MS BGD
MDI 720 : Ridge / Tikhonov

Joseph Salmon
http://josephsalmon.eu
Télécom ParisTech

Plan

Sommaire

Rappel

$$\boxed{\mathbf{y} = X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}}$$

- $\mathbf{y} \in \mathbb{R}^n$ est le vecteur des observations
- $X \in \mathbb{R}^{n \times p}$ est la matrice des variables explicatives
- $\boldsymbol{\theta}^\star \in \mathbb{R}^p$ est le **vrai** paramètre du modèle que l'on veut retrouver.
- $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ est le bruit

Rem: possiblement une variable supplémentaire pour la constante

La décomposition en valeur singulières

Théorème : Golub et Van Loan (2013)

Pour toute matrice $X \in \mathbb{R}^{n \times p}$, il existe deux matrices orthogonales $U = [\mathbf{u}_1, \dots, \mathbf{u}_n] \in \mathbb{R}^{n \times n}$ et $V = [\mathbf{v}_1, \dots, \mathbf{v}_p] \in \mathbb{R}^{p \times p}$, telles que

$$U^\top X V = \mathrm{diag}(s_1, \dots, s_{\mathrm{rg}(X)}) = \Sigma \in \mathbb{R}^{n \times p}$$

avec $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_{\mathrm{rg}(X)} > 0$, avec $\mathrm{rg}(X) = \mathrm{rang}(X)$.

$$X = U \Sigma V^\top \Leftrightarrow X = \sum_{i=1}^{\mathrm{rg}(X)} s_i \mathbf{u}_i \mathbf{v}_i^\top$$

Une solution des moindres carrés est alors :

$$\hat{\boldsymbol{\theta}}^{\mathrm{MCO}} = X^+ \mathbf{y} = \sum_{i=1}^{\mathrm{rg}(X)} \frac{1}{s_i} \mathbf{v}_i \mathbf{u}_i^\top \mathbf{y}$$

Retour sur les problèmes numériques

$$\hat{\boldsymbol{\theta}}^{\mathrm{MCO}} = X^{+}\mathbf{y} = \sum_{i=1}^{\mathrm{rg}(X)} \frac{1}{s_i}\mathbf{v}_i\mathbf{u}_i^{\top}\mathbf{y}$$

Si les plus petites valeurs singulières s_i s'approchent de zéro alors la solution numérique de la SVD n'est pas stable !

Rem: le défaut n'est pas propre aux moindres carrés, mais inhérent aux problèmes difficiles (on dit aussi "mal posé" en analyse numérique et en traitement du signal)

Les équations normales

Une solution $\boldsymbol{\theta}$ des moindres carrés doit vérifier :

$$X^\top X\boldsymbol{\theta} = X^\top \mathbf{y} \Leftrightarrow V\Sigma^\top\Sigma V^\top\boldsymbol{\theta} = V\Sigma^\top U^\top\mathbf{y}$$

et si l'on cherche $\boldsymbol{\theta}$ sous la forme $\boldsymbol{\theta} = V\boldsymbol{\beta}$, c'est équivalent à

$$\Sigma^\top\Sigma\boldsymbol{\beta} = \Sigma^\top U^\top\mathbf{y}$$

$\Sigma^\top\Sigma$ diagonale avec $r = \mathrm{rang}(X)$ éléments non nuls qui sont les s_i^2

$$\Sigma^\top\Sigma = \left[\begin{array}{ccc|c} s_1^2 & & 0 & \\ & \ddots & & \mathbf{0} \\ 0 & & s_r^2 & \\ \hline & \mathbf{0} & & \mathbf{0} \end{array}\right] \in \mathbb{R}^{p\times p}$$

Les équations normales (suite)

Alternative régularisée : résoudre les équations normales

$$\left[\begin{array}{ccc|c} s_1^2 & & 0 & \\ & \ddots & & 0 \\ 0 & & s_r^2 & \\ \hline & 0 & & 0 \end{array}\right] \text{ remplacé par } \left[\begin{array}{ccc|c} s_1^2 & & 0 & \\ & \ddots & & 0 \\ 0 & & s_r^2 & \\ \hline & 0 & & 0 \end{array}\right] + \lambda \operatorname{Id}_p$$

De manière synthétique cela s'écrit : $(\lambda \operatorname{Id}_p + \Sigma^\top \Sigma)\boldsymbol{\beta} = \Sigma^\top U^\top \mathbf{y}$

i.e., on ajoute à toutes les valeurs propres de $X^\top X$ un terme $\lambda > 0$ "petit", λ est nommé **paramètre de régularisation**

$$\boldsymbol{\beta} = (\lambda \operatorname{Id}_p + \Sigma^\top \Sigma)^{-1} \Sigma^\top U^\top \mathbf{y}$$

et donc

$$\boldsymbol{\theta} = V(\lambda \operatorname{Id}_p + \Sigma^\top \Sigma)^{-1} \Sigma^\top U^\top \mathbf{y}$$

Ridge forme explicite

Avec la SVD, l'équation suivante se simplifie :

$$\boldsymbol{\theta} = V(\lambda \operatorname{Id}_p + \Sigma^\top \Sigma)^{-1} \Sigma^\top U^\top \mathbf{y}$$

Cela donne une première forme de l'estimateur *Ridge*

$$\boxed{\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} = (\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \mathbf{y}}$$

Rappel : sous l'hypothèse de plein rang $\hat{\boldsymbol{\theta}}^{\mathrm{MCO}} = (X^\top X)^{-1} X^\top \mathbf{y}$

Rem:

$$\lim_{\lambda \to 0^+} \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} = \hat{\boldsymbol{\theta}}^{\mathrm{MCO}}$$

$$\lim_{\lambda \to +\infty} \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} = 0 \in \mathbb{R}^p$$

Astuce du noyau

Astuce du noyau (🇬🇧 : *Kernel trick*) : Selon si $n > p$ ou $n \leqslant p$, une méthode qui cherche à trouver une solution de Ridge par inversion peut préférer l'une des deux formulations suivantes :

$$\boxed{X^\top (XX^\top + \lambda \operatorname{Id}_n)^{-1} \mathbf{y} = (X^\top X + \lambda \operatorname{Id}_p)^{-1} X^\top \mathbf{y}}$$

