① Understand conventional regression model (personal perspective)

Result: the basic of Mean Regression is exactly our model setting !

Firstly, we have $\begin{cases} \text{dependent var: } Y \\ \text{independent var: } X \end{cases}$

→ when X is fixed, consider $Y|X=x$ r.v.,

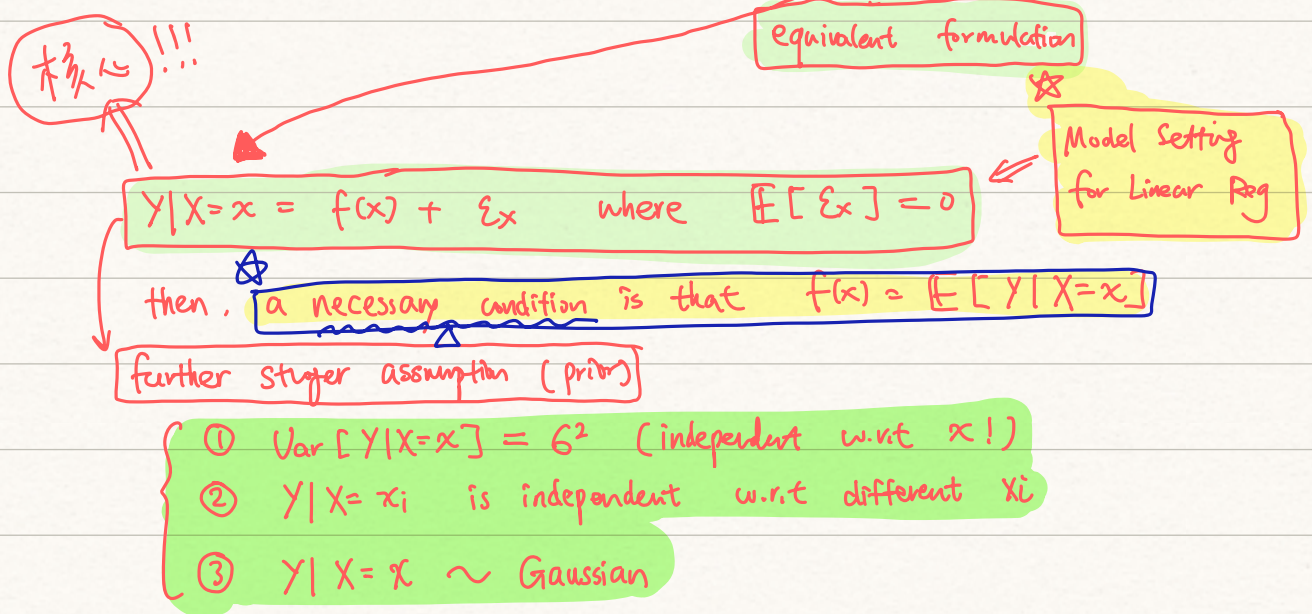
it can be decomposed into $Y|X=x = \mathbb{E}[Y|X=x] + \varepsilon_x$

generally speaking, the decomporition guarantees that $\mathbb{E}[\varepsilon_x]=0$ ← necessary condition

Issue:

① ε_x's variance might be different

② ε_x's distribution might be different



Since Under this Model Setting, $f(x) = \mathbb{E}[Y|X=x]$

our aim is to construct an Estimator $\hat{f}(x)$ to appximate $f(x)$

↓

the given method is to use MLE under our assumed prob model

⟹ finally, what we have done is Regression towards Mean

Linear Regression

① Model :

$$y = X\beta + \varepsilon \quad \begin{cases} 1.\ \varepsilon \sim N(0, \sigma^2 I_n) \\ 2.\ \text{rank}(X) = p+1 \end{cases}$$

② MLE $\Rightarrow$ $\begin{cases} \hat{\beta} = (X^TX)^{-1}X^Ty \Rightarrow \hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty := H\cdot y \\ \hat{\sigma}^2 = \frac{1}{n} y^T(I-H)y \end{cases}$

