

Apprentissage: cours 8

Estimateur de Stein, régularisation et pénalisation

Guillaume Obozinski

18 avril 2013

Estimateur de James et Stein

Soit $X = (X^{(1)}, \dots, X^{(p)})^\top$ un vecteur Gaussien d'espérance $\boldsymbol{\mu} \in \mathcal{R}^p$ et de matrice de covariance $\sigma^2 \mathbf{I}$. L'estimateur du maximum de vraisemblance est $\bar{X} \sim \mathcal{N}(\boldsymbol{\mu}, \frac{\sigma^2}{n})$.

On s'intéresse à la possibilité de produire un estimateur dont le risque quadratique est uniformément meilleur de celui de \bar{X} .

Comme $\bar{X} \sim \mathcal{N}(\boldsymbol{\mu}, \bar{\sigma}^2 \mathbf{I})$ avec $\bar{\sigma}^2 = \sigma^2/n$ on peut se restreindre au cas de l'échantillon de taille 1.

Définition 1 (Estimateur de James et Stein). L'estimateur de James et Stein pour l'espérance $\boldsymbol{\mu}$ d'une variable aléatoire X est défini par :

$$\hat{\boldsymbol{\mu}}^{(JS)} = \left(1 - \frac{\sigma^2(p-2)}{\|X\|^2}\right) X$$

Théorème 1. Pour $p \geq 3$, l'excès de risque de l'estimateur de James-Stein est

$$\mathbb{E}[\mathcal{R}(\hat{\boldsymbol{\mu}}^{(JS)}) - \mathcal{R}(\boldsymbol{\mu})] = \mathbb{E}[(\hat{\boldsymbol{\mu}}^{(JS)} - \boldsymbol{\mu})^2] = p\sigma^2 - (p-2)^2\sigma^4\mathbb{E}[\|X\|^{-2}] < p\sigma^2 = \mathbb{E}[(X - \boldsymbol{\mu})^2]$$

Exercice 1. Montrer à partir du résultat du théorème que :

$$\mathbb{E}[(\hat{\boldsymbol{\mu}}^{(JS)} - \boldsymbol{\mu})^2] \leq 4\sigma^2 + \frac{p\sigma^2\|\boldsymbol{\mu}\|^2}{p\sigma^2 + \|\boldsymbol{\mu}\|^2} \leq 4\sigma^2 + \sigma^2 \min\left(p, \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}\right)$$

On peut en fait avec une analyse plus fine montrer que

$$\mathbb{E}[(\hat{\boldsymbol{\mu}}^{(JS)} - \boldsymbol{\mu})^2] \leq 2\sigma^2 + \frac{(p-2)\sigma^2\|\boldsymbol{\mu}\|^2}{(p-2)\sigma^2 + \|\boldsymbol{\mu}\|^2} \leq \sigma^2 \min\left(p, 2 + \frac{\|\boldsymbol{\mu}\|^2}{\sigma^2}\right)$$

Paradoxe de Stein

Supposons qu'on cherche à estimer

- la vitesse de la lumière μ_1 ,
- la consommation de thé annuelle en Chine μ_2 ,
- la quantité de précipitation annuelle moyenne à Paris μ_3 .

L'estimateur de Stein permet d'obtenir des estimateurs $\hat{\mu}_1^{(JS)}$, $\hat{\mu}_2^{(JS)}$ et $\hat{\mu}_3^{(JS)}$ de ces trois quantités telles que l'erreur quadratique globale $\sum_{j=1}^3 \mathbb{E}[(\hat{\mu}_j^{(JS)} - \mu_j)^2]$ soit *strictement plus faible* que celle de la moyenne empirique!

De plus si on est en dimension $p \geq 4$ on a la variante de l'estimateur de James-Stein du à Lindley.

$$\boldsymbol{\mu}^{(L)} = \left(1 - \frac{\bar{\sigma}^2(p-3)}{\|\bar{X} - \bar{X}\mathbf{1}\|_2^2}\right) (\bar{X} - \bar{X}\mathbf{1}) + \bar{X}\mathbf{1}$$

avec $\bar{X} = \frac{1}{p} \sum_p \bar{X}^{(j)}$.

Etude à design fixe

En apprentissage supervisé, nous avons défini jusqu'ici le risque d'un prédicteur comme l'espérance sur les valeurs possible d'un nouveau couple entrée-sortie :

$$\mathcal{R}(f) = \mathbb{E}[\ell(f(X), Y)].$$

Nous allons maintenant modifier légèrement ce cadre et considérer le cadre du design fixe.

Dans ce cadre, on considère un ensemble d'entraînement $D_n = \{(x_1, y_1, \dots, x_n, y_n)\}$, et on souhaite faire en sorte que la prédiction soit la plus exacte possible pour les valeurs d'entrées x_i considérées, mais pour de nouvelles valeurs y'_i tirée de la même loi conditionnelle que les y_i , c'est-à-dire selon $\mathbb{P}(Y = \cdot | X = x_i)$.