- membre de gauche : on résout un système $n \times n$
- membre de droite : on résout un système $p \times p$

Rem: cette propriété est aussi très utile pour les méthodes à noyaux de type SVM (*cf.* cours de *Machine Learning*)

Exo: Démontrer la propriété précédente avec la SVD

Sommaire

Ridge / Tikhonov : la définition pénalisée

$$\boxed{\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p} \left(\underbrace{\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\text{attache aux données}} + \underbrace{\lambda \|\boldsymbol{\theta}\|_2^2}_{\text{régularisation}} \right)}$$

- Noter que l'estimateur *Ridge* est **unique** pour un λ fixé
- On retrouve de nouveau les cas limites :

$$\begin{aligned} \lim_{\lambda \to 0} \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} &= \hat{\boldsymbol{\theta}}^{\mathrm{MCO}} \text{(solution de norme } \|\cdot\|_2 \text{ minimale)} \\ \lim_{\lambda \to +\infty} \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} &= 0 \in \mathbb{R}^p \end{aligned}$$

- Lien des deux formulations par les CNO : pour

$$f(\boldsymbol{\theta}) = \frac{\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}{2} + \frac{\lambda \|\boldsymbol{\theta}\|_2^2}{2}$$

$$\nabla f(\boldsymbol{\theta}) = X^\top (X\boldsymbol{\theta} - \mathbf{y}) + \lambda \boldsymbol{\theta} = 0 \Leftrightarrow (X^\top X + \lambda \operatorname{Id}_p)\boldsymbol{\theta} = X^\top \mathbf{y}$$

Interprétation contrainte

Un problème de la forme "Lagrangienne" suivante :

$$\boxed{\underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg\min} \left(\underbrace{\frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\text{attache aux données}} + \underbrace{\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2}_{\text{régularisation}} \right)}$$

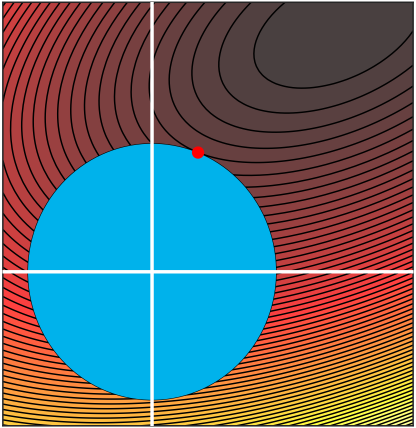
admet pour un certain $T > 0$ la même solution que :

$$\begin{cases} \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg\min} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 \\ \text{t.q. } \|\boldsymbol{\theta}\|_2^2 \leqslant T \end{cases}$$

<u>Rem</u>: le lien $T \leftrightarrow \lambda$ n'est pas explicite !

- Si $T \to 0$ on retrouve le vecteur nul : $0 \in \mathbb{R}^p$
- Si $T \to \infty$ on retrouve $\hat{\boldsymbol{\theta}}^{\mathrm{MCO}}$ (non contraint)

Lignes de niveau et ensemble de contraintes



Optimisation sous contraintes ℓ_2

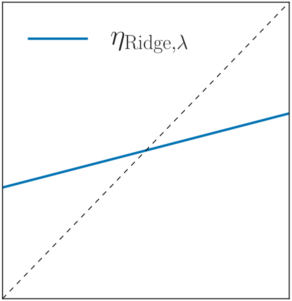
Le cas orthogonal

Retour sur un cas simple $X^\top X = \mathrm{Id}_p$

$$\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} = (\lambda \,\mathrm{Id}_p + X^\top X)^{-1} X^\top \mathbf{y}$$

$$\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} = (\lambda \,\mathrm{Id}_p + \mathrm{Id}_p)^{-1} X^\top \mathbf{y} = \frac{1}{\lambda + 1} X^\top \mathbf{y}$$

$$\hat{\mathbf{y}} = \frac{1}{\lambda + 1}\mathbf{y} = \left(\eta_{\mathrm{rdg},\lambda}(\mathbf{y}_i)\right)_{i=1,\dots,n}$$



Rem: la fonction réelle $\eta_{\mathrm{rdg},\lambda}$ est une contraction linéaire (shrinkage)

Prédiction associée

Partant du coefficient *Ridge* :

$$\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} = (\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \mathbf{y}$$

la prédiction associée s'obtient ainsi :

$$\hat{\mathbf{y}} = X \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} = X(\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \mathbf{y}$$

Rem: l'estimateur $\hat{\mathbf{y}}$ est toujours linéaire en $\mathbf{y}$

Rem: l'équivalent de la matrice chapeau (: *hat matrix*) est

$$H_\lambda = X(\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top = \sum_{j=1}^{\mathrm{rg}(X)} \frac{s_j^2}{s_j^2 + \lambda} \mathbf{u}_j \mathbf{u}_j^\top$$

Attention : si $\lambda \neq 0$, on n'a plus $H_\lambda^2 = H_\lambda = \sum_{j=1}^{\mathrm{rg}(X)} \mathbf{u}_j \mathbf{u}_j^\top$, *i.e.*, H_λ n'est donc pas un projecteur

Point normalisation et centrage

Rappel : normaliser les p variables de la même manière pour que la pénalisation contraigne de manière similaire toutes les variables

- centrer l'observation et les variables explicatives $\Rightarrow$ pas de coefficient pour la variable constante (donc pas de contrainte)
- ne pas centrer les variables explicatives $\Rightarrow$ ne pas mettre de contrainte sur la variable constante (: *bias/intercept*),

$$\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg\min} \|\mathbf{y} - X\boldsymbol{\theta} - \theta_0 \mathbf{1}_n\|^2 + \lambda \sum_{j=1}^p \theta_j^2$$

Alternative (si l'on n'a pas normalisé) : changer la pénalité en

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg\min} \|\mathbf{y} - X\boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^p \alpha_j \boldsymbol{\theta}_j^2 \quad (\textit{e.g.}, \ \ \alpha_j = \|\mathbf{x}_j\|^2)$$

Rem: pour la validation croisée on utilisera plus naturellement $\frac{\|\mathbf{y}-X\boldsymbol{\theta}\|^2}{2n}$ que $\frac{\|\mathbf{y}-X\boldsymbol{\theta}\|^2}{2}$ pour conserver l'amplitude de λ

Point normalisation et centrage (bis)

