MAP 433
Statistique

Christophe Giraud

Université Paris Sud et Ecole Polytechnique

PC4

Quels métiers en Mathématiques Appliquées?

The worst and best jobs in 2014 (from CareerCast, 2014)

The Best	The Worst
1. Mathematician	200. Lumberjack
2. University Professor	199. Newspaper Reporter
3. Statistician	198. Enlisted Military Personnel
4. Actuary	197. Taxi Driver
5. Audiologist	196. Broadcaster

Un panorama partiel des metiers en Mathématiques Appliquées

Secteur	Exemples d'employeurs	Métiers
Business Analytics et optimisation de la production	Cabinets de consultants (Capgemini, Accenture, etc), PME / start-up (fifty-five, vertica, Mu Sigma, Eurodecision, MFG labs, Palantir, spatialytics, etc) la plupart des grands groupes (en interne)	• Marketing • Gestion des ressources (approvisionnement, ressources humaines) • Gestion des tarifications • Optimisation de la conception et des procédés • Recherche opérationnelle
Services web et logiciels	Services web (Google, Yahoo, etc), Logiciels (Microsoft, IBM, Dassault-system, Xerox, etc) SSII et start-up (IBM, Logica, CSC, GFI informatique, etc) Paiement électronique (Visa, E-commerce, etc)	• Moteurs de recherche, fonctionnalités web, etc • Logiciels génériques ou solutions spécialisées • Cryptographie (services sécurisés)
R&D réseaux et communication	Opérateurs mobile (Orange, Bouygues, Free, SFR, etc), Constructeurs (Alcatel-Lucent, Huawei, Ericsson, Sagem, etc),	• Planification réseaux • Prospective technologie et équipements • Qualité de service
Analyste statisticien	Industrie pharmaceutique (Sanofi, Servier, etc), Biointelligence, Toutes les branches de l'industrie / agroalimentaire, Organismes parapubliques (sécurité sanitaire, surveillance d'épidémie / pollution, services sociaux et de santé, etc)	• Biostatistiques • Production d'indices et prévision (trafic, consommation, ozone, coûts, marché, etc)
R&D signal & images	Thales, Safran, Dassault-system, Matra, General Electric, etc	• Traitement du signal et images • Guidage et contrôle • Imagerie médicale
R&D énergie, transport et environnement	RTE, EDF, Areva, Veolia, SNCF, Schlumberger, Industrie pétrolière (Total, etc), Michelin, Renault, PSA, EADS, Dassault-aviation, Air-France, Altran, Akka technologies, etc	• Analyse, prévision, prospective • Gestion des risques • Modélisation, dimensionnement, conception • Simulation numérique

http://www.cmap.polytechnique.fr/~giraud/MetiersMaths.html

Information de Fisher

Optimalité

Optimalité uniforme: minimiser $\mathbb{E}_\theta[(\widehat{\theta} - \theta)^2]$ pour tout θ : impossible!

Optimalité minimax: minimiser $\sup_{\theta \in \Theta} \mathbb{E}_\theta[(\widehat{\theta} - \theta)^2]$.

Cadre asymptotique: si $\sqrt{n}(\widehat{\theta} - \theta) \overset{\text{loi}}{\to} \mathcal{N}(0, v(\theta))$, chercher $v(\theta)$ minimal.

Modèle régulier

$(X_1, \dots, X_n) \sim \mathbb{P}_{\theta_0} \in (\mathbb{P}_\theta)_{\theta \in \mathbb{R}^d}$ avec $d\mathbb{P}_\theta(x) = p(\theta, x)\, d\mu(x)$ tel que

- $\ell_x(\theta) := \log p(\theta, x)$ est D^2 en θ
- hypothèses assurant "$\partial_\theta \int = \int \partial_\theta$" (voir cours)

Information de Fisher

Pour un modèle régulier:

- Si $\theta \in \mathbb{R}$:

$$I_X(\theta) = \mathbb{E}_\theta\left[(\ell'_X(\theta))^2\right] = -\mathbb{E}_\theta\left[\ell''_X(\theta)\right]$$

- Si $\theta \in \mathbb{R}^d$:

$$I_X(\theta) = \mathbb{E}_\theta\big[\underbrace{(\nabla \ell_X(\theta))(\nabla \ell_X(\theta))^T}_{[\partial_i \ell \partial_j \ell]_{i,j}}\big] = -\mathbb{E}_\theta\big[\underbrace{H_X(\theta)}_{[\partial_{i,j} \ell]_{i,j}}\big] \in \mathcal{S}^+(\mathbb{R}^d)$$

Cadre i.i.d.

Si $d\mathbb{P}_\theta(x_1, \dots, x_n) = f(\theta, x_1) \dots f(\theta, x_n)\, d\mu(x_1) \dots d\mu(x_n)$, on a $I_X(\theta) = n I_{X_1}(\theta)$.

Cadre i.i.d.

Z-estimateur

- ϕ telle que $\mathbb{E}_\theta[\phi(\theta, X)] = 0$
- $\widehat{\theta}$ solution de $\sum_i \phi(\widehat{\theta}, X_i) = 0$.

