Note 2 Multivariate analysis

- Logic
  - On Note 1,
    - modeling one kind of stock's price from the raw data, normal distribution, student t distribution, and general t distribution, and we can valid it, KS test etc.
    - some index from the data, E, Var, f(X) < - moment estimate, interval estimate, MLE, 蒙特卡洛模拟
    - 当然这些，尤其是蒙特卡洛模拟都是建立在样本量足够大的情况下，样本量很小的情况下怎么解决？bootstrap
  - 但仅仅研究一只股票是不够的，我们最终要构建的就是一个portfolio，并且股票之间很可能是related的。这样我们就不能用之前的方法认为他们是iid的
  - On Note 2,
    - 我们将考虑多个股票可能服从的分布，joint distribution, 期望，方差（normal， t），同时还有多元的蒙特卡洛模拟
    - 对于多元函数来讲，我们还有另外的方法来模拟股票收益的走势，线性回归
      - CAPM
      - Fama-French three-factor model
- Basic concepts
  - joint distribution:
    - If $X_i s$ are discrete random variables, the joint distribution function can be expressed as *probability mass function*. That is,
      $$f_{\vec{X}}(x_1, x_2 \dots, x_n) = P(X_1 = x_1, X_2 = x_2 \dots, X_n = x_n)$$
    - If $X_i s$ are continuous random variables, the joint distribution function can be defined as *probability density function* such that
      $$P\big(X_1 \in (a_1, b_1), X_2 \in (a_2, b_2), \dots, X_n \in (a_n, b_n)\big) = \int_{a_n}^{b_n} \int_{a_{n-1}}^{b_{n-1}} \dots \int_{a_1}^{b_1} f_{\vec{X}}(x_1, \dots, x_n) dx_1 \dots dx_n.$$
      for any real numbers a_i, b_i where $a_i < b_i$.
  - marginal distribution:
    - If X_i is discrete random variable, then we have
      $$f_{X_i}(x_i) = P(X_i = x_i) = P\big(X_i = x_i \ \ and \ \ X_j \in (-\infty, \infty) \ \ for \ \ all \ j \neq i\big) = \sum_{all\ x_1} \cdots \sum_{all\ x_{i-1}} \sum_{all\ x_{i+1}} \cdots \sum_{all\ x_n} f_X(x_1, x_2, \dots, x_n).$$
    - If X_i is continuous random variable, we consider
      $$F_{X_i}(x) = P(X_i \leq x) = P\big(X_i = x \ \ and \ \ X_j \in (-\infty, \infty) \ \ for \ \ all \ j \neq i\big) = \int_{-\infty}^{x} \left(\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\vec{X}}(x_1, x_2, \dots, x_n) dx_1 \dots dx_{i-1} dx_{i+1} dx_n \right) dx_i$$
      Then the probability density function of X_i can be computed as
      $$f_{X_i}(x) = F'_{X_i}(x) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{\vec{X}}(x_1, \dots, x_{i-1}, x, x_{i+1} \dots, x_n) dx_1 \dots dx_{i-1} dx_{i+1} dx_n$$
  - change of variable formula

- g 是将x转变为y的函数，本质上是关于x的函数，h是g的反函数，y= Ax, 雅各比矩阵保证了将x转化为y的坐标系

$$f_Y(y_1, y_2, \dots, y_n) = f_X\big(h_1(y_1, \dots, y_n), \dots, h_n(y_1, \dots, y_n)\big)\left(\frac{1}{|J|}\right).$$

- Expected value =mean, covariance matrix= cov, 其实也就是之前的Sigma
  - property:
    - $E[AX + b] = AE[X] + b$
    - $Cov(AX + b) = ACov(X)A^T$

- Multi- normal distribution
  - 其实也就是Gaussian vector

We let $\vec{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$ be a vector of independent standard normal random variables (i.e. $Z_i \sim N(0,1)$). We define

$$\vec{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \mu_1 + a_{11}Z_1 + \cdots + a_{1n}Z_n \\ \mu_2 + a_{21}Z_1 + \cdots + a_{2n}Z_n \\ \vdots \\ \mu_n + a_{n1}Z_1 + \cdots + a_{nn}Z_n \end{pmatrix} = \underbrace{\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}}_{\vec{\mu}} + \underbrace{\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}}_{\vec{Z}}.$$

Then we say $\vec{Y} = \vec{\mu} + A\vec{Z}$ follows multivariate normal distribution with mean $\mathbb{E}[\vec{Y}] = \vec{\mu}$ and covariance matrix $\Sigma = Cov(\vec{Y}) = ACov(\vec{Z})A^T = AA^T$. We write $\vec{Y} \sim N_n(\vec{\mu}, \Sigma)$.

  - 这里注意，多元正态除了Gaussian vector的定义还有一种是可以写成$X = \mu + AZ$的形式，$AA^T = \Sigma$,其中Z是互相独立的，但是如果不是gaussian vector，而仅仅只是两个任意的正态分布，那么是不能找到这样的A的，例子： X正态, ϵX正态（LC chapter4 的例子），不能找到这样的A，因为两者是相互独立的
  - 同样，如果(X, X)按照Gaussian vector的定义确实是Gaussian vector， 那能不能写成那种形式呢，可以，按照后面给的公式就可以，一个自由变量，将它设为0
  - Properties
    - if X服从于多元正态分布，那么$a_1X_1 + a_2X_2 + ... + a_nX_n$ 服从正态分布
    - if X服从$N_n(\mu, \Sigma), Y = AX + b$服从$N(A\mu + b, A\Sigma A)$
    - if X服从$N_n(\mu, \Sigma), and \Sigma$ is invertible, $Y = (X - \mu)^T\Sigma^{-1}(X - \mu)$ 服从χ^2 distribution with df = n
    - Gaussian vector(大前提！！！): independent < = > cov=0

We let $\vec{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix}$ and $\vec{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ be two vectors of random variables such that

$\vec{W} = \begin{pmatrix} \vec{X} \\ \vec{Y} \end{pmatrix}$ follows multivariate normal distribution with $\vec{W} \sim N\left(\underbrace{\begin{pmatrix} \vec{\mu}_X \\ \vec{\mu}_Y \end{pmatrix}}_{\vec{\mu}}, \underbrace{\begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{XY} & \Sigma_{YY} \end{pmatrix}}_{\Sigma}\right)$

Then $\vec{X}$ and $\vec{Y}$ are independent if and only if $\Sigma_{XY} = 0$ (zero matrix).

    - 第四个性质并没有说X和Y必须独立，但是他们可以服从multi normal distribution，但需要额外的条件，什么条件未知, 如果我们由他们独立，那么他们就满足服从mulit normal distribution

We let $\vec{X}_1$ and $\vec{X}_2$ be two independent random variables such that $\vec{X}_1 \sim N_p(\vec{\mu}_1, \Sigma_{11})$ and $\vec{X}_2 \sim N_q(\vec{\mu}_2, \Sigma_{22})$. Then $\vec{X} = \begin{pmatrix} \vec{X}_1 \\ \vec{X}_2 \end{pmatrix}$ follows multivariate normal distribution with mean $\vec{\mu} = \begin{pmatrix} \vec{\mu}_1 \\ \vec{\mu}_2 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$.