Ici $X = [\frac{\mathbb{1}_{a_1}}{\sqrt{n_1}}, \dots, \frac{\mathbb{1}_{a_K}}{\sqrt{n_K}}]$ ($\mathbf{x}_k = \frac{\mathbb{1}_{a_j}}{\sqrt{n_k}}$,*i.e.*, on), avec $\hat{\mu}_k = \frac{1}{n_k} \sum_{x_i = a_k} y_i$ et $n_k = \#\{x_i = a_k\}$

L'estimateur Ridge (sans pénalité sur la constante) est solution de

$$(\hat{\theta}_0, \hat{\theta}_1, \dots, \hat{\theta}_K)^\top = (\hat{\theta}_0, \tilde{\boldsymbol{\theta}})^\top \in \underset{\theta_0, \dots, \theta_p}{\arg\min}\, f(\theta_0, \dots, \theta_p)$$

$$\text{avec} \quad f(\theta_0, \dots, \theta_K) = \left\| \mathbf{y} - \theta_0 \mathbf{1}_n - \sum_{j=1}^K \theta_j \mathbf{x}_j \right\|^2 + \lambda \sum_{j=1}^K \theta_j^2$$

Partant de $X^\top X = \mathrm{Id}_K$ et $X^\top \mathbf{y} = (\sqrt{n_1}\hat{\mu}_1, \dots, \sqrt{n_K}\hat{\mu}_K)$, on a :

$$\tilde{\boldsymbol{\theta}} = (X^\top X + \lambda \, \mathrm{Id}_K)^{-1} X^\top (\mathbf{y} - \hat{\theta}_0 \mathbf{1}_n) = \frac{1}{1+\lambda} \begin{pmatrix} \sqrt{n_1}(\hat{\mu}_1 - \hat{\theta}_0) \\ \vdots \\ \sqrt{n_K}(\hat{\mu}_K - \hat{\theta}_0) \end{pmatrix}$$

Suite

$$\text{CNO de Ridge : } \begin{cases} \hat{\theta}_0 = \frac{1}{n}\langle \mathbf{1}_n, \mathbf{y} - X\tilde{\boldsymbol{\theta}}\rangle & (1) \\ \tilde{\boldsymbol{\theta}} = \frac{1}{\lambda}X^\top(\mathbf{y} - X\tilde{\boldsymbol{\theta}} - \mathbf{1}_n\hat{\theta}_0) & (2) \end{cases}$$

Avec $e = (\sqrt{n_1}, \ldots, \sqrt{n_K})$, alors $Xe = \mathbf{1}_n$ et $\mathbf{1}_n^\top X = e^\top$. Avec (2)

$$\langle e, \tilde{\boldsymbol{\theta}}\rangle = \frac{1}{1+\lambda} e^\top X^\top(\mathbf{y} - \mathbf{1}_n\hat{\theta}_0 - X\tilde{\boldsymbol{\theta}}) = \frac{1}{1+\lambda}\mathbf{1}_n^\top(\mathbf{y} - X\tilde{\boldsymbol{\theta}} - \mathbf{1}_n\hat{\theta}_0)$$

$$\langle e, \tilde{\boldsymbol{\theta}}\rangle = \frac{1}{1+\lambda}(n\hat{\theta}_0 - n\hat{\theta}_0) = 0$$

Avec (1) on déduit que : $\boxed{\hat{\theta}_0 = \bar{y}_n}$ puis que pour une nouvelle observation $x_{n+1} = a_k$ on a : $\boxed{y_{n+1} = \frac{1}{1+\lambda}(\lambda\bar{y}_n + \hat{\mu}_k)}$

Rem: λ permet d'osciller entre le prédicateur globale ($\bar{y}_n$, si $\lambda = +\infty$) et le prédicateur par modalité ($\hat{\mu}_k$, si $\lambda = 0$)

Rem: si l'on ne prend pas en compte la constante, $y_{n+1} = \frac{1}{1+\lambda}\hat{\mu}_k$, et donc pour λ grand on prédit 0 ! ! !

Sommaire

Le biais dans le cas général

Sous l'hypothèse de bruit "blanc" $\mathbf{y} = X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}$ avec $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$:

$$
\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) =& \mathbb{E}\left((\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \mathbf{y}\right) \\
=& \mathbb{E}\left((\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top X \boldsymbol{\theta}^\star + (\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}\right) \\
=& (\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top X \boldsymbol{\theta}^\star \\
=& \sum_{i=1}^{\mathrm{rg}(X)} \frac{s_i^2}{s_i^2 + \lambda} \mathbf{v}_i \mathbf{v}_i^\top \boldsymbol{\theta}^\star
\end{aligned}
$$

Rem: on retrouve $\mathbb{E}\left(\hat{\boldsymbol{\theta}}^{\mathrm{MCO}}\right) = \sum_{i=1}^{\mathrm{rg}(X)} \mathbf{v}_i \mathbf{v}_i^\top \boldsymbol{\theta}^\star$ quand $\lambda \to 0$

Rem: le biais vaut $-\lambda(X^\top X + \lambda \operatorname{Id}_p)^{-1} \boldsymbol{\theta}^\star$ (grâce à la 3e ligne)

Sommaire

Variance dans le cas général

Sous l'hypothèse de bruit "blanc" (*i.e.*, $\mathbb{E}(\varepsilon) = 0$) et de modèle homoscédastique : $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \operatorname{Id}_n$

Variance / Covariance

$$V_\lambda^{\mathrm{rdg}} = \mathbb{E}\left((\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \mathbb{E}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}))(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \mathbb{E}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}})^\top \right)$$

Calcul explicite :

$$\begin{aligned} V_\lambda^{\mathrm{rdg}} &= \mathbb{E}\left((\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X (\lambda \operatorname{Id}_p + X^\top X)^{-1} \right) \\ &= (\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \mathbb{E}(\varepsilon\varepsilon^\top) X (\lambda \operatorname{Id}_p + X^\top X)^{-1} \end{aligned}$$

Variance dans le cas général

Sous l'hypothèse de bruit "blanc" (*i.e.*, $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$) et de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Variance / Covariance

$$V_\lambda^{\mathrm{rdg}} = \mathbb{E}\left((\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \mathbb{E}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}))(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \mathbb{E}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}})^\top \right)$$