Définition 2. Le *risque à design fixe* est défini comme

$$\mathcal{R}_n^{DF}(f) := \mathbb{E}_n[\mathbb{E}[\ell(f(x_i), Y) | X = x_i]] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\ell(f(x_i), Y) | X = x_i]$$

De façon équivalente, étant donné notre définition du risque empirique $\widehat{\mathcal{R}}_n(f) = \mathbb{E}_n[\ell(f(X), Y)] = \frac{1}{n} \sum_{i=1}^n \ell(f(X_i), Y_i)$, le risque à design fixe est l'espérance du risque empirique lorsqu'on conditionne sur les données d'entrée :

$$\mathcal{R}_n^{DF}(f) = \mathbb{E}[\widehat{\mathcal{R}}_n | X_1, \dots, X_n].$$

Une troisième manière d'envisager \mathcal{R}_n^{DF} est que c'est le risque si l'on suppose que la distribution des données d'entrées est uniforme sur les valeurs déjà observées.

Comme on s'en doute, l'étude à design fixe est plus adaptée au débruitage qu'à la prédiction.

Dans le cadre à design fixe, comme on ne cherche pas à généraliser à de nouvelles valeurs en entrée (mais seulement à bien généraliser sur de nouvelles valeurs en sortie), le prédicteur est entièrement spécifié par l'ensemble des valeurs qu'il prend sur l'ensemble x_1, \dots, x_n . Par conséquent, plutôt que de raisonner sur des prédicteurs qui sont des fonctions, étant donné un ensemble de données d'entrées (x_1, \dots, x_n) , il suffit de raisonner sur les vecteurs de valeurs $\mathbf{f} = (f(x_1), \dots, f(x_n))^\top$. On notera $\mathbf{y} = (y_1, \dots, y_n)^\top$ le vecteur dont les entrées sont les valeurs de sortie de l'ensemble d'entraînement.

De même que dans le cadre initial, on peut définir dans ce cadre une fonction cible $f^* = \operatorname{argmin}_f \mathcal{R}_n^{DF}(f)$ (lorsque ce minimum existe) et l'excès de risque d'un prédicteur f :

$$\mathcal{E}_n^{DF}(f) = \mathcal{R}_n^{DF}(f) - \mathcal{R}_n^{DF}(f^*).$$

Estimation linéaire

Définition 3 (Estimateur par transformation linéaire). Un estimateur par transformation linéaire (ETL) est un estimateur dont les valeurs sont obtenues en appliquant une transformation linéaire (dépendant en général des données d'entrée) aux données de sortie.

$$\hat{\mathbf{f}} = \mathbf{A}\mathbf{y} \quad \text{pour} \quad \mathbf{A} \in \mathcal{R}^{n \times n}.$$

Définition 4 (Estimateur par projection). Un estimateur par projection est un estimateur par transformation linéaire pour lequel A est une matrice de projection.

Exemples d'ETL :

- les estimateurs par moyennage local (histogrammes, Nadaraya-Watson, k -p.p.v.),
- l'estimateur de la régression linéaire ordinaire, de la régression ridge et de la régression dans un RKHS (voir cours prochain).

Exemples d'estimateurs par projection :

- estimateurs par histogrammes,
- estimateur de la régression linéaire ordinaire.

Pénalité C_L de Mallows

On s'intéresse aux estimateurs par projection pour la perte des moindres carrés. Pour un prédicteur \mathbf{f} fixe, le risque empirique et le risque à design fixe sont respectivement

$$\widehat{\mathcal{R}}_n(\mathbf{f}) = \frac{1}{n} \|\mathbf{y} - \mathbf{f}\|_2^2 \quad \text{et} \quad \mathcal{R}_n^{DF} = \frac{1}{n} \mathbb{E}[\|\mathbf{y} - \mathbf{f}\|_2^2] = \frac{1}{n} \|\mathbf{f} - \mathbf{f}^*\|_2^2 + \frac{1}{n} \mathbb{E}[\|\mathbf{y} - \mathbf{f}^*\|_2^2]$$

Supposons que $\mathbf{y} = \mathbf{f}^* + \boldsymbol{\varepsilon}$ avec $\varepsilon_1, \dots, \varepsilon_n$ i.i.d. $\mathbb{E}[\varepsilon_i] = 0$ et $\mathbb{E}[\varepsilon_i^2] = \sigma^2$.

