Complétion de matrice 0-1 : Étude PAC-Bayésienne et Approximation variationnelle

Vincent Cottet (avec P. Alquier)
Directeur de thèse : N. Chopin

JPS - 18-22 avril 2016 - Les Houches

Introduction

Cadre général

Les données $(X_i, Y_i), i \in \{1, \ldots, n\}$

- $X_i \in \{1, \ldots, m_1\} \times \{1, \ldots, m_2\}$
- $Y_i \in \{-1, 1\}$

$\Rightarrow$ But : prédire Y.

Hypothèse Courante

Structure de faible rang.

$\Rightarrow$ **Dans ce cas, on peut arriver à faire quelque chose.**

Utilisation : Système de recommandation

Exemple Jouet : Recommandation de film

	Film					
Utilisateur	James Bond	Toy Story	Batman	Heat	Psycho	...
Michel		-1	1	1		...
Vincent	-1		1		-1	...
Pierre			1			...
Keefe	1			1	-1	...
Gosia		-1			1	...
Emma		-1		1	-1	...
...	...	...	...	...	...	...

Introduction - Modélisation habituelle

Loi de génération des données

$$Y|X=(k,l)=\begin{cases}1 & \text{avec proba } f\left(M^0_{k,l}\right)\\ -1 & \text{avec proba } 1-f\left(M^0_{k,l}\right)\end{cases}$$

$f:\mathbb{R}\to[0,1]$ est la fonction de lien.

Hypothèse et Résultat

- Hypothèse : M^0 est de faible rang.
- Estimateur :

$$\widehat{M}=\arg\min_M -\log\mathcal{L}(M)+\lambda\|M\|_*$$

- Résultat : Reconstruction de M^0

Plan

Cadre différent : Classification (Machine Learning)

- Risque pertinent : 0-1

$$r_n(M) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(Y_i \neq \operatorname{sign}(MX_i)\right),$$

$$R(M) = \mathbb{P}(Y \neq \operatorname{sign}(MX))$$

Cadre différent : Classification (Machine Learning)

- Risque pertinent : 0-1

$$r_n(M) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(Y_i \neq \operatorname{sign}(MX_i)\right),$$
$$R(M) = \mathbb{P}(Y \neq \operatorname{sign}(MX))$$

- Risque Charnière (Hinge Loss)

$$r_n^h(M) = \frac{1}{n}\sum_{i=1}^{n}(1 - Y_i M X_i)_+,$$
$$R^h(M) = \mathbb{E}\left[(1 - YMX)_+\right]$$