Loi asymptotique

Si $\widehat{\theta}$ est un Z-estimateur régulier associé à ϕ on a

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{\text{loi}} Z_\phi \sim \mathcal{N}(0, v_\phi(\theta)) \quad \text{avec} \quad v_\phi(\theta) = \frac{\mathbb{E}_\theta[\phi(\theta, X_1)^2]}{\mathbb{E}_\theta[\partial_\theta \phi(\theta, X_1)]^2}.$$

Optimalité

Si le modèle est **régulier**:

- $v_\phi(\theta) \geq I_{X_1}(\theta)^{-1}$
- $\sqrt{n}(\hat{\theta}_{MV} - \theta) \xrightarrow{\text{loi}} Z \sim \mathcal{N}\left(0, I_{X_1}(\theta)^{-1}\right)$

Interprétation heuristique

- **La fonction** $H_{\theta_0}(\alpha) = \mathbb{E}_{\theta_0}[\ell_{X_1}(\alpha)]$ est maximale en $\alpha = \theta_0$ donc

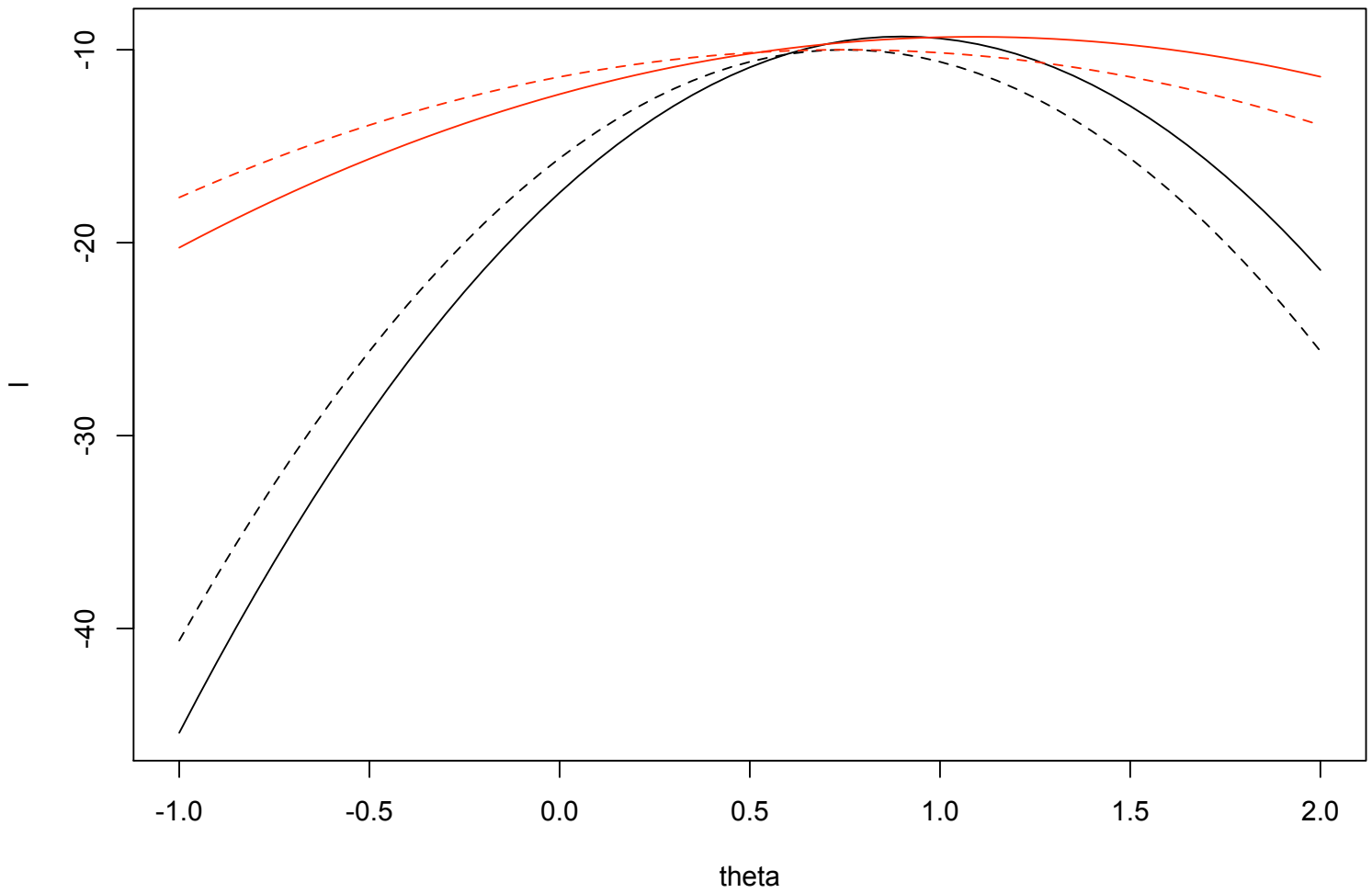
$$\begin{aligned} H_{\theta_0}(\alpha) &\approx H_{\theta_0}(\theta_0) + \frac{(\theta_0 - \alpha)^2}{2} H''_{\theta_0}(\theta_0) \quad \text{pour } \alpha \approx \theta_0 \\ &\approx H_{\theta_0}(\theta_0) - \frac{(\theta_0 - \alpha)^2}{2} I_{X_1}(\theta_0) \end{aligned}$$

