u\sqrt{r_k}\widehat{u}_{ku} = 0$$

利用 $F_k{}^{\intercal}F_k = T \cdot I_T$，合并同类项得

$$\frac{-1}{NT}F_k{}^{\intercal}\left(y_u - X_u\beta - \sum_{j \ne k} F_j \widehat{\lambda}_{ju}\right) + \frac{1}{N}\widehat{\lambda}_{ku} + \lambda_2^u\sqrt{r_k}\widehat{u}_{ku} = 0$$

thus, for $k = 1, \ldots, S$, $u = 1, \ldots, N$,

$$\begin{aligned}\widehat{\lambda}_{ku} = 0 &\Leftrightarrow \|\widehat{u}_{ku}\| \le 1 \\ &\Leftrightarrow \tau_2^u\sqrt{r_k} \ge \left\|\lambda_2^u\sqrt{r_k}\widehat{u}_{ku}\right\| = \frac{1}{NT}\left\|F_k{}^{\intercal}\left(y_u - X_u\beta - \sum_{j \ne k} F_j \widehat{\lambda}_{ju}\right)\right\|\end{aligned}$$

and

$$\begin{aligned}
\widehat{\lambda}_{ku} \neq 0 \Leftrightarrow \widehat{u}_{ku} &= \widehat{\lambda}_{ku} / \left\| \widehat{\lambda}_{ku} \right\| \\
&\Leftrightarrow \left(\frac{1}{N} + \frac{\lambda_2^u \sqrt{r_k}}{\left\| \widehat{\lambda}_{ku} \right\|} \right) \widehat{\lambda}_{ku} = \frac{1}{NT} F_k^\intercal \left(y_u - X_u \beta - \sum_{j \neq k} F_j \widehat{\lambda}_{ju} \right) \\
&\Leftrightarrow \widehat{\lambda}_{ku} = \left(\frac{1}{N} + \frac{\lambda_2^u \sqrt{r_k}}{\left\| \widehat{\lambda}_{ku} \right\|} \right)^{-1} \frac{1}{NT} F_k^\intercal \left(y_u - X_u \beta - \sum_{j \neq k} F_j \widehat{\lambda}_{ju} \right)
\end{aligned}$$

然后我们需要给出 item 4 的工作. 我还不知道 $\widehat{G}_i$ 是什么样的，但是 $\widehat{\beta}$ 与 $\widehat{g}_i$ 可以是这样的

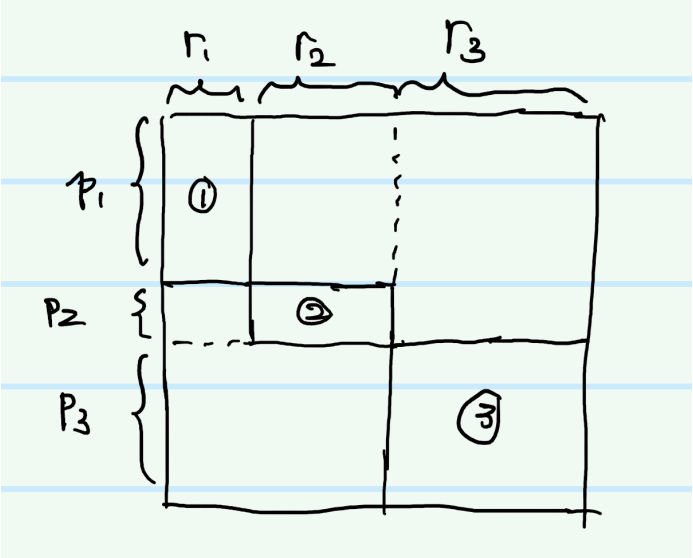
$$\frac{-1}{NT} \sum_{i=1}^{N} X_i^{(k)\intercal} \left(y_i - \sum_{\ell \neq k} X_i^{(\ell)} \widehat{\beta}_\ell - \sum_{j \in \widehat{G}_i} F_j \lambda_{ji} \right) + \left(\frac{1}{NT} \sum_{i=1}^{N} X_i^{(k)\intercal} X_i^{(k)} \right) \widehat{\beta}_k + \lambda_1 \widehat{s}_k = 0, \quad k = 1, \ldots, p$$

$$\widehat{g}_i = \arg\max_{j \in \widehat{G}_i} \| F_j^\intercal (y_i - X_i \widehat{\beta} - \sum_{\substack{\ell \in \widehat{G}_i \\ \ell \neq j}} F_\ell \lambda_{\ell i}) \| / \sqrt{r_j}$$

4.4 Sparsity of loading matrix

根据张老师 8 月 2 日的建议，需要注意到载荷矩阵 Λ 有两个层面的稀疏性，如Figure 4.1所示，$p = \sum p_j$ 项资产可以分划成 3 个组，每个组分别有 $r_j, j = 1,2,3$ 个因子；尽管 Λ 是一个 $p \times r$ 的矩阵，但是只有 $\sum p_j \cdot r_j$ 个非零元素. 回到我们的

Figure 4.1: Sparse structure of loading matrix



模型，由于每项资产只属于 S 组中的某一组，因此中只有与真实分组相应的载荷非零；每个分组中只含有一部分资产，因此大部分资产在某一个组因子上的载荷卫零. 鉴于此，我们采用如下的估计方法

$$\frac{1}{2NT}\sum_{i=1}^{N}\|y_i - X_i\beta - \sum_{j=1}^{S} F_j\lambda_{ji}\|^2 + \tau_1\|\beta\|_1 + \sum_{i=1}^{N}\tau_2^i\sum_{j=1}^{S}\tau_3^j\|\lambda_{ji}\| \tag{4.12}$$

其中，τ_2^i 可以控制 Λ 中每一列的稀疏性，τ_3^j 可以控制 Λ 中每一行的稀疏性. 若遵照Yuan and Lin (2006) 的做法，可以取 $\tau_3^j = \sqrt{r_j}$ ，则这时需要先对每个组因子进行定阶.

2021-8-25, 可能按照下面的方式去说明我们的分类方法（grouping/clustering method）是一个适用于 the true number of groups(clusters) $K^\star$ 依赖于资产数 N 且 $K^\star \to \infty$ 的情形：

1. choose a pre-specified (and sufficiently large) number of groups K, and maybe we still need other conditions on K, then we need to show that β, (estimated, not true) factor structure F_k, $k = 1,\ldots,K$ and corresponding loadings λ_{ki} are consistent under prediction error, that is

$$\frac{1}{NT}\sum_{i=1}^{N}\left\|X_i(\beta^\star - \widehat{\beta}) + F^0_{g_i^\star}\lambda^0_{g_i^\star i} - \sum_{k=1}^{K}F_k\lambda_{ki}\right\|^2 \le \text{const}\cdot\text{tuning parameters}$$

2. and further, we can show that $\frac{1}{NT}\sum_{i=1}^{N}\left\|X_i(\beta^\star - \widehat{\beta})\right\|^2$ and $\frac{1}{NT}\sum_{i=1}^{N}\left\|F^0_{g_i^\star}\lambda^0_{g_i^\star i} - \sum_{k=1}^{K}F_k\lambda_{ki}\right\|^2$ are bounded from above, and their upper bound will converge to 0.

3. If the above consistency-like theorem holds, we need to show that the grouping with sufficiently large K will is very good in the sense that every group (amoung K groups) is very pure.

4. If every group is very pure, we need to show that when we merge groups (among K gourps) with LAR-like method, the number of groups/clusters will converges to the true number of groups, and the group factor is consistent under average norm.

5. Lastly, we will show that grouping is consistent.

Note. 当 K 相比于真实的组数 $K^\star$ 相当地大时，因子是不会有相合性的. 只有我们通过合并，使得同一类的样本糅合在一起共同去恢复 group specific factors，才有因子结构的相合性

5 The Lasso

5.1 The motivation for lasso

In linear regression setting, we are given N samples $\{(x_i, y_i)\}_{i=1}^N$ where x_i is a p-dimensional vector of predictors and each $y_i \in \mathbb{R}$ is the associated continuous response. Construct a linear regression model,

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \tag{5.1}$$

$$= \beta_0 + \boldsymbol{x}_i^\intercal \beta + \epsilon_i \,. \tag{5.2}$$

The usual LSE for the regression weights β and an intercept termβ_0 is based on minimizing squared-error loss (or called RSS, residual squared sum),

$$(\widehat{\beta}_0, \widehat{\beta}) = \arg\min_{\beta_0,\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - \boldsymbol{x}_i^\intercal \beta)^2 \tag{5.3}$$

There are two reasons why we might consider an alternative to the LSE

1. Prediction accuracy (as measured in terms of the MSE[16]). LSE is an unbiased estimator, 而它的方差与解释变量的个数成正比，thus it often has low bias but large variance[17]. 因为 MSE 由偏差与方差组成，we hope we can decrease the MSE by introducing some bias but reducing the variance of the predicted values. Lasso will arrive this goal by shrinking the values of the regression weights, that is setting some components of th exactly zero.
   Remark in this way, lasso provides an automatic way for doing model/variable selection in linear regression[18]

2. Interpretaton improvement: Lasso will construct a sparse model using a smaller subset of predictors that exhibit the strongest effects (more large weights). LSE will construct a model using all of the predictors, 并且许多变量的系数较小.

There are two forms of optimization objectives. Let $y = (y_1, \ldots . y_N)^\intercal$, $X = (x_1, \ldots, x_N)^\intercal$.

- Constraint form

$$(\widehat{\beta}_0, \widehat{\beta}) = \arg\min_{\beta_0,\beta} \frac{1}{2N} \|y - \mathbf{1}_N\beta_0 - X\beta\|^2 \tag{5.4}$$

$$\text{subject to} \|\beta\|_1 \le t$$

- Penalty form (Lagrangian form)

$$(\widehat{\beta}_0, \widehat{\beta}) = \arg\min_{\beta_0, \beta} \frac{1}{2N} \|y - \mathbf{1}_N \beta_0 - X\beta\|^2 + \lambda \|\beta\|_1 \tag{5.5}$$

可以去证明：there is a one-to-one correspondence between the constrained form and lagrangian form. 利用 lagrangian duality 粗粗地来说就是：从 constraint form $\|\beta\|_1 \leq t$ 算出一个 $\widehat{\beta}$，代入 KKT 条件 $\frac{-1}{N}\left\langle x_j, y - X\widehat{\beta}\right\rangle + \lambda s_j$ 能算出 λ. 而从 penalty form 可以算出 $\widehat{\beta}$，令 $t = \left\|\widehat{\beta}\right\|_1$ 即可以得到 constraint form.

