Minimiser le risque empirique pour des pertes à queue lourde

Emilien Joly

Université Paris Sud

sous la direction de Gábor Lugosi et de Gilles Stoltz

11 avril 2014

Quelle question ?

Un statisticien (même le vendredi matin) travaille avec un échantillon $Z_1, \dots, Z_n$ i.i.d. dans un espace plus ou moins complexe $\mathcal{Z}$.
Et on optimise...

Exemples :

- Si $Z_i = (X_i, Y_i)$, on cherche a minimiser $\sum_{i=1}^n |Y_i - aX_i - b|$. Plus particulièrement, on cherche les valeurs de a et de b optimales.
- Moindre carrés : $\min_{a,b} \sum_{i=1}^n (Y_i - aX_i - b)^2$
- Maximum de vraisemblance : $\max_\theta \sum_{i=1}^n \log f_\theta(Z_i)$

Quelle question ?

Dans tout ces cas, on cherche à minimiser une quantité empirique "s'approchant" d'une espérance. Dans l'idéal, le choix du minimum serait $\operatorname{argmin}_\ell \mathbb{E}\left[\ell(Z)\right]$.

1. Cadre de l'estimation robuste
2. Estimateur de Catoni
3. Résolution sur une classe $\mathcal{F}$
4. Un exemple de perte

Risque et pertes

Soit Z une variable aléatoire réelle. Z peut être une variable réelle, à valeur dans $\mathbb{R}^d$ ou même dans un espace mesurable quelconque Ω.

- Une fonction de risque est $\ell : \Omega \to [0, +\infty[$. On l'appelle certaines fois fonction de coût ou fonction de perte.
- Choix de ℓ : $\mathcal{F}$ ensemble des candidats.

On se réfère à la perte optimale théorique ℓ_{opt} définie par :

$$\ell_{opt} \in \operatorname*{argmin}_{\ell \in \mathcal{F}} \mathbb{E}\left[\ell(Z)\right]$$

Pour toute fonction de perte ℓ, le risque prend la forme :

$$R(\ell) = \mathbb{E}\left[\ell(Z)\right] - \mathbb{E}\left[\ell_{opt}(Z)\right]$$

Objectif : Trouver une estimation $\hat{\ell}$ rendant $R(\hat{\ell})$ le plus petit possible.

Minimisation du risque empirique classique

Le statisticien a accès à $(Z_1, \ldots, Z_n)$ un échantillon de la loi de Z.
Un estimateur de risque empirique $\hat{\ell}_{erm}$ est défini par :

$$\hat{\ell}_{erm} \in \operatorname*{argmin}_{\ell \in \mathcal{F}} \sum_{i=1}^{n} \ell(Z_i)$$

Le risque de cet estimateur dépend fortement du comportement de $\sum_{i=1}^{n} \ell(Z_i)$.

Deux types d'hypothèses de concentration :

- $\mathrm{Var}(\ell(Z)) < \infty \implies \sum_{i=1}^{n} \ell(Z_i)$ à queue sous-quadratique
- ℓ bornée $\implies \sum_{i=1}^{n} \ell(Z_i)$ à queue sous-Gaussienne

Peut-on faire mieux ? On ne s'autorise que des hypothèses faibles sur la distribution Z.

Queue sous-Gaussienne

Une inégalité de concentration a la forme :

$$\mathbb{P}(Z > u) \leq g(u)$$

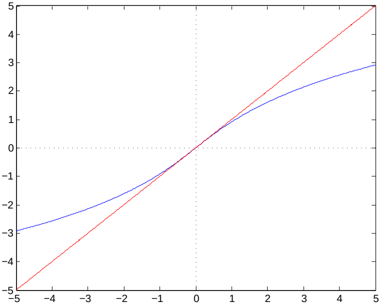
où g est une fonction décroissant plus ou moins rapidement en u.
La concentration sous-Gaussienne est donnée par
$g(u) = K \exp(-\frac{u^2}{v})$, plus utilisée sous la forme

$$Z \leq \sqrt{c \ln(\frac{K}{\varepsilon})} \text{ avec probabilité au moins } 1 - \varepsilon$$

Un estimateur robuste pour les queues lourdes

On définit une fonction ϕ de troncature.

$$\phi(x) = \mathbb{1}_{\{x \geq 0\}} \log(1 + x + \frac{x^2}{2}) - \mathbb{1}_{\{x < 0\}} \log(1 - x + \frac{x^2}{2})$$



Le nouvel estimateur est définit comme l'unique solution $\hat{\mu}_\ell$ de :

$$\sum_{i=1}^{n} \phi\Big(\alpha\big(\ell(Z_i) - \mu\big)\Big) = 0$$

α est un paramètre à optimiser.

Concentration autour de la moyenne

Pour **chaque** ℓ on obtient une concentration quasi-sous-Gaussienne.

