Apprentissage statistique multi-tâches

Matthieu Solnon [1]

Directeurs de thèse : Sylvain Arlot et Francis Bach

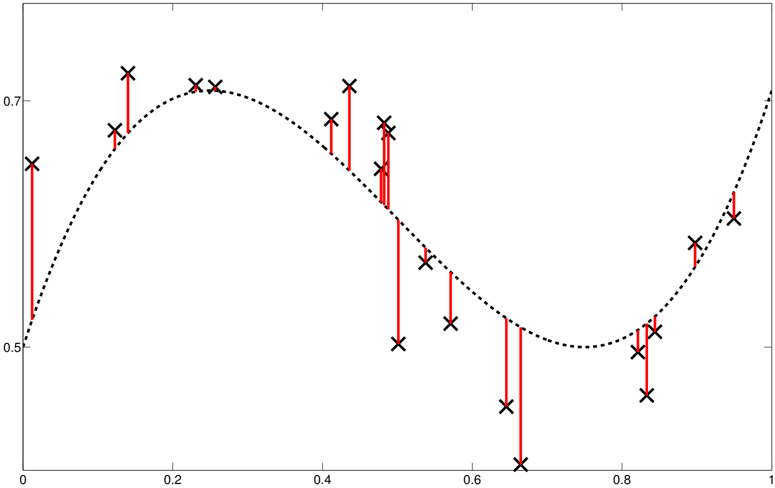
Équipe Sierra
Département d'Informatique de l'École normale supérieure
(CNRS/ENS/INRIA UMR 8548)

Lundi 25 novembre 2013

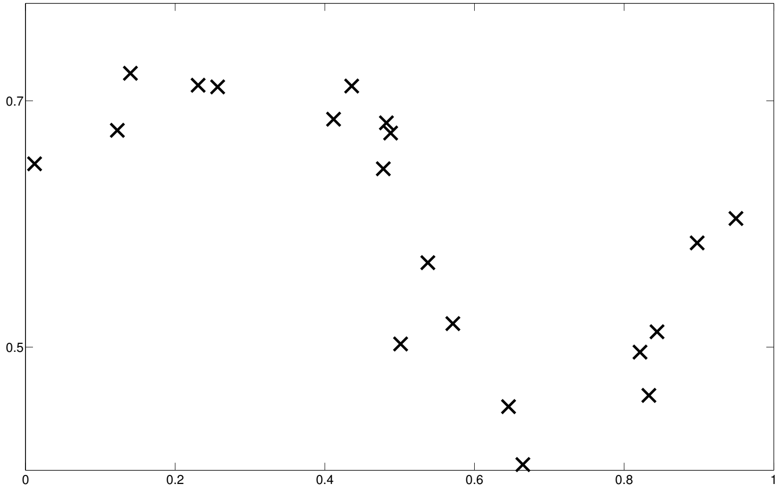
1. http://www.di.ens.fr/∼solnon/

matthieu.solnon@ens.fr

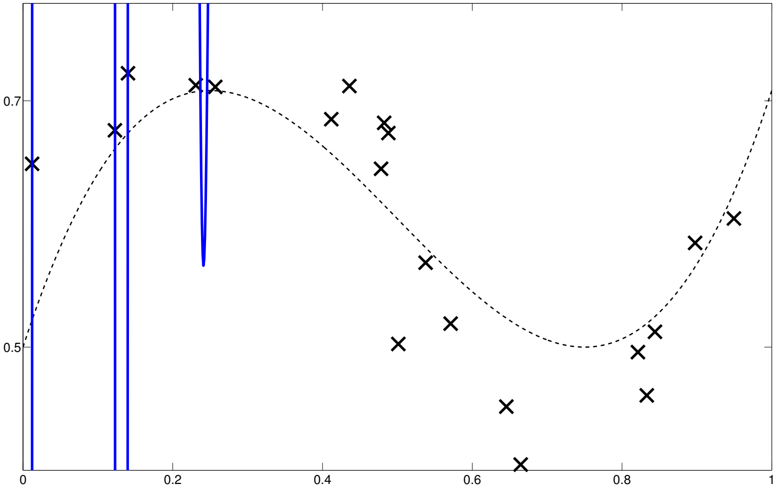
Lissage par des *splines* cubiques.



