La régression Ridge

Chargé du cours
Prof. Mustapha Rachdi

Université Grenoble Alpes
UFR SHS, BP. 47
38040 Grenoble cedex 09
France
Bureau : C08 du Bât Michel Dubois
(ex BSHM)
e-mail : mustapha.rachdi@univ-grenoble-alpes.fr

- Les tests multiples ou à grande échelle présentent également des problèmes en grande dimension. Ils représentent, bien sûr, un domaine important et nous en reviendrons ultérieurement !

- Cependant, pour le reste du cours, nous aborderons le problème de la régression en grande dimension : utiliser les co-variables (variables explicatives) pour prédire/expliquer la réponse (outcome).

- Comme nous l'avons vu dans Chap. 1 : la méthode des moindres carrés ordinaires (MCO) est problématique en grande dimension.

- Réduire la dimensionnalité grâce/à travers la sélection de modèles permet un certain progrès, mais présente plusieurs lacunes !

- Plus généralement, cela peut être considéré comme un échec des méthodes basées sur la vraisemblance.

- Dans ce chapitre, nous utiliserons la notation L pour désigner la log-vraisemblance négative :

$$L(\theta|Data) = -\log \ell(\theta|Data) = -\log p(Data|\theta)$$

- Ici, L est connue comme étant la *fonction de perte* et nous recherchons des estimations avec une “faible” perte. Cela équivaut à trouver une valeur (ou un intervalle de valeurs) avec une forte/grande vraisemblance.

- Dans le contexte de la régression linéaire, la fonction de perte est :

$$L(\beta|\mathbf{X}, y) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_i (y_i - {}^t x_i\,\beta)^2$$

- C'est seulement la différence des fonctions de perte entre deux valeurs, $L(\beta_1|\mathbf{X}, y) - L(\beta_2|\mathbf{X}, y)$, c'est-à-dire le rapport (ratio) de vraisemblances, qui est pertinente dans l'inférence statistique basée sur la vraisemblance. Ainsi, le premier terme peut être ignoré.

- Dans le but de trouver le MLE, le facteur $(2\sigma^2)^{-1}$ peut également être ignoré, bien que nous devrions en tenir compte lors de la construction d'intervalles basés sur la vraisemblance.

Compte tenu des problèmes susmentionnés avec les méthodes de vraisemblance, considérons plutôt la modification suivante :

$$Q(\beta|\mathbf{X}, \mathbf{y}) = L(\beta|\mathbf{X}, \mathbf{y}) + P_\lambda(\beta),$$

où

- $P_\cdot$ est une *fonction de pénalisation*, qui pénalise ce que l'on pourrait considérer comme des valeurs moins réalistes des paramètres inconnus.
- λ est un *paramètre de régularisation*, qui contrôle le compromis entre les deux composants.
- La fonction Q est connue comme étant la *fonction objective*.

- Que voulons-nous dire exactement par des valeurs "moins réalistes" ?
- L'utilisation la plus courante de la pénalisation consiste à imposer le fait que les petits coefficients de régression sont plus probables que les grands ; c'est-à-dire que nous ne serions pas surpris si β_j était 1.2 ou 0.3 ou 0, mais serions très surpris si β_j était 9.7×10^4.
- Plus tard dans le cours, nous considérons d'autres utilisations de la pénalisation pour refléter le fait que les vrais coefficients peuvent être regroupés de façon hiérarchique, ou configurés selon un modèle spatial tel que β_j est susceptible d'être proche de $\beta_j + 1$.

- Il faut être prudent dans l'application de l'idée que de petits coefficients dans la régression sont plus préférables que les grands.
    - Tout d'abord, il n'est généralement pas logique d'appliquer ce raisonnement à l'intercep. Par conséquent, β_0 n'est pas inclus dans la pénalité.
    - Deuxièmement, la taille/grandeur d'un coefficient dans le modèle de régression dépend de l'échelle avec laquelle la covariable associée a été mesurée. Selon les unités pour lesquelles x_j est mesurée, par exemple, la valeur $\beta_j = 9.7 \times 10^4$ pourrait, en fait, être réaliste.