Calcul explicite :

$$\begin{aligned} V_\lambda^{\mathrm{rdg}} &= \mathbb{E}\left((\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X (\lambda \operatorname{Id}_p + X^\top X)^{-1} \right) \\ &= (\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) X (\lambda \operatorname{Id}_p + X^\top X)^{-1} \\ &= \sigma^2 (\lambda \operatorname{Id}_p + X^\top X)^{-2} X^\top X \end{aligned}$$

Variance dans le cas général

Sous l'hypothèse de bruit "blanc" (*i.e.*, $\mathbb{E}(\varepsilon) = 0$) et de modèle homoscédastique : $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \operatorname{Id}_n$

Variance / Covariance

$$V_\lambda^{\mathrm{rdg}} = \mathbb{E}\left((\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \mathbb{E}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}))(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \mathbb{E}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}})^\top\right)$$

Calcul explicite :

$$\begin{aligned} V_\lambda^{\mathrm{rdg}} &= \mathbb{E}\left((\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \varepsilon\varepsilon^\top X (\lambda \operatorname{Id}_p + X^\top X)^{-1}\right) \\ &= (\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \mathbb{E}(\varepsilon\varepsilon^\top) X (\lambda \operatorname{Id}_p + X^\top X)^{-1} \\ &= \sigma^2 (\lambda \operatorname{Id}_p + X^\top X)^{-2} X^\top X \\ &= \sum_{i=1}^{\operatorname{rg}(X)} \frac{s_i^2 \sigma^2}{(s_i^2 + \lambda)^2} \mathbf{v}_i \mathbf{v}_i^\top \end{aligned}$$

Variance dans le cas général

Sous l'hypothèse de bruit "blanc" (*i.e.*, $\mathbb{E}(\varepsilon) = 0$) et de modèle homoscédastique : $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \operatorname{Id}_n$

Variance / Covariance

$$V_\lambda^{\mathrm{rdg}} = \mathbb{E}\left((\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \mathbb{E}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}))(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \mathbb{E}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}})^\top \right)$$

Calcul explicite :

$$\begin{aligned}
V_\lambda^{\mathrm{rdg}} &= \mathbb{E}\left((\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \varepsilon\varepsilon^\top X (\lambda \operatorname{Id}_p + X^\top X)^{-1}\right) \\
&= (\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \mathbb{E}(\varepsilon\varepsilon^\top) X (\lambda \operatorname{Id}_p + X^\top X)^{-1} \\
&= \sigma^2 (\lambda \operatorname{Id}_p + X^\top X)^{-2} X^\top X \\
&= \sum_{i=1}^{\operatorname{rg}(X)} \frac{s_i^2 \sigma^2}{(s_i^2 + \lambda)^2} \mathbf{v}_i \mathbf{v}_i^\top
\end{aligned}$$

Rem: on retrouve $V^{\mathrm{MCO}} = \sum_{i=1}^{\operatorname{rg}(X)} \frac{\sigma^2}{s_i^2} \mathbf{v}_i \mathbf{v}_i^\top$ quand $\lambda \to 0$

Rem: on retrouve une variance nulle quand $\lambda \to \infty$

Variance dans le cas général

Sous l'hypothèse de bruit "blanc" (*i.e.*, $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$) et de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Variance / Covariance

$$V_\lambda^{\mathrm{rdg}} = \mathbb{E}\left((\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \mathbb{E}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}))(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \mathbb{E}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}})^\top \right)$$

Calcul explicite :

$$\begin{aligned}
V_\lambda^{\mathrm{rdg}} =& \mathbb{E}\left((\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X (\lambda \operatorname{Id}_p + X^\top X)^{-1}\right) \\
&= (\lambda \operatorname{Id}_p + X^\top X)^{-1} X^\top \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) X (\lambda \operatorname{Id}_p + X^\top X)^{-1} \\
&= \sigma^2 (\lambda \operatorname{Id}_p + X^\top X)^{-2} X^\top X \\
&= \sum_{i=1}^{\mathrm{rg}(X)} \frac{s_i^2 \sigma^2}{(s_i^2 + \lambda)^2} \mathbf{v}_i \mathbf{v}_i^\top
\end{aligned}$$

<u>Rem</u>: on retrouve $V^{\mathrm{MCO}} = \sum_{i=1}^{\mathrm{rg}(X)} \frac{\sigma^2}{s_i^2} \mathbf{v}_i \mathbf{v}_i^\top$ quand $\lambda \to 0$

<u>Rem</u>: on retrouve une variance nulle quand $\lambda \to \infty$

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque (quadratique) de prédiction $\mathbb{E}\|X\boldsymbol{\theta}^\star - X\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)\right]$$

Calcul explicite (début identique) :

$$\begin{aligned} R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) &= \mathbb{E}\left[(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)\right] \\ &= \mathbb{E}[(X(X^\top X + \lambda \operatorname{Id}_p)^{-1} X^\top \boldsymbol{\varepsilon})^\top (X(X^\top X + \lambda \operatorname{Id}_p)^{-1} X^\top \boldsymbol{\varepsilon})] \\ &\quad + \boldsymbol{\theta}^{\star\top} \lambda^2 (X^\top X)(X^\top X + \lambda \operatorname{Id}_p)^{-2} \boldsymbol{\theta}^\star \end{aligned}$$

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\varepsilon\varepsilon^\top) = \sigma^2 \operatorname{Id}_n$

Risque (quadratique) de prédiction $\mathbb{E}\|X\boldsymbol{\theta}^\star - X\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)\right]$$

Calcul explicite (début identique) :

$$\begin{aligned}
R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) =& \mathbb{E}\left[(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)\right] \\
=& \mathbb{E}\big[(X(X^\top X + \lambda \operatorname{Id}_p)^{-1} X^\top \boldsymbol{\varepsilon})^\top (X(X^\top X + \lambda \operatorname{Id}_p)^{-1} X^\top \boldsymbol{\varepsilon})\big] \\
& + \boldsymbol{\theta}^{\star\top} \lambda^2 (X^\top X)(X^\top X + \lambda \operatorname{Id}_p)^{-2} \boldsymbol{\theta}^\star \\
=& \sum_{i=1}^{\mathrm{rg}(X)} \frac{s_i^4 \sigma^2}{(s_i^2 + \lambda)^2} + \boldsymbol{\theta}^{\star\top} \lambda^2 (X^\top X)(X^\top X + \lambda \operatorname{Id}_p)^{-2} \boldsymbol{\theta}^\star
\end{aligned}$$