③ Property of estimator :

$$\begin{cases} \mathbb{E}[\hat{\beta}] = (X^TX)^{-1}X^T \cdot \mathbb{E}[y] = \beta \\ \text{var}[\hat{\beta}] = \text{var}[(X^TX)^{-1}X^Ty] = (X^TX)^{-1}X^T\text{var}[y]X(X^TX)^{-1} \\ \qquad\qquad = \sigma^2(X^TX)^{-1} \\ \hat{e} = y - \hat{y} = (I-H)y, \quad \text{cov}(\hat{\beta}, \hat{e}) \\ \qquad\qquad = \text{cov}((X^TX)^{-1}X^Ty, (I-H)y) \\ \qquad\qquad = (X^TX)^{-1}X^T\text{var}(y)(I-H) \\ \qquad\qquad = \sigma^2 \cdot 0 = 0 \end{cases}$$

$\text{var}(\hat{e}) = \sigma^2(I-H)$

$\mathbb{E}[SSE] = \text{tr}(\text{var}[\hat{e}]) = (n-p-1)\sigma^2$

④ Best Linear Unbiased Estimator $\Rightarrow \hat{y} \xrightarrow{\text{BLUE}} X^T\hat{\beta}$

→ for $\forall c$, $c^T\hat{\beta} \xrightarrow{\text{BLUE}} c^T\beta$

Pf Sketch : Suppose α^Ty is an estimator for $c^T\beta$

then $\mathbb{E}[\alpha^Ty] = \alpha^TX\beta = c^T\beta \quad \forall\beta$

$\Rightarrow \alpha^TX = c^T \Leftrightarrow X^T\alpha = c$

Consider $\mathrm{Var}[\alpha^T y] - \mathrm{Var}[c^T \hat{\beta}]$

$= \sigma^2 \alpha^T \alpha - \sigma^2 c^T (X^T X)^{-1} c$

$= \sigma^2 \cdot \alpha^T \underbrace{(I - X(X^T X)^{-1} X^T)}_{\text{幂等 } (I-H)} \alpha \geq 0$

⑤ Hypothesis Testing

$H_0: \beta_1 = \cdots = \beta_p = 0$

a) $\dfrac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} \sim F(p, n-p-1)$

$HX = X \Rightarrow H \cdot \mathbb{1} = \mathbb{1}$

$SSR = y^T [\underbrace{H - \frac{1}{n}\mathbb{1}\cdot\mathbb{1}^T}_{\text{对称幂等}}] y$ $\quad \mathrm{tr}(\cdot) = p+1-1 = p$ $\quad \beta_0 \cdot \mathbb{1}$

$(H - \frac{1}{n}\mathbb{1}\mathbb{1}^T) X \begin{pmatrix} \beta_0 \\ 0 \end{pmatrix} = 0$

$SSE = y^T [\underbrace{I-H}_{\text{对称幂等}}] y$ $\quad \underbrace{(I-H)\cdot X\beta = 0}$ $\quad \mathrm{tr}(\cdot) = n-p-1$

$SSE \perp\!\!\!\perp SSR$

b) $\hat{\beta} \sim \mathrm{Gaussin}(\beta, \sigma^2 (X^T X)^{-1})$

$\mathrm{Cov}(\hat{\beta}, \hat{e}) = 0 \Rightarrow \hat{\beta} \perp\!\!\!\perp \hat{e} \Rightarrow \hat{\beta} \perp\!\!\!\perp SSE$

$H_0: \beta_j = 0$

$\Rightarrow \dfrac{\frac{\hat{\beta}_j}{\sigma\sqrt{c_{jj}}}}{\sqrt{\frac{\frac{SSE}{\sigma^2}}{n-p-1}}} = \dfrac{\hat{\beta}_j}{\hat{\sigma}_{\text{残差}} \cdot \sqrt{c_{jj}}} \sim t(n-p-1)$

c) $R^2 = \dfrac{SSR}{SST} = 1 - \dfrac{SSE}{SST}$

SST

n ↑ R^2 ↑ ⁂

Quantile Regression ⟶ A different perspective

① First Problem-Setting

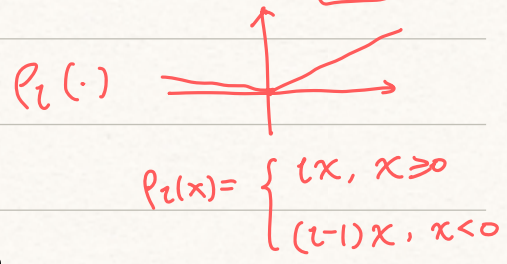
$Y|X=x = g_\tau(x) + \varepsilon_\tau|X=x$ where $Q_\tau(\varepsilon_\tau|X=x) = 0$