5.1.1 Bias-variance Trade-off Perspective

References

- Tutorial on Lasso-Honglang Wang: The subsection 1.1
- Linear Methods for Regression-Lijun Zhang: 我觉得他的 expected prediction error 定义不太合适，应该用Hastie et al. (2019)

5.2 Centering and scaling

数据分析的一开始往往要对响应变量 y 和协变量 $\boldsymbol{x}$ 做预处理. 常见的预处理包括

1. 对响应变量 y 与协变量 $\boldsymbol{X}$ 进行中心化，即
2. 对协变量 $\boldsymbol{X}$ 进行标准化
3. 对协变量 $\boldsymbol{X}$ 进行正交化

以下我们来讨论各种预处理对模型估计的影响.

对数据进行中心化意味着对响应变量和协变量都进行了中心化. 相比于原始数据，利用中心化的数据建立的线性模型可以不考虑截距项，或者说 $\widehat{\beta} = 0$. 在利用中心化数据估计出线性模型后，we can discover the optimal solutions for the uncentered data. 这一点也可以从一般的线性回归中推出来.

设 $\{x_i, y_i\}_{i=1}^N$ 是未进行中心化的原始数据，而 $\{\widetilde{x}_i, \widetilde{y}_i\}_{i=1}^N$ 是中心化数据，即满足

$$\widetilde{y}_i = y_i - \frac{1}{N}\sum_{i=1}^N y_i =: y_i - \overline{y}$$

$$\widetilde{x}_i = x_i - \frac{1}{N}\sum_{i=1}^N x_i =: x_i - \overline{x}$$

其中 $\overline{x} = (\overline{x}_1, \ldots, \overline{x}_p)^\intercal$. 对于这两种数据，分别建立如下两种线性模型

$$y_i = \beta_0 + x_i^\intercal \beta + \epsilon_i$$
$$\widetilde{y}_i = \widetilde{x}_i^\intercal \beta + \epsilon_i$$

然后，在相同的调节系数 λ 下，通过 lasso 分别得到两个线性模型的估计 $\left(\widehat{\beta}_0, \widehat{\beta}\right)$ 与 $\widetilde{\beta}$，即

$$y_i = \widehat{\beta}_0 + x_i^\intercal \widehat{\beta} + \epsilon_i \tag{5.6}$$
$$\widetilde{y}_i = \widetilde{x}_i^\intercal \widetilde{\beta} + \epsilon_i \tag{5.7}$$

那么我们可以证明如下结论

Proposition 5.1.

$$\widehat{\beta}_0 = \overline{y} - \sum_{j=1}^{p} \overline{x}_j \widehat{\beta}_j \tag{5.8}$$
$$\widehat{\beta} = \widetilde{\beta} \tag{5.9}$$

Proof: For uncentered data, consider penalty formEquation 5.5. For easy intuition, we rewrite it as

$$(\widehat{\beta}_0, \widehat{\beta}) = \arg\min_{\beta_0, \beta} \frac{1}{2N} (y - \mathbf{1}_N \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \tag{5.10}$$

分别对 β_0, β 求导，得到

$$\frac{\partial}{\partial \beta_0} = \frac{-1}{N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right) = 0$$
$$\frac{\partial}{\partial \beta_j} = \frac{-1}{N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right) x_{ij} + \lambda s_j = 0, \quad j = 1, \ldots, p$$

其中，s_j 是关于 β_j 的次微分，s_j 的取值依赖于 β_j. $(\widehat{\beta}_0, \widehat{\beta})$ 就是满足上述方程组的解，即有

$$\frac{\partial}{\partial \widehat{\beta}_0} = \frac{-1}{N} \sum_{i=1}^{N} \left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} x_{ij} \widehat{\beta}_j \right) = 0 \tag{5.11}$$
$$\frac{\partial}{\partial \widehat{\beta}_j} = \frac{-1}{N} \sum_{i=1}^{N} \left(y_i - \widehat{\beta}_0 - \sum_{j=1}^{p} x_{ij} \widehat{\beta}_j \right) x_{ij} + \lambda \widehat{s}_j = 0, \quad j = 1, \ldots, p \tag{5.12}$$

To get the estimate for β_0, consider the first equation, 化简得

$$\begin{aligned}\widehat{\beta}_0 &= \frac{1}{N}\sum y_i - \frac{1}{N}\sum_{i,j} x_{ij}\widehat{\beta}_j \\ &= \bar{y} - \sum_j \bar{x}_j\widehat{\beta}_j\end{aligned}$$

which implies that if we get the estimate $\widehat{\beta}_j$, $j = 1,\ldots,p$ for regression coefficient β, then we can derive the estimate $\widehat{\beta}_0$. And after replacing the $\widehat{\beta}_0$ with $\widehat{\beta}$ in the second equation, the estimate $\widehat{\beta}$ comes from the iterations.

For centered data, the objective function is

$$\widetilde{\beta} = \arg\min_{\beta}\frac{1}{2N}\sum_{i=1}^{N}[\widetilde{y}_i - \sum_{j=1}^{p}\widetilde{x}_{ij}\beta_j]^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

关于 $\widetilde{\beta}$ 求导后可以得到如下方程

$$\frac{\partial}{\partial\widetilde{\beta}_j} = \frac{-1}{N}\sum_{i=1}^{N}\left(\widetilde{y}_i - \sum_{j=1}^{p}\widetilde{x}_{ij}\widetilde{\beta}_j\right)\widetilde{x}_{ij} + \lambda\widetilde{s}_j = 0, \quad j = 1,\ldots,p$$

由于 $\frac{-1}{N}\sum_{i=1}^{N}\left(\widetilde{y}_i - \sum_{j=1}^{p}\widetilde{x}_{ij}\beta_j\right)\overline{x}_j = 0, j = 1,\ldots,p$ ，因此用 $\{x_i, y_i\}_{i=1}^{N}$ 代替 $\{\widetilde{x}_i, \widetilde{y}_i\}$ 可得

$$\frac{\partial}{\partial\beta_j} = \frac{-1}{N}\sum_{i=1}^{N}\left(y_i - \overline{y} - \sum_{j=1}^{p}(x_{ij} - \overline{x}_j)\beta_j\right)x_{ij} + \lambda s_j = 0, \quad j = 1,\ldots,p$$

由于上述方程与Equation 5.12相同，因此命题成立 □

标准化一定在中心化之后，它的目标是使得转化后的数据 $x_{ij}^*, i = 1,\ldots,N$ 满足 $\frac{1}{N}\sum_{i=1}^{N} {x_{ij}^*}^2 = 1$，不妨设 raw data (y, X) 已经是中心化数据，那么只需要考虑 β 的估计在标准化和为标准化下有无差异. 由于 β 是有约束的，因此从 Lagrangian form 去看，对5.5求微分（次微分）后的 KKT 条件就是解的充要条件，即，

$$-\frac{1}{N}\langle x_j, y - X\beta\rangle + \lambda s_j = 0, \tag{5.13}$$

但是这不太容易携程 scaled case，因此将它改写成 sum 的形式

$$\frac{1}{N}\sum_{i=1}^{N} x_{ij}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j) \tag{5.14}$$

而 scaled case 为

$$\frac{1}{N}\sum_{i=1}^{N}\frac{x_{ij}}{\sigma_j}(y_i-\sum_{j=1}^{p}\frac{x_{ij}}{\sigma_j}\beta_j) \tag{5.15}$$

where $\sigma_j^2=\frac{1}{N}\sum_{i=1}^{N}x_{ij}^2$

接下来讨论 RSS 中的常数因子 $\frac{1}{2N},\frac{1}{2}$ 或 1 时对解的影响. 显然，他们对于 constraint form 是没有影响的，而对于 penalty form，所得的 $\widehat{\beta}$ 没有影响，但是改变了调节系数 λ 的尺度，其中，采用 $\frac{1}{2N}$ 具有更良好的含义，并且也方便 RSS 中平方项的求导.$\frac{1}{N}$ 可以看成是调节系数关于 sample size 的标准化，这对后面的 cross validation 有帮助，并且可以解释绝对值相关系数.

至此，完成了对 lasso 问题的几个基本细节的讨论.

5.3 Computation of the Lasso Solution

5.3.1 KKT Condition

然后我们给出 lasso 的求解，它的理论求解要利用次梯度，它的数值求解可以从 single predictor 推广到 multiple predictors，体现了 cyclic coordinate descent（当然，也可以从次梯度下的 KKT 条件直接写出来）.

通过对 penalty form 求微分（次微分），由凸分析知，penalty form 的解的充要条件是

$$-\frac{1}{N}\langle x_j,y-X\beta\rangle+\lambda s_j=0,\quad j=1,\dots,p, \tag{5.16}$$

其中 s_j 是一个依赖于 β_j 的未知数：若 $\beta_j\neq 0$，则 $s_j=\beta_j/|\beta_j|$; 若 $\beta_j=0$, 则 $s_j\in[-1,1]$. 对这个式子稍一改写得，

$$\lambda s_j=\frac{1}{N}\langle x_j,y-X\beta\rangle. \tag{5.17}$$

那么 $\beta_j=0$ 当且仅当 $\left|\frac{1}{N}\langle x_j,y-X\beta\rangle\right|\leq\lambda$, 因此可以将 λ 看成衡量 predictor 与模型残差相关性的绝对值大小的阈值（门限值）.

下面我们来介绍 computation of the lasso solution.

5.3.2 Coordinatewise Gradient Decent Algorithm

5.3.3 Least Angle Regression

5.4 Consistency Results for the Lasso

这一部分我参考了Hastie et al. (2019)

5.4.1 Various ℓ_2-norm

在讨论 lasso estimator 的相合性之前，我们先交代依托这些相合性结果的标准定义和符号. 我们基于最小化残差平方和,即采用 MSE 损失函数估计线性模型.Thus, based on ℓ_2-norm we can assess the quality of $\widehat{\beta}$ in 3 ways:

- ℓ_2-error, that is

$$\|\widehat{\beta} - \beta^\star\|^2 \tag{5.18}$$

  We also call it parameter estimation loss, lasso error, lasso ℓ_2-error

- Prediction loss function, that is

$$\frac{1}{N}\left\|X\widehat{\beta} - X\beta^\star\right\|^2 \tag{5.19}$$

  we also call is prediction error. This is the MSE of $\widehat{\beta}$ over the given samples.