- C'est un problème particulier si, les différentes co-variables sont mesurées avec différentes échelles, car la pénalité n'aurait pas le même effet (effet n'est pas égal d'une co-variable à l'autre) sur toutes les estimations des coefficients.

- Pour éviter ce problème et assurer une invariance par rapport à l'échelle (variance de la variable), les co-variables sont habituellement normalisées (centrées-réduites) avant l'ajustement du modèle :

$$\bar{x}_j = 0 \quad \text{et} \quad {}^t x_j \, x_j = n \ \text{ pour tout } \ j = 1, \ldots, p$$

- Ceci peut être accompli sans aucune perte de généralité, car tous les décalages/translations de $\mathbf{X}$ sont absorbés par l'intercept, et les changements d'échelle (à cause de la réduction de $\mathbf{X}$) peuvent être inversés après que le modèle ait été ajusté.

Avantages supplémentaires de la normalisation

Centrer et réduire les co-variables (variables explicatives) va ajouter des avantages en termes d'économies de calcul et de simplicité conceptuelle :

- Les co-variables sont maintenant orthogonales à l'intercept, ce qui signifie que dans l'espace des co-variables normalisées, $\widehat{\beta}_0 = \bar{y}$, quel que soit le reste du modèle.

- En outre, la normalisation simplifie les solutions. Pour une illustration avec une régression linéaire simple :

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}\bar{x} \quad \text{et} \quad \widehat{\beta}_1 = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$

  Cependant, si nous centrons et réduisons x et centrons y, alors nous obtenons une expression beaucoup plus simple

$$\widehat{\beta}_0 = 0 \quad \text{et} \quad \widehat{\beta}_1 = {}^t x\, y/n$$

- Si la régression pénalisée consiste à imposer l'hypothèse que les petits coefficients dans le modèle de régression sont plus préférables que les plus grands, nous devrions choisir donc une pénalité qui décourage/déssuade les grands coefficients de régression.

- Un choix naturel est de pénaliser la somme des carrés des coefficients de régression par :

$$P_\lambda(\beta) = \frac{1}{2\tau^2} \sum_{j=1}^{p} \beta_j^2$$

- Utiliser cette pénalité dans le contexte de la régression pénalisée est connue sous le nom de *Régression Ridge (de crête)*, qui a une longue histoire en statistiques, qui remonte à 1970

- La fonction objective dans la régression ridge est :

$$Q(\beta|\mathbf{X}, y) = \frac{1}{2\sigma^2}\sum_i (y_i - {}^t x_i\,\beta)^2 + \frac{1}{2\tau^2}\sum_{j=1}^{p}\beta_j^2$$

- Il est souvent commode de multiplier la fonction objectif ci-dessus par σ^2/n. Comme nous le verrons, cela permettrait de simplifier les expressions impliquées dans la régression pénalisée :

$$Q(\beta|\mathbf{X}, y) = \frac{1}{2n}\sum_i (y_i - {}^t x_i\,\beta)^2 + \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2$$

où $\lambda = \sigma^2/(n\tau^2)$

- Pour la régression linéaire, la pénalité ridge est particulièrement attrayante car l'estimateur du maximum de vraisemblance pénalisé a une solution de forme fermée simple.

- Cette fonction objective est différentiable, et il est facile de montrer que son minimum se situe à

$$\widehat{\beta} = \left(\frac{1}{n} {}^t\mathbf{X}\mathbf{X} + \lambda \mathbf{I} \right)^{-1} \frac{1}{n} {}^t\mathbf{X}\mathbf{y}$$

- La solution est similaire à la solution des moindres carrés (MCO), mais avec l'ajout d'une "ridge/crête" le long de la diagonale de la matrice à inverser

- Noter que la solution ridge/crête est une simple fonction des solutions marginales MCO $\left(\frac{1}{n} {}^t\mathbf{X}\mathbf{y} \right)$ et de la matrice des corrélations $\left(\frac{1}{n} {}^t\mathbf{X}\mathbf{X} \right)$.