Théorème (Catoni 2011)

Soit $\varepsilon > 0$, $\alpha = \sqrt{\frac{2\log\varepsilon^{-1}}{vn}}$. *On suppose que* $n > 2\log\varepsilon^{-1}$ *et que* $Var(\ell(Z)) \leq v$. *L'estimateur* $\hat{\mu}_\ell$ *vérifie avec probabilité plus grande que* $1 - 2\varepsilon$,

$$|\mathbb{E}[\ell(Z)] - \hat{\mu}_\ell| \leqslant \sqrt{\frac{2v\log\varepsilon^{-1}}{n}}$$

Le choix de ε dépend de n.

Par α l'estimateur dépend de la variance v et du niveau de confiance ε. Il faut donc connaître une borne supérieure de la variance *à priori*.

Entropie

Le cadre non paramétrique demande un contrôle de l'espace $\mathcal{F}$. On défini l'entropie :

Definition

On munit l'espace $\mathcal{F}$ d'une distance d. On appelle Entropie la quantité :

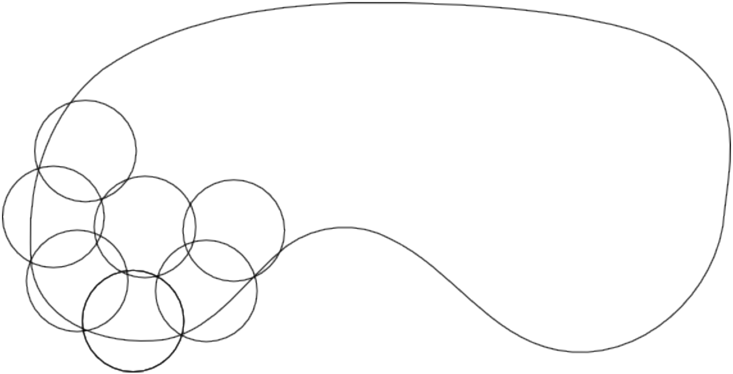
$$\gamma_d = \int_0^{diam_d \mathcal{F}} \sqrt{\log N(\mathcal{F}, d, \varepsilon)} d\varepsilon$$

où $N(\mathcal{F}, d, \varepsilon)$ est le nombre minimal de boules de rayon ε recouvrant $\mathcal{F}$.

Cette quantité mesure la complexité de l'espace $\mathcal{F}$.
On notera γ_2 l'entropie pour la distance L_2.

Un aperçu du chaining

L'idée du chaining est de contrôler le $\mathbb{E}\left[\sup_{f \in \mathcal{F}} X_f\right]$.



Un aperçu du chaining

Soient X_f des variables aléatoires d'espérance nulle. On suppose que $\mathbb{P}(X_f - X_{f'} > t) \leq \exp(-\frac{t^2}{d(f,f')^2})$

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} X_f\right] \leq K\gamma_d(\mathcal{F})$$

On récupère aussi une concentration sous-Gaussienne pour $\sup_{f \in \mathcal{F}} X_f$

Résultat principal

De la même façon on peut définir : $\hat{\ell} = \operatorname{argmin} \hat{\mu}_\ell$.
On désigne par δ_2 la norme quadratique sur $\mathcal{F}$. Le risque de $\hat{\ell}$ est contrôlé par une queue sous-gaussienne :

Theorem

Soit $\varepsilon > 0$. On choisit $\alpha = \sqrt{2 \log \varepsilon^{-1}/(vn)}$. On suppose que pour tout $\ell \in \mathcal{F}$, $Var(\ell(Z)) < v$. Enfin on suppose que γ_{δ_2} est finie. Alors avec probabilité plus grande que $1 - 6\varepsilon$,

$$\mathbb{E}\left[\hat{\ell}(Z)\right] - \mathbb{E}\left[\ell_{opt}(Z)\right] \leq \square \sqrt{\log \varepsilon^{-1}} \left(\sqrt{\frac{v}{n}} + \sqrt{\frac{\gamma_{\delta_2}^2}{n}} + o\left(\frac{1}{\sqrt{n}}\right) \right)$$

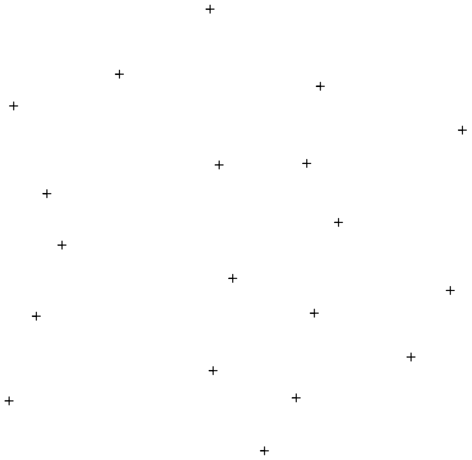
où $\square$ désigne une constante universelle.