- 同样，如果独立，也可以获得两者相加也满足multi normal distribution, 但要注意他们的变量个数必须是相同的

- Central Limit Theorem

We let $\vec{X}_1, \vec{X}_2, \dots, \vec{X}_n$ (where $\vec{X}_i \in \mathbb{R}^p$) be n independent and identically distributed random variables with $\mathbb{E}[\vec{X}_i] = \vec{\mu}$ and $Cov(\vec{X}_i) = \Sigma$. Then when $n \to \infty$, $\frac{1}{n}(\vec{X}_1 + \vec{X}_2 + \cdots + \vec{X}_n)$ is approximately multivariate normal with mean $\vec{\mu}$ and covariance matrix $\frac{1}{n}\Sigma$. $\sim N(\mu, \frac{1}{n}\Sigma)$

- Drawbacks
  - 如果我们选择利用多元正态分布来model stock return，就可以得到每个stock都必须满足正态分布，但这个事情是非常questionable的，因为会有tail很大的stock，也就不满足正态分布，所以multi normal distribution is not a good fit

- Multi t distribution
  - 定义的来源
    - 我们在上面提到了，并不是每个股票都满足normal distribution，所以我们需要考虑另一种model，在topic 1 我们已经学到了t distribution其实是蛮好的，所以我们猜想能不能有multi t distribution 呢?
      - 然后我们直觉来看，模仿multi normal distribution, 我们可以将multi t distribution 写成$Y = \mu + AT_v$的形式，但注意，T_v很麻烦，他每一个都有不同的自由度，首先我们并不能确实这个的covariance是什么，其次还记得我们之前求df吗，ad-hoc， it's so tedious!!!

        - 于是我们转化了一下multi t distribution 的形式，其中μ称为location vector, v is the degree of freedom, 要求每个股票的自由度必须相同，控制tail, v越小，tail 越大， Z，本来应该写成AZ，为了方便，我们现在将其搞成一个基础分布，服从normal distribution(0,Λ), Z 之间independent

$$\vec{Y} = \vec{\mu} + \sqrt{\frac{v}{\chi^2_v}}\vec{Z},$$

  - properties
    - mean= μ, cov(Y) = $\frac{v}{v-2}\Lambda, v > 2$
    - wY 也服从 t distribution: 这就对我们求portfolio return 有了帮助，如果股票的return 服从multi t distribution, 那么portfolio return 服从t 分布

- monte Carlo simulation
  - 和single stock 的monte Carlo 的步骤类似，这里我们要choose independent的$X = (X_1, X_2, ..., X_n)$,如果以实例来看，我们需要抽取n个股票的m个数据，我们一直知道n个股票服从于multi t distribution ,而且他们之间很可能存在correlation, 但如果我们只是简单的抽取sample， 计算机是会independent抽取的，这就和我们的假设不符，所以我们要想一个办法独立抽取然后组成可能相关的multi t distribution
  - 这就用到了之前的multi t Y= $\mu + \sqrt{\frac{v}{\chi_v^2}}Z$, Z服从multi normal , 所以我们按照multi normal 的方法抽取Z，然后抽取独立的χ_v^2, 就可以获得Y了（这对multi normaldistribution也同样适用）
  - process
    - multi normal distribution
      - 我们已知一个sample data, 假设X服从于multi normal , 我们可以用sample data 计算出应该服从的μ, Σ, 然后我们可以重新抽取样本，$X = \mu + AZ, AA^T = \Sigma$,可以计算出A, 然后只要抽取Z,每个$Z_i$服从独立标准正态分布，就可以计算了
    - multi t distribution
      - 和multi normal 一样，我们已知μ, Σ, v, 如何计算multi t 的A？只要抽取一个随机的χ_v^2 , 就可以知道A， 然后抽取Z， Z服从multi normal distribution, 抽取multi normal 我们在前面讲了，只是这里的Z服从multi normal $(0, \Lambda)$ ，然后可以计算X_i, 再继续抽取多个X_i,就可以monte Carlo simulation
- Linear regression model
  - 我们在上面确实推出了一个较为良好的model， 但是市场是在变化的，所以model也应该变化，而且每个model和每个model之间也应该是correlated的，所以其实multi t 也并不是一个perfect 的model，我们提出了linear regression model；对于一个模型来说，我们在构建的过程中会先猜想很多影响因素，然后我们需要首先选择到底需要哪些变量，真正有意义的变量有哪些？有意义之后，但是如果数量很多怎么办，可不可以简化计算（注意和减少变量数量不是一个概念！！！）-->主成分分析
  - 我们有学到，stock return 应该和很多因素有关系，eg. interest rate, inflation, GDP etc. 而且会有一些unexpected events, 于是我们建立了stock return 和多个因子之间的model，为了简单，我们将其定为线性模型；
    - response variable, 因变量
    - predictor variables, 自变量
    - ϵ_t ， random error, 我们规定$E[\epsilon_t] = 0, Var[\epsilon_t] = \sigma^2$,为什么规定期望为0呢，因为这种unexpected events有好有坏，期望为0代表了我们对这种unexpected events have no information
    - β_i coefficients
  - 有几种特殊的模型:
    - CAPM: r_i是我们想要预测的stock return,

      $$r_i - r_f = \beta(r_M - r_f) + \varepsilon_i,$$

- 如何determine the coefficients?
  - 我们选择让RSS最小，y_i是真实值，$\hat{y}_i = \sum \beta_i y_i$是我们预测的值(不包含$\epsilon$)

$$RSS_p = \sum_{i=1}^{n} \left(\underbrace{y_i}_{\substack{actual \\ value}} - \underbrace{(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}_{\substack{predicted\ value \\ (Linear\ regression)}} \right)^2$$

  - 如何求解?
    - 求导

To express the equations in a more convenient way, we define the following matrices:

$$\hat{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1}, \quad \vec{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \in \mathbb{R}^{n\times(p+1)}.$$

Then the above system of equations can be expressed as

$$\begin{cases} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0 \\ \sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0 \\ \vdots \\ \sum_{i=1}^{n} x_{ip}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0 \end{cases} \Rightarrow \begin{cases} (1 \quad 1 \quad \cdots \quad 1)(\vec{Y} - X\hat{\beta}) = 0 \\ (x_{11} \quad x_{21} \quad \cdots \quad x_{n1})(\vec{Y} - X\hat{\beta}) = 0 \\ \vdots \\ (x_{1p} \quad x_{2p} \quad \cdots \quad x_{np})(\vec{Y} - X\hat{\beta}) = 0 \end{cases}$$

$$\Rightarrow X^T(\vec{Y} - X\hat{\beta}) = \vec{0} \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T \vec{Y}.$$

  - 但实际上，仅仅只有一个point estimator 是非常有限的，point estimator 确实对我们目前的sample 的拟合in-sample是非常好的，但对于out of sample的估计有可能并不好，在这种情况下，我们就需要寻找一个interval estimate，这就涉及了他的四个性质
    - 方程:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1$$
$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2$$
$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \\ & & & & x_{np} \end{pmatrix} = n \times (p+1)$$