- Pour comprendre l'effet de la pénalité ridge sur l'estimateur $\widehat{\beta}$, il est utile de considérer le cas particulier d'une *design matrice*, ou matrice de conception, qui soit orthonormale $\left(\frac{1}{n}{}^t\mathbf{X}\mathbf{X} = I_p\right)$.

- Dans ce cas, on trouve

$$\widehat{\beta}_j = \frac{\widehat{\beta}_j^{MCO}}{1+\lambda}$$

- Ceci illustre le fait que la caractéristique essentielle de la régression ridge est *le rétrécissement (shrinkage)* : c'est-à-dire que le principal effet de la pénalité ridge, dans l'application, est de réduire/rétricir les estimations à zéro

Régression Ridge : comprendre l'effet de la pénalité

Exemple simple

- Les avantages de la régression ridge sont plus frappants quand on est en présence de multicolinéarité.

- Considérons l'exemple simulé très simple suivant :

```
> x1 <- rnorm (20)
> x2 <- rnorm (20, mean = x1, sd = 0.01)
> y <- rnorm (20, mean = 3 + x1 + x2)
> lm (y ~ x1 + x2)
...
(Intercept)          x1          x2
2.582064   39.971344   -38.040040
```

- Bien qu'il n'y ait que deux co-variables, la forte corrélation entre X_1 et X_2 cause beaucoup de problèmes pour l'estimateur du maximum de vraisemblance.

- Le problème ici est que la surface de vraisemblance est très plate le long de $\beta_1 + \beta_2 = 2$, conduisant à une incertitude énorme

- Quand nous introduisons l'hypothèse supplémentaire que les petits coefficients sont plus préférables que les grands en utilisant une pénalité ridge, cependant, cette incertitude est résolue :

```
> lm.ridge(y ~ x1+x2,lambda=1)
                     x1           x2
2.6214998    0.9906773    0.8973912
```

```
> ridge(y ~ x1+x2,lambda=0.1)
                     x1           x2
3.0327231    0.9575176    0.9421784
```

Régression Ridge : propriétés

La régression Ridge a toujours des solutions uniques

- L'estimateur du maximum de vraisemblance n'est pas toujours unique : Si $\mathbf{X}$ n'est pas de rang plein, la matrice ${}^t\mathbf{X}\,\mathbf{X}$ n'est pas inversible et un nombre infini de valeurs β maximise la vraisemblance.

- Ce problème ne se produit pas avec la régression ridge.

Théorème

Pour toute design matrix $\mathbf{X}$*, la quantité*

$$\frac{1}{n}{}^t\mathbf{X}\,\mathbf{X} + \lambda\mathbf{I}$$

est toujours inversible à condition que $\lambda > 0$*. Donc, il y a toujours une solution unique* $\widehat{\beta}$*.*

Régression Ridge : proriétés

Est-ce que la régression ridge est meilleure que le maximum de vraisemblance ?

- Dans notre exemple simple, précédent, l'estimateur de la régression ridge était beaucoup plus proche de la vérité/réalité que de MLE.
- Une question évidente est de savoir si les estimateurs de la régression ridge sont systématiquement plus proches de la vérité/réalité que les MLEs, ou si cet exemple n'est dû qu'au hasard.
- Pour répondre à cette question, commençons par déterminer le biais et la variance de la régression ridge.

- La variance de l'estimateur de la régression ridge est :

$$var(\widehat{\beta}) = \frac{\sigma^2}{n}\mathbf{W}\left(\frac{1}{n}{}^t\mathbf{X}\,\mathbf{X}\right)\mathbf{W},$$

où $\mathbf{W} = \left(\frac{1}{n}{}^t\mathbf{X}\,\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}$.