k-means

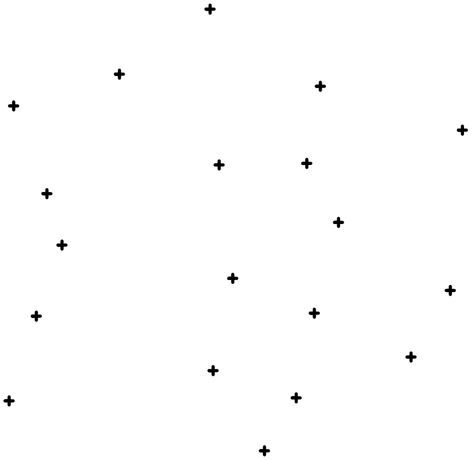
Prenons l'exemple d'une classe de fonctions vivant dans un espace de dimension finie d.
k-means : Soit Z une variable aléatoire de $\mathbb{R}^d$ de variance finie V. A l'aide de k éléments c_i de l'espace $\mathbb{R}^d$, on veut "décrire" la distribution de Z.

$$\ell(Z) = \min_{1 \leq i \leq k} ||Z - c_i||$$

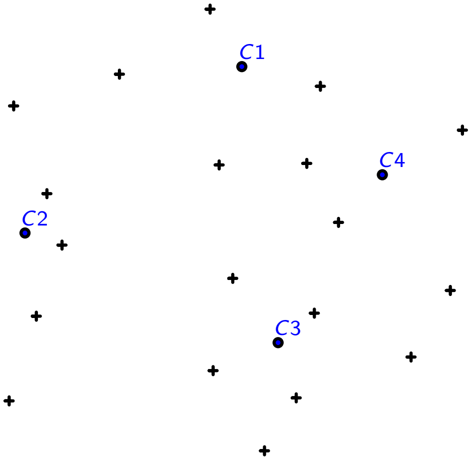
k-means



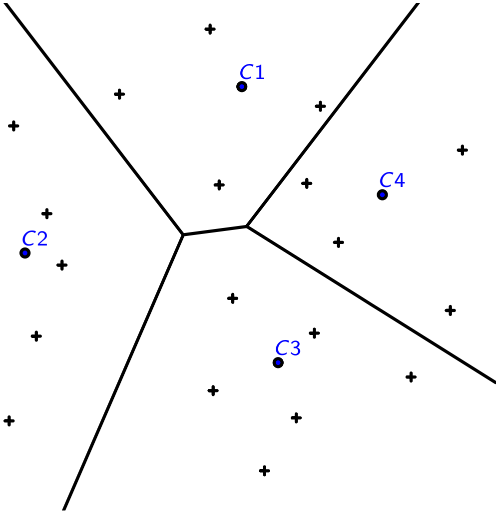
k-means



k-means



k-means



k-means

Prenons l'exemple d'une classe de fonctions vivant dans un espace de dimension finie d.
k-means : Soit Z une variable aléatoire de $\mathbb{R}^d$ de variance finie V.
A l'aide de k éléments c_i de l'espace $\mathbb{R}^d$, on veut "décrire" la distribution de Z.

$$\ell(Z) = \min_{1 \leq i \leq k} ||Z - c_i||$$

k-means

Prenons l'exemple d'une classe de fonctions vivant dans un espace de dimension finie d.
k-means : Soit Z une variable aléatoire de $\mathbb{R}^d$ de variance finie V. A l'aide de k éléments c_i de l'espace $\mathbb{R}^d$, on veut "décrire" la distribution de Z.

$$\ell(Z) = \min_{1 \leq i \leq k} ||Z - c_i||$$

On peut montrer qu'avec "grande probabilité" $(\hat{c}_1, \ldots, \hat{c}_k) \in B(0, R)$. R ne dépendant que de la distribution de Z.

Il faut calculer la variance de $\ell(Z)$! Elle est majorée par $2Vk$.
Il faut calculer γ_{δ_2} ! Nous avons donc besoin de recouvrir la boule de rayon R par des boules de rayon ε. On a un majorant bien connu :

$$N_{d_2}(B(0,R),\epsilon) \leq \left(\frac{4R}{\epsilon}\right)^d$$

donc $\gamma_{\delta_2} \leq C\sqrt{d}$
La borne est donc de la forme
$\mathbb{E}\left[\hat{\ell}(Z)\right] - \mathbb{E}\left[\ell_{opt}(Z)\right] \leq C\sqrt{\log\frac{1}{\varepsilon}}(\sqrt{\frac{Vk}{n}} + \sqrt{\frac{d}{n}})$

Conclusion

- Simplifier le résultat lorsque l'on se trouve en fait dans un cadre paramétrique.
- Qu'est-ce que ça donne lorsque les données sont sparses ?
- Peut-on appliquer cette technique à des estimations robustes alternatives ?

Bibliographie I

- S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities : A nonasymptotic theory of independence*, Oxford University Press, 2013.
- P.L. Bartlett and S. Mendelson, *Empirical minimization*, Probability Theory Related Fields **135** (2006), 311–334.
- O. Catoni, *Challenging the empirical mean and empirical variance : a deviation study*.
- T. Linder, *Learning-theoretic methods in vector quantization*, Principles of nonparametric learning, Springer-Verlag, 2002.
- M. Talagrand, *The generic chaining*, Springer, 2005.
- S. van de Geer, *Empirical processes in M-estimation*, Cambridge University Press, Cambridge, UK, 2000.

Merci !

Des questions ?