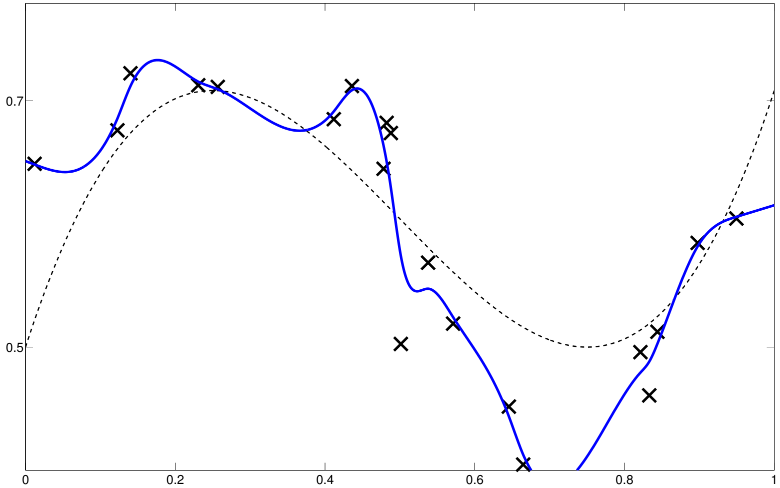
Lissage par des *splines* cubiques.



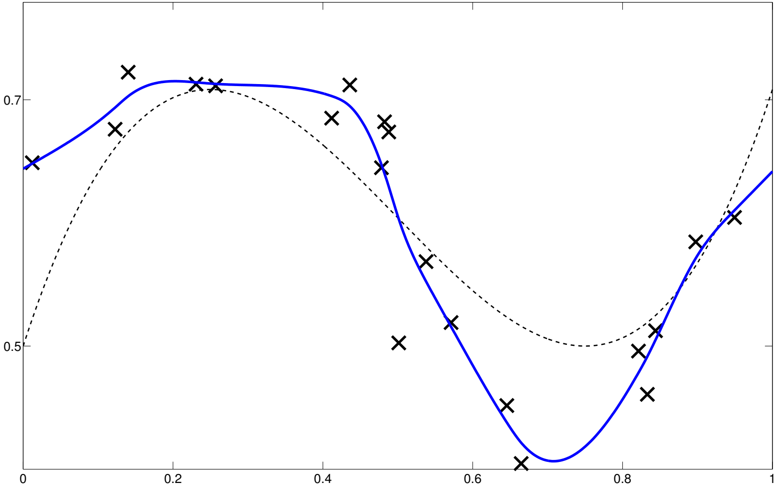
Lissage par des *splines* cubiques.



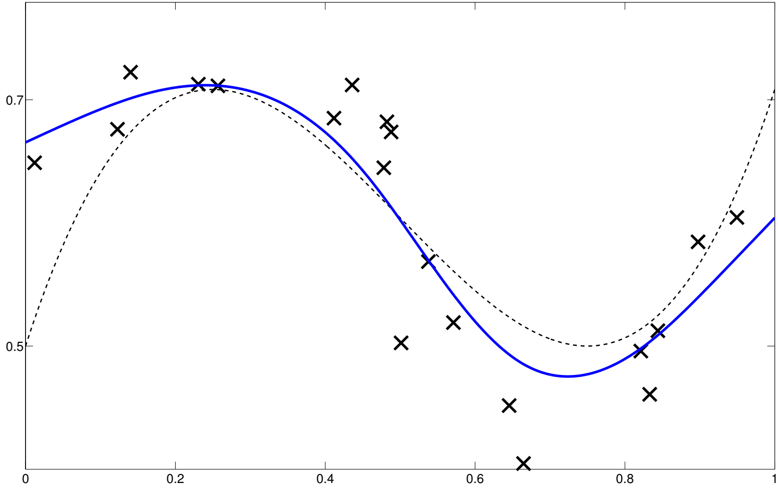
Lissage par des *splines* cubiques.



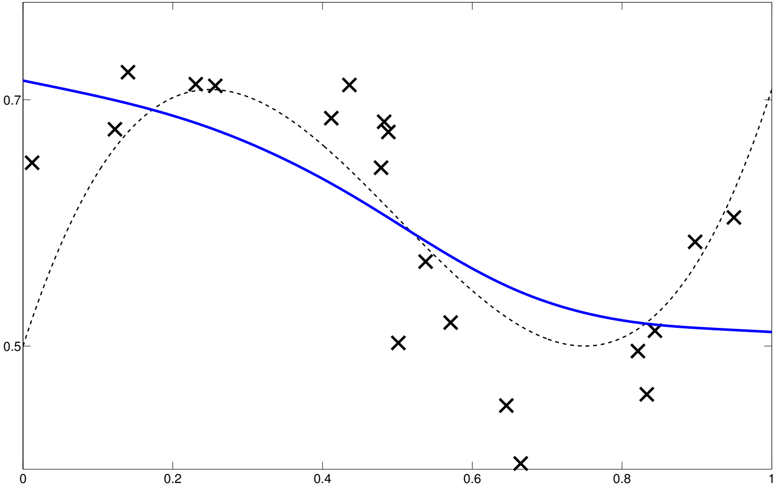
Lissage par des *splines* cubiques.



