Multicollinearity

(1) Transformation:

① Linear Model:

<u>Thm.</u> For $Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I_n)$

If $C(X_1) \perp C(X_2)$. Then LSE of β_1, β_2: $\hat{\beta}_1, \hat{\beta}_2$

satisfies $X_1\hat{\beta}_1 = M_1 Y$. $X_2\hat{\beta}_2 = M_2 Y$. $M_1 = P_{C(X_1)}$. $M_2 = P_{C(X_2)}$

Moreover. if $C(X)$ is full rank. then β_1, β_2 are estimable. $\hat{\beta}_i = (X_i^T X_i)^{-1} X_i^T Y$.

<u>Pf:</u> $M = M_1 + M_2^* = M_1^* + M_2$.

$C(X_1) \perp C(X_2) \Rightarrow C(M_2^*) = C((I - M_1)X_2) = C(X_2)$

$\therefore M = M_1 + M_2$. $X_i\hat{\beta}_i = M_i X\hat{\beta} = M_i Y$.

$C(X)$ is full rank $\Rightarrow \beta$ is estimable.

So as $(I_t \ 0)\beta$, $(0 \ I_q)\beta$.

<u>Rmk:</u> Apply on general case: ($C(X)$ full rank)

$$Y = X\beta + \varepsilon = X_1\beta_1 + M_1 X_2\beta_2 + (I - M_1)X_2\beta_2 + \varepsilon$$

$$= X_1(\beta_1 + (X_1^T X_1)^{-1} X_1^T X_2\beta_2) + (I - M_1)X_2\beta_2 + \varepsilon.$$

$$=: X_1\delta_1 + (I - M_1)X_2\beta_2 + \varepsilon.$$

satisfies the condition of Thm. So that:

$$\left\{ \begin{array}{l} \hat{\delta}_1 = (X_1^T X_1)^{-1} X_1^T Y \\ \hat{\beta}_2 = (X_2^T (I - M_1) X_2)^{-1} X_2^T (I - M_1) Y \end{array} \right. \Rightarrow \text{Solve } \hat{\beta}_1$$

we obtain: $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} (X_1^T X_1)^{-1} X_1^T (Y - X_2 \hat{\beta}_2) \\ (X_2^T (I - M_1) X_2)^{-1} X_2^T (I - M_1) Y \end{pmatrix}$

<u>prop.</u> For $Y = (J_n \ X) \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} + \varepsilon = J_n \beta_0 + X\beta + \varepsilon$. $P_n = M_{J_n}$

$\varepsilon \sim N(0, \sigma^2 I_n)$. $r(J_n \ X) = p+1$ (full rank)

$\Rightarrow \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \frac{1}{n} J_n^T Y - \frac{1}{n} J_n^T X (X^T (I - P_n) X)^{-1} X^T (I - P_n) Y \\ (X^T (I - P_n) X)^{-1} X^T (I - P_n) Y \end{pmatrix}$

<u>pf:</u> Set $X_1 = J_n$. $X_2 = X$ in the remark above.

<u>Cor.</u> Set $\tilde{Y} = (Y - J_n c_0)/\lambda_0$. $\tilde{X} = (X - J_n c^T) \Lambda^{-1}$.

$c = (c_1 \cdots c_p)^T$. $\Lambda = \text{diag}\{\lambda_1 \cdots \lambda_p\}$. $r(J_n \ \tilde{X}) = p+1$.

For $\tilde{Y} = J_n \tilde{\beta}_0 + \tilde{X} \tilde{\beta} + \tilde{\varepsilon}$. $\tilde{\varepsilon} \sim N(0, \tilde{\sigma}^2 I)$.

$\Rightarrow \begin{pmatrix} \hat{\tilde{\beta}}_0 \\ \hat{\tilde{\beta}} \end{pmatrix} = \begin{pmatrix} \dfrac{\hat{\beta}_0 - c_0 + c^T \hat{\beta}}{\lambda_0} \\ \Lambda \hat{\beta} / \lambda_0 \end{pmatrix}$ where

$\hat{\beta}_0$. $\hat{\beta}$ are LSE's of initial: $Y = X\beta + J_n \beta_0 + \varepsilon$.