$p+1 \le n$

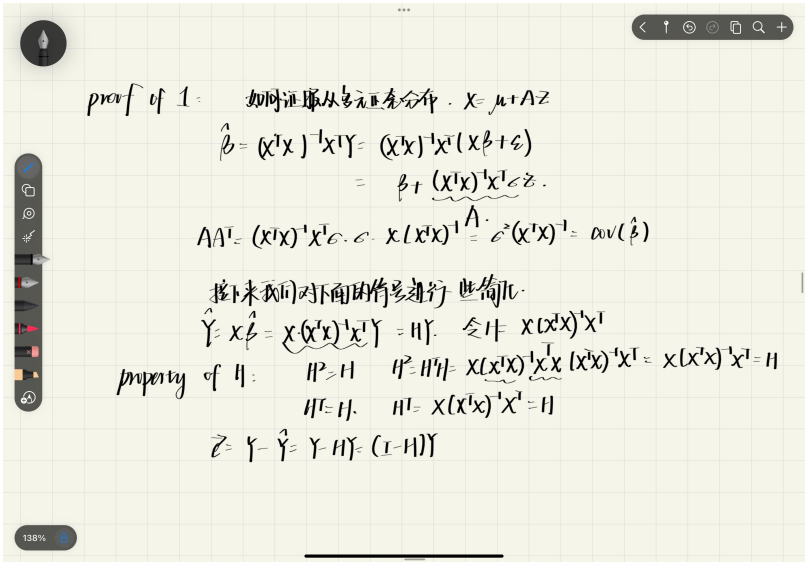
如果2个式子，有3个X因素？有多个解，其中1个X可确定用不到

$$\text{rank}(X^T X) = \begin{cases} (p+1)\times(p+1) \\ (p+1)\times n \times n\times(p+1) \end{cases} = p+1$$

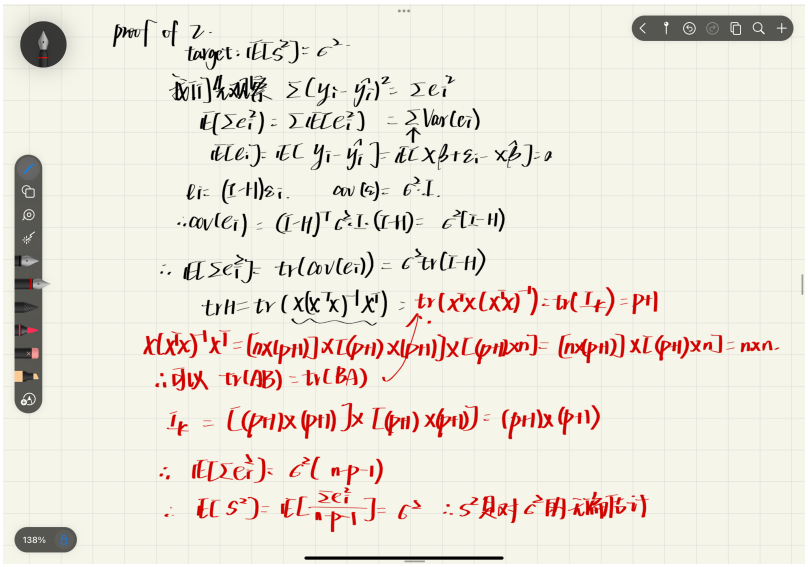
is invertible

    - 但在此之前，我们对性质的推导过程进行了一些假设
      - ϵ_i are independent ,服从于$N(0, \sigma^2)$--> mean=0 意味着stay neutral, 为什么independent呢？其实当然可以有cov，而且有cov的情况还很常见，但是为了simplify，我们将其设为independent的
      - X has full rank p+1
    - property 1: $\hat{\beta} \sim N_{p+1}(\vec{\beta}, \sigma^2 (X^T X)^{-1})$

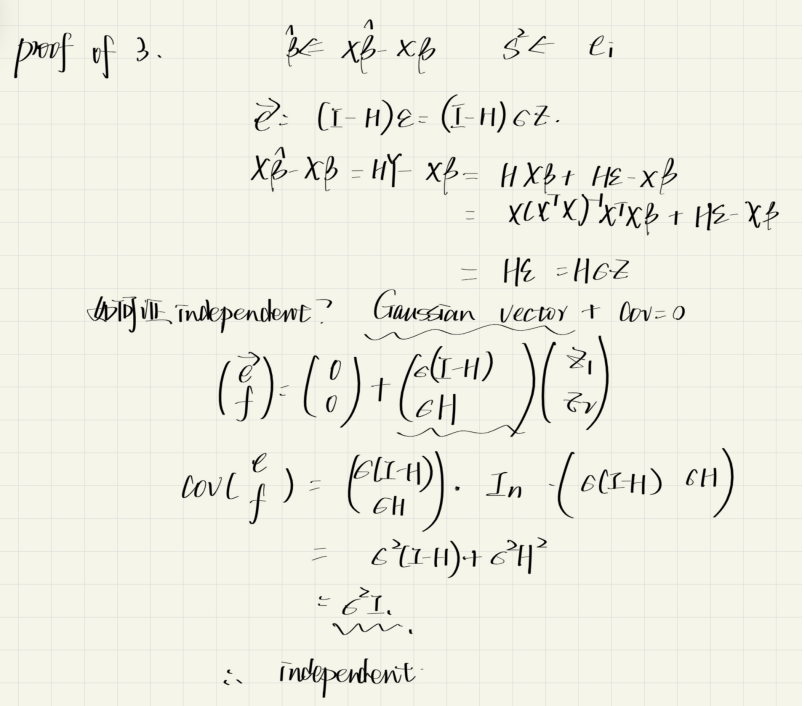
- proof



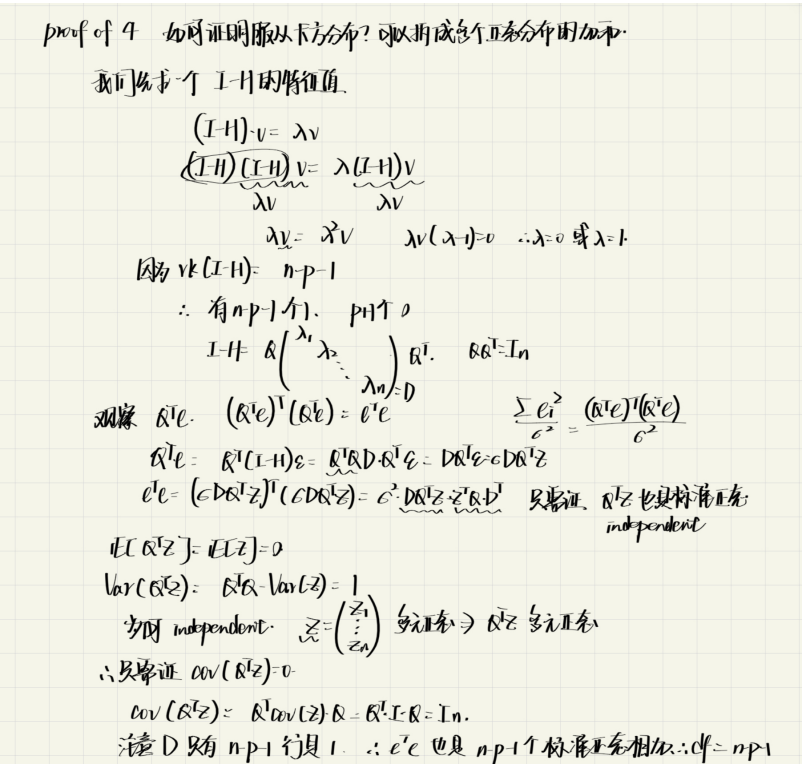
- property 2: s^2 is the unbiased estimator of σ^2, $s^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - 注意这里的s2不再是样本差了，因为分母不再是数目了而是n-p-1
  - proof