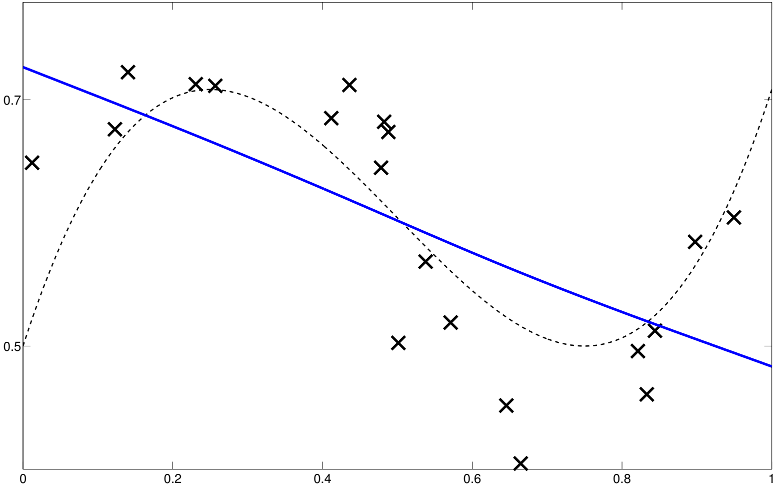
Lissage par des *splines* cubiques.



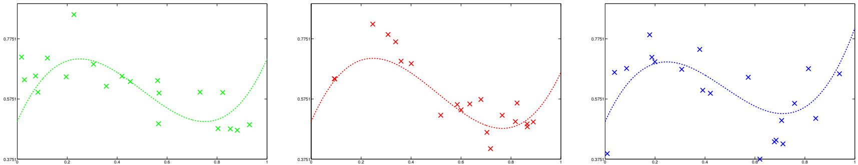
Lissage par des *splines* cubiques.



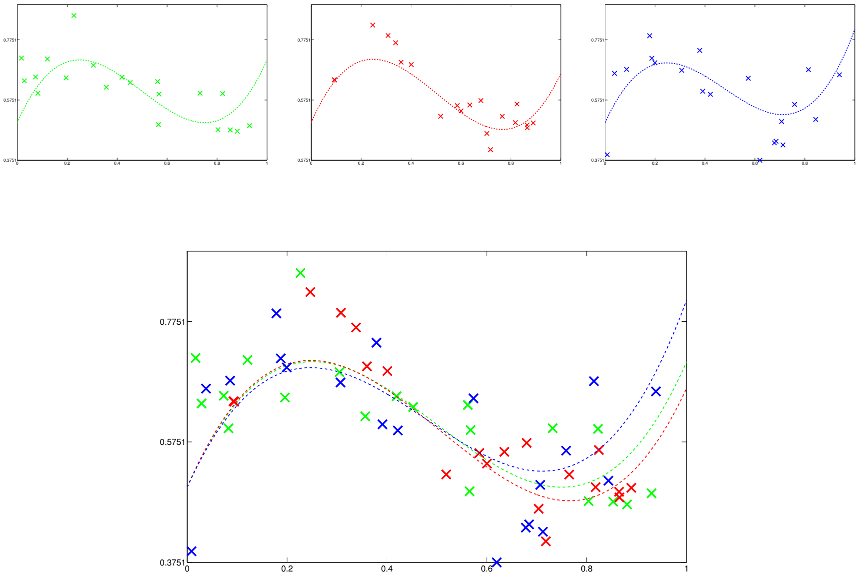
Lissage par des *splines* cubiques.



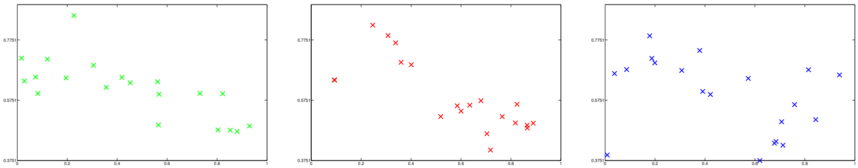
But du multi-tâches.