<u>Rmk:</u> $\tilde{Y}^T (I - P_n) \tilde{Y} = Y^T (I - P_n) Y / \lambda_0^2$

$\tilde{Y}^T (I - \tilde{M}) \tilde{Y} = Y^T (I - M) Y / \lambda_0^2 \Rightarrow$

For test $H_0: \beta = 0$ or $\tilde{H}_0: \tilde{\beta} = 0$

on initial or transformed model. Then:

$F = \tilde{F}$. (test statistics.)

② <u>Centralization:</u>

For $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots \hat{\beta}_p x_p$. $\hat{\beta}_0 = \bar{Y} - \sum_1^p \beta_k \bar{x}_k$

cross $(\bar{x}_1 \cdots \bar{x}_p, \bar{Y})$. set $\tilde{X} = (X - J_n c^T)\Lambda^{-1}$. and

$\tilde{Y} = (Y - c_0 J_n)/\lambda_0$. $c^T = \frac{1}{n} J_n^T X$. $\Lambda = I$. $c_0 = \frac{1}{n} J_n^T Y$. $\lambda_0 = 1$

$$\Rightarrow \hat{\tilde{Y}} = \sum_1^p \hat{\tilde{\beta}}_k \tilde{X}_k, \quad \begin{pmatrix} \hat{\tilde{\beta}}_0 \\ \hat{\tilde{\beta}} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 + c^T\hat{\beta} - c_0/\lambda_0 \\ \Lambda \hat{\beta}/\lambda_0 \end{pmatrix} = \begin{pmatrix} 0 \\ \hat{\beta} \end{pmatrix}$$

③ <u>Standardlization:</u>

For $Y = X\beta + \varepsilon$. set $X^* = (X - J_n \frac{J_n^T X}{n}) L_X^{-1}$. and

$Y^* = (Y - \frac{1}{n} J_n^T Y)/L_Y$. $L_X = \mathrm{diag}\{L_{11} \cdots L_{pp}\}$. $L_Y = Y^T(I - P_n)Y$.

$L_{ii} = X_j^T (I - P_n) X_j$. $X = (X_1 \cdots X_p)$.

$$\text{i.e.} \begin{cases} x_{ij}^* = x_{ij} - \bar{x}_j / \sqrt{L_{jj}} \\ Y_i^* = Y_i - \bar{Y} / \sqrt{L_Y} \end{cases} \Rightarrow \begin{pmatrix} \hat{\beta}_0^* \\ \hat{\beta}^* \end{pmatrix} = \begin{pmatrix} 0 \\ \hat{\beta}_1^* \\ \vdots \\ \hat{\beta}_p^* \end{pmatrix}$$

$\hat{\beta}_i^* = \frac{\sqrt{L_{ii}}}{\sqrt{L_Y}} \hat{\beta}_i$. by the Formula.

(2) <u>Background of Collinearity:</u>

<u>Def.</u> $\{\vec{X}_k\}_1^p$ is multicollinear if $\exists \vec{C} \neq \vec{0}_{p+1}$ s.t.

$\sum_{j=1}^p c_j x_{ij} + c_0 = 0$. $1 \le i \le n$.

<u>Rmk</u>: i) It's not common that multicollinearity exists. But it's common: $c_0 + \sum_1^p c_j x_{ij} \doteq 0$. Which's called complex MCL.

ii) Note that $Y = \beta_0 + \sum_1^p \beta_k X_k + \varepsilon$. MCL of X means: $r(X) < p$. $(X^T X)^{-1}$ doesn't exist.

If complex MCL happened $\Rightarrow r(X) = p$. but $|X^T X| \doteq 0$. $\Rightarrow D(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$ has large elements $\Rightarrow \hat{\beta}$ isn't accuracy estimation.

e.g. $(p=2)$ $\hat{Y} = \hat{\beta}_1 \tilde{x}_1 + \hat{\beta}_2 \tilde{x}_2$. $\tilde{x}_i = x_i - \bar{x}_i$.

$$\Rightarrow (X^T X)^{-1} = \frac{1}{L_{11}L_{22}(1-r_{12}^2)} \begin{pmatrix} L_{22} & -L_{12} \\ -L_{12} & L_{11} \end{pmatrix}$$

$r_{12} = L_{12} / \sqrt{L_{11}L_{22}}$. if X_1, X_2 are related a lot, then $r_{12} \uparrow 1$, $D(\hat{\beta}_1) + D(\hat{\beta}_2) \uparrow \infty$.