- property 3: $\hat{\beta} and s^2$ is independent
  - 首先肯定不可以直接求probability function, 太麻烦而且我们目前是求不出来的，那怎么做呢，我们可以找俩个等价的，其中分别都只和其中一个相关
  - $\hat{\beta}$和谁相关？X，Y，f；s和谁相关？e，所以我们只需要证明$\hat{\beta} and e$independent
  - 证明他们两个independent也不能用probability function
  - proof

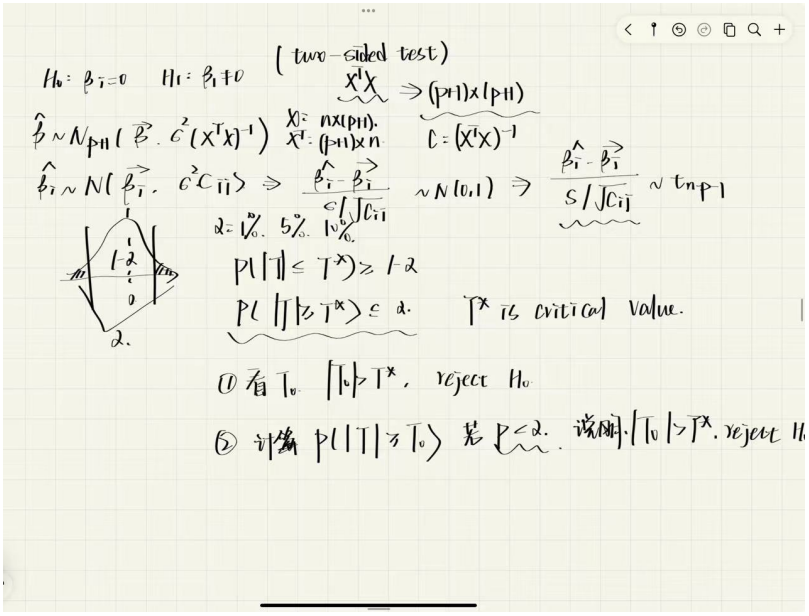


- property 4 : $\frac{RSS_p}{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sigma^2} \sim \chi^2_{n-p-1}$
  - 如何证明服从卡方分布呢？证明是可以写成多个正态分布的平方求和的形式，有多少个正态分布求和这个卡方分布的自由度就是多少
  - proof

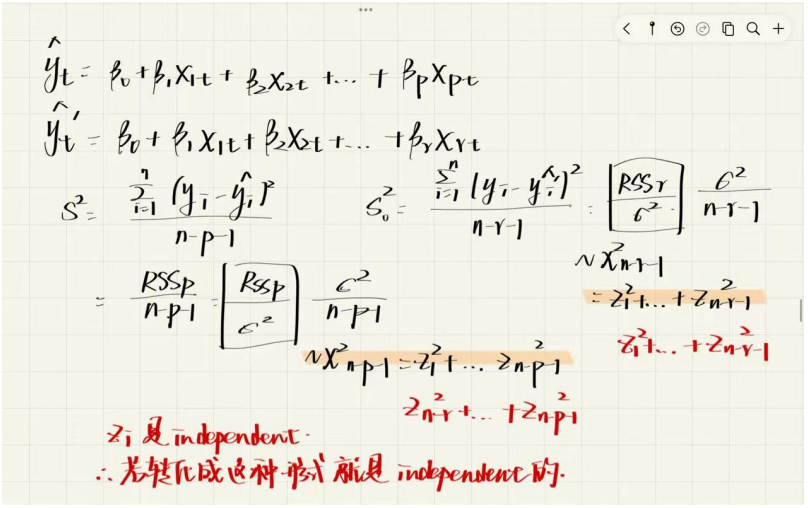


- 如何选择变量?
  - 如果变量太多，多一个变量就多一个需要估计的β,就会是error 变大； 但少一个变量就可能少了信息，所以如何选择一个最合适的变量呢?
  - 这个变量有没有意义<--$\beta = 0$？<-- hypothesisi test: t-test/F-test
  - Method 1: T- test

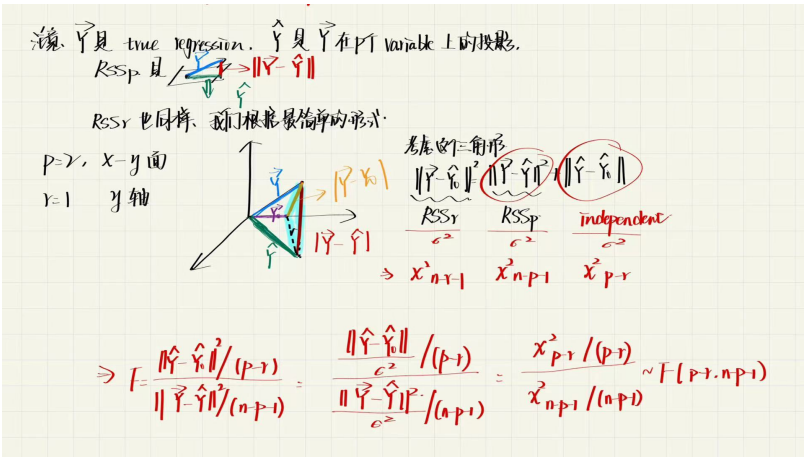
- process



- Method 2: F-test
  - 不再是一个一个检验了，我们之前不是有p个变量吗，我用p个variable做了一个regression, 然后再用r个variable 做了一个regression,如果可以简化成r个变量，那么这两套的方差应该是相近的，方差服从什么分布？χ^2, 两个χ^2相比服从什么分布？F分布，趋于1
  - 真的吗？当然是假的！！！两个独立的χ^2分布相比才是F分布，怎么让他们独立呢？需要modify