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque (quadratique) de prédiction $\mathbb{E}\|X\boldsymbol{\theta}^\star - X\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}\|^2$

Sous l'hypothèse de modèle homoscédastique :

$$R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)\right]$$

Calcul explicite (début identique) :

$$\begin{aligned}
R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) =& \mathbb{E}\left[(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)\right] \\
=& \mathbb{E}\big[(X(X^\top X + \lambda \operatorname{Id}_p)^{-1} X^\top \boldsymbol{\varepsilon})^\top (X(X^\top X + \lambda \operatorname{Id}_p)^{-1} X^\top \boldsymbol{\varepsilon})\big] \\
& + \boldsymbol{\theta}^{\star\top} \lambda^2 (X^\top X)(X^\top X + \lambda \operatorname{Id}_p)^{-2} \boldsymbol{\theta}^\star \\
=& \sum_{i=1}^{\mathrm{rg}(X)} \frac{s_i^4 \sigma^2}{(s_i^2 + \lambda)^2} + \boldsymbol{\theta}^{\star\top} \lambda^2 (X^\top X)(X^\top X + \lambda \operatorname{Id}_p)^{-2} \boldsymbol{\theta}^\star
\end{aligned}$$

<u>Rem</u>: $\lim_{\lambda \to 0} R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) = \mathrm{rg}(X)\sigma^2$, $\lim_{\lambda \to \infty} R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) = \|X\boldsymbol{\theta}^\star\|_2^2$

Risque de prédiction

Hypothèse de modèle homoscédastique : $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \operatorname{Id}_n$

Risque (quadratique) de prédiction $\mathbb{E}\|X\boldsymbol{\theta}^\star - X\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}\|^2$

Sous l'hypothèse de modèle homoscédastique :

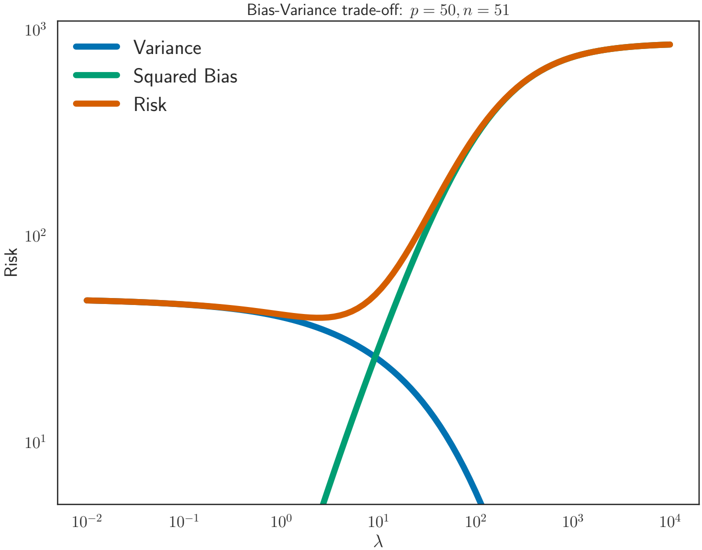
$$R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) = \mathbb{E}\left[(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)\right]$$

Calcul explicite (début identique) :

$$\begin{aligned}
R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) =& \mathbb{E}\left[(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)^\top (X^\top X)(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}} - \boldsymbol{\theta}^\star)\right] \\
=& \mathbb{E}\big[(X(X^\top X + \lambda \operatorname{Id}_p)^{-1} X^\top \boldsymbol{\varepsilon})^\top (X(X^\top X + \lambda \operatorname{Id}_p)^{-1} X^\top \boldsymbol{\varepsilon})\big] \\
& + \boldsymbol{\theta}^{\star\top} \lambda^2 (X^\top X)(X^\top X + \lambda \operatorname{Id}_p)^{-2} \boldsymbol{\theta}^\star \\
=& \sum_{i=1}^{\mathrm{rg}(X)} \frac{s_i^4 \sigma^2}{(s_i^2 + \lambda)^2} + \boldsymbol{\theta}^{\star\top} \lambda^2 (X^\top X)(X^\top X + \lambda \operatorname{Id}_p)^{-2} \boldsymbol{\theta}^\star
\end{aligned}$$

Rem: $\lim_{\lambda \to 0} R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) = \mathrm{rg}(X)\sigma^2$, $\lim_{\lambda \to \infty} R_{\mathrm{pred}}(\boldsymbol{\theta}^\star, \hat{\boldsymbol{\theta}}_\lambda^{\mathrm{rdg}}) = \|X\boldsymbol{\theta}^\star\|_2^2$

Biais / Variance : exemple de simulation

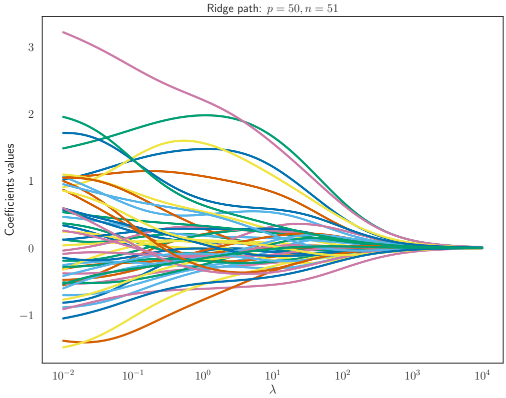


$$X \in \mathbb{R}^{50\times 50}, \boldsymbol{\theta}^\star = (2, 2, 2, 2, 2, 0, \ldots, 0)^\top$$

Sommaire

Choix de λ

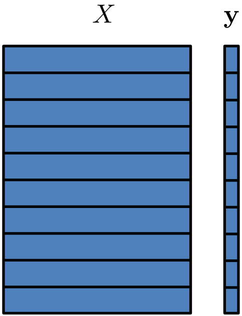
```
n_features = 50; n_samples = 51
X = np.random.randn(n_samples, n_features)
theta_true = np.zeros([n_features, ])
theta_true[0:5] = 2.
y_true = np.dot(X, theta_true)
y = y_true + 1. * np.random.rand(n_samples,)
```