Estimateur PAC-Bayésien

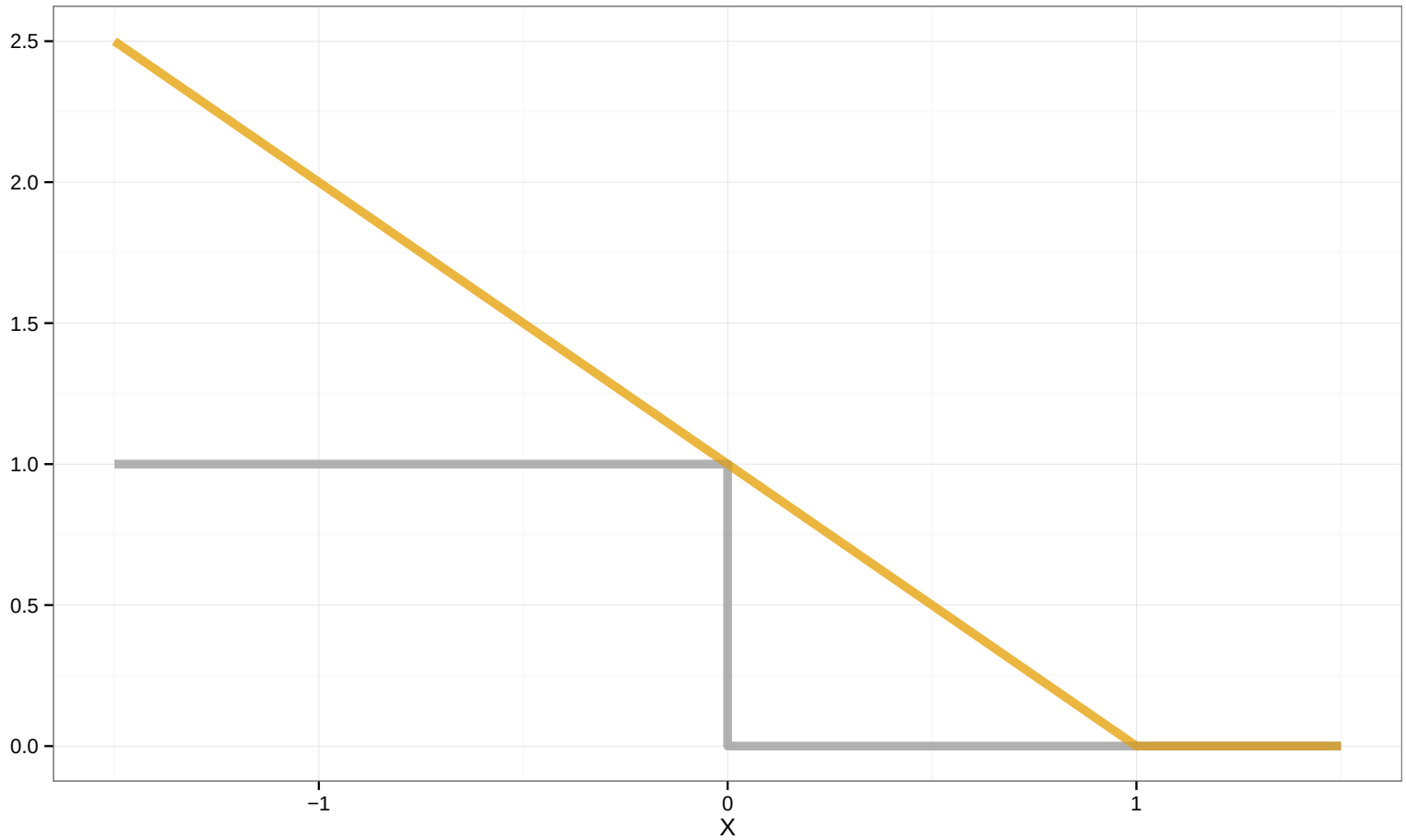

FIGURE : risque 0-1 vs risque charnière (Y=1)

Loi a priori sur l'ensemble des matrices

Décomposition $M = LR^\top, \dim(L) = m_1 \times K, \dim(R) = m_2 \times K$ avec :

$$L_{i,k}, R_{j,k} | \gamma_k \sim \mathcal{N}(0, \gamma_k)$$

$$\frac{1}{\gamma_k} \sim \Gamma(a, b)$$

On note $\theta = (L, R, \gamma)$.

Estimateur PAC-Bayésien

Loi a priori sur l'ensemble des matrices

Décomposition $M = LR^\top, \dim(L) = m_1 \times K, \dim(R) = m_2 \times K$ avec :

$$L_{i,k}, R_{j,k} | \gamma_k \sim \mathcal{N}(0, \gamma_k)$$

$$\frac{1}{\gamma_k} \sim \Gamma(a, b)$$

On note $\theta = (L, R, \gamma)$.

Objet d'intérêt : la loi a posteriori

$$p(d\theta) \propto \mathcal{L}(\theta)\pi(d\theta)$$

Estimateur PAC-Bayésien

Loi a priori sur l'ensemble des matrices

Décomposition $M = LR^\top, \dim(L) = m_1 \times K, \dim(R) = m_2 \times K$ avec :

$$L_{i,k}, R_{j,k} | \gamma_k \sim \mathcal{N}(0, \gamma_k)$$

$$\frac{1}{\gamma_k} \sim \Gamma(a, b)$$

On note $\theta = (L, R, \gamma)$.

Objet d'intérêt : la loi pseudo-posterior

$$p(d\theta) \propto \exp[-\lambda r_n^h(LR^\top)]\pi(d\theta)$$

Problème : Difficile à calculer

Estimateur PAC-Bayésien

Méthode *Rapide*

Recherche d'une distribution approchée dans une famille $\mathcal{F}$.

$$\widetilde{\rho_\lambda} = \arg\min_{\rho \in \mathcal{F}} \mathcal{KL}(\rho, p))$$

Note : $\mathcal{KL}(\rho, p) = \int \rho \log \frac{\rho}{p}$

Aspects pratiques

- Choix de la famille $\mathcal{F}$ pour un problème simple.
- Problème d'optimisation sur un paramètre de dim. fini.
- Optimisation bi-convexe, pas de garantie de convergence.

Plan

Résultats Théoriques

Meilleur prédicteur

prédicteur de Bayes :

$$\forall (k,l), M^B_{k,l} = \operatorname{sign}\left(\mathbb{E}[Y|X=(k,l)]\right)$$

Contrôle du risque dans un cas restrictf

- Y est observé sans bruit.
- M^B est de rang r (potentiellement petit)

Alors

$$\int R d\widetilde{\rho_\lambda} \leq C \left[\frac{r(m_1+m_2)(\log n + \ell) + \log \frac{1}{\epsilon}}{n} \right]$$

avec probabilité $1-\epsilon$.

Résultats Théoriques

Meilleur prédicteur

prédicteur de Bayes :

$$\forall (k,l), M^B_{k,l} = \text{sign}\left(\mathbb{E}[Y|X=(k,l)]\right)$$

Contrôle du risque dans un cas restrictf

- Y est observé avec un bruit switch, de proba p.
- M^B est de rang r (potentiellement petit)

Alors

$$\int R d\widetilde{\rho_\lambda} \leq C \left[\frac{r(m_1+m_2)(\log n + \ell) + \log \frac{1}{\epsilon}}{n} \right]$$

avec probabilité $1-\epsilon$.