  - modify 之后，趋于0



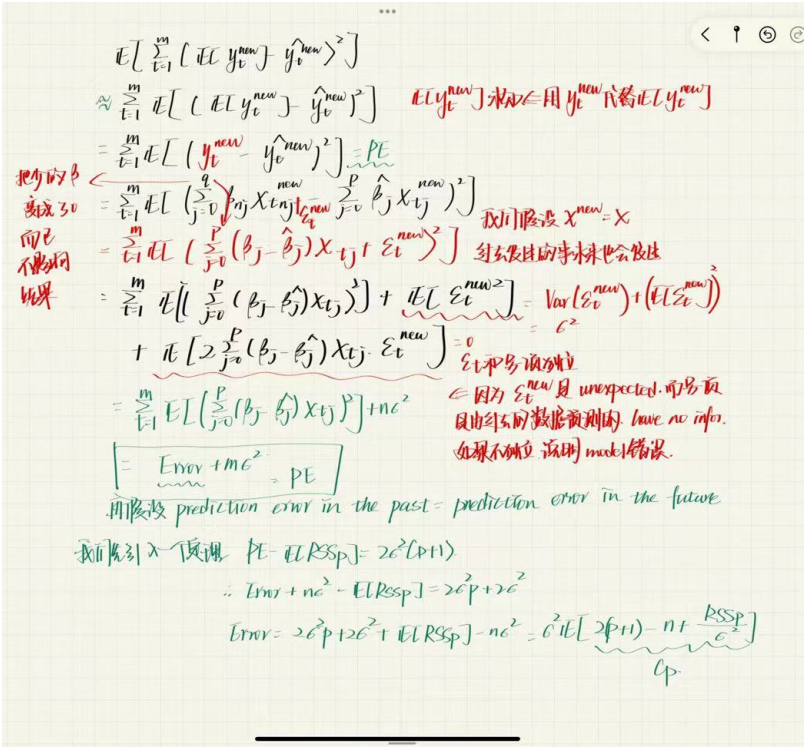
- hypothesis

The null hypothesis and alternative hypothesis are set to be

$$\begin{cases} H_0\colon\ \beta_{r+1} = \beta_{r+2} = \cdots = \beta_p = 0 \\ H_1\colon\ otherwise \end{cases} \Leftrightarrow \text{Some of these factors are irrelevant}$$

- H_0的含义是，所有的都等于0，H_1指只要有某些r+1到p不等于0就可以，但我们不知道具体是哪些，所以其实还需要T-test
- 如果F足够大，则拒绝H_0

- Method 3: Mallow's C_p statistics
  - 其实是另一种方法来判断哪些变量的参数等于0，哪些不等于0，这个方法和T-test、F-test 是独立的
  - 我们做回归的目的是什么？predict，我们希望predict的error最小，所以如果我们用一部分数据做了回归之后，在用X_{new}做预测时，希望$\min \sum\{y_{new} - \hat{y_{new}}\}^2$

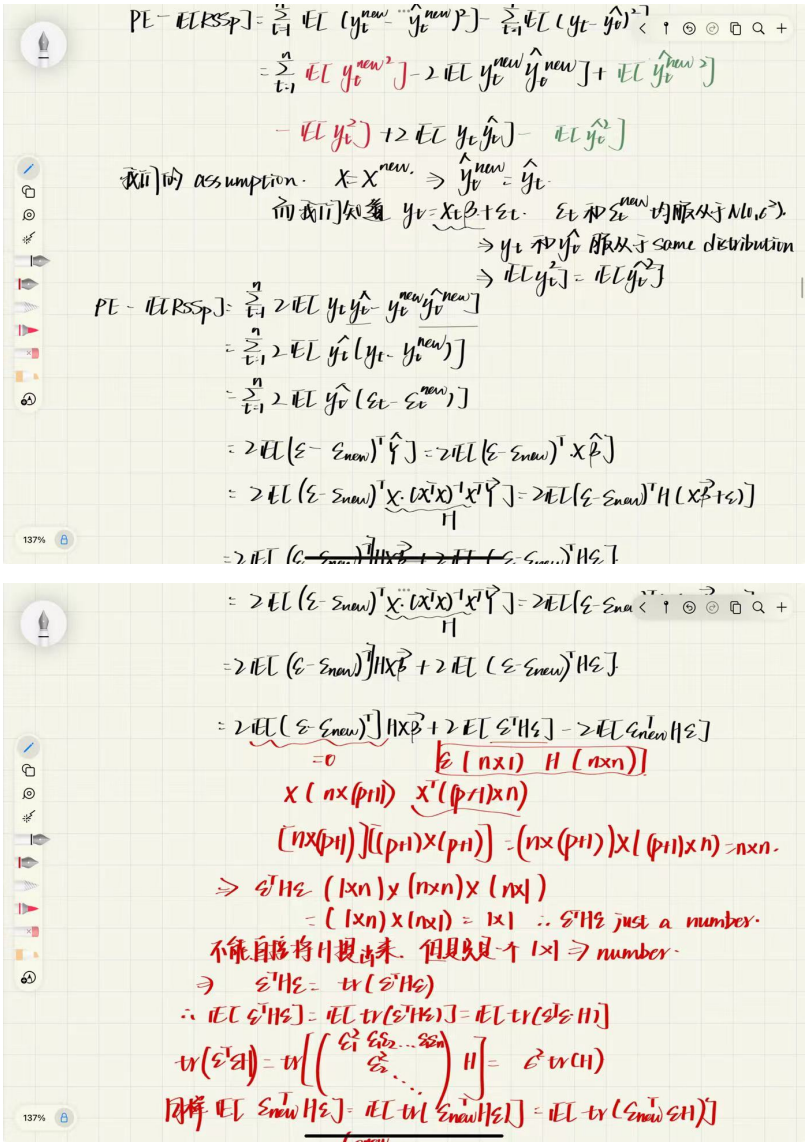


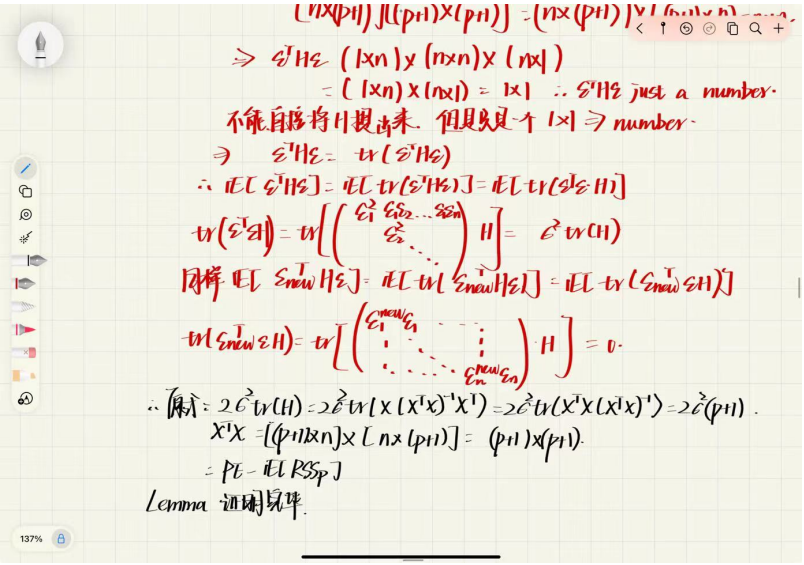
  - Remark:
    - σ^2未知，我们可以用s_k^2来代替
    - 最好的model就是最小的C_p statistics