- **TCL:** $n^{-1}\ell_X(\alpha) = \mathbb{E}_{\theta_0}[\ell_{X_1}(\alpha)] + \xi_n(\alpha)$
avec $\mathbb{E}_{\theta_0}[\xi_n(\alpha)] = 0$ et $\mathrm{var}_{\theta_0}[\xi_n(\alpha)] = \mathrm{var}_{\theta_0}[\ell_{X_1}(\alpha)]/n$

- **Conclusion:** pour α dans un voisinage de θ_0

$$n^{-1}\ell_X(\alpha) \approx H_{\theta_0}(\theta_0) - \frac{(\theta_0 - \alpha)^2}{2} I_{X_1}(\theta_0) + O(n^{-1/2})$$

Illustration (loi Gaussienne)



Log-vraisemblance (plein) versus sa moyenne $H_{\theta_0}(\theta)$ (pointillé) pour $\sigma = 1$ et $\sigma = 2$.

Estimation bayésienne

Choisir au mieux un estimateur

Observation: $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ de loi $\mathbb{P}_{\theta_0} \in \{\mathbb{P}_\theta \ : \ \theta \in \mathbb{R}\}$

Un estimateur: à tout $\widehat{\theta} : \mathbb{R}^n \to \mathbb{R}$ mesurable on associe l'estimateur $\widehat{\theta}(X)$ de θ

Qualité

$$R_{\theta_0}(\widehat{\theta}) = \mathbb{E}_{\theta_0}\left[(\theta_0 - \widehat{\theta}(X))^2\right]$$

Meilleur estimateur?

Risques minimax et bayésien

Cadre fréquentiste

On ne connait rien sur θ_0:

$$\text{Risque minimax}: \quad R^*(\widehat{\theta}) = \sup_{\theta \in \mathbb{R}} R_\theta(\widehat{\theta})$$

Cadre bayésien

θ_0 issu d'un tirage selon une loi π connue (avec $\int \theta^2 \, d\pi(\theta) < +\infty$)

$$\text{Risque bayésien}: \quad R^\pi(\widehat{\theta}) = \int R_\theta(\widehat{\theta}) \, d\pi(\theta).$$

$\pi =$ loi a priori

Exemple illustratif

Indexation sémantique automatique de pages web

1. une page est tirée au hasard
2. une analyse du texte est réalisée
3. la page est classifiée automatiquement dans diverses catégories (cuisine, sport, news, littérature, etc)

Avec quelle information classifier?

- l'analyse sémantique
- les proportions (connues) de chaque catégories

A piori: Si $\pi(\theta)$ est la proportion de page de catégorie θ, avant même d'analyser la page on a l'a priori que la page a $\pi(\text{news})/\pi(\text{littérature})$ plus de chance d'être des news que de la littérature.

Echantillonnage

1. θ est tiré selon π
2. $X = (X_1, \ldots, X_n)$ est tiré selon la loi $\mathbb{P}_\theta$

Le couple (θ, X) est donc tiré selon la loi $dp(\theta, x) = d\pi(\theta)\, d\mathbb{P}_\theta(x)$

Risque bayésien

$$\begin{aligned} R^\pi(\hat{\theta}) &= \int_{\mathbb{R}} \mathbb{E}_\theta \left[(\widehat{\theta} - \theta)^2 \right] \, d\pi(\theta) \\ &= \mathbb{E}_p \left[(\tilde{\theta} - \widehat{\theta}(\tilde{X}))^2 \right] \end{aligned}$$

où $(\tilde{\theta}, \tilde{X})$ est distribué selon la loi p.

Estimateur bayésien

Hypothèses

Supposons que

- $d\pi(\theta) = \pi(\theta)\, d\theta$ avec $\theta \in \mathbb{R}$
- $d\mathbb{P}_\theta(x) = F(\theta, x)\, dx$ avec $x \in \mathbb{R}^n$

d'où

$$dp(\theta, x) = F(\theta, x)\pi(\theta)\, d\theta\, dx.$$

Questions

1. Expliciter $R^\pi(\hat{\theta})$ en fonction de F et π.
2. Quel est le meilleur estimateur $\widehat{\theta}^\pi$ au sens du risque bayésien?

Bayesian or not Bayesian?

Estimateur bayésien: $\widehat{\theta}^{\pi} = \int_{\theta} \theta \, d\pi(\theta|X).$

Pour and contre

Pour	Contre
• Incorporation de connaissances a priori • Formule explicite pour l'estimateur optimal	• π? • difficulté de calcul en grande dimension: méthodes MCMC