(3) Diagnosis:

① Variance Inflation Factor:

· $X^* = X - \frac{1}{n} J_n J_n^T X$. $(X^{*T} X^*) = (r_{ij}) = R > 0$ is the correlated matrix of X. Denote: $C = (c_{ij}) = R^{-1}$

Def: $VIF_i = c_i$ is the variance inflation factor of variable x_i.

prop. i) $Var(\hat{\beta}_i) = c_{ii}\sigma^2 / L_{ii}$. $L_{ii} = X_i^T (I - P_n) X_i$.

ii) $c_{ii} = 1/(1 - R_{i|(-i)}^2)$. $R_{i|(-i)}^2 = \dfrac{X_{(j)}^T (P_n - P_{((J_n X_{(-j)}))}) X_{(j)}}{X_{(j)}^T (I - P_n) X_{(j)}}$

Pf: i) Directly $\hat{\beta} = (X^T (I - P_n) X)^{-1} X^T (I - P_n) Y$

$$\Rightarrow Var(\hat{\beta}) = \sigma^2 (X^T (I - P_n) X)^{-1}$$

$$= \sigma^2 (\Lambda (X^{*T} X^*) \Lambda)^{-1}$$

ii) $c_{ii}^{-1} = X_{(i)}^{*T} (I - P_{((X^*_{(-i)}))}) X^*_{(i)}$

Criteria: If $\exists i$, $VIF_i \geq 10$. Then it means: x_i is heavily collinear with other x_k's.

Or if $\overline{VIF} = \sum_1^p VIF_i / p \geq 1 \Rightarrow$ Problem exists.

② By eigenvalues:

· Note that if $|X^TX| \approx 0$. Then it (X^TX) has at least one eigenvalue $\lambda_0 \approx 0$ (corresp. c.)

$\Rightarrow \quad X^TXc = \lambda_0 c \quad \therefore \quad c^TX^TXc \approx 0 \Rightarrow Xc \approx 0$

which implies: multicollinearity exists!

Rmk: Denote eigenvalues of X^TX: $\lambda_1 \geq \cdots \geq \lambda_{p+1}$.

$k_i = \frac{\lambda_1}{\lambda_i}$ is called condition index of λ_i.

Set $k = \max_i k_i$. if $k > 100$. $\Rightarrow$ Problem exists.

(4) Correction:

① Ridge Estimate:

· The idea: Since $|X^TX| \approx 0$. Add kI_p on X^TX

$\Rightarrow X^TX + kI$ may be far away from singular.

Def: $\hat{\beta}(k) = (X^TX + kI)^{-1}X^TY$ is ridge estimator.

Rmk: $\{\hat{\beta}(k)\}_{k\in\mathbb{R}}$ is a family with para. k.

Denote: $\phi = (l_1 \cdots l_p)$ is matrix of orthonormal eigenfun's of X^TX. i.e. $\phi^TX^TX\phi = \text{diag}\{\lambda_i\}_1^p \triangleq \Lambda$. Set $Z = X\phi$. $\alpha = \phi^T\beta$.

Rmk: $Y = Z\alpha + \varepsilon$. $\hat{\alpha}(k) = (Z^TZ + kI)^{-1}Z^TY$.

$\hat{\beta}(k) = \phi\hat{\alpha}(k) \Rightarrow \|\hat{\beta}(k)\| = \|\hat{\alpha}(k)\|$

$\|\cdot\|$ is: $\|\vec{x}\| = (\sum_i x_i^2)^{\frac{1}{2}}$.

i) <u>properties</u>:

i) $\hat\beta(k)$ is biased estimator of β. if $k \neq 0$

ii) If k is indept with Y. Then $\hat\beta(k)$ is linear transform of $\hat\beta$ and Y

<u>Pf</u>: $\hat\beta(k) = (X^TX + kI)^{-1} X^T X \hat\beta$.