$Q_\tau(Y|X=x) = g_\tau(x)$

↓

we want to seek for an estimator for $g_\tau(x)$

对勾 function



② Ground-Truth (Amazing observation)

$Q_{Y|X=x}(\tau) = \inf\{y : F_{Y|X=x}(y) \geq \tau\}$

then $Q_{Y|X=x}(\tau) = \underset{\theta}{\operatorname{argmin}}\ \mathbb{E}[\rho_\tau(Y|X=x-\theta)]$

↓

$Q_{Y|X}(\tau) = \underset{g\in G}{\operatorname{argmin}}\ \mathbb{E}[\rho_\tau(Y-g(X))] \quad (X,Y)\sim P$

$= \underset{g\in G}{\operatorname{argmin}}\ T(P,g)$

③ Plug-in estimator

$\hat{Q}_{Y|X}(\tau) = \underset{g\in G}{\operatorname{argmin}}\ T(\hat{P}_n, g)$

$= \underset{g \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_2(Y_i - g(X_i))$

(X, Y) random

for fixed X, only random Y

similar derivation also holds!

Generally, $\mathcal{G} = \{g : g : x \longrightarrow \beta^T x, \forall \beta\}$

Why Mean Regression? (Another Interpretation)

Observation: (optimality of conditional expectation w.r.t MSE)

$$\mathbb{E}[Y|X=x] := g(x) = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \mathbb{E}[(Y - g(X))^2]$$

should be similar!

plug-in estimator

More related to Non-parametric Framework (like quantile regression)

$$\hat{g} = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - g(X_i))^2$$

choose $\mathcal{G} = \{g : g(x) = \beta^T x, \beta \in \mathbb{R}^n\}$,

then this is exactly **Linear Regression** framework

Update at 11/18

→ plug-in estimator is just one (view) understanding of Linear Regression Algo

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1} (y_i - x_i^T\beta)^2 = (X^TX)^{-1}X^Ty$$

$$\hat{g}(x) = \mathbb{E}[y|X=x] \leftarrow \tilde{g} = \underset{g \in \mathcal{G}}{\operatorname{argmin}} \mathbb{E}_{(X,Y)}[(y - g(x))^2]$$

→ However, it is more reasonable to use **Parametric Estimation Framework**

(under Gaussian assumption, infer Gaussian $(x^T\beta, \sigma^2) \rightsquigarrow \hat{\beta}$)

$\hat{y} \rightarrow X\hat{\beta}$ estimator (when we achieve $y \sim \text{Gaussian}(x^T\hat{\beta}, \hat{\sigma}^2)$, estimate y!)

#

Previously, mainly talk about: (LR)

① derivation of $\hat{\beta} = (X^TX)^{-1}X^Ty$

② Basic property of $\hat{\beta}$ (estimator of β)

1) unbiased

2) $\text{Var}[\hat{\beta}] = (X^TX)^{-1}X^T \text{var}(y) \times (X^TX)^{-1}$

$= \sigma^2(X^TX)^{-1}$ ✡

$\Rightarrow \hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$ $\Rightarrow$ **t test for $\hat{\beta}_i$**

3) $\hat{e} = y - X\hat{\beta} = (1-H)y$ $\quad H = X(X^TX)^{-1}X^T$

$\begin{cases} \mathbb{E}[\hat{e}] = 0 \\ \text{Var}[\hat{e}] = \sigma^2(1-H) \end{cases}$