- χ^2分布的均值是df，方差是2*df，$E[RSS_p] = (n-p-1)*\sigma^2, E[C_p] = p+1$,也就是说，我们选择$C_p$最小的，就可以解决overfitiing（变量冗余）的情况
- 以上所有我们的结论都是基于一个最基本的假设来讲的：这个model包含所有的reliable的变量，如果有遗漏的变量，那么上面的$E[RSS_p] = (n-p-1)*\sigma^2, E[C_p] = p+1$并不成立，而变成了下面的公式：（这里不证明）

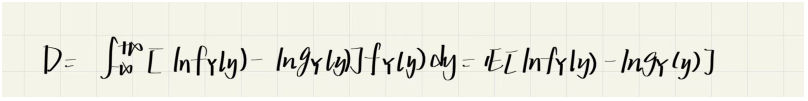
$$\mathbb{E}[RSS_p] = \sum_{i=1}^{n}\left(\mathbb{E}[y_t - \hat{\beta}^T x_t]\right)^2 + (n-p-1)\sigma^2.$$

- 我们的判断标准是最小的C_p，但实际上我们在这里做的只是计算C_p,你看error的公式其实是C_p的期望的最小值，而不是我们真正用的仅仅只是选择最小的C_p，所以有可能我们得到的model再进行T-test，是可能会有irrelevant的variable存在的-->保证了包含了所有的relevant的variable，但没有保证不包含所有irrelevant的variable
- Lemma的证明

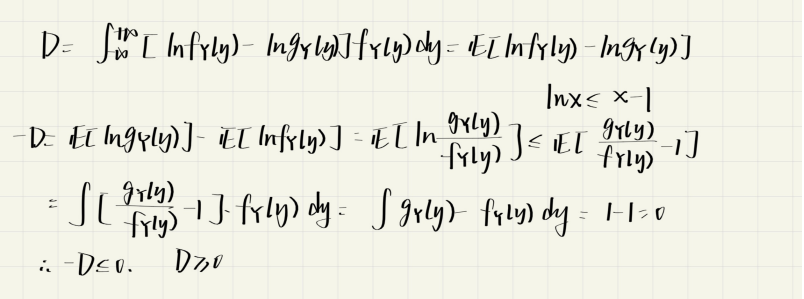




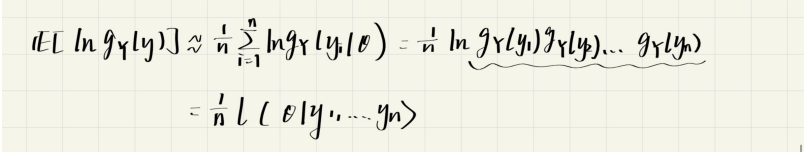
- Method4: AIC
  - C_p实际上只考虑了expectation，AIC和C_p不同的点在于它考虑的是distribution，而不是简单的expectation，$|g_Y(y) - f_Y(y)|$,但这个公式是non-feasible的，所以我们选择了average distance



  - Remark: 为什么取了ln?
    - 可以simplify calculation
    - guarantee positive



    - 这两条也就说明了这个D是个reasonable indicator
  - 我们希望D min， 而$E[\ln f_Y(y)]$是一定的，不会发生改变，所以希望g max，g是什么？是我们估计的y的distribution，让distribution的期望最大又是什么？--> MLE
  - $E[\ln g_Y(y)]$这个很熟悉吧，蒙特卡洛模拟！



  - 经过证明，但实际上，这并不是一个无偏估计，所以经过证明我们能得到，这里的k是参数的数量，也就是p+2, p+1个β +一个σ

$$\mathbb{E}\left[\frac{1}{n}l(\hat{\theta}|y_1,\dots,y_n)\right] \approx \mathbb{E}\left[\ln g_Y(y|\vec{\theta})\right]_{\vec{\theta}=\hat{\theta}} + \frac{k}{n}$$
$$\Rightarrow \mathbb{E}\left[\frac{1}{n}l(\hat{\theta}|y_1,\dots,y_n) - \frac{k}{n}\right] = \mathbb{E}\left[\ln g_Y(y|\vec{\theta})\right]_{\vec{\theta}=\hat{\theta}}.$$

- $\max g \to \max E[\frac{1}{n}l(\theta) - \frac{k}{n}] \to \max E[-\frac{1}{n}(-l(\theta)+k)] \to \min(-2l(\theta)+2k) \to \min AIC = \min(-2l(\theta)+2k)$
- 接下来我们计算着p+2个参数的理论值：， β 就是我们之前的理论值，σ 通过求导取最值

$$L(\beta_0,\dots,\beta_p,\sigma^2) = \prod_{i=1}^{n} \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2}{2\sigma^2}}}_{g_{y_t}(y_i|\beta_0,\dots,\beta_p,\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\sum_{i=1}^{n}\frac{\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2}{2\sigma^2}}$$

$$\Rightarrow l(\beta_0,\dots,\beta_p,\sigma^2) = \ln L(\cdot) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \sum_{i=1}^{n}\frac{\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2}{2\sigma^2}.$$

$$\hat{\beta} = (X^T X)^{-1} X^T \vec{Y}.$$

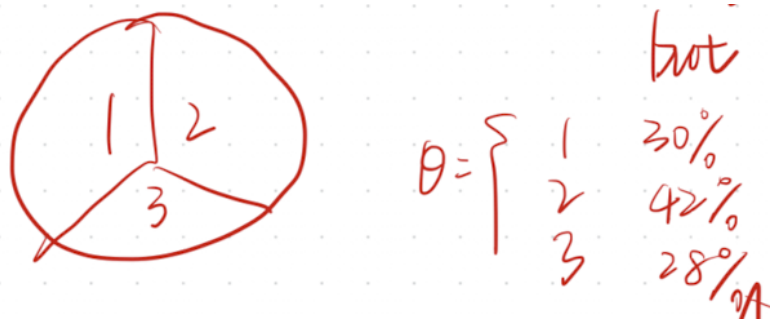
$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2}{\sigma^3} = 0 \Rightarrow \hat{\sigma} = \sqrt{\frac{RSS_p}{n}}.$$

$$l(\hat{\theta}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\frac{RSS_p}{n} - \frac{RSS_p}{2\left(\frac{RSS_p}{n}\right)} = -\frac{n}{2}\ln(2\pi) - \frac{n}{2} - \frac{n}{2}\ln\frac{RSS_p}{n}.$$

$$\begin{aligned} AIC &= -2\left(-\frac{n}{2}\ln(2\pi) - \frac{n}{2} - \frac{n}{2}\ln\frac{RSS_p}{n}\right) + 2(p+2) \\ &= n\ln(2\pi) + n + n\ln\frac{RSS_p}{n} + 2(p+2) \\ &= n\ln(2\pi) + n + n\left(\ln\frac{RSS_p}{n} + \frac{2(p+2)}{n}\right) \sim \ln\frac{RSS_p}{n} + \frac{2(p+1)}{n} \end{aligned}$$

  - 这里为什么吧p+2写成p+1， 是因为σ是一直都有的，但其他的β可能会存在irrelevant 的factor，所以不一定都存在，--> 只有p+1个free parameters