Sommaire

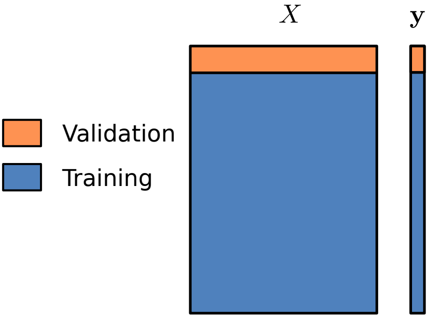
Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \ldots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs ( : *fold*) :



Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \ldots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs (: *fold*) :

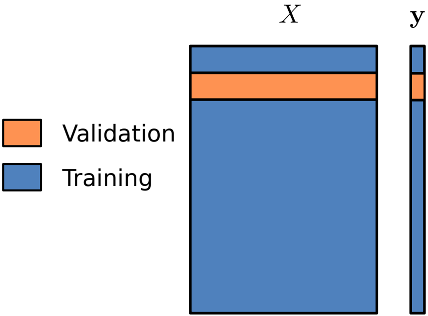


$$k = 1$$

1. Calculer avec les données d'entraînement les estimateurs pour $\lambda_1, \ldots, \lambda_r$: $\hat{\boldsymbol{\theta}}^{\lambda_1}, \ldots, \hat{\boldsymbol{\theta}}^{\lambda_r}$
2. Calculer les erreurs de prédiction $\mathrm{Error}_1^k, \ldots, \mathrm{Error}_r^k$ sur l'ensemble de validation,

Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \dots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs (: *fold*) :

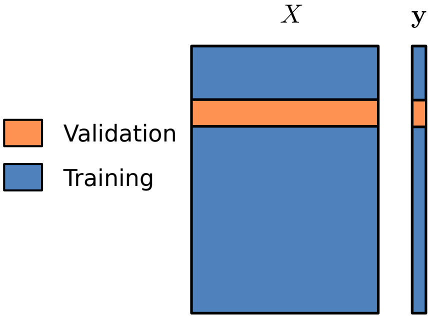


$$k = 2$$

1. Calculer avec les données d'entraînement les estimateurs pour $\lambda_1, \dots, \lambda_r$: $\hat{\boldsymbol{\theta}}^{\lambda_1}, \dots, \hat{\boldsymbol{\theta}}^{\lambda_r}$
2. Calculer les erreurs de prédiction $\mathrm{Error}_1^k, \dots, \mathrm{Error}_r^k$ sur l'ensemble de validation,

Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \ldots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs (🇬🇧 : *fold*) :

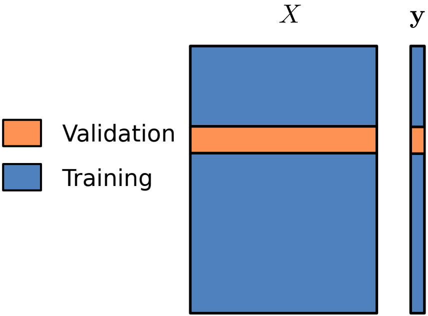


$k = 3$

1. Calculer avec les données d'entraînement les estimateurs pour $\lambda_1, \ldots, \lambda_r$: $\hat{\boldsymbol{\theta}}^{\lambda_1}, \ldots, \hat{\boldsymbol{\theta}}^{\lambda_r}$
2. Calculer les erreurs de prédiction $\mathrm{Error}_1^k, \ldots, \mathrm{Error}_r^k$ sur l'ensemble de validation,

Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \dots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs (: *fold*) :

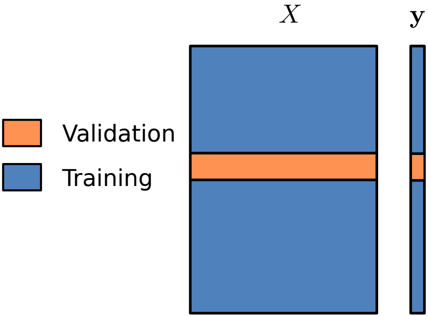


$$k = 4$$

1. Calculer avec les données d'entraînement les estimateurs pour $\lambda_1, \dots, \lambda_r$: $\hat{\boldsymbol{\theta}}^{\lambda_1}, \dots, \hat{\boldsymbol{\theta}}^{\lambda_r}$
2. Calculer les erreurs de prédiction $\mathrm{Error}_1^k, \dots, \mathrm{Error}_r^k$ sur l'ensemble de validation,

Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \ldots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs (: *fold*) :

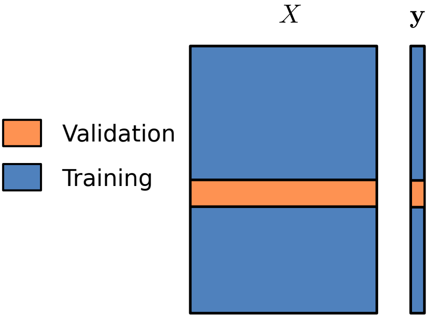


$$k = 5$$

1. Calculer avec les données d'entraînement les estimateurs pour $\lambda_1, \ldots, \lambda_r$: $\hat{\boldsymbol{\theta}}^{\lambda_1}, \ldots, \hat{\boldsymbol{\theta}}^{\lambda_r}$
2. Calculer les erreurs de prédiction $\mathrm{Error}_1^k, \ldots, \mathrm{Error}_r^k$ sur l'ensemble de validation,

Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \dots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs (: *fold*) :

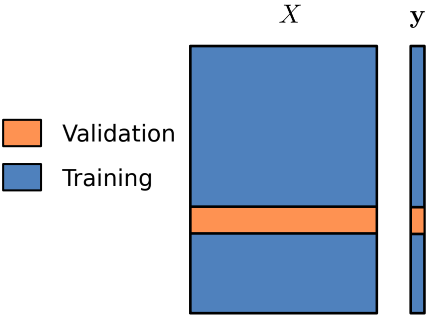


$k = 6$

1. Calculer avec les données d'entraînement les estimateurs pour $\lambda_1, \dots, \lambda_r$: $\hat{\boldsymbol{\theta}}^{\lambda_1}, \dots, \hat{\boldsymbol{\theta}}^{\lambda_r}$
2. Calculer les erreurs de prédiction $\mathrm{Error}_1^k, \dots, \mathrm{Error}_r^k$ sur l'ensemble de validation,

Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \ldots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs (: *fold*) :

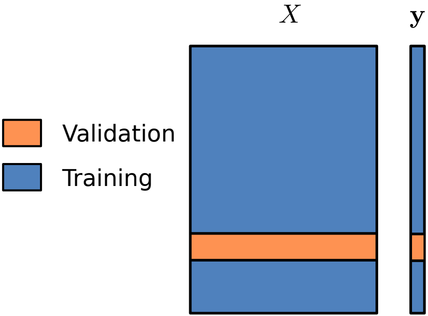


$k = 7$

1. Calculer avec les données d'entraînement les estimateurs pour $\lambda_1, \ldots, \lambda_r$: $\hat{\boldsymbol{\theta}}^{\lambda_1}, \ldots, \hat{\boldsymbol{\theta}}^{\lambda_r}$
2. Calculer les erreurs de prédiction $\mathrm{Error}_1^k, \ldots, \mathrm{Error}_r^k$ sur l'ensemble de validation,

Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \ldots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs (: *fold*) :

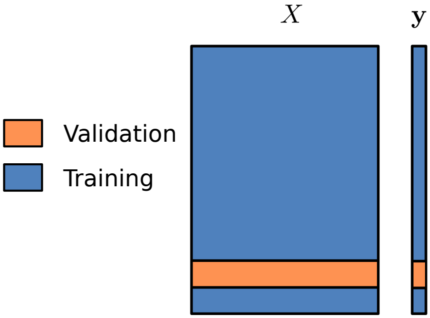


$$k = 8$$

1. Calculer avec les données d'entraînement les estimateurs pour $\lambda_1, \ldots, \lambda_r$: $\hat{\boldsymbol{\theta}}^{\lambda_1}, \ldots, \hat{\boldsymbol{\theta}}^{\lambda_r}$
2. Calculer les erreurs de prédiction $\mathrm{Error}_1^k, \ldots, \mathrm{Error}_r^k$ sur l'ensemble de validation,

Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \dots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs (: *fold*) :

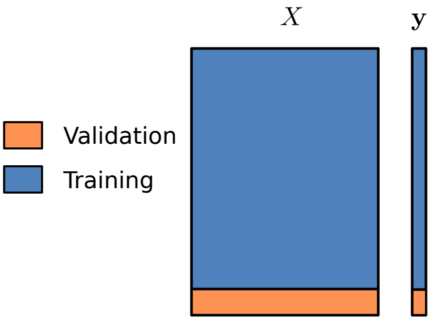


$k = 9$

1. Calculer avec les données d'entraînement les estimateurs pour $\lambda_1, \dots, \lambda_r$: $\hat{\boldsymbol{\theta}}^{\lambda_1}, \dots, \hat{\boldsymbol{\theta}}^{\lambda_r}$
2. Calculer les erreurs de prédiction $\mathrm{Error}_1^k, \dots, \mathrm{Error}_r^k$ sur l'ensemble de validation,

Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \ldots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs (: *fold*) :

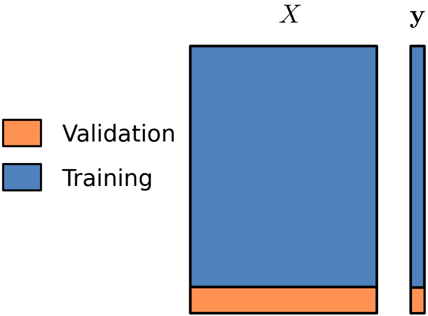


$k = 10$

1. Calculer avec les données d'entraînement les estimateurs pour $\lambda_1, \ldots, \lambda_r$: $\hat{\boldsymbol{\theta}}^{\lambda_1}, \ldots, \hat{\boldsymbol{\theta}}^{\lambda_r}$
2. Calculer les erreurs de prédiction $\mathrm{Error}_1^k, \ldots, \mathrm{Error}_r^k$ sur l'ensemble de validation,

Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \ldots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs (: *fold*) :



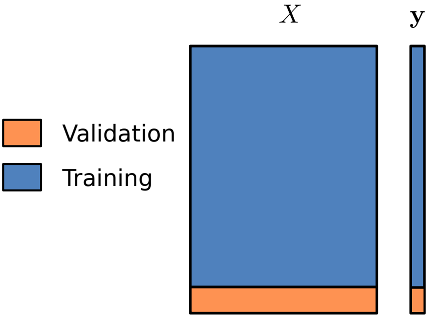
$k = 10$

1. Calculer avec les données d'entraînement les estimateurs pour $\lambda_1, \ldots, \lambda_r$: $\hat{\boldsymbol{\theta}}^{\lambda_1}, \ldots, \hat{\boldsymbol{\theta}}^{\lambda_r}$
2. Calculer les erreurs de prédiction $\mathrm{Error}_1^k, \ldots, \mathrm{Error}_r^k$ sur l'ensemble de validation,

Choix du paramètre : calculer $\widehat{\mathrm{Error}}_1, \ldots, \widehat{\mathrm{Error}}_r$, moyennes des erreurs et choisir $\hat{i}^{\mathrm{CV}} \in [\![1, r]\!]$ atteignant la plus petite

Validation croisée K-fold ($K = 10$)

- Choisir une grille de taille r de λ à tester : $\lambda_1, \dots, \lambda_r$
- Diviser $(X, \mathbf{y})$ selon les observations en K blocs (: *fold*) :



$k = 10$

1. Calculer avec les données d'entraînement les estimateurs pour $\lambda_1, \dots, \lambda_r$: $\hat{\boldsymbol{\theta}}^{\lambda_1}, \dots, \hat{\boldsymbol{\theta}}^{\lambda_r}$
2. Calculer les erreurs de prédiction $\mathrm{Error}_1^k, \dots, \mathrm{Error}_r^k$ sur l'ensemble de validation,

Choix du paramètre : calculer $\widehat{\mathrm{Error}}_1, \dots, \widehat{\mathrm{Error}}_r$, moyennes des erreurs et choisir $\hat{i}^{\mathrm{CV}} \in [\![1, r]\!]$ atteignant la plus petite

Re-calibration : calculer $\hat{\boldsymbol{\theta}}^{\lambda_{\hat{i}^{\mathrm{CV}}}}$ sur toutes les observations $(X, \mathbf{y})$