- Support recovery loss function, that is

$$\sum_{i=1}^{p} \mathbb{1}_{(sign(\widehat{\beta}_i)\neq sign(\beta_i^\star))} \tag{5.20}$$

  This loss function is designed for variable selection consistency and signed recovery consistency.

5.4.2 Bounds on Lasso ℓ_2-error

Consider the estimation of the following k-sparse linear regression

$$y = X\beta^\star + \epsilon,$$

where $\beta^\star$ is k-sparse, supported on the subset S.

Theorem 5.1. *Assume that the true linear model is k-sparse. If the design matrix satisfies the restricted eigenvalue condition, and we estimate the model through constrained lasso, then the estimate of regression slope is bounded by nosie. We can summarize the conditions and results as*
Conditions:

1. $|S| = k$
2. $\gamma-RE$ *conditions on* $\mathcal{C}(S, 1)$

3. $\|\beta\|_1 \leq R = \left\|\beta^\star\right\|_1$

Results:

$$\left\|\widehat{\beta} - \beta^\star\right\| \leq \frac{4}{\gamma}\sqrt{\frac{k}{N}}\left\|\frac{X^\intercal \epsilon}{\sqrt{N}}\right\|_\infty \tag{5.21}$$

Note. This is Theorem 11.1 (a) on Hastie et al. (2019).

Proof: Let $\widehat{\nu} = \widehat{\beta} - \beta^\star$ and the true DGP is $y = X\beta^\star + \epsilon$. Since $\widehat{\beta}$ is the minimizer of the objective function, then

$$\left\|y - X\widehat{\beta}\right\|^2 \leq \left\|y - X\beta^\star\right\|^2.$$

Replace the y by the noise ϵ by the true DGP, we arrive that

$$\|X\widehat{\nu}\|^2 \leq 2\epsilon^\intercal X\widehat{\nu}$$

I have question about this inequality[19]. Rewrite it as

$$\frac{\|X\widehat{\nu}\|^2}{2N} \leq \frac{\left\langle \widehat{\nu}, X^\intercal \epsilon \right\rangle}{N}$$

By the Holder inequality at $\ell_\infty-$norm and ℓ_1-norm, we get

$$\frac{\|X\widehat{\nu}\|_2^2}{2N} \leq \frac{\left\|X^\intercal \epsilon\right\|_\infty \|\widehat{\nu}\|_1}{N}$$

Then the hard sparsity plays a role. Since condition 3, we have

$$\widehat{\nu} \in \mathcal{C}(S, 1)$$

then

$$\begin{aligned}\|\widehat{\nu}\|_1 &= \|\widehat{\nu}_S\|_1 + \|\widehat{\nu}_{S^c}\|_1 \leq 2\|\widehat{\nu}_S\|_1 \\ &= 2\left\|\mathbb{1}_{(i\in S)}\widehat{\nu}\right\|_1 \leq 2\sqrt{k}\|\widehat{\nu}\|\end{aligned}$$

then

$$\frac{\gamma}{2} \leq \frac{\|X\widehat{\nu}\|^2/N}{2\|\widehat{\nu}\|^2} \leq \frac{\left\|X^\intercal \epsilon\right\|_\infty \cdot 2\sqrt{k}}{N\|\widehat{\nu}\|}$$

then we arrive that

$$\|\widehat{\nu}\| \leq \frac{4}{\gamma}\sqrt{\frac{k}{N}}\left\|\frac{X^\intercal \epsilon}{N}\right\|_\infty$$

□

Note. 张老师指出，在 time series 中，当模型有结构变化时，RE conditon 可能会不成立. Refer to Chan et al. (2014)

Note. 关于两种惩罚的等价性，参考这两篇文章Candes, Tao, et al. (2007) 和Bickel et al. (2009)

Theorem 5.2. *If we estimate the model through lagrangian lasso and the other keep the same, then the estimate of regression slope is bounded by the tunning parameter. We can summarize the detailes as*
Conditions:

1. $|S| = k$
2. $\lambda_N \geq 2\left\|X^\intercal \epsilon\right\|_\infty / N > 0$
3. $\gamma-RE$ *conditon on* $\mathcal{C}(S,3)$

Results:

$$\left\|\widehat{\beta} - \beta^\star\right\| \leq \frac{3}{\gamma}\sqrt{k}\lambda_N \tag{5.22}$$

Note. This is Theorem 11.1 (b) on Hastie et al. (2019)

Proof: Let $G(\nu) \coloneqq \frac{1}{2N}\|y - X(\beta^* + \nu)\|^2 + \lambda_N\|\beta^* + \nu\|_1$, then since the estimator is the minimizer of the objective function,

$$G(\widehat{\nu}) \leq G(0)$$

我们希望得到用噪声 ϵ 表达的式子，可以像之前那样用 true DGP 代入式子，或者我们采用加一项减一项：

$$\frac{1}{2N}\left\|y - X\beta^* + X\beta^* - X\widehat{\beta}\right\|^2 + \lambda_N\|\widehat{\beta}\|_1 \leq \frac{1}{2N}\|y - X\beta^*\|^2 + \lambda_N\|\beta^*\|_1$$

两边消掉 $\frac{1}{2N}\|y - X\beta^*\|^2$ 和 $\frac{1}{2N}\|\epsilon\|^2$，并由 $\epsilon = y - X\beta^*$ 以及 $\left\|\beta^\star\right\|_1 = \left\|\beta^\star_S\right\|$, 得到

$$\begin{aligned}\frac{\|X\widehat{\nu}\|^2}{2N} &\leq \frac{\epsilon^\intercal X\widehat{\nu}}{N} + \lambda_N(\|\beta^*\|_1 - \|\widehat{\beta}\|_1) \\ &= \frac{\langle\widehat{\nu}, X^\intercal\epsilon\rangle}{N} + \lambda_N(\|\beta^\star_S\|_1 - \|\beta^\star + \widehat{\nu}\|_1)\end{aligned} \tag{5.23}$$

this is the *Modified Basic Inequality*, which is very important. 利用 k-sparse 和三角不等式，得到

$$\begin{aligned}\|\beta^* + \widehat{\nu}\|_1 &= \left\|\beta^*_S + \widehat{\nu}_S\right\|_1 + \|\widehat{\nu}_{S^c}\| \\ &\geq \left\|\beta^*_S\right\|_1 - \|\widehat{\nu}_S\|_1 + \|\widehat{\nu}_{S^c}\|_1\end{aligned}$$

where the second step follows by $\|\beta^*\|_1 > 0$ and $\|\widehat{\nu}_S\|_1 = 0$ is possible. Substituting these relations into modified basic inequality yields

$$\begin{aligned}\frac{\|X\widehat{\nu}\|^2}{2N} &\leq \frac{\left\langle \widehat{\nu}, X^\intercal \epsilon \right\rangle}{N} + \lambda_N(\|\widehat{\nu}_S\|_1 - \|\widehat{\nu}_{S^c}\|_1) \\ &\leq \frac{\left\|X^\intercal \epsilon\right\|_\infty}{N}\|\widehat{\nu}\|_1 + \lambda_N(\|\widehat{\nu}_S\|_1 - \|\widehat{\nu}_{S^c}\|_1)\end{aligned}$$

where the second step follows by applying Holder's inequality with ℓ_1 and ℓ_∞ norms.Since condition 1, we arrive that

$$\begin{aligned}\frac{\|X\widehat{\nu}\|^2}{2N} &\leq \frac{\lambda_N}{2}(\|\widehat{\nu}_S\|_1 + \|\widehat{\nu}_{S^c}\|_1) + \lambda_N(\|\widehat{\nu}_S\|_1 - \|\widehat{\nu}_{S^c}\|_1) \\ &\leq \frac{3}{2}\sqrt{k}\lambda_N\|\widehat{\nu}\| \,. \end{aligned} \tag{5.24}$$

This is another basic inequality. Notice that we haven't use the $\gamma-$RE condition. Before use this condition we have to prove that $\widehat{\nu} \in \mathcal{C}(S,3)$. This is the Lemma 11.1 in Hastie et al. (2019) in pg 298. We can show this by the first step in the "another basic inequality". So $\gamma-$RE condition can be applied to $\widehat{\nu}$, then we get

$$\gamma\|\widehat{\nu}\|^2 \leq \frac{1}{N}\|X\widehat{\nu}\|^2$$

The remaining is obvious. □

Note. 1. 可以考虑用 quantile，LAD 作为新的 loss function，其中 quantile 可以降低对噪声的矩的存在性的要求.

2. 参考 Runze Li, 史成春的工作

Example 5.1. Gaussian Linear Regression
Conditions:

1. $\left\|x_j\right\|_2/\sqrt{N} \leq 1, j = 1, \ldots, p$
2. $\epsilon \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right)$

Results:

1. 当 $\frac{\log p}{N} \to 0$ 时，$\widehat{\beta}$ 是相合估计，速度是 $\sqrt{\frac{\log p}{N}}$
2. $\lambda_N = c\sigma\sqrt{\frac{\log p}{N}}$, where c is a constant

我们想去证，$P\left(\left\|\widehat{\beta} - \beta^\star\right\|_2 \le \epsilon\right) \ge 1 - \delta$, 其中 ϵ 中包含样本量的信息 n 与 p, 样本量反映了收敛速率. 若找到好的上界，即 $\left\|\widehat{\beta} - \beta^\star\right\|_2$ ≤upperbound, 那么自然有 $P\left(\left\|\widehat{\beta} - \beta^\star\right\|_2 \le \epsilon\right) \ge P(\text{upperbound} \le \epsilon)$. Thm. 5.2在 $\lambda_N \ge 2\left\|X^\intercal\epsilon\right\|_\infty/N$ 的条件下（以及 γ−RE condition）给出了 upper bound. 直观上我们希望这个“条件”能以较大的概率发生. 这时需要去计算 tail probability.

