

TD 7 : RÉGULARISATION

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE, PRINTEMPS 2013

Rémi Lajugie
remi.lajugie@ens.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 $\theta \in \mathbb{R}$. Pour les simulations numériques on pourra prendre $n = 500$ et $p = 40$

- 1) Quel est l'estimateur du maximum de vraisemblance de θ , $\hat{\theta}_{MV}$? Quel est le risque quadratique moyen qui lui est associé?
- 2) Rappelez la définition de l'estimateur de James-Stein $\hat{\theta}_{JS}$ vu en cours ainsi que le paradoxe qui lui est associé.
- 3) Générez un vecteur de moyenne θ et simulez des données suivant une loi normale de moyenne θ et de variance unité. Tracez la courbe donnant le risque quadratique évalué empiriquement sur les données simulées en fonction de la norme de θ .
- 4) Expliquez le comportement de l'estimateur quand $\|\theta\|$ tend vers 0.
- 5) BONUS : (à faire à la fin) On rappelle le résultat suivant, vu en cours concernant l'estimateur de James-Stein :

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

. A l'aide de simulations illustrez cette borne.

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 que vous les avez 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)

Tout d'abord, quelques rappels sur cette statistique. On appelle dans la suite σ^2 la variance globale associée au prédicteur de la régression linéaire complète (avec tous les régresseurs), on appelle ensuite, pour un modèle impliquant seulement p variables explicatives, $SSE(M_p)$, SSE , la somme des erreurs quadratiques entre les observations y_i et les prédictions faites par le modèle $\hat{\mathbf{f}}$. On définit alors la quantité :

$$C_p = SSE/n + 2 * \sigma^2 p/n.$$

6) Sur des données de dimension $p = 8$, implémentez une méthode de sélection des variables pour les prédicteurs de la régression linéaire de modèle en se fondant sur C_p . Pour chaque $i \leq p$ vous chercherez le meilleur modèle impliquant i variable et le C_{i^*} associé. Vous tracerez l'évolution de ces C_{i^*} en fonction du nombre de régresseurs.

3. EXERCICE : RÉGRESSION RIDGE

Dans cet exercice on considère la régression linéaire des moindres carrés pénalisée 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.

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

7) Rappelez l'expression analytique de l'estimateur de la régression ridge ci-dessus. (Savez-vous le démontrer ?)

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

a) Montrez 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), Exprimez l'excès de risque quadratique du prédicteur avec seulement σ^2 , le paramètre de régularisation λ et les valeurs singulières d_k .

9) En notant $\tilde{\beta}$ la représentation de β dans une base bien choisie, le biais de la régression ridge s'exprime en fonction de $\tilde{\beta}$, σ^2 , λ et les coefficients de la svd de A .

10) On considère maintenant $p = 8$. Générez 100 points dans \mathbb{R}^p suivant une loi uniforme. Générez un vecteur y de réponses dans \mathbb{R} en les générant suivant une fonction déterministe des données d'entrée que vous augmenterez d'un bruit.

11) En faisant varier la valeur du paramètre de régularisation (sur une échelle logarithmique par exemple), représentez 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").

12) Représentez l'évolution des coefficients de la régression le long du chemin de régularisation (utilisez l'expression analytique des questions précédentes), que constatez vous ?