# TD 11 : RÉGULARISATION

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE, 4 DÉCEMBRE 2015

Jean-Baptiste Alayrac
jean-baptiste.alayrac@inria.fr

RÉSUMÉ. Dans ce TP/TD, on va commencer par chercher à illustrer le paradoxe de James-Stein. Ensuite on s'intéressera à la régression ridge, déjà vue lors du TP4 du point de vue de l'optimisation. On essaiera ici de s'intéresser aux aspects plus statistiques de cette régularisation.

## 1. Exercice : Estimateur de James-Stein

Dans cet exercice, on cherche à illustrer le paradoxe de James-Stein évoqué en cours. Pour cela on considère un $n$ échantillon gaussien de dimension $p$, de variance unité et de moyenne $\theta \in \mathbb{R}^p$. Pour les simulations numériques on pourra prendre $n = 500$ et $p = 40$.

**1**) Quel est l'estimateur du maximum de vraisemblance de $\theta$, $\widehat{\theta}_{MV}$ ? Quel est le risque quadratique moyen qui lui est associé ?

**2**) Rappeler la définition de l'estimateur de James-Stein $\widehat{\theta}_{JS}$ vu en cours et le paradoxe qui lui est associé.

**3**) Générer un vecteur de moyenne $\theta$ et simuler des données suivant une loi normale de moyenne $\theta$ et de variance unité. Tracer la courbe donnant le risque quadratique de James-Stein évalué empiriquement sur les données simulées en fonction de la norme de $\theta$. Comparer au risque quadratique de l'estimateur du maximum de vraisemblance.

**4**) Expliquer le comportement de l'estimateur quand $\|\theta\|$ tend vers 0.

**5**) BONUS : (à faire à la fin) A l'aide de simulations illustrer la borne suivante, vue en cours concernant l'estimateur de James-Stein :

$$\mathbb{E}[(\widehat{\theta}_{JS} - \theta)^2] \leq 2\sigma^2 + \frac{\sigma^2(p-2)\|\theta\|^2}{(p-2)\sigma^2 + \|\theta\|^2}.$$

## 2. Exercice : Sélection de modèle avec le $C_p$ de Mallows

On considère le cas de la régression linéaire multivariée de $\mathbb{R}^p$ dans $\mathbb{R}$. On appelle $X$ la matrice de design des données. On va chercher à utiliser les idées de la sélection de modèle telles qu'elles ont été vues en cours. On se place dans le même cadre que le cours, à savoir que l'on considère la classe des ETL (Estimateurs par transformation linéaire).

On appelle dans la suite $\sigma^2$ la variance associée au modèle vu en cours où $y = f^* + \epsilon$ avec $\mathbb{E}[\epsilon] = 0$ et $\mathbb{E}[\epsilon^2] = \sigma^2$ . On appelle, pour un modèle $\widehat{\mathbf{f}}$, $SSE(\widehat{f})$ la somme des erreurs quadratiques entre les observations $y_i$ et les prédictions faites par le modèle $\widehat{\mathbf{f}}$. Si on note $p$ le nombre de variables explicatives utilisées dans la construction de $\widehat{f}$, on peut alors définir le risque régularisé suivant :

$$R_n(\widehat{f}) = SSE(\widehat{f}) + 2\sigma^2 p/n.$$

La pénalité $2\frac{\sigma^2 p}{n}$ s'appelle le $C_p$ de Mallows. Elle correspond à la différence qui existe, à design fixe, entre la minimisation de l'espérance (sur tous les échantillons possibles) du risque empirique et la minimisation de l'espérance de l'excès de risque empirique pour un estimateur par transformation linéaire.

**6**) Générer des données (quelques centaines suffisent) en dimension $p = 8$ suivant un modèle du type régression linéaire $y_i = x_i^T \beta + \epsilon$ avec $\epsilon$ un bruit Gaussien centré de variance unité.

**7**) Pour chaque $i \leq p$ chercher (en testant toutes les combinaisons possibles) le meilleur modèle impliquant $i$ variable et calculer le $R_n(\widehat{f})$ associé. Tracer l'évolution de cette quantité en fonction de $p$.

## 3. Exercice : Régression ridge

Dans cet exercice on considère la régression linéaire des moindres carrés pénalisée (régression ridge) de $\mathbb{R}^p$ dans $\mathbb{R}$, connue aussi sous le nom de régression ridge. $X$ désigne la matrice de design des données, qui comporte $n$ lignes et $p$ colonnes.

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \|y - Xw\|_2^2 + \lambda \|w\|_2^2.$$

**Biais et variance de la régression ridge**

**8**) Rappeler l'expression analytique de l'estimateur de la régression ridge ci-dessus.

**9**) On suppose que les données $y_i$ sont générées selon $y_i = x_i^T \beta + \epsilon_i$ où $\epsilon_i$ sont des bruits i.i.d de variance $\sigma^2$. On se place dans le cadre du design fixe.

a) Montrer que l'estimateur de la régression ridge, sous ces hypothèses est un ETL. Explicitez la matrice associée que l'on notera $A$.

b) En utilisant la Décomposition en Valeurs Singulières (SVD en Anglais et `svd` en MATLAB/Octave) de la matrice de design (on écrira $X = UDV^T$ avec $D$ la matrice des valeurs singulières, $U$ et $V$ respectivement unitaires sur $\mathbb{R}^n$ et $\mathbb{R}^p$). Exprimer la variance du prédicteur avec seulement $\sigma^2$, le paramètre de régularisation $\lambda$ et les valeurs singulières $d_k$.

**10**) En notant $\tilde{\beta}$ la représentation de $\beta$ dans une base bien choisie, le biais de l'estimateur de la régression ridge s'exprime en fonction de $\tilde{\beta}$, $\sigma^2$, $\lambda$ et les coefficients de la svd de $A$.

**Choix de $\lambda$**

**11**) On considère maintenant $p = 8$. Générer 300 points dans $\mathbb{R}^p$ suivant une loi uniforme. Générer un vecteur $y$ de réponses dans $\mathbb{R}$ suivant une fonction déterministe des données d'entrée que vous augmenterez d'un bruit.

**12**) En faisant varier la valeur du paramètre de régularisation (en utilisant le `logspace` de Matlab par exemple), représenter en fonction de lambda, la norme du vecteur donné par la régression ridge (on dit aussi que l'on représente l'évolution de la norme sur le "chemin de régularisation").

**13**) Représenter l'évolution des coefficients de la régression le long du chemin de régularisation, que constate-t-on ?

**14**) Générer maintenant des données de test sur lesquelles on testera la capacité de généralisation de l'estimateur de la régression ridge.

**15**) Implémenter la méthode de sélection de $\lambda$ par validation croisée. Si le temps manque, on peut se contenter d'implémenter la méthode de validation simple (le $K$ de la validation croisée fixé à 1).

**16**) En supposant que la variance du bruit est connue (ce qui est le cas ici), comment peut on utiliser le $C_L$ de Mallows pour choisir le $\lambda$ de la régression ridge, au moins de manière numérique ? Implémenter cette méthode et comparer le $\lambda$ avec celui de la validation croisée.