令 $\lambda_N = 2t, x_j{}^\intercal\epsilon/N \sim \mathcal{N}\left(0, \left\|x_j\right\|_2^2\sigma^2/N^2\right)$. 由 Gaussian tail probability[15]得：

$$P\left(\left|x_j{}^\intercal\epsilon/N\right| > t\right) \le 2\exp\left(\frac{-t^2N^2}{2\left\|x_j\right\|_2^2\sigma^2}\right) \le 2\exp\left(\frac{-t^2N}{2\sigma^2}\right), j = 1, \ldots, p \tag{5.25}$$

and then

$$P\left(\left\|X^\intercal\epsilon\right\|_\infty/N > t\right) = P\left(\cup\left\{\left|x_j{}^\intercal\epsilon/N\right| > t\right\}\right) \tag{5.26}$$

$$\le \sum_{j=1}^{p} 2\exp\left(\frac{-t^2N^2}{2\left\|x_j\right\|_2^2\sigma^2}\right) \le 2p\exp\left(\frac{-t^2N}{2\sigma^2}\right) \tag{5.27}$$

其实，LHS = $P\left(2\left\|X^\intercal\epsilon\right\|_\infty/N > \lambda_N\right)$, 因此我们说明了 $\left\{\lambda_N \ge 2\left\|X^\intercal\epsilon\right\|_\infty/N\right\}$ 是一个大概率事件. 接下来我们通过条件概率，将事件 $\left\{\lambda_N \ge 2\left\|X^\intercal\epsilon\right\|_\infty/N\right\}$, $\left\|\widehat{\beta} - \beta^\star\right\|_2$ 的上界以及 tuning parameter$\lambda_N = 2t$ 结合起来，能反映出 $\widehat{\beta}$ 的收敛速率:

$$P\left(\left\|\widehat{\beta} - \beta^\star\right\|_2 \le \frac{3}{\gamma}\sqrt{k}2t\right) \ge P\left(\left\|\widehat{\beta} - \beta^\star\right\|_2 \le \frac{3}{\gamma}\sqrt{k}2t \middle| \left\|X^\intercal\epsilon\right\|_\infty \le t\right) P\left(\left\|X^\intercal\epsilon\right\|_\infty \le t\right) \tag{5.28}$$

$$= P\left(\left\|\widehat{\beta} - \beta^\star\right\|_2 \le \frac{3}{\gamma}\sqrt{k}\lambda_N \middle| \left\|X^\intercal\epsilon\right\|_\infty \le t\right) P\left(\left\|X^\intercal\epsilon\right\|_\infty \le t\right) \tag{5.29}$$

$$= 1 \cdot P\left(\left\|X^\intercal\epsilon\right\|_\infty \le t\right) \tag{5.30}$$

$$= 1 - 2\exp\left(\frac{-t^2N}{2\sigma^2} + \log p\right) \tag{5.31}$$

[15]对于 sub-gaussian noise，利用 subG 的线性和仍然是 subG，并且可以被 larger variance proxy 控制，可以类似地证明

其中 2exp 项扮演着 δ 的角色，令 $2\exp = \delta$,

$$\frac{-t^2 N}{2\sigma^2} + \log p = \log \frac{\delta}{2} \tag{5.32}$$

$$t = \sqrt{\frac{2\sigma^2(\log p - \log \delta/2)}{N}} \tag{5.33}$$

$$= \sigma\sqrt{\frac{2(\log p - \log \delta/2)}{N}} \approx c\sigma\sqrt{\frac{\log p}{N}} \tag{5.34}$$

where c is a constant, and the approximation holds when p and N go to infinity. 尽管 2exp 包含了样本量 p,N 的信息，但这可以归结到 λ_N 中，从而 2exp 只是体现 δ 的作用 (若 p 也同时与 N 一起趋于 ∞，则不容易写成随机有界的形式，但从主项，也就是约等于部分，的角度看，$\widehat{\beta}$ 的收敛速度是 $\sqrt{\frac{\log p}{N}}$).

5.4.3 Bounded on Prediction Error

Theorem 5.3. *If the true regression slope is bounded, then the lagrangian lasso with specific tuning parameter produces consistent estimate under prediction error. They can be summarized as*
Conditions:

1. $\frac{\lambda_N}{2} \geq \left\|X^\intercal \epsilon\right\|_\infty / N$
2. $\exists R_1 \ s.t \ \|\beta^\star\|_1 \leq R_1$

Results:

1. $$\frac{\|X\widehat{\nu}\|^2}{N} \leq 6R_1\lambda_N$$

Note. This is the theorem 11.2 (a) on Hastie et al. (2019).

Proof: From modified basic inequalityEquation 5.23, we have

$$0 \leq \frac{\left\|X^\intercal \epsilon\right\|_\infty}{N}\|\widehat{\nu}\|_1 + \lambda_N(\|\beta^\star\|_1 - \|\widehat{\beta}\|_1) \tag{5.35}$$

在之前我们将 $\|\beta^\star\|_1 - \|\widehat{\nu}\|_1$ 作为 $\|\widehat{\beta}\|_1$ 的下界，然而这里，我们要去求 $\|\widehat{\nu}\|_1$ 的上界，因此 $\|\widehat{\nu}\|_1$ 可以比较大，表示估计不好. 因此这时以 $-\|\beta^\star\|_1 + \|\widehat{\nu}\|_1$ 作为下界更合适，

那么

$$\begin{aligned} 0 &\leq \frac{\left\|X^{\intercal}\epsilon\right\|_{\infty}}{N}\|\widehat{\nu}\|_1 + \lambda_N(\left\|\beta^{\star}\right\|_1 + \left\|\beta^{\star}\right\|_1 - \|\widehat{\nu}\|_1) \\ &= \left(\frac{\left\|X^{\intercal}\epsilon\right\|_{\infty}}{N} - \lambda_N\right)\|\widehat{\nu}\|_1 + 2\lambda_N\|\beta^{\star}\|_1 \end{aligned}$$

由于 condition 1，得

$$\begin{aligned} 0 &\leq (\frac{1}{2}\lambda_N - \lambda_N)\|\widehat{\nu}\|_1 + 2\lambda_N\|\beta^{\star}\|_1 \\ &= \frac{1}{2}(4\|\beta^{\star}\|_1 - \|\widehat{\nu}\|_1) \end{aligned}$$

因此, $\|\widehat{\nu}\|_1 \leq 4\|\beta^{\star}\|_1 \leq 4R_1$

Returning again to the modified basic inequality Equation 5.23 and ignoring the negative term $-\lambda_N\|\widehat{\beta}\|_1$, we have

$$\begin{aligned} \frac{\|X\widehat{\nu}\|^2}{2N} &\leq \frac{\left\|X^{\intercal}\epsilon\right\|_{\infty}}{N}\|\widehat{\nu}\|_1 + \lambda_N\|\beta^{\star}\| \\ &= \frac{\lambda_N}{2}\cdot(4R_1 + 2R_1) = 3\lambda_N R_1\,, \end{aligned}$$

then we arrive the result 1. □

5.4.4 Variable Selection Consistency

Note. Lasso 往往是多估的，即真实的变量是选入变量的真子集，但是符号的估计是相合的. 为了处理多估，可以参考 adaptive Lasso.

5.4.5 Remained Issues

1. Hastie et al. (2019) 的第 11 章主要在 iid 正态白噪声的线性模型中讨论了回归系数的相合性，其中关键的一步是通过 iid 正态白噪声的 Bernstein 不等式求出尾概率的上界. 张老师指出，在更一般的情形中也有相应的 Bernstein-type 不等式，比如鞅的 Bernstein 不等式，Mixing 的 Bernstein 不等式. 这方面的工作可以参考

2. 高维领域的相合性问题相对比较容易，在高维的假设检验中往往会把高维的统计量转化为低维的形式，然后再进行研究. 目前高危险性模型方面已经有大量的、较完善的工作，但是非线性模型中还有许多问题有待解决. 这方面的工作可以参考 Peter Bickel 和 Aad W. van der Vaart 的工作（张老师原话是 Perter Bickel 和经验过程若收敛的一个作者的工作）.

3. 目前，许多高维线性模型的损失函数都是 MSE，如果换成 MLE 又可以怎么处理？再比如换成 Quantile 又该如何处理 heavy tailed 数据中的高维问题.

4. 可以尝试将 Lasso 用到 Functional data, spatial data 中. 这时数据中复杂的相依性结构可能无法通过一层的 lasso 捕捉出来，而是多要多层的惩罚函数.

5.5 Standard rate $\sqrt{\log p/n}$

我参考了这一篇High-dimensional regression: why is $\log p/n$ special? the answer by mweylandt，作者试图直观地解释 $\log p$ 和 n 的原因以及它们的作用，并且给出了 simulations.

我们发现，$\frac{\log p}{n}$ 经常出现在 rate of convergence for high-dimensional regression estimator，比如说

$$\frac{1}{n}\left\|X\widehat{\beta}-X\beta\right\|_2^2=\mathcal{O}_p(1)(\sigma\sqrt{\frac{\log p}{n}}\|\beta\|_1) \tag{5.36}$$

. 为了让误差趋向于 0，通常要求 $\log p$ 小于 n，即 $p<e^n$，或者说增长速率较小. 我们希望解释这两个问题：1. $\frac{\log p}{n}$ 的直觉；2. 的问题往往无法解决，这个原因是什么？

$\sqrt{\log p}$ 来自于 gaussian/subgaussian random variables 的 concentration of measure. 设有 p 个 IID gaussian r.v., 它们的 maximum is on the order of with high probability. 这个定理的形式我还没看到，但是否与下面这个 uniform bounds on expted value 有联系：

$$\mathbf{E}\left([]\max_{1\le i\le n}X_i\le\sigma\sqrt{2\log n}\right]. \tag{5.37}$$

$1/n$ 来自于所考虑的是 average prediction error $\left\|X(\widehat{\beta}-\beta)\right\|_2^2\Big/n$.

在直观上看，penalty term 控制了数据中两方面的信息：

1. goodness from more data ($n\to\infty$)

2. badness from the extra irrelevant features (想要得到一个稀疏模型)

由于 classical statistics 考虑的是 fixed p，let $n\to\infty$ 的情形 (例如 traditional linear regression)，但现实中常常有高维的问题，因此必须要考虑 large p problem. 但是 p 与 n 的增长速率是需要有限制的. 直观上可以理解 infinite features and finite data 是没有办法估计的，因此我们往往考虑 p 与 n 共同趋于无穷的情形. 这时候处

于方便[16]，令 $p = f(n)$，其中 $f(\cdot)$ 是某一确定的函数，让 p 在 $n \to \infty$ 时间接地趋于无穷. 由于 $f(\cdot)$ 的形式决定了 p 与 n 趋于无穷的速率，从而影响了估计的好坏. 一般而言，badness from the extra features only grows as $\log p$ while the goodness from the extra data grows as n. 具体指的是：

1. if $\frac{\log p}{n}$ stays constant, error stays fixed asymptotically, this is called "boarder-line" ultra-high-dimensional;

2. if $\frac{\log p}{n} \to 0$, we asymptotically achieve zero error, this is called "tractable" high-dimensional;

3. if $\frac{\log p}{n} \to \infty$, the error eventually goes to infinity. 这往往被称为 ultra-high-dimensional (UHD)

在实际中我们也发现，UHD 并非完全没有办法去处理，但这时要用到的技巧比用 a simple max of gaussian r.v. to control the error 要更加复杂. 在 UHD 领域，Jianqing Fan and Jinchi Lv 做了很多出色的工作.