Résultats Théoriques

Meilleur prédicteur

prédicteur de Bayes :

$$\forall (k,l), M^B_{k,l} = \operatorname{sign}\left(\mathbb{E}[Y|X=(k,l)]\right)$$

Contrôle du risque dans un cas restrictf

- Y est observé avec un bruit switch, de proba p.
- M^B est de rang r (potentiellement petit)

Alors

$$\int Rd\widetilde{\rho_\lambda} \leq C \left[\frac{r(m_1+m_2)(\log n + \ell) + \log\frac{1}{\epsilon}}{n}\right] + (2+\delta)p$$

avec probabilité $1-\epsilon$.

Plan

Illustration

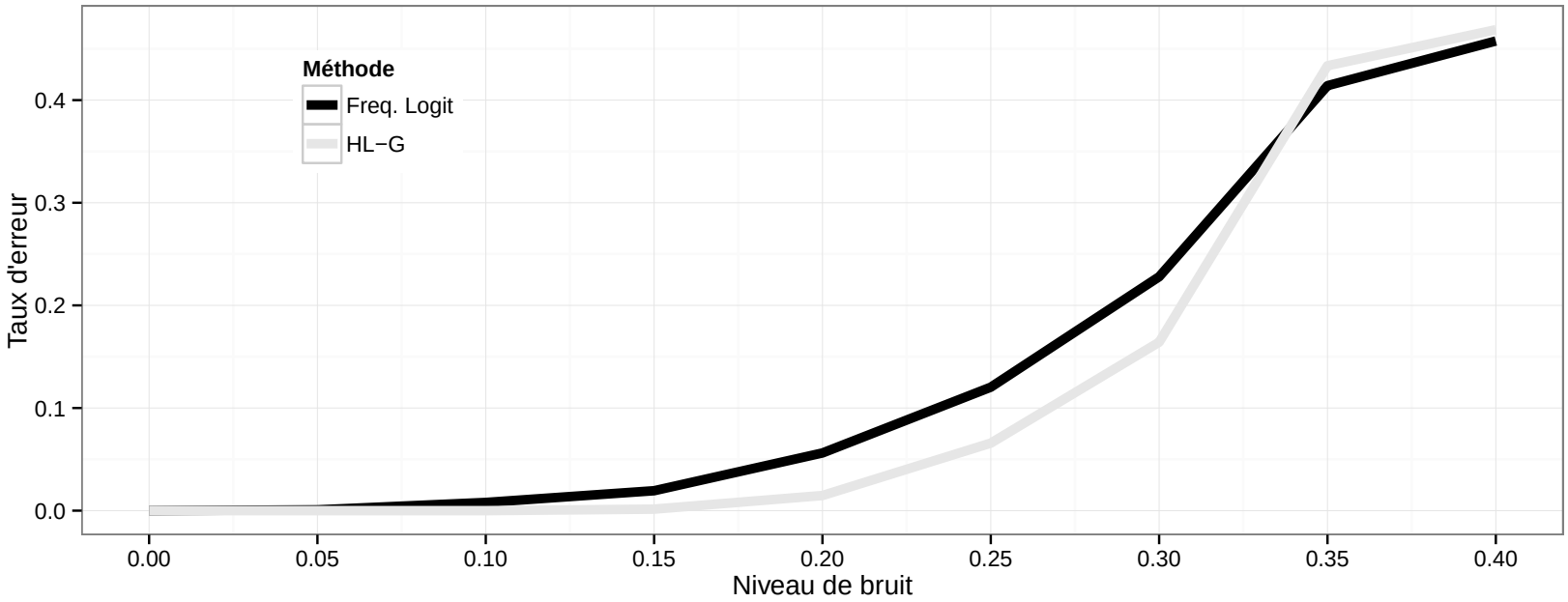

FIGURE : Résultats avec différents niveaux de bruit "switch"

Références Bibliographiques

- **Résultats probabilistes sur la complétion de matrice :**
  Candès, Tao, *The power of convex relaxation : near-optimal matrix completion*, 2010
- **Complétion de matrice 0-1 :**
  Lafond, Klopp, Moulines, Salmon, *Probabilistic low-rank matrix completion on finite alphabets*, 2014
- **Borne sur les estimateurs Bayésiens variationnels :**
  Alquier, Ridgway, Chopin *On the properties of variational approximations of Gibbs posteriors*, 2015
- **Complétion de matrice 0-1 par des méthodes PAC-Bayésiennes :**
  Cottet, Alquier, *1-bit Matrix Completion : PAC-Bayesian Analysis of a Variational Approximation*, 2016

Merci pour votre attention

Questions ?