<u>Rmk</u>: Commonly. k should depend on data Y.

iii) For $\|\hat\beta\| \neq 0$. $k>0 \Rightarrow \|\hat\beta(k)\| < \|\hat\beta\|$.

<u>Pf</u>: $\|\hat\beta(k)\| = \|\hat\alpha(k)\| = \|(\Lambda + kI)^{-1}\Lambda\hat\alpha\|$
$< \|\hat\alpha\| = \|\hat\beta\|$.

<u>Rmk</u>: $R_k(\beta) = \beta(k)$ is a contraction
$k \to \infty \to \hat\beta(k) \to \vec{0}$.

iv) $\exists\, k>0$. $MSE(\hat\beta(k)) < MSE(\hat\beta)$ where
$MSE_\theta(\hat\theta) = E\|\hat\theta - \theta\|^2 = E((\hat\theta-\theta)^T(\hat\theta-\theta))$

<u>Pf</u>: $MSE(\hat\beta(k)) = MSE_\alpha(\hat\alpha(k))$
$= tr(Var(\hat\alpha(k))) + \|E(\hat\alpha(k)) - \alpha\|^2$

$$\begin{cases} Var(\hat\alpha(k)) = \sigma^2(\Lambda+kI)^{-1}\Lambda(kI+\Lambda)^{-1} \\ E(\hat\alpha(k)) = (\Lambda+kI)^{-1}\Lambda\alpha \end{cases}$$

$$\Rightarrow MSE(\hat\beta(k)) = \sigma^2 \sum_1^p \frac{\lambda_i}{(\lambda_i+k)^2} + k^2\sum_1^p \frac{\alpha_i^2}{(\lambda_i+k)^2}.$$

ii) <u>Choice for k</u>:

i) By trace of $\hat\beta(k)$ in Plot:

choose k to make sign of $\hat{\beta}(k)_j$ reasonable.

and SSE won't increase so much. (Since $\hat{\beta}(k)$ deviates $\hat{\beta}$ a lot for large $|k|$.)

ii) By VIF:

$$Var(\hat{\beta}(k)) = \sigma^2 (X^T X + kI)^{-1} X^T X (X^T X + kI)^{-1}.$$
$$\triangleq \sigma^2 (C_{ij}(k))$$

when $k \uparrow$. $C_{ii}(k) = VIF_i(k) \downarrow$.

Choose k st. $VIF_i(k) \leq 10.\ \forall i$.

iii) By SSE:

Set a const: c. st. $SSE(k) < c\,SSE.\ (c > 1)$.

iv) By Hoerl-Kennad Formula:

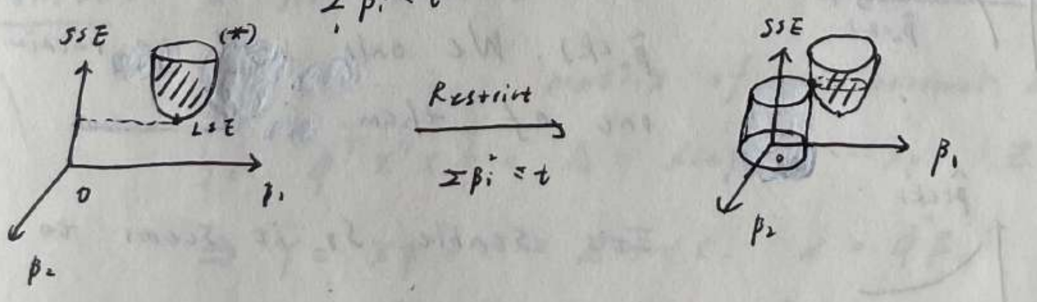
Denote $f(k) = MSE(\hat{\beta}(k))$. $f'(k) = 2\sum \frac{\lambda_i}{(\lambda_i + k)^3}(k\alpha_i^2 - \sigma^2)$

Choose $\hat{k} = [\sigma^2 / \max_i \alpha_i^2]$. Since $f(k) \downarrow$ when $k \uparrow$ in $[0, \hat{k}]$. $\Rightarrow$ minimize $MSE_p(\hat{\beta}(k))$

iii) <u>Geometric Intepretion:</u>

· Generally. fix $c > 1$. st. $(Y - X\hat{\beta}(k))^T (Y - X\hat{\beta}(k)) < c\, Y^T (I - M) Y$. (Guarantee $SSE(k)$ won't be large)