[16]math generally lacks languages and tools for discussing limits with two "degrees of freedom"，从而我们希望通过某个形式联结 p 和 n. 在真实世界中，没有任何证据支持 p 与 n 之间有联系，但是若让 obj fun 中有 2 个 df，即 p 与 n，则会难以处理

6 Generalizations of the Lasso Penalty

In the following chapters, we will discuss the various generalizations of the Lasso method. In section 7 we will discuss the Elastic net method, and in section 8 we will talk about the group Lasso.

7 The Elastic Net

8 The Group Lasso

考虑如下线性模型：

$$\mathbf{E}\left(y \,\middle|\, Z_j, j = 1, \ldots, J\right) = \sum_{j=1}^{J} Z_j{}^{\intercal}\theta_j\,,$$

其中 $y \in \mathbb{R}$ 是响应变量，$Z_j \in \mathbb{R}^{p_j}$ 是 jth group of covariates, θ_j 是相应的回归系数

Given a collection of N samples $\left\{y_i, Z_{ij}\right\}_{i\geq 1}^{N}$. A linear model for the regression function takes the form

$$y_i = \sum_{j=1}^{J} Z_{ij}{}^{\intercal}\theta_j + \epsilon_i\,,$$

and we can rewrite it as the matrix form

$$\begin{aligned} y &= \sum_{j=1}^{J} Z_j\theta_j + E \\ &= Z\,\mathrm{vec}\,(\Theta) + E \end{aligned}$$

where $Z = (Z_1, \ldots, Z_J)$ is a $N \times p$ matrix, $Z_j = \left(Z_{1j}, \ldots, Z_{Nj}\right)^{\intercal}$ is a $N \times p_j$ matrix and $p = \sum_{j=1}^{J} p_j$. The group lasso solves the convex problem

$$\begin{aligned} \widehat{\theta}_j &= \arg\min \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{J} z_{ij}{}^{\intercal}\theta_j\right)^2 + \lambda\sum_{j=1}^{J}\left\|\theta_j\right\| \\ &= \arg\min \frac{1}{2N}\|y - \sum_{j=1}^{J} Z_j\theta_j\|^2 + \lambda\sum_{j=1}^{J}\left\|\theta_j\right\| \end{aligned}$$

Expand it as

$$\frac{1}{2N}\left[y^{\intercal}y - 2y^{\intercal}\sum_{j=1}^{J} Z_j\theta_j\right] + \left(\sum_{j=1}^{J}\theta_j{}^{\intercal}Z_j{}^{\intercal}\right)\left(\sum_{j=1}^{J} Z_j\theta_j\right)\lambda\sum_{j=1}^{J}\left\|\theta_j\right\|$$

为了避免引起下标的混淆，对上式关于 θ_k 求导，得：

$$\frac{-1}{N}Z_k{}^{\intercal}y + \frac{1}{N}Z_k{}^{\intercal}\left(\sum_{j=1}^{J} Z_j\widehat{\theta}_j\right) + \lambda\widehat{s}_k = 0$$

单独将 $\widehat{\theta}_k$ 提出来，合并同类项，整理得

$$\frac{-1}{N}Z_k{}^{\intercal}\left(y - \sum_{j\neq k} Z_j\widehat{\theta}_j\right) + \frac{1}{N}Z_k{}^{\intercal}Z_k\widehat{\theta}_k + \lambda\widehat{s}_k = 0$$

因此我们得到: 对于 $\widehat{\theta}_k = 0$ 的情形，有

$$\begin{aligned}\widehat{\theta}_k = 0 &\Leftrightarrow \|\widehat{s}_k\| \leq 1 \\ &\Leftrightarrow \lambda \geq \|\lambda \widehat{s}_k\| = \frac{1}{N}\left\| Z_k^{\intercal}\left(y - \sum_{j\neq k} Z_j \widehat{\theta}_j\right)\right\|\end{aligned}$$

对于 $\widehat{\theta}_k \neq 0$ 的情形，有

$$\begin{aligned}\widehat{\theta}_k \neq 0 &\Leftrightarrow \widehat{s}_k = \widehat{\theta}_k / \|\widehat{\theta}_k\| \\ &\Leftrightarrow N\lambda \leq \left\| Z_k^{\intercal}\left(y - \sum_{j\neq k} Z_j \widehat{\theta}_j\right)\right\| \\ &\Leftrightarrow \left(\frac{1}{N} Z_k^{\intercal} Z_k + \frac{\lambda}{\left\|\widehat{\theta}_k\right\|}\right)\widehat{\theta}_k = \frac{1}{N} Z_k^{\intercal}\left(y - \sum_{j\neq k} Z_j \widehat{\theta}_j\right)\end{aligned}$$

其中，$Z_k^{\intercal} Z_k$ 是一个 $p_k \times p_k$ 的矩阵，$\frac{\lambda}{\left\|\widehat{\theta}_k\right\|}$ 是一个数. 为了使 $\widehat{\theta}_k$ 的表达式不依赖于 $\widehat{\theta}_k$ 本身，需要先计算 $\left\|\widehat{\theta}_k\right\|$. 从而对上式两边求 ℓ_2-norm. 而由于 $Z_k^{\intercal} Z_k$ 是一个 $p_k \times p_k$ 的矩阵，$\frac{\lambda}{\left\|\widehat{\theta}_k\right\|}$ 是一个数，从而只有当 $Z_k^{\intercal} Z_k = I_{p_k}$ 时，即每一组协变量均为 orthonormal 时，可以求出 $\widehat{\theta}_k$ 的表达式. 先假设 $Z_j^{\intercal} Z_j = I_{p_j}$, $j = 1, \ldots, J$，那么上式两边取 ℓ_2-norm 有

$$\left(\frac{1}{N} + \frac{\lambda}{\|\widehat{\theta}_k\|}\right)\left\|\widehat{\theta}_k\right\| = \frac{1}{N}\left\| Z_k^{\intercal}\left(y - \sum_{j\neq k} Z_j \widehat{\theta}_j\right)\right\|$$

因此得到

$$\left\|\widehat{\theta}_k\right\| = \left\| Z_k^{\intercal}\left(y - \sum_{j\neq k} Z_j \widehat{\theta}_j\right)\right\| - N\lambda$$

故有

$$\begin{aligned}\widehat{\theta}_k &= \left(\frac{1}{N} + \frac{\lambda}{\left\| Z_k^{\intercal}\left(y - \sum_{j\neq k} Z_j \widehat{\theta}_j\right)\right\| - N\lambda}\right)^{-1} \frac{1}{N} Z_k^{\intercal}\left(y - \sum_{j\neq k} Z_j \widehat{\theta}_j\right) \\ &= \left(\frac{\left\| Z_k^{\intercal}\left(y - \sum_{j\neq k} Z_j \widehat{\theta}_j\right)\right\| - N\lambda}{\left\| Z_k^{\intercal}\left(y - \sum_{j\neq k} Z_j \widehat{\theta}_j\right)\right\|}\right) Z_k^{\intercal}\left(y - \sum_{j\neq k} Z_j \widehat{\theta}_j\right) \\ &= \left(1 - \frac{N\lambda}{\left\| Z_k^{\intercal}\left(y - \sum_{j\neq k} Z_j \widehat{\theta}_j\right)\right\|}\right) Z_k^{\intercal}\left(y - \sum_{j\neq k} Z_j \widehat{\theta}_j\right)\end{aligned}$$

将 $\widehat{\theta}_k = 0$ 与 $\widehat{\theta}_k \neq 0$ 两种情形合并，得到

$$\widehat{\theta}_k = \left(1 - \frac{N\lambda}{\left\|Z_k^{\mathsf{T}}\left(y - \sum_{j\neq k} Z_j\widehat{\theta}_j\right)\right\|}\right)_+ Z_k^{\mathsf{T}}\left(y - \sum_{j\neq k} Z_j\widehat{\theta}_j\right)$$

Yuan and Lin (2006) 中的 proposition 1 也给出了同样的结果. 在那篇文章中考虑了 group size p_j, $j = 1, \ldots, J$.

I mainly refer to Hastie et al. (2019). Hear are related posts on stack:

- Derivation of Group Lasso sebp offered a detailed solution for the group lasso estimate in the case of orthonormal group covariates.
- Derivation of the Group SCAD Solution

8.1 Group LAR selection

This was proposed by Yuan and Lin (2006)

9 Oracle inequality

为了先理解 oracle inequality 的动机，我在 stack 上看了(https://stats.stackexchange.com/users/203 gogolashvili) (n.d.)，并且摘要如下.

在一般的 linear regression 中（即 $p < n$ 或者 $n \gg p$），prediction error $\left\|X(\widehat{b}-\beta^\star)\right\|_2^2$ 在经过噪声方差的尺度化后服从 $\chi^2(p)$，即 $\left\|X(\widehat{b}-\beta^\star)\right\|_2^2/\sigma^2 \sim \chi^2(p)$. 那么可以计算得：

$$\frac{\mathbf{E}\left(\left\|X(\widehat{b}-\beta^\star)\right\|^2\right)}{n} = p\frac{\sigma^2}{n}. \tag{9.1}$$

这有良好的意义：每个协变量都“贡献”了在平均意义上高达 $\frac{\sigma^2}{n}$ 的预测精度.

进一步，考虑高维情形 ($p \gg n$)，这时不仅仅是 LSE 不存在，而且我们先验地知道只有一小部分协变量对 response 有影响（这是后续 $\mathcal{M}(\beta) \leq s$ 的现实依据）. 那么倘若我们完全清楚这 large p 个协变量中哪 small k 个起作用，则最后的预测精度在平均意义上就是 $k\frac{\sigma^2}{n}$. 正是由于我们并不清楚在 p 个 covariates 中究竟是哪 k 个变量起作用，因此会有众多的 regulatization penalty methods （这还依赖于 model selection，也就是选一个合适的调节系数）. 与之前一样，我们要计算 regularized estimator 的预测精度. Oracle inequality 就是要解决这件事：它将给出大概率下 (with probability close to 1) (平均的) 预测误差的上界，同时也还是预测精度的上界. 那样就可以说明某一种压缩惩罚的有效性与优劣.