- Alors que le biais est donné par :

$$Biais(\widehat{\beta}) = -\lambda\mathbf{W}\beta$$

- Le biais et la variance contribuent à la précision globale, mesurée par l'erreur quadratique moyenne (MSE) :

$$MSE(\widehat{\beta}) = \mathbb{E}\left|\left|\widehat{\beta} - \beta\right|\right|^2 = \sum_j var(\widehat{\beta}_j) + \sum_j Biais(\widehat{\beta}_j)^2$$

Régression Ridge : propriétés

Théorème d'existence

- Question : est ce que la régression ridge est meilleure que le maximum de vraisemblance (MCO) ?
-

Théorème

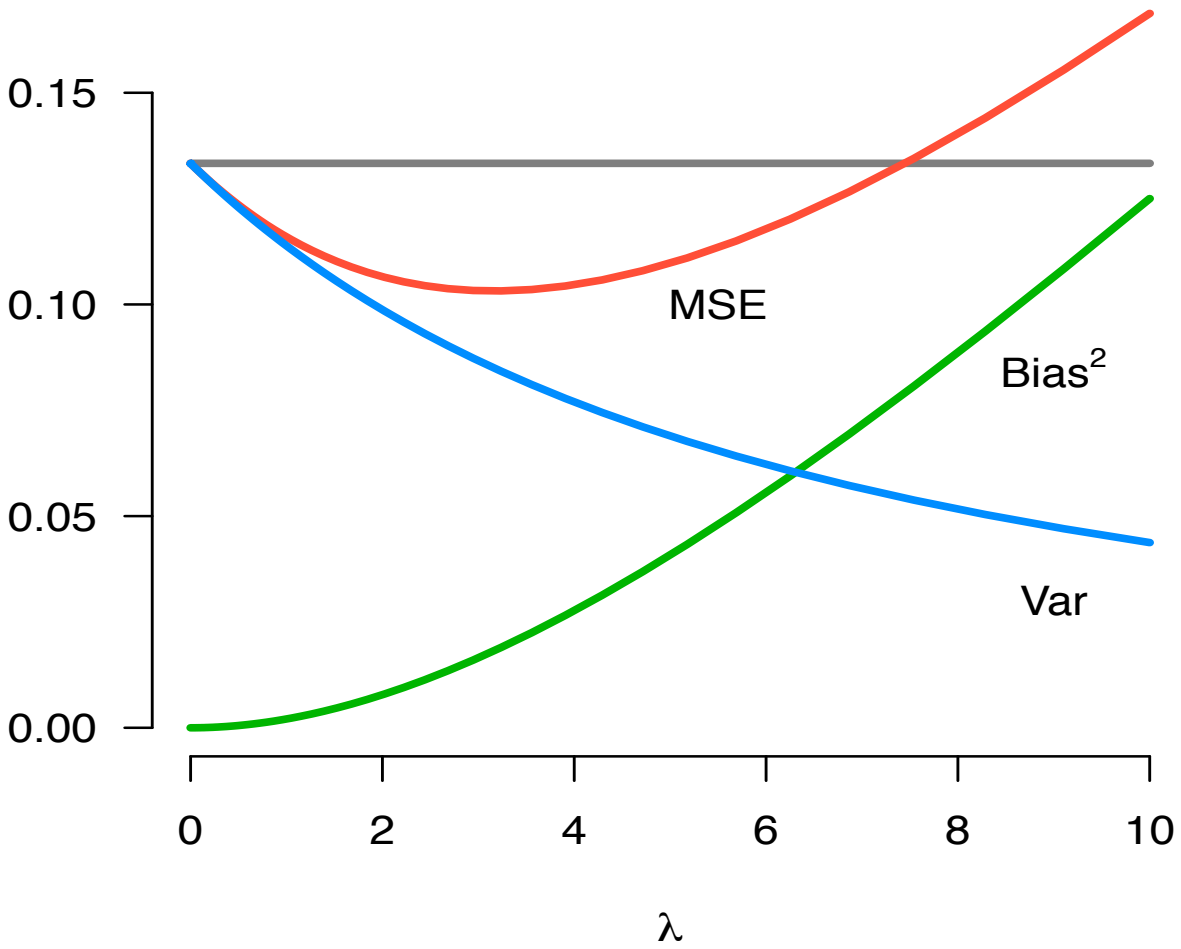
Il existe toujours une valeur de λ pour laquelle