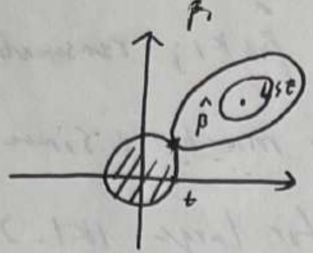
$$\Rightarrow \hat{\beta}^{ridge} =: \underset{\sum_1^p \beta_i^2 \leq t}{\arg\min} \sum_1^n (y_i - \beta_0 - \sum \beta_j x_{ij})^2$$



(*): SSE is quadratic func. of $\hat{\beta}$.

Consider the cylinder $\sum \beta_i^2 \leq t$ intersects $f(\vec{\beta})$

i.e.



$\hat\beta$ is the ridge estimate under restrict of $\sum \beta_i^2 = t$.

If we see $\hat\beta(k)$ is contraction of $\hat\beta$.

$\Rightarrow \hat\beta^{ridge} = \underset{\|\beta\| = c\|\hat\beta\|}{\arg\min} \|Y - X\beta\|^2 \quad (\|\cdot\| = \|\cdot\|_{L^2})$

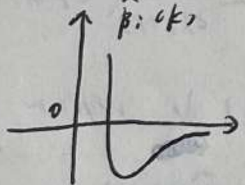
$= \underset{\|\beta\| = c\|\hat\beta\|}{\arg\min} \|X\hat\beta - X\beta\|^2$ (By Lagrange)

Rmk: Another form: $\hat\beta^{ridge} = \underset{\beta}{\arg\min} (\|Y - X\beta\|^2 + \lambda\|\beta\|^2)$

iv) Select variables:

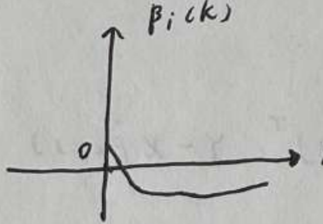
Analysis by trace:

i)



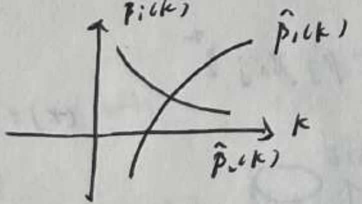
When $k\uparrow$, $\hat\beta_i(k) \Downarrow$ rapidly and changes its sign. It's unstable. Besides, $\hat\beta_i(k) \to 0$ means it loses predictable ability.

ii)



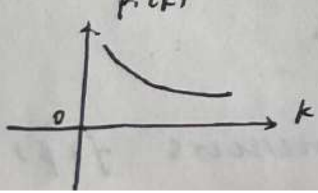
when $k\uparrow$. Its influential ability $\uparrow$.

iii)



It means: There's a strong relation between $\hat\beta_1(k)$ and $\hat\beta_2(k)$. We only need to retain one of them.

iv)



It's stable. So it seems to be reasonable.

v) <u>Generalization:</u>

Consider $\hat{\beta}(k) = \phi\,\hat{\alpha}(k)$. $\hat{\alpha}(k) = (Z^TZ+k)^{-1}Z^TY$.

for general matrix $k = \text{diag}\{k_1 \cdots k_p\}$. $k_i > 0$.

It retains the properties of $\hat{\beta}^{(ridge)}$.

<u>Rmk:</u> $\hat{k}_i = \hat{\sigma}^2/\hat{\alpha}_i^2 - \hat{\sigma}^2/\lambda_i$ is common chooices.