$\text{tr}(H) = \text{tr}(1_{p+1}) = p+1$

$\Rightarrow \mathbb{E}[\hat{e}^T\hat{e}] = \text{tr}[\text{Var}[\hat{e}]]$

$= \text{tr}[\sigma^2(1-H)]$

$= \sigma^2(n-p-1)$

$\Rightarrow \hat{e}^T\hat{e}/(n-p-1)$ is an unbiased estimator for σ^2

4) $\hat{\beta} \perp\!\!\!\perp \hat{e}$ $\Leftarrow \text{cov}(\hat{\beta}, \hat{e})$

$= \text{cov}((X^TX)^{-1}X^Ty, (1-H)y)$

$= 0$

characterization of chi-square dist

KEY POINT is:

$y^T \Lambda y \sim \chi^2(p)$

$\Leftrightarrow$ { $\Lambda \cdot \mathbb{E}[y] = 0$; Λ 对称幂等 ; $tr(\Lambda) = p$

5) ① t-test : $H_0 : \hat{\beta}_i = 0$ $H_a : \hat{\beta}_i \neq 0$

$\Leftarrow$ $SSE = \hat{e}^T \hat{e} = y^T (1-H) y \sim \sigma^2 \chi^2(n-p-1)$

② $SSR = y^T \left(X(X^TX)^{-1}X^T - \frac{1}{n} \mathbb{1}\mathbb{1}^T \right) y$ → 对称幂等

(also need cochran theorm to charcterize the independency)

$\Rightarrow$ F-test $\Rightarrow$ $H_0 : \beta_1 = \cdots = \beta_p = 0$ if $\beta_1 = \cdots = \beta_p = 0$

Notation

$(y_i - \bar{y})$ $(\hat{y}_i - \bar{y})$ $(y_i - \hat{y}_i)$

$SST = SSR + SSE$

↓ ↓ ↓

TSS ESS RSS

H_a : o/w

& $\dfrac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} \sim F(p, n-p-1)$

③ A more general statement : (Constrainted LR)

→ Problem Setting (Model)

{ $y = X\beta + \varepsilon$

$\varepsilon \sim N(0, \sigma^2 I_n)$

$H\beta = c$ → Linear Constraints

→ Hypothesis testing :

{ $H_0 : H\beta = c$

$H_A : H\beta \neq c$

$\Longrightarrow$ $\dfrac{\frac{SSE_H - SSE}{q}}{\frac{SSE}{n-m-1}} \sim F(q, n-m-1)$

Rmk: Previous $\dfrac{SST - SSE}{SSE} \sim F$ is a special case of this result !

Here, we summarize on the scenarios that violate assumptions

1) Briefly discuss

① different variances → [e.g.] { poor / rich } Consumption

② co-variance ≠ 0 → observations are co-related

③ Multi-collinearity → features are too many

Property: (with respect to Design Matrix X)

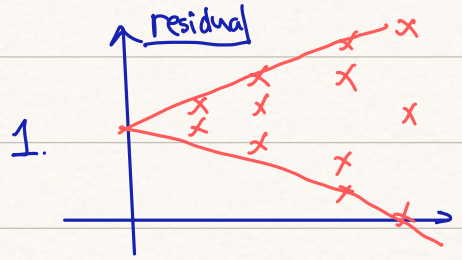
Row-wise Issue:	① + ②	→ observation-wise
Column-wise Issue:	③	→ attribute-wise

2) Details understanding (TBC 11/14)

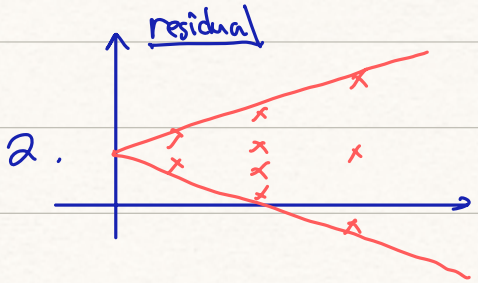
12/12

① Heterogeneity (different Variance) ★ $var[\varepsilon_i] \neq \sigma^2$

a) Residual-plot



if this situation exists for some x_i



b) Rank-test around same variance

(Non-parametric Statistical Testing)

→ Spearman Test

c) Recipe → WLS (weighted least square)