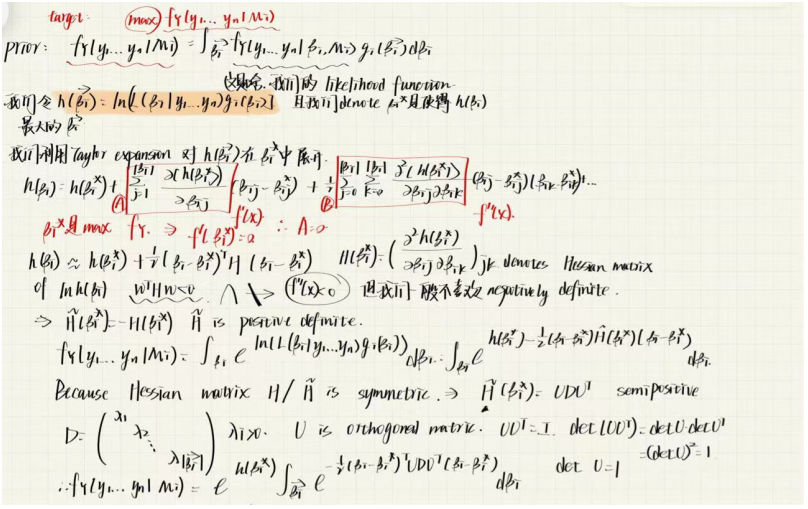
- Method 5 : BIC
  - BIC 和AIC的不同：
    - 还记得吗，我们在AIC中应用了蒙特卡洛模拟，应用蒙特卡洛模拟的前提是必须有一个large sample，如果没有这个前提，比如IPO数据，结果就不再reliable
    - 之前在数据很少的情况下我们怎么做的？find some similar stocks to indicate something about this stock --> 但注意，这个$\theta_i \neq \theta$,我们treat θ as a random variable--> prior distribution, 例如



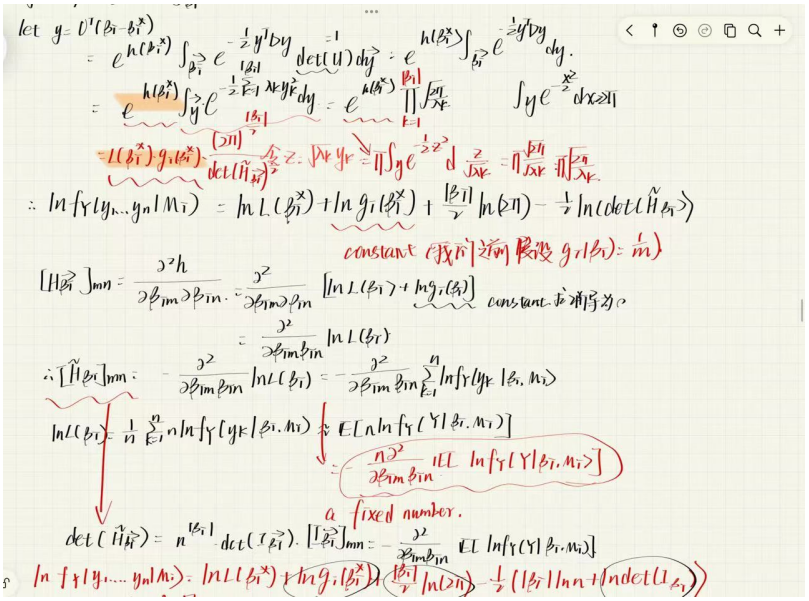
- 一旦有了新的信息，我们就updata这个distribution --> information updating
- process:
  - 我们需要一个prior distribution, 这个distribution感觉包含了所有的结果，且之后所有的结果可能不会再增加新的事件 --> $\pi(\theta) = 1(40\%), 2(30\%), 3(30\%)$, $f_{X^*}(x) = \int_{-\infty}^{+\infty} f_X(x|\theta)\pi(\theta)d\theta$,这个函数是概率密度函数，也就是$X^* = x$的概率是这个，x是提前已知的，是我们的输入量，所以是对θ进行积分
  - 之后，经过了一段时间之后，我们可以获得一些新信息，这时就要更新我们的参数的分布，进而更新我们的新的概率密度函数，此时计算概率密度函数的公式不变，只是$\pi(\theta)$变了

$$\pi(\theta|X_1, X_2, \dots, X_n) = \frac{\pi(\theta) f_X(X_1, X_2, \dots X_n|\theta)}{\int_{-\infty}^{\infty} \pi(\theta) f_X(X_1, X_2, \dots X_n|\theta) d\theta}.$$

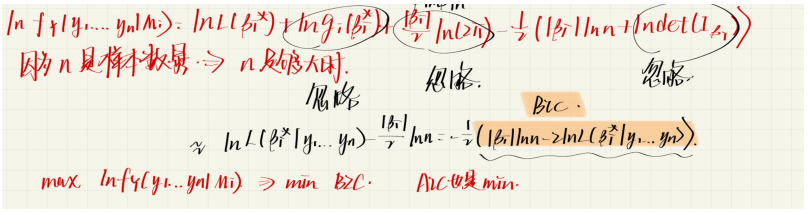
- 我们怎么利用这个process 来进行model selection?
  - 假设我们可以假设m个model，我们suppose每个model （其实每个model就是不同的parameter组合），也就是每个θ的初始概率是$\frac{1}{m}$
  - 这里我们更改一下denotion, $\pi(\theta) = g_i(\beta_i)$
  - BIC score:
    - 1