② <u>Stein estimate:</u>

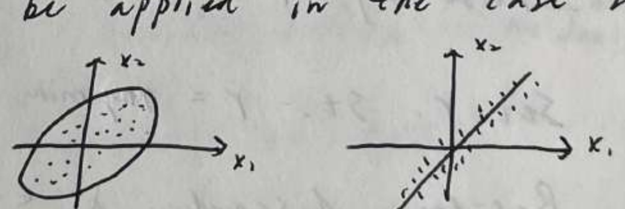
· $\hat{\beta}_s(c) = c\hat{\beta}$. $0<c<1$. ($\hat{\beta}$ is LSE). a contraction

$\exists\ 0 \le c \le 1$. st. $MSE_\beta(\hat{\beta}_s(c)) \le MSE_\beta(\hat{\beta})$

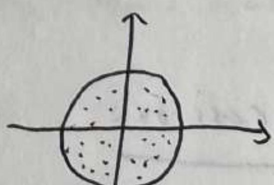
③ <u>PCA:</u>

· We want to get information from linear combination of $X = (X_1 \cdots X_p)$. So that reduces the dim of datas. (but retain most of information).

<u>Rmk:</u> It only can be applied in the case $\{X_i\}$ are correlated:



when $\{X_i\}$ is uncorrelated
PCA doesn't work:



<u>Denote:</u> X has been standardized. (so $X^TX = R$)

$\phi = (l_1 \cdots l_p)$ matrix of orthonormal eigenfunc.

st. $\phi^T X^T X \phi = \Lambda = \text{diag}\{\lambda_1 \cdots \lambda_p\}$. $Z = X\phi$.

$\Rightarrow Y = X\beta + \varepsilon = Z\alpha + \varepsilon$. $\alpha = \phi^T\beta$.

Since $|X^T X| \approx 0$. $\exists r$. s.t. $\lambda_{r+1} \cdots \lambda_p \approx 0$

i.e. $Z^T_{(p-r)} Z_{(p-r)} \approx 0$. $Z_{(p-r)} = (Z_{r+1} \cdots Z_p)$

$\Rightarrow$ Simplify the model:

$$Y = Z\alpha + \varepsilon \approx Z_{(r)} \alpha_{(r)} + \varepsilon. \quad \alpha_{(r)} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_r \end{pmatrix}$$

$$\Rightarrow \hat{\alpha}_{(r)} = (Z^T_{(r)} Z_{(r)})^{-1} Z^T_{(r)} Y = \begin{pmatrix} \sum z_{i1} y_i / \lambda_1 \\ \vdots \\ \sum z_{ir} y_i / \lambda_r \end{pmatrix}. \quad \phi \hat{\alpha}_{(r)} = \hat{\beta}_{(r)}$$

<u>prop.</u> $\sigma^2 tr(\Lambda^{-1}_{(p-r)}) \geq \| \alpha_{(p-r)} \|^2 \Rightarrow MSE(\hat{\beta}) > MSE(\hat{\beta}_{(r)})$

<u>Pf:</u> $E \| \hat{\beta}_{(r)} - \beta \|^2 = E \| \begin{pmatrix} \hat{\alpha}_{(r)} \\ 0 \end{pmatrix} - \alpha \|^2$

$$= tr(Var(\hat{\alpha}_{(r)})) + \| E\begin{pmatrix} \hat{\alpha}_{(r)} \\ 0 \end{pmatrix} - \alpha \|^2$$

$$= tr(\sigma^2 \Lambda^{-1}_{(r)}) + \| \alpha_{(p-r)} \|^2.$$

$$E(\| \hat{\beta} - \beta \|^2) = tr(\sigma^2 \Lambda^{-1}).$$

<u>Rmk:</u> Choise of r: Fix $c \in (0,1)$.

Set r. s.t. $r = \arg\min \sum_1^r \lambda_i / \sum_1^p \lambda_i \geq c$

Besides, discard $\lambda_i < 0.01$.

④ <u>Lasso Regression:</u>

$$\hat{\beta} = \arg\min_{\beta} \sum_1^n (y_i - \beta_0 - \sum \beta_j x_{ij})^2 + \lambda \sum |\beta_i| . \text{ fix } \lambda.$$

<u>Rmk:</u> Modification:

A-lasso: $\hat{\beta} = \arg\min \| Y - X\beta \|_2^2 + \lambda \sum w_i |\beta_i|$

Elastic-net: $\hat{\beta} = \arg\min \| Y - X\beta \|_2^2 + \lambda_1 \sum |\beta_i| + \lambda_2 \sum \beta_i^2$.