Soit $\hat{\mathbf{f}} = \mathbf{A}\mathbf{y}$ un estimateur par transformation linéaire. Si l'on suppose que l'on connaît la variance du bruit σ^2 , peut-on utiliser cette information pour "corriger" le risque empirique.

$$n \mathbb{E}[\widehat{\mathcal{R}}_n(\hat{\mathbf{f}})] = \mathbb{E}[\|\mathbf{y} - \hat{\mathbf{f}}\|_2^2] = \mathbb{E}[\|(\mathbf{I} - \mathbf{A})\mathbf{y}\|_2^2] = \|(\mathbf{I} - \mathbf{A})\mathbf{f}^*\|_2^2 + \sigma^2 \|\mathbf{I} - \mathbf{A}\|_F^2$$

et comme $\mathcal{E}_n^{DF}(\hat{\mathbf{f}}) = \frac{1}{n} \|\hat{\mathbf{f}} - \mathbf{f}^*\|_2^2$, on a :

$$n \mathbb{E}[\mathcal{E}_n^{DF}] = \mathbb{E}[\|\hat{\mathbf{f}} - \mathbf{f}^*\|_2^2] = \mathbb{E}[\|\mathbf{A}\mathbf{y} - \mathbf{f}^*\|_2^2] = \mathbb{E}[\|(\mathbf{A} - \mathbf{I})\mathbf{f}^* + \mathbf{A}\boldsymbol{\varepsilon}\|_2^2] = \mathbb{E}[\|(\mathbf{A} - \mathbf{I})\mathbf{f}^* + \mathbf{A}\boldsymbol{\varepsilon}\|_2^2] = \|(\mathbf{A} - \mathbf{I})\mathbf{f}^*\|_2^2 + \sigma^2 \|\mathbf{A}\|_F^2$$

Nous avons encore

$$\mathbb{E}[\mathcal{R}_n^{DF}(\hat{\mathbf{f}})] = \mathbb{E}[\mathcal{E}_n^{DF}] + \frac{\sigma^2}{n} \|\mathbf{I}\|_F^2$$

En comparant les expressions

$$\mathbb{E}[\widehat{\mathcal{R}}_n(\hat{\mathbf{f}})] = \frac{1}{n} \|(\mathbf{I} - \mathbf{A})\mathbf{f}^*\|_2^2 + \frac{\sigma^2}{n} \|\mathbf{I} - \mathbf{A}\|_F^2 \quad \text{et} \quad \mathbb{E}[\mathcal{R}_n^{DF}(\hat{\mathbf{f}})] = \frac{1}{n} \|(\mathbf{I} - \mathbf{A})\mathbf{f}^*\|_2^2 + \frac{\sigma^2}{n} (\|\mathbf{I}\|_F^2 + \|\mathbf{A}\|_F^2)$$

on obtient

$$\boxed{\mathbb{E}[\mathcal{R}_n^{DF}(\hat{\mathbf{f}})] = \mathbb{E}[\widehat{\mathcal{R}}_n(\hat{\mathbf{f}})] + 2 \frac{\sigma^2}{n} \text{tr}(\mathbf{A})}$$

Si $\text{tr}(\mathbf{A}) > 0$, ce qui typiquement le cas en pratique, cela montre que le risque empirique sous-estime en espérance le risque de l'estimateur. On voit également que le biais est estimé correctement et que c'est la variance de l'estimateur qui n'est pas juste en espérance. L'estimation de la variance est biaisée.

Définition 5 (Degrés de liberté, pénalité C_L de Mallows).

On appelle *nombre de degrés de liberté* de l'estimateur linéaire (ETL) basé sur \mathbf{A} la quantité $\text{tr}(\mathbf{A})$.

On appelle *pénalité C_L de Mallows* le terme $2 \frac{\sigma^2}{n} \text{tr}(\mathbf{A})$.

Le résultat ci-dessous montre que la quantité obtenue en pénalisant le risque empirique avec le C_L de Mallows, soit

$$\widehat{\mathcal{R}}_n(\hat{\mathbf{f}}) + 2 \frac{\sigma^2}{n} \text{tr}(\mathbf{A}),$$

est un estimateur sans biais du risque à design fixe.

Pénalité C_p de Mallows

La pénalité C_p de Mallows est un cas particulier de la pénalité C_L lorsque, $\hat{\mathbf{f}}$ est un estimateur par projection.

Dans ce cas \mathbf{A} est un projecteur et $\text{tr}(\mathbf{A})$ est la dimension du sous-espace sur lequel on projette. Soit $p = \text{tr}(\mathbf{A})$ cette dimension, le C_p de Mallows est la quantité $2p \frac{\sigma^2}{n}$. La pénalisation de Mallows est alors

$$\widehat{\mathcal{R}}_n(\hat{\mathbf{f}}) + 2p \frac{\sigma^2}{n}$$

Exercice 2. Montrer que le prédicteur de la régression linéaire ordinaire est un estimateur par projection. Quelle est la dimension p correspondante ?

Exercice 3. Supposer que les données sont générées selon $y_i = x_i^\top \beta + \varepsilon_i$. Montrer que le prédicteur de la régression ridge est un ETL. Montrer qu'à design fixe et sous les hypothèses de bruit i.i.d., on peut exprimer la variance du prédicteur seulement en fonction de σ^2 , des valeurs singulières de la matrice de design λ_i et du paramètre de régularisation λ . Montrer que s'il on note $\tilde{\beta}$ la représentation de β dans une base orthonormale bien choisie, le biais de la régression ridge s'exprime en fonction de $\tilde{\beta}$, σ^2 , λ et les valeurs singulières de \mathbf{A} .

Exercice 4. Comment utiliser le C_L de Mallows pour choisir λ pour la régression ridge si l'on connaît σ^2 ?