10 Bickel et al. (2009)

10.1 Abstract and Intro.

学者们先回顾了 lasso 与 Dantzig 这两方面的工作. 在 lasso 方面：

1. sparsity oracle inequalities for the prediction loss;
2. minimax estimation aggregation of estimators;
3. lasso for density estimation;
4. modified versions of lasso (nonquadratic term?) in random design.

而在 Dantzig 方面：

1. Dantzig 对于估计线性模型具有 optimal ℓ_2 rate properties;
2. sparsity oracle inequalities in random design.

我认为我需要去学习和复习：model average, density estimation, random design.

本文要做的是这两方面的工作：

1. 在稀疏的设定下（即控制 nonzero coefficients 的个数），lasso 与 Dantzig 在 prediction loss 的意义下是几乎等价的；
2. 两种方法的 oracle inequalities.

其中学者们的主要贡献是：给出了更一般的假设 (RE)，并且将 ℓ_2 下的结果推广到 ℓ_p loss, $1 \le p \le 2$.

10.2 Definitions and notations

设 $(Z_i^\intercal, Y_i), i = 1, \dots, n$ 为一组 independent random pairs，它们来自于：

$$Y_i = f(Z_i) + W_i, \quad i = 1, \dots, n, \tag{10.1}$$

其中，$f : \mathcal{Z} \to \mathbb{R}$ 是未知待估的回归函数，$\mathcal{Z}$ 是 $\mathbb{R}^d$ 的 Borel 子集，且 Z_i 是 $\mathcal{Z}$ 中的 fixed elements（这里我并不很清楚 fixed design 与 random design 之间的区别）. 在为这个数据生成过程建立估计方法（模型）之前，我们要先定义 a finite dictionary of functions[17] $\mathcal{F}_M = \{f_1, \dots, f_M\}$，其中 $f_j : \mathcal{Z} \to \mathbb{R}$.

对上面这个数据生成过程和估计模型，要做以下几点额外的补充：

[17]要么是 fixed functions，要么是一些 data-driven methods

1. 在本文中，当得到 $Z_1, \cdots, Z_n$ 这些协变量的观测后，可以从选定的 $\mathcal{F}_M$ 直接算出 $f_j(Z_i), i, j$;

2. 高维是由 M，即字典中的函数个数，而非 d，即可观测协变量的个数，所决定的；

3. $\mathcal{F}_M$ 在统计问题上有多个方向的意义：

   - 比如在非参数回归中，$f_j, j = 1, \ldots, M$ 是 a collection of basis functions；
   - 再比如，在 aggregation problem（有点像模型选择）中，f_j 可以是来自于 M 种方法/模型对 f 的估计；或者说，f_j 是在用一种方法/模型（比如 lasso）下选取 M 种 tuning parameters 时对 f 的估计.

4. 客观的说，$\mathcal{F}_M$ 的选取对于估计 f 是很有影响的. 在论文中学者们假设是可以通过 the span of $\mathcal{F}_M$ 被很好地近似的（作者并没有提出一个量去度量这个假设的偏离程度，在后面的证明中得到某一项为 $\|f_\beta - f\|_2^2$，那么如果真实的回归函数确为线性的，则这一项总是很小的）.

再接下来，学者们交代一些记号，主要是 $n \times M$ 的设计矩阵 X:

$$X = (f_j(Z_i))_{n \times M} = \begin{bmatrix} f_1(Z_1) & \ldots & f_M(Z_1) \\ \vdots & & \vdots \\ f_1(Z_n) & \ldots & f_M(Z_n) \end{bmatrix}_{n \times M}, \tag{10.2}$$

f 是 $n \times 1$ 的向量，表示没有被噪声污染的未知回归函数的真实值：

$$f = \begin{bmatrix} f(Z_1) \\ \vdots \\ f(Z_n) \end{bmatrix}_{n \times 1} \tag{10.3}$$

对 $\forall \beta \in \mathbb{R}^M$，本文用 $f_\beta(Z) = \sum_{j=1}^M \beta_j f_j(Z)$ 去近似 f，因此我们需要从 lasso/Dantzig 方法中估计出 β. 本文就是来探究这两种方法估计得到的 β 的效果.

还有一个让 M 维向量 δ 带有以指标集 $J = \{1, \ldots, M\}$ 为脚标的记号 $\delta_J . \delta_J$ 依然是 M 维向量，但在 J^c 分量上均为 0，而在 J 上与 δ 中分量相同.

10.3 Restricted eigenvalue assumptions

定义 $n \times M$ 的 design matrix X 的 $M \times M$ 的 Gram matrix:

$$\Psi_n = \frac{1}{n} X^\intercal X, \tag{10.4}$$

其中，这里定义的 gram matrix 在其他文献中有可能也被称为 scatter matrix，因此主要得看是怎么计算的. 在有了这一个记号后，下面提出的限制性特征值假设是得到 lasso 与 Dantzig selectors 的统计性质的关键假设. 其基本动机是：通过限制可行域使得 Gram matrix 再次具有正定性（文中称为 restricted positive definiteness）

当 $M \gg n$ 时极可能出现 Ψ_n 非正定，用 matrix ℓ_2-norm 来表示，也就是：

$$\min_{\delta\in\mathbb{R}^M:\delta\neq 0} \frac{(\delta^\intercal \Psi_n \delta)^{1/2}}{|\delta|_2} = \min_{\delta\in\mathbb{R}^M:\delta\neq 0} \frac{|X\delta|_2}{\sqrt{n}|\delta|_2} = 0, \tag{10.5}$$

其中，$\frac{(\delta^\intercal \Psi_n \delta)^{1/2}}{|\delta|_2}$ 在不受约束的 δ 下的取值范围是 $\left[\sqrt{\lambda_1}, \sqrt{\lambda_M}\right]$. 那么退化就意味着存在着零特征根. 而 OLS 的有效性需要 $X^\intercal X$ 可逆，即上式需要严格大于 0. 在全部可行域，即 $\delta \in \mathbb{R}^M : \delta \neq 0$ 上均成立，这一点太强了.lasso 与 Dantzig 所要求的假定会弱很多，并且分母 $|\delta|_2$ 也改为 only a part of δ.

设 $\delta_L = \widehat{\beta}_L - \beta, \delta_D = \widehat{\beta}_D - \beta$，而 δ 为 δ_L 与 δ_D 共用的记号.δ 表示 lasso/Dantzig 方法估计后计算出的 residual. 以接近概率 1（with probability close to 1）成立下面这个不等式：

$$\left|\delta_{J_0{}^c}\right|_1 \leq c_0 \left|\delta_{J_0}\right|_1, \tag{10.6}$$

其中 $J_0 = J(\beta)$ 为 the index set[18] of nonzero coefficients of the true parameter β. 这个不等式表示稀疏部分的残差的距离被非零部分控制. “接近于概率 1” 可以用一个 large number M 表示，c_0 依赖于所采取的模型：Dantzig 时为 1，lasso 时为 3.

进一步考虑，若 Ψ 为十分接近于 Ψ_n 的一个 $M \times M$ 的正定阵，即 $\varepsilon_n \overset{def}{=} \max_{ij}\left|(\Psi_n - \Psi)_{i,j}\right|$ 十分小. 尝试用 Ψ 去表示的下界：

$$\begin{aligned}
\frac{\delta^\intercal \Psi_n \delta}{|\delta|_2^2} &= \frac{\delta^\intercal \Psi \delta}{|\delta|_2^2} - \frac{\delta^\intercal (\Psi_n - \Psi)\delta}{|\delta|_2^2} && (10.7)\\
&\geq \frac{\delta^\intercal \Psi \delta}{|\delta|_2^2} - \frac{\varepsilon_n |\delta|_1^2}{|\delta|_2^2} && (10.8)\\
&\geq \frac{\delta^\intercal \Psi \delta}{|\delta|_2^2} - \varepsilon_n \left(\frac{(1+c_0)\left|\delta_{J_0}\right|_1}{\left|\delta_{J_0}\right|_2}\right)^2 && (10.9)\\
&\geq \frac{\delta^\intercal \Psi \delta}{|\delta|_2^2} - \varepsilon_n (1+c_0)^2 |J_0| . && (10.10)
\end{aligned}$$

注意，这个不等式的证明用到这几个东西：

1. 将 ℓ_1-norm 进行如下改写：

$$\begin{aligned}
|\delta|_1 &= \left|\delta_{J_0{}^c}\right|_1 + \left|\delta_{J_0}\right|_1 && (10.11)\\
&\leq (1+c_0)\left|\delta_{J_0}\right|_1, && (10.12)
\end{aligned}$$

[18]指标集

2. 还需要证明 $\left|\delta_{J_0}\right|_1^2 \le |J_0|\left|\delta_{J_0}\right|_2^2$，这在 2 维的时候是显然的：$(|x_1|+|x_2|)^2 \le 2(x_1^2+x_2^2)$. 我本来考虑用数学归纳法，但是其实可以直接证明：

$$|x_i|\left|x_j\right| \le \frac{1}{2}(x_i^2+x_j^2), \tag{10.13}$$

$$\sum_{i=1}^n\sum_{j=1}^n \le \sum_{i=1}^n\sum_{j=1}^n \frac{1}{2}(x_i^2+x_j^2) = \sum_{i=1}^n x_i^2 . \tag{10.14}$$

10.4 Appendix B

先给出第一个重要的 lasso 情形下地 lemma B.1: 取 tuning parameter 为 [19]. 那么以接近于 1 的概率成立下面的 3 个不等式：

$$\left\|\widehat{f}_L - f\right\|_n^2 + r\sum_{j=1}^M \|f_j\|_n\left|\widehat{\beta}_{j,L}-\beta_j\right| \tag{10.15}$$

$$\le \|f_\beta - f\|_n^2 + 4r\sum_{j\in J(\beta)} \|f_j\|_n\left|\widehat{\beta}_{j,L}-\beta_j\right| \tag{10.16}$$

$$\le \|f_\beta - f\|_n^2 + 4r\sqrt{|J(\beta)|}\sqrt{\sum_{j\in J_0}\|f_j\|_n^2\left|\widehat{\beta}_{j,L}-\beta_j\right|^2} . \tag{10.17}$$

式 (10.16) 是关键的，而后面的一步式 (10.17) 不过是利用 cauchy 不等式的改写. 对这个不等式的理解：$\left\|\widehat{f}_L - f\right\|_n^2$ 刻画的是 lasso prediction loss，它的上界可以被 $\|f_\beta - f\|_n^2$ 与 nonzero coefficients 的 ℓ_1-vector-norm 给控制住，其中 $\|f_\beta - f\|_n^2$ 表示我们假设的（最优的）线性模型与真实的回归函数的差距，这一误差是来自于模型假设的优劣，而非模型估计的好坏.$4r\sum_{j\in J_0}\|f_j\|_n\left|\widehat{\beta}_{j,L}-\beta_j\right|$ 这部分才是依赖于模型估计.

$$\left|\frac{1}{n}X^\intercal(\mathbf{f} - X\widehat{\beta}_L)\right|_\infty \le 3r f_{\max}/2 \tag{10.18}$$

第二个不等式刻画了所有观测中最大的绝对值相关系数的上界.

$$\left|J(\widehat{\beta}_L)\right| \le 4\phi_{\max} f_{\min}^{-2}(\left\|\widehat{f}_L - f\right\|_n^2/r^2) . \tag{10.19}$$

第三个不等式刻画了 lasso coefficients $\widehat{\beta}_L$ 的稀疏性.