    - 2



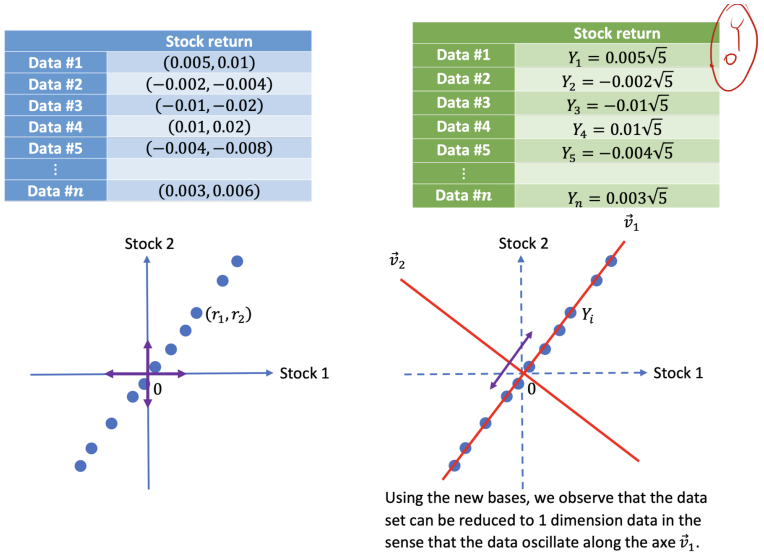
- 3



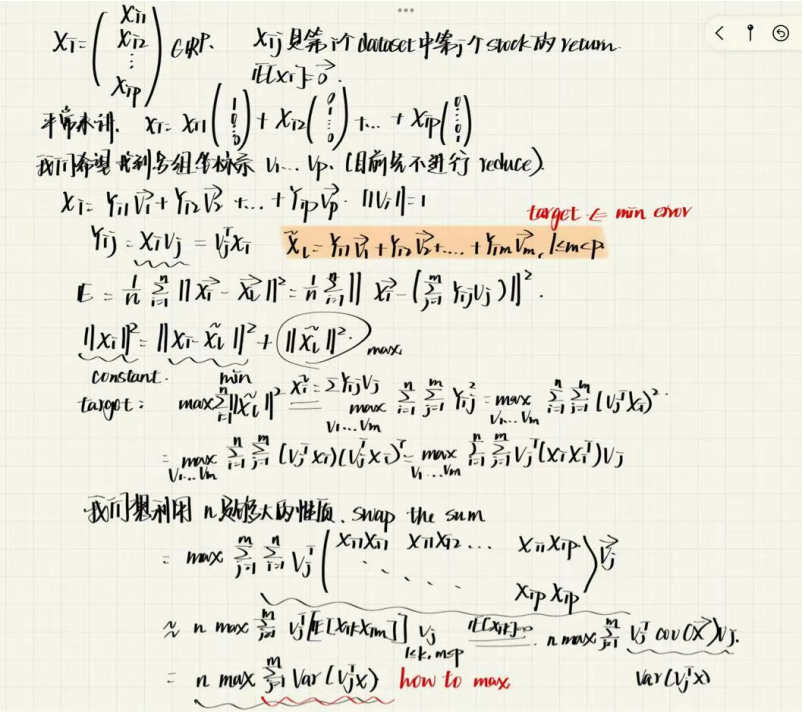
- Model selection
  - Method 1: Forward selection
    - 刚开始考虑null model，然后一个一个加variable
    - 刚开始先加p最小的，即拒绝H_0（$\beta \neq 0$）的，也就是F越大
  - Method 2:Backward Selection
    - 刚开始考虑full model，然后一个一个减variable
    - 刚开始减去$\beta = 0$，也就是不能拒绝H_0的，即p最大的，F最小的
- Principal Component Analysis
  - reduce the dimension of the data
  - 比如这个数据，stock 1 和stock 2 的每组数据都是有关系的，都是1:2的关系，所以我们只需要每组数据之间的系数

	Stock 1	Stock 2
Data #1	0.005	0.01
Data #2	-0.002	-0.004
Data #3	-0.01	-0.02
Data #4	0.01	0.02
Data #5	-0.004	-0.008
⋮	⋮	⋮
Data #n	0.003	0.006

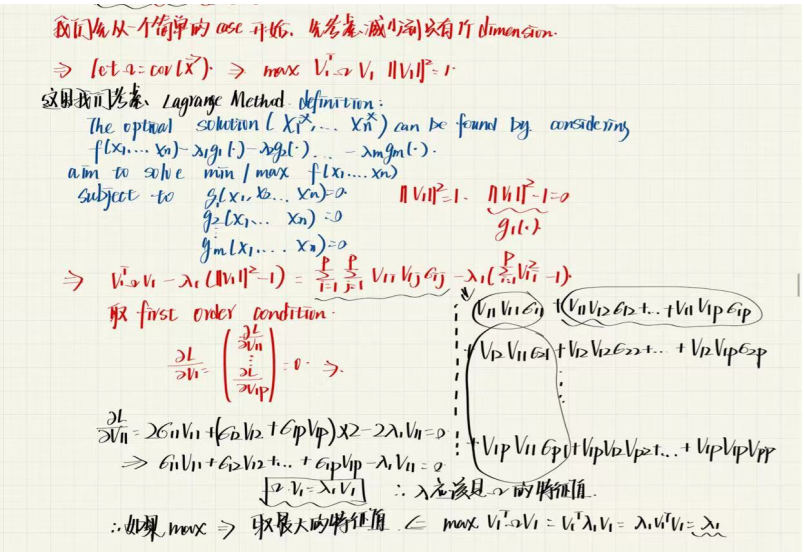
- result



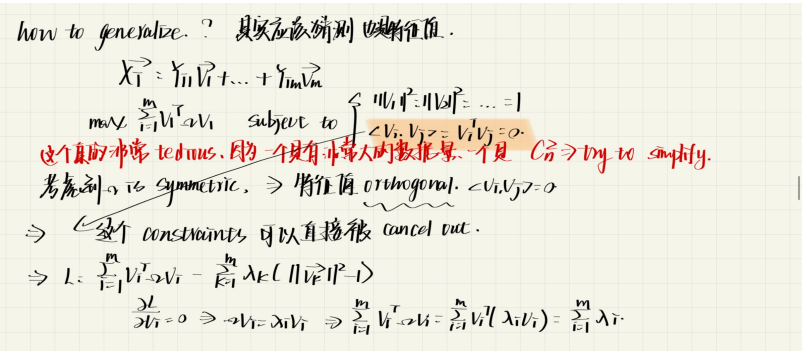
- 但问题是，实际上并没有这样perfect的dataset，如果我们reduce the dimension,就一定会产生error,所以我们reduce 的target就是使得误差最小
- assumption： 我们需要是的$E[X_i] = 0$, 也就是每个dataset的期望必须是0， 如果期望不是0的话，他在坐标系上的中心点就不是坐标原点，我们是没有办法仅仅通过旋转坐标轴来使得所有的数据在更小的维度上
- process
  - 1



- 2



- 3



- Remark:
  - Y_{ij} principal components
  - v_j factor loading
  - Property

- $Var(Y_{ij}) = \lambda_j$
- $Cov(Y_{ij}, Y_{ik}) = 0$

- proof

  Proof of the properties

  We let $U = (\vec{v}_1, \vec{v}_2, \dots, \vec{v}_p)$ be a $p \times p$ matrix which the columns are the eigenvectors of Ω. Since Ω is symmetric, so $\vec{v}_i$s are orthogonal and $UU^T = I_p$.

  Then the data $\vec{X}_i$ (with no approximation) can be expressed as

  $$\vec{X}_i = Y_{i1}\vec{v}_1 + Y_{i2}\vec{v}_2 + \cdots + Y_{ip}\vec{v}_p = U\vec{Y}_i.$$

  This implies that $\vec{Y}_i = U^T\vec{X}_i$.

  Using the properties of covariance, we deduce that

  $$Cov(\vec{Y}_i) = U^T \underbrace{Cov(\vec{X}_i)}_{=\Omega} (U^T)^T = U^T \left(U \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_p \end{pmatrix} U^T \right) U = \underbrace{\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_p \end{pmatrix}}_{=D}$$

  This implies that $Var(Y_i) = Cov(Y_i, Y_i) = \lambda_i$ and $Cov(Y_i, Y_j) = 0$.

- total variance of $X_i = \sum_{i=1}^{p} \lambda_i = tr(D)$
- total variance of $\hat{X}_l = \sum_{i=1}^{m} \lambda_i = tr(D)$

- parallel shift component: affect in the same direction
- tilt component: short term is opposite with the long term
- curvature: mid -term is different from else