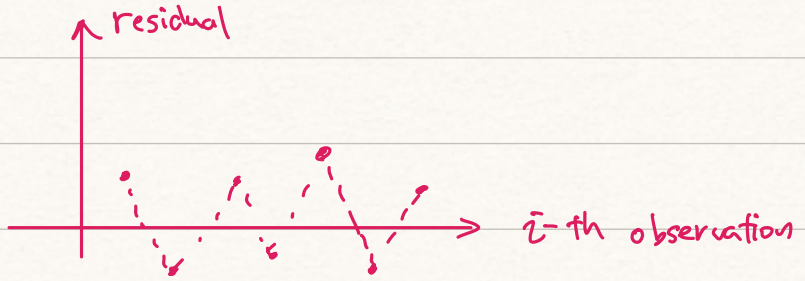
we should estimate the variance (weight)

Power Function $\boxed{w_i = \frac{1}{\sigma_i^2} \approx \frac{1}{k x_{ji}^2}}$ ← weight

Here, x_{ji} → j-th attribute for i-th observation

② Cor-related issue → $cov(\varepsilon_i, \varepsilon_j) \neq 0$ ★

a) Residual-Plot



b) Auto-correlation Coefficient (measure) → first-order correlation

$$\rho = \frac{\sum_{t=2}^{n} \varepsilon_t \cdot \varepsilon_{t-1}}{\sqrt{\sum_{t=2}^{n} \varepsilon_t^2} \sqrt{\sum_{t=2}^{n} \varepsilon_{t-1}^2}} \in [-1, 1]$$

Recap: if $E[X] = E[Y] = 0$

then $corr(X, Y) = \frac{E[X] \cdot E[Y]}{\sqrt{E[X^2]} \cdot \sqrt{E[Y^2]}} \in [-1, 1]$

$\varepsilon_t \leftarrow e_t$

c) DW test $\longrightarrow$ $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$

$$\begin{cases} H_0: \rho = 0 \\ H_A: \rho \neq 0 \end{cases}$$

consider $DW = \dfrac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=2}^{n} \varepsilon_t^2} \approx 2(1-\rho)$

Rmk: Under Null Hypothesis, $DW = \dfrac{\sum_{t=2}^{n} (u_t - u_{t-1})^2}{\sum_{t=2}^{n} u_t^2}$

$\{u_t\}_{t=1}^{n} \longrightarrow$ i.i.d Gaussian

interpretation of the effectiveness

d) Recipe

1. add attributes

2. Assume it is first-order correlated.

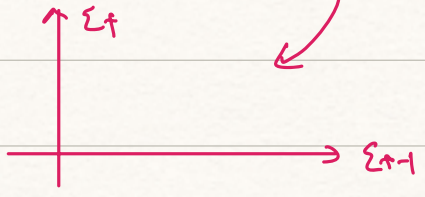
then apply some models to handle this issue!

keep on iterating until pass DW Test!

2. one example:

Assume $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$ (suppose it is correct)

① find out ρ



although in real-application, it is highly likely that this is incorrect!

② modify the Regression Model:

$$Y_t - \rho Y_{t-1} = (\beta_0 - \rho\beta_0) + \beta_1 (X_t - \rho X_{t-1}) + (\varepsilon_t - \rho \varepsilon_{t-1})$$

$$\Updownarrow$$

$$\underbrace{Y_t - \rho Y_{t-1}}_{Y_t'} = (1-\rho)\beta_0 + \beta_1 \underbrace{(X_t - \rho X_{t-1})}_{X_t'} + u_t$$

(adjusted model)
$$\begin{cases} X_t' = X_t - \rho X_{t-1} \\ Y_t' = Y_t - \rho Y_{t-1} \end{cases}$$

③ DW-test for Residual term

③ Col-linearity Issue (X^TX is not full-rank)

1. Judgement
   - 方差扩大因子 VIF
   - eigenvalue of X^TX (Condition Number)
     - the measure of ill-condition for one matrix

2. Recipe → Generally speaking, 2 ways
   - ↑ sample size
   - feature selection

Details :

a) PCA + LR

Idea: Thru PCA, select those 'important' principal component and conduct LR for them (subset of all PCs)
- make sure the Design Matrix is well-conditioned