$$MSE\left(\widehat{\beta}_{\lambda}\right) < MSE\left(\widehat{\beta}^{MCO}\right)$$

- C'est un résultat plutôt surprenant avec des conséquences qui sont quelque peu radicales : malgré les propriétés théoriques qui sont typiquement impressionnantes du maximum de vraisemblance (MCO) et de la régression linéaire, nous pouvons **toujours** obtenir un meilleur estimateur en rétrécissant la MLE vers zéro

Régression Ridge : propriétés

Justification/Sketch de la preuve



- D'un point de vue bayésien, on peut penser que la pénalité résulte, a priori, d'une distribution formelle sur les paramètres

- Soit $p(y|\beta)$ la distribution conditionnelle de y sachant β et $p(\beta)$ la loi a priori de β. Alors, la densité a postériori est

$$p(\beta|y) = \frac{p(y|\beta)p(\beta)}{p(y)} \propto p(y|\beta)p(\beta),$$

ou

$$\log p(\beta|y) = \log p(y|\beta) + \log p(\beta) + \textit{constante},$$

sur l'échelle logarithmique : ce qui est exactement la forme générique d'une vraisemblance pénalisée.

- En optimisant la fonction objective, nous trouvons le *mode* de la distribution a posteriori de β. Ceci est connu comme l'estimateur du *maximum a posteriori*, ou MAP

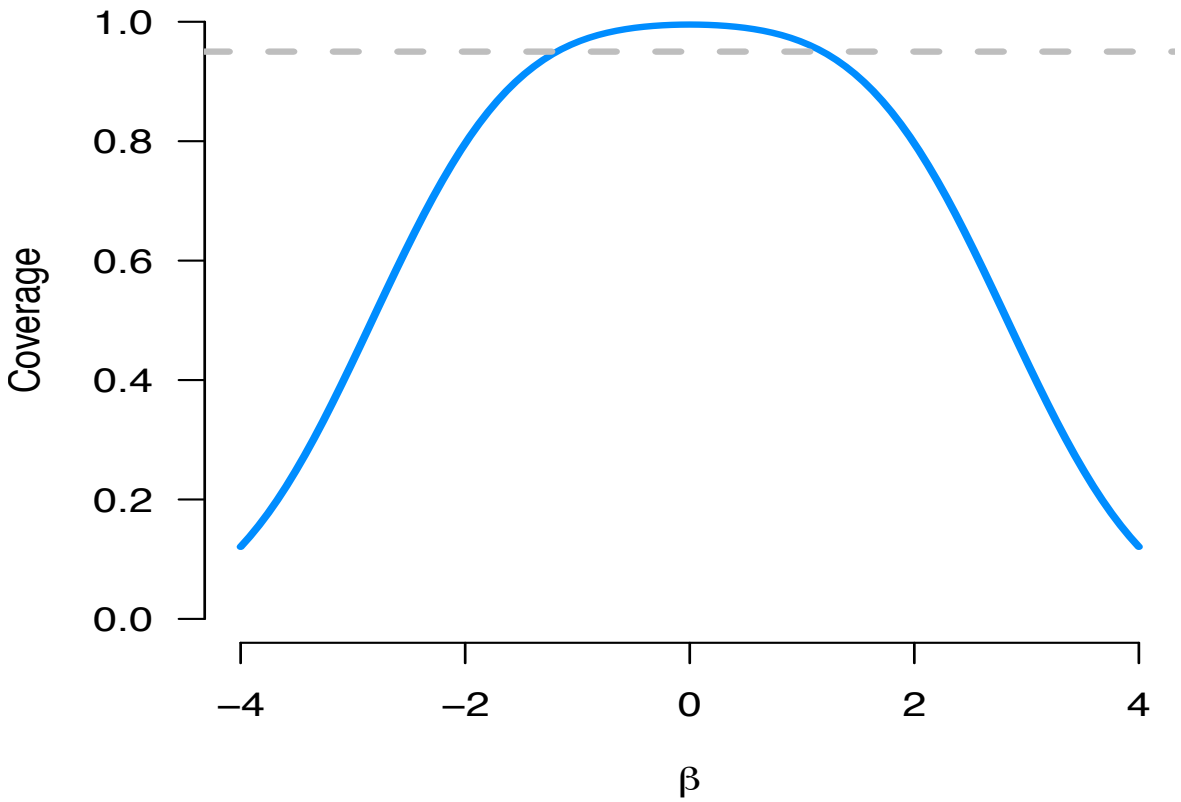
- Plus précisément, si l'on suppose, a priori, que :