CV en pratique

Cas extrême de validation croisée (: *cross-validation*)

- $K = 1$: impossible, au moins $K = 2$
- $K = n$: stratégie "*leave-one-out*" (*cf.* **Jackknife**) : autant de blocs que de variables

Rem: $K = n$: calcul efficace pour Ridge mais assez instable

Conseils pratiques :

- "randomiser les observations" : observations dans un ordre aléatoire, évite des blocs de données trop similaires (chaque sous-bloc doit être représentatif de l'ensemble)
- choix habituels : $K = 5, 10$

Rem: en prédiction on peut aussi moyenner les meilleurs estimateurs obtenus plutôt que de re-calibrer sur toutes les données

Variantes de CV et `sklearn`

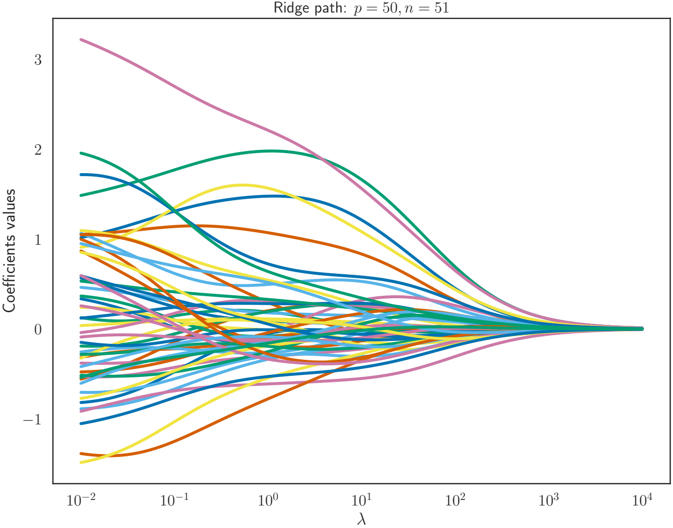
Alternatives classiques :

- partition aléatoire entre ensemble d'apprentissage et validation ($K = 2$ en gros,*cf.* `train_test_split`)
- variante pour séries temporelles : `TimeSeriesSplit`
- variante pour la classification et des cas des classes déséquilibrées `StratifiedKFold`

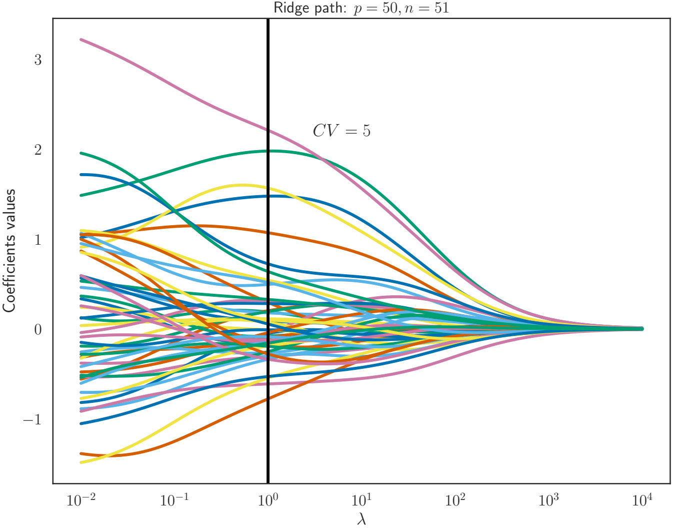
Plus de détails :

http://scikit-learn.org/stable/modules/cross_validation.html

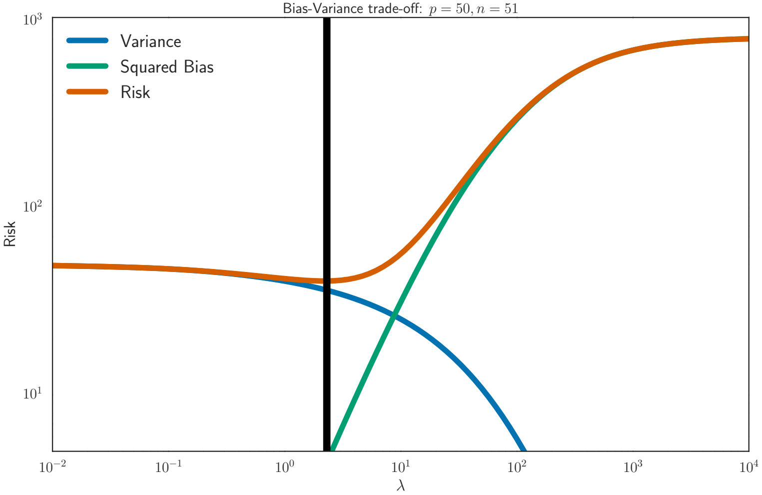
Choix de λ : exemple avec $CV = 5$ (I)



Choix de λ : exemple avec $CV = 5$ (I)



Choix de λ : exemple avec $CV = 5$ (II)



Régularisation et colonne de zéro

ATTENTION : utilisation de CV avec des données catégorielles
En effet : si on enlève toute les occurrences d'une modalité de la partie apprentissage, on crée une colonne de zéro (les MCO ne peuvent plus marcher...)

Remèdes :

- "régularisation" : on peut obtenir une solution
- faire une séparation apprentissage/test plus poussée pour équilibrer les *folds*

Algorithmes pour la méthode *Ridge*

- `'svd'` : méthode la plus stable, avantageuse pour calculer plusieurs λ car on ne “paye” la SVD qu'une fois
- `'cholesky'` : décomposition matricielle proposant une formule fermée `scipy.linalg.solve`
- `'sparse_cg'` : gradient conjugué utile dans les cas creux (: *sparse*) et de grande dimension (baisser `tol`/`max_iter`)
- approche de type gradient stochastique si n est très grand

cf. le code des fonctions `Ridge`, `ridge_path`, `RidgeCV` dans le module `linear_model` de `sklearn`

Rem: on calcule rarement l'estimateur *Ridge* pour un λ, en général on en calcule plusieurs $(10, 100, \dots)$ et on cherche le meilleur

Rem: enjeu crucial de calculer des SVD de grandes tailles

Références I

- G. H. Golub and C. F. van Loan.
  Matrix computations.
  Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.