[19]我还不清楚这有什么深意，或者说这一形式是来自于哪里的直觉？但是 $\frac{\log M}{n}$ 确实是有意义的

10.4.1 Proof

下面给出经过我补充之后的证明细节.

利用 cauchy 不等式进行改写，因为

$$\sum_{i=1}^{n}|x_i| \le \sqrt{n(x_1^2+\cdots+x_n^2)} \tag{10.20}$$

$$(\sum_{i=1}^{n}|x_i|\cdot 1)^2 \le (\sum 1^2)(\sum x_i^2) \tag{10.21}$$

那么就有

$$\sum_{j\in J(\beta)} \|f_j\|_n \left|\widehat{\beta}_{j,L}-\beta_j\right| \le \sqrt{|J(\beta)|}\sqrt{\sum_{j\in J_0}\|f_j\|_n^2\left|\widehat{\beta}_{j,L}-\beta_j\right|^2}. \tag{10.22}$$

$$\widehat{\beta}_L = \arg\min_{\beta}\{\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-f_\beta(Z_i)\right]^2 + 2r\sum_{j=1}^{M}\|f_j\|_n|\beta_j|\}, \tag{10.23}$$

其中 $\|f_j\|_n$ 有点类似于标准差. 令 $\widehat{S}(\beta)$ 为平均的 RSS，即 $\widehat{S}=\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-f_\beta(Z_i)\right]^2$，那么根据 lasso solution 的定义，就有：

$$\widehat{S}(\widehat{\beta}_L)+2\sum_{j=1}^{M}r\|f_j\|_n\left|\widehat{\beta}_{j,L}\right| \le \widehat{S}(\beta)+2\sum_{j=1}^{M}r\|f_j\|_n|\beta_j|, \tag{10.24}$$

相比残差中的 $Y_i-f_\beta(Z_i)$，我们更关心没有噪声的估计函数与真实的未知回归函数的差距，利用 $Y_i=f(Z_i)+W_i$，改写 $\widehat{S}(\beta)$:

$$\widehat{S}(\widehat{\beta}_L) = \frac{1}{n}\sum_{i=1}^{n}\left[f(Z_i)-\widehat{f}_L(Z_i)+W_i\right]^2 \tag{10.25}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\left[f(Z_i)-\widehat{f}_L(Z_i)\right]^2+2\left[f(Z_i)-\widehat{f}_L(Z_i)\right]W_i+W_i^2\right) \tag{10.26}$$

$$= \left\|f-\widehat{f}_L\right\|_n^2+\frac{2}{n}\sum_{i=1}^{n}\left[f(Z_i)-\widehat{f}_L(Z_i)\right]W_i+\sum_{i=1}^{n}W_i^2 \tag{10.27}$$

同理

$$\widehat{S}(\beta) = \|f-f_\beta\|_n^2+\frac{2}{n}\sum_{i=1}^{n}\left[f(Z_i)-f_\beta(Z_i)\right]W_i+\sum_{i=1}^{n}W_i^2 \tag{10.28}$$

整理并约去 $\sum W_i^2$ 后可得：

$$\begin{aligned}&\left\|\widehat{f}_L - f\right\|_n^2 + 2\sum_{j=1}^{M} r\|f_j\|_n\left|\widehat{\beta}_j\right| \\ &\qquad \le \|f - f_\beta\|_n^2 + 2\sum_{j=1}^{M} r\|f_j\|_n|\beta_j| + \frac{2}{n}\sum_{i=1}^{n} W_i(\widehat{f}_L - f_\beta)(Z_i)\end{aligned} \tag{10.29}$$

然后，定义随机变量 $V_j = \frac{1}{n}\sum_{i=1}^{n} f_j(Z_i)W_i, 1 \le j \le M$，以及事件 $\mathcal{A}$:

$$\mathcal{A} = \bigcap_{j=1}^{M}\left\{2|V_j| \le r\|f_j\|_n\right\}, \tag{10.30}$$

从而

$$\mathcal{A}^c = \bigcup_{j=1}^{M}\left\{2|V_j| > r\|f_j\|_n\right\} \tag{10.31}$$

注意 V_j 的形式其实和式 (10.29) 的 RHS 的最后一项有联系，因为

$$(\widehat{f}_L - f_\beta)(Z_i) = \sum_{j=1}^{M}(\widehat{\beta}_{j,L} - \beta_j)f_j(Z_i) \tag{10.32}$$

而事件 $\mathcal{A}$ 所指向的是之后会算出的一个不等式的成立，其含义是控制了绝对值相关系数的大小.

下面利用 Gaussian distribution 的尾概率不等式去写出 $\mathcal{A}^c$ 的概率的上界.

$$P(()\,\mathcal{A}^c) \le \sum_{j=1}^{M} P(()\,2|V_j| > r\|f_j\|_n) \tag{10.33}$$

为了方便起见，先计算的期望和方差，再将它标准化:

$$\mathbf{E}(V)_j = 0 \tag{10.34}$$

$$\mathrm{Var}(()\,V_j) = \frac{1}{n^2}\sum_{i=1}^{n} f_j^2(Z_i)\sigma^2 = \frac{1}{n}\|f_j\|_n^2\sigma^2 \tag{10.35}$$

$$\frac{\sqrt{n}V_j}{\|f_j\|_n\sigma} \sim \mathcal{N}(0,1) \tag{10.36}$$

将该标准正态随机变量记为 η. 在直接利用 gaussian 分布的尾概率不等式之前，还应确保能使得 $\sqrt{\frac{2}{\pi}}\frac{1}{t} \le 1, t = \frac{r\sqrt{n}}{2\sigma}$，这就等价于 $\sqrt{\pi \log 2} \ge 1$，显然是没问题的.

11 Sparseness of lasso solution

Sextus Empiricus在What is the smallest λ that gives a 0 component in lasso?探讨了 lasso solution 中调节系数 λ 与 $\widehat{\beta}$ 的稀疏性的具体关系：what's the value of λ at which any component of $\widehat{\beta}$ is initially zero? 当 λ 充分小时，$\widehat{\beta}$ 的所有分量均非零. 那么这里感兴趣的是求出：

$$\lambda_{\min} = \min_{\substack{\exists j, \widehat{\beta}_j = 0 \\ \forall i \neq j, \widehat{\beta}_i \neq 0}} \lambda, \tag{11.1}$$

这里的 min 可以理解为符合条件的调节系数的集合的下确界. 同时我也想知道能否确定这时 j 的取值.

在提问中提问者也阐述了 $\lambda_{\min}$ 不可能有基于 y, x 的 closed form solution. 否则，在 LARS 算法的 R 包中应该会利用这一点，（结合当 $\lambda \geq \frac{1}{n}\left\|X^{\intercal} y\right\|_{\infty}$ 时，$\widehat{\beta} = 0$）直接给出画图中调节系数的范围.

一般在考虑 lasso 回归的优化目标时都写成 lagrange 形式，而写成对偶形式则可以更好地解释其几何意义：

$$\min \frac{1}{2n}\|y - X\beta\|_2^2, \quad \text{subject to} \|\beta\|_1 \leq t, \tag{11.2}$$

从几何的角度可以审视不同稀疏程度的 lasso solution $\widehat{\beta}$：$\|y - X\beta\|_2^2$ 可以理解为以 n 维向量 y 为球心，$X\beta$ 为球面上一点的 n 维空间中的球体的半径平方，而 $\|\beta\|_1$ 则是从 design matrix X 的 p 个列向量所张成的 p 维空间映射到 y 所在的 n 维标准正交空间的多面体.n 维球体很好理解，而 p 维的多面体有着较为隐蔽的含义：令 $\|\beta\|_1 = const.$，那么以 $\beta_1, \ldots, \beta_p$ 为坐标轴则容易画出 p 维正方体. Answer 的作图是以 $y_1, \ldots, y_n$ 为轴的 n 维空间，那么这里存在一个 p 维到 n 维的线性映射，即 $X : \beta \mapsto X\beta$，这样就得到了 n 维空间中的 p 维多面体. 这时需要将 X 看成由 p 个列向量 $\ell_1, \ldots, \ell_p$ 组成[20]，那么：

$$X\beta = \ell_1\beta_1 + \cdots + \ell_p\beta_p, \tag{11.3}$$

而 $\ell_1, \ldots, \ell_p$ 即为图中的黑色向量，$\beta_1, \ldots, \beta_p$ 则可以看成是这些坐标轴上的坐标. 这里还可以理解 lasso solution $\widehat{\beta}$ 不是 OLS solution 的某个倍数，a multiple of the OLS solution. 因为：整个实心多面体是解的可行域，在多面体上任取一点与球心相连，当两者相切时即为优化目标取得最小[21]. 而相切的位置可能在多面体的顶点、脊[22]和面上.

[20]回答中的记号是 $x_1, \ldots, x_p$

[21]但是这还没有揭示 lasso solution 的非唯一性

[22]ridge: a raised line on the surface of sth

这里有必要提一提 OLS solution 在图像上的位置：若考虑将多面体放大到整个 n 维空间，这时从优化目标中得到的 OLS solution 即落在一个较小的球体的球面上，当多面体随着 t 减小而缩小时，即可行域缩小，则 OLS solution 必将落在可行域之外.