$$\beta_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^2)$$

  le log-posteriori résultant est exactement la fonction objective de la régression ridge (à une constante près)

- En outre,
  - L'estimateur de la régression ridge $\widehat{\beta}$ est la *moyenne* a posteriori (en plus d'être le *mode* a posteriori)
  - Le paramètre de régularisation λ est le rapport de la précision a priori $(1/\tau^2)$ sur l'information (n/σ^2).

- Nous arrivons donc au même estimateur $\widehat{\beta}$ que nous considérions comme un estimateur du maximum de vraisemblance modifié ou comme un estimateur bayésien
- À d'autres égards, cependant, la similarité entre Bayésien et Fréquentiste se détériore.
- Deux aspects, en particulier, méritent d'être mentionnés.
  - Le premier a un but inférenciel i.e., est de construire des intervalles de confiance pour β
  - Quelles propriétés de tels intervalles devraient avoir.
- Des intervalles de confiance fréquentistes sont requis pour maintenir un certain niveau de couverture pour toute valeur fixe de β.
- D'autre part, les intervalles bayésiens, a posteriori, peuvent avoir une couverture beaucoup plus élevée de certaines valeurs de β que d'autres.



- La couverture de Bayes pour un intervalle a posteriori de 95% en $\beta_j \simeq 0$ est $> 99\%$, mais seulement $\simeq 20\%$ pour $\beta_j \simeq 3.5$.
- L'intervalle maintient néanmoins une couverture de 95% sur une collection de valeurs β_j, qui est intégrée a priori.

- L'autre aspect dans lequel un clivage clair émerge entre les perspectives Bayésienne et Fréquentiste est en ce qui concerne la valeur spécifique $\beta = 0$.
- D'un point de vue bayésien, la probabilité a posteriori que $\beta = 0$ est 0 parce que sa distribution a posteriori est continue.
- D'un point de vue fréquentiste, cependant, la notion de tester si $\beta = 0$ est toujours significative. En effet, elle est souvent intéressante dans une analyse.

- La littérature sur la régression pénalisée adopte généralement la perspective de la théorie du maximum de vraisemblance, bien que l'apparition d'une pénalité dans la vraisemblance brouille quelque peu les limites entre les idées Bayésiennes et Fréquentistes.

- La grande majorité des recherches sur les méthodes de régression pénalisées se sont concentrées sur l'estimation ponctuelle et ses propriétés, de sorte que ces différences inférentielles entre les perspectives bayésiennes et fréquentistes sont relativement inexplorées

- Néanmoins, le développement de méthodes inférentielles pour la régression pénalisée est un domaine actif dans la recherche actuelle, et nous reviendrons sur certains de ces problèmes lorsque nous discuterons de l'inférence pour les modèles en grande dimension.