Model: ① Standard LR:

$$y = \beta_0 \cdot \mathbb{1}_n + X\beta + \varepsilon$$

② PCA LR:

$$y = \beta_0 \cdot \mathbb{1}_n + X \cdot Q \cdot Q^T \beta + \varepsilon$$
$$:= \beta_0 \mathbb{1}_n + Z \cdot \alpha + \varepsilon$$

Then $Z^T Z = Q^T X^T X Q = \text{diag}(\lambda_1, \ldots, \lambda_p)$

idea: if $(Z^T Z)_{ii} = \lambda_i$ is small.

then it means the i-th attribute is almost constant for all observation $\Rightarrow$ $\hat{\alpha}_i$ can be set to 0

Conclusion: $\hat{\alpha} = \begin{pmatrix} \hat{\alpha}_1 \\ 0 \end{pmatrix} \Rightarrow \beta^* = Q\hat{\alpha}$

assume $\hat{\alpha}_1 \in \mathbb{R}^r$

$Z = (Z_1, Z_2)$

$Z_1 \in \mathbb{R}^{n \times r}$

$$= (Q_1 \; Q_2) \begin{pmatrix} \hat{\alpha}_1 \\ 0 \end{pmatrix}$$
$$= Q_1 \hat{\alpha}_1$$
$$= Q_1 (Z_1^T Z_1)^{-1} Z_1^T y$$
$$= Q_1 \Lambda_r^{-1} Z_1^T y$$
$$= Q_1 \Lambda_r^{-1} (X \cdot Q_1)^T y$$
$$= Q_1 \Lambda_r^{-1} Q_1^T X^T y$$
$$= Q_1 \Lambda_r^{-1} Q_1^T X^T X \hat{\beta}$$
$$= Q_1 \Lambda_r^{-1} Q_1^T Q \Lambda Q^T \hat{\beta}$$
$$= Q_1 \Lambda_r^{-1} (I_r \; 0) \Lambda Q^T \hat{\beta}$$
$$= Q_1 Q_1^T \hat{\beta}$$

$(I_r \; 0)$

Recap: $\hat{\beta} = (X^T X)^{-1} X^T y$

$\Rightarrow (X^T X) \cdot \hat{\beta} = X^T y$

Collary { 1. $\|\beta^*\| < \|\hat{\beta}\|$

2. $\exists r$. s.t $MSE(\beta_r^*) < MSE(\hat{\beta})$

b) Ridge Regression

idea: 1. bias - variance Trade-off

LR estimator → { unbiased

extremely large variance if X^TX is ill-conditioned

Surrogation → estimator that { small bias

small variance

2. Small component-wise value prior

→ we prefer the estimator that $\|\beta\|$ is small

Details: $\min_w \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$

penalty term

$\Rightarrow \hat{\beta}_{RR}(\lambda) = (X^TX + \lambda I)^{-1} X^T y$.

Property: ① $\|\hat{\beta}_{RR}(\lambda)\| \leq \|\hat{\beta}\|$ for $\forall \lambda \geq 0$

② $\hat{\beta}_{RR}(\lambda)$ is a biased estimator for β

③ $\exists \lambda$ s.t $MSE(\hat{\beta}_{RR}(\lambda)) < MSE(\hat{\beta})$

Rmk: The Proof for Ridge Regression Property is strongly associated with the canonical form of Regression:

That is, **Standard form**

$$y = \beta_0 \mathbb{1}_n + X\beta + \varepsilon$$

Canonical form Consider $Z = X \cdot Q$

$$\alpha = Q^T \beta$$

where $Q^T (X^T X) Q = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$

$$y = \beta_0 \mathbb{1}_n + Z\alpha + \varepsilon$$

Here ① $\hat{\alpha}_{RR}(\lambda) = Q^T \hat{\beta}_{RR}(\lambda)$

② $$\begin{aligned} \hat{\alpha}_{RR}(\lambda) &= (Z^T Z + \lambda I)^{-1} Z^T y \\ &= (Z^T Z + \lambda I)^{-1} Z^T Z \cdot \hat{\alpha} \\ &= (\Lambda + \lambda I)^{-1} \Lambda \hat{\alpha} \end{aligned}$$

#