Answer 中讨论了 case 1 与 case 3，其中 case 1 讨论在 λ 取何值时能在多面体的顶点上取到优化目标的最小值，即 $\widehat{\beta}$ 中只有一个分量非零，最后一个 reserved variable 即为 the associated vector ℓ_i has the highest absolute value of the covariance with y，这也对应于 $\left\|X^{\intercal}y\right\|_{\infty}$.

12 Clustering

12.1 Hierarchical Clustering

12.2 K-Means Clustering

According to the decomposition of sum of squares, we have

$$\sum_{i=1}^{N} \|X_i - \mu\|^2 = \sum_{k=1}^{K} \sum_{i \in C_k} \|X_i - \mu_k\|^2 + \sum_{k=1}^{K} N_k \|\mu_k - \mu\|^2$$

where $\mu = \frac{1}{N} \sum_{i=1}^{N} X_i$, $N_k = \sum_{i=1}^{N} \mathbb{1}_{(i \in C_k)}$. In James et al. (2013) and R function kmeans(), the total sum of squares (totalSS) is

$$\sum_{i=1}^{N} \|X_i - \mu\|^2$$

and the within-cluster sum of squares(withinSS) is

$$\sum_{i \in C_k} \|X_i - \mu_k\|^2$$

13 Sparse Clustering

13.1 Sparse Hierarchical Clustering

13.2 Sparse K-Means Clustering

13.3 Convex clustering

14 Appendix

14.1 Cauchy-Schwartz inequality

从一元的 Cauchy-Schwartz 不等式说起，

$$\begin{aligned}\left|\sum_{i=1}^{n} x_i y_i\right| = |\langle x, y\rangle| = \left|\boldsymbol{x}^\intercal \boldsymbol{y}\right| &\leq \left(\sum x_i^2\right)^{1/2}\left(\sum y_i^2\right)^{1/2} \\ &= \|x\|\|y\|\end{aligned} \tag{14.1}$$

因此，一元的 Cauchy-Schwartz 不等式可以用 $n\times 1$ 向量的内积改写. 内积形式有更大的便利，不仅体现在内积的范数 $\leq$ 各自范数的乘积，更体现在可以推广到矩阵的内积. 将上述向量的内积形式的 Cauchy-Schartz 不等式推广到矩阵中，需要将内积和 ℓ_2-norm 推广到 $\mathrm{tr}(\cdot)$ 和 F 范数. 令 $A^\intercal = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix}, B = \begin{bmatrix} b_1 & \dots & b_n \end{bmatrix}$，它们是为可以相乘的矩阵，并且 AB 为方阵. 有

$$\left|\sum_{i=1}^{n}\langle a_i, b_i\rangle\right| = \left|\sum_{i=1}^{n} a_i^\intercal b_i\right| = |\mathrm{tr}\,(AB)| \leq \|A\|_F\|B\|_F = \left(\sum_{i=1}^{n}\|a_i\|^2\right)^{1/2}\left(\sum_{i=1}^{n}\|b_i\|^2\right)^{1/2}$$

compare the above with Equation 14.1, we find that $\mathrm{tr}(\cdot)$ 我们可以从 F-norm 的 Cauchy-Schwartz 不等式证明 F-norm 具有次可乘性，这是因为

$$\begin{aligned}\|AB\|_F = (\mathrm{tr}\,(ABB'A'))^{1/2} &= (\mathrm{tr}\,(A'ABB'))^{1/2} \\ &\leq (\|A'A\|_F\|BB'\|_F)^{1/2} \\ &= \|A\|_F\|B\|_F\end{aligned}$$

其中，要利用到 $\|A'A\|_F = \|A\|_F^2$，这是因为: 令 $A = (\ell_1, \dots, \ell_n)$，那么

$$\begin{aligned}\|A'A\|_F &= \left(\sum_{i,j}(\ell_i'\ell_j)^2\right)^{1/2} \\ &= (\sum_{i,j}(\ell_i'\ell_j\ell_j'\ell_i))^{1/2} \\ \|A\|_F^2 &= \left(\sum_i \ell_i'\ell_i\right) \\ &= \left(\left(\sum_i \ell_i'\ell_i\right)\left(\sum_j \ell_j'\ell_j\right)\right)^{1/2} \\ &= (\sum_{i,j}(\ell_i'\ell_j\ell_j'\ell_i))^{1/2}\end{aligned}$$

证明了：

- Bai and Ng (2002) theorem 1,

最后，我们可以从Equation 14.2得到更一般的 cs inequality[20]

$$\left|\sum_{i=1}^{n} \operatorname{tr}\left(A_i B_i\right)\right| = \left\|\operatorname{tr}\begin{bmatrix} A_1 \\ \vdots \\ A_n \end{bmatrix}\begin{bmatrix} B_1, \ldots, B_n \end{bmatrix}\right\| =\leq \left(\sum_{i=1}^{n} \|A_i\|_F{}^2\right)^{1/2}\left(\sum_{i=1}^{n} \|B_i\|_F{}^2\right)^{1/2} \tag{14.2}$$

$$= \left\|\begin{bmatrix} A_1 \\ \vdots \\ A_n \end{bmatrix}\right\|_F \left\|\begin{bmatrix} B_1, \ldots, B_n \end{bmatrix}\right\|_F \tag{14.3}$$

References

(https://stats.stackexchange.com/users/17023/a-donda), A. D. (n.d.). *What is the relationship between orthogonal, correlation and independence?* Cross Validated. URL:https://stats.stackexchange.com/q/171347 (version: 2017-04-13). eprint: `https://stats.stackexchange.com/q/171347`. URL: `https://stats.stackexchange.com/q/171347`.

(https://stats.stackexchange.com/users/203885/dato-gogolashvili), D. G. (n.d.). *Oracle Inequality : In basic terms*. Cross Validated. URL:https://stats.stackexchange.com/q/3423 (version: 2018-04-23). eprint: `https://stats.stackexchange.com/q/342338`. URL: `https://stats.stackexchange.com/q/342338`.

Ando, T. and Bai, J. (2016). "Panel data models with grouped factor structure under unknown group membership." In: *Journal of Applied Econometrics* 31.1, pp. 163–191.

Bai, J. (1994). "Least squares estimation of a shift in linear processes." In: *Journal of Time Series Analysis* 15.5, pp. 453–472.

— (2003). "Inferential theory for factor models of large dimensions." In: *Econometrica* 71.1, pp. 135–171.

— (2009). "Panel data models with interactive fixed effects." In: *Econometrica* 77.4, pp. 1229–1279.

Bai, J. and Ng, S. (2002). "Determining the number of factors in approximate factor models." In: *Econometrica* 70.1, pp. 191–221.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. (2009). "Simultaneous analysis of Lasso and Dantzig selector." In: *The Annals of statistics* 37.4, pp. 1705–1732.

Candes, E., Tao, T., et al. (2007). "The Dantzig selector: Statistical estimation when p is much larger than n." In: *The Annals of Statistics* 35.6, pp. 2313–2351.

Chan, N. H., Yau, C. Y., and Zhang, R.-M. (2014). "Group LASSO for structural break time series." In: *Journal of the American Statistical Association* 109.506, pp. 590–599.

Connor, G. and Korajczyk, R. A. (1986). "Performance measurement with the arbitrage pricing theory: A new framework for analysis." In: *Journal of Financial Economics (JFE)* 15.3.

Hastie, T., Tibshirani, R., and Wainwright, M. (2019). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

James, G. et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). "A modified principal component technique based on the LASSO." In: *Journal of computational and Graphical Statistics* 12.3, pp. 531–547.

Newey, W. K. and McFadden, D. (1994). "Large sample estimation and hypothesis testing." In: *Handbook of econometrics* 4, pp. 2111–2245.

Stock, J. H. and Watson, M. W. (2002). "Forecasting using principal components from a large number of predictors." In: *Journal of the American statistical association* 97.460, pp. 1167–1179.

Yuan, M. and Lin, Y. (2006). "Model selection and estimation in regression with grouped variables." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.1, pp. 49–67.

Zou, H., Hastie, T., and Tibshirani, R. (2006). "Sparse principal component analysis." In: *Journal of computational and graphical statistics* 15.2, pp. 265–286.

高惠璇 (2005). *应用多元统计分析 (北京: 北京大学出版社) Gao HX 2005 Applied Multivariate Statistical Analysis*.

Notes

Notes for section 2

1. 我觉得需要去看一些前沿的论文，需要更具体的例子
2. APCA 是吗？
3. 需要找一个对 n 降维的例子
4. 只需要利用 F 的信息，还是需要同时利用 A 的信息？找一个例子. 如果说只是利用 F 对样品进行分类，那么和 PCA 十分类似，有没有两种方法的对比？
5. 这一假设和 model identificaton 是怎样联系起来的？

Notes for section 3

6. 再考虑一下投影的直观含义
7. 张老师提出引文：这的不相等到底用在哪里？吴潇然师姐认为，会不会是选择因子个数时，利用到了不同 eigenvalues 之间的差必须是正的，那么根据这个差可以来确定因子个数，如果出现相等的特征根，则差为 0，多选或者少选一个因子都是一样的；但是张老师不这么认为，maybe 选因子的信息准则是根据 eigenvalue 的累积和来确定的. 所以这一块要再看一看
8. 这两篇文章尽量快点看一下
9. 这个问题需要考虑下，看一下线性代数的书
10. F 与 X 允许在多大程度上相关？
11. 这样做合理吗？
12. 我还没有证明这一个式子
13. 我觉得这好像说的是：一个确定的正定阵差不多可以看成一个有界的正数，若这个正定阵乘上一个矩阵依概率收敛到 0，那么乘上的这个矩阵一定要依概率收敛到 0
14. 为什么不是更强的 $\frac{1}{T^2}\sum_{ts}\tau_{ts} \leq M$
15. why this magnification (Cauchy-Schwartz inequality) is appropriate?

Notes for section 5

16. standard error? variance?
17. linear regression 中的结果要自己算一算
18. the difference between model selection and variable selection

19. 解释为什么右边没有绝对值. Consider that

$$y = X\widehat{\beta} + \widehat{\epsilon} \tag{14.4}$$

$$y = X\beta^{\star} + \epsilon\,, \tag{14.5}$$

then $X(\widehat{\beta} - \beta^{\star}) = \epsilon - \widehat{\epsilon}$. 但是我们并不能说明 $X(\widehat{\beta} - \beta^{\star})$ 的分量与 ϵ 的分量同正同负

Notes for section 14

20. 我暂时还没有证明