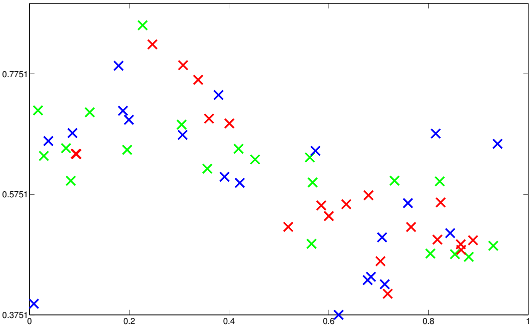


But du multi-tâches.



But du multi-tâches.





Construction d'un estimateur multi-tâches

Modèle.

Design : n points $(X_i)_{i=1}^n \in \mathcal{X}^n$.

Cible : p fonctions $(F^j)_{j=1}^p \in \mathcal{F}^p$.

Design fixe : On veut estimer $(F^j(X_i))_{i,j} = (f_i^j)_{i,j}$, comme si les $(X_i)_{i=1}^n$ étaient fixés.

Observations : $Y_i^j = F^j(X_i) + \varepsilon_i^j$, $\forall i\ (\varepsilon_i^j)_{j=1}^p \sim \mathcal{N}(0, \Sigma)$.

Risque quadratique : On voudrait minimiser, sur $(G^j)_{j=1}^p \in \mathcal{F}^p$,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{np}\sum_{i,j}(F^j(X_i) - G^j(X_i))^2\right] &= \mathbb{E}\left[\frac{1}{np}\sum_{i,j}(f_i^j - g_i^j)^2\right] \\
&= \mathbb{E}\left[\frac{1}{np}\left\|f - g\right\|^2\right]
\end{aligned}$$

Modèle.

Design : n points $(X_i)_{i=1}^n \in \mathcal{X}^n$.

Cible : p fonctions $(F^j)_{j=1}^p \in \mathcal{F}^p$.

Design fixe : On veut estimer $(F^j(X_i))_{i,j} = (f_i^j)_{i,j}$, comme si les $(X_i)_{i=1}^n$ étaient fixés.

Observations : $Y_i^j = F^j(X_i) + \varepsilon_i^j$, $\forall i \ (\varepsilon_i^j)_{j=1}^p \sim \mathcal{N}(0, \Sigma)$.

Risque quadratique : On voudrait minimiser, sur $(G^j)_{j=1}^p \in \mathcal{F}^p$,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{np}\sum_{i,j}(F^j(X_i) - G^j(X_i))^2\right] &= \mathbb{E}\left[\frac{1}{np}\sum_{i,j}(f_i^j - g_i^j)^2\right] \\ &= \mathbb{E}\left[\frac{1}{np}\left\|f - g\right\|^2\right]\end{aligned}$$

Notation des vecteurs.

- Pour chaque tâche $j \in \{1, \ldots, p\}$, on note

$$f^j = \begin{pmatrix} f^j(X_1) \\ \vdots \\ f^j(X_n) \end{pmatrix} = \begin{pmatrix} f_1^j \\ \vdots \\ f_n^j \end{pmatrix}, \text{ de taille } n \times 1 \ .$$

- On note

$$f = \begin{pmatrix} f^1 \\ \vdots \\ f^p \end{pmatrix}, \text{ de taille } np \times 1 \ .$$

- Idem pour y, $\widehat{f}_M$, ε, etc.

Procédure simple-tâche.

$\mathcal{F}$ est un RKHS, avec une matrice de noyau K.

L'estimateur *ridge* en régression.

- Pour chaque tâche $j \in \{1, \ldots, p\}$, minimiser

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i^j - G(X_i))^2}_{\text{Risque Empirique}} + \lambda^j \|G\|_{\mathcal{F}}^2 \ .$$

- Estimateur linéaire : $\forall j \in \{1, \ldots, p\}, \widehat{f}_\lambda^j = A_\lambda y^j$ avec $A_\lambda = K(K + n\lambda I_n)^{-1}$.
- La calibration du paramètre λ^j a été étudiée dans [AB11] [2], [Wah90] [3], [Gu02] [4].

---

2. S. Arlot et F. Bach, Data-driven calibration of linear estimators with minimal penalties, 2011.
3. G. Wahba, Spline Models for Observational Data, 1990.
4. C. Gu, Smoothing spline ANOVA models, 2002.

Généralisation de la pénalisation simple-tâche.

Équivalent à p problèmes simple-tâches indépendants.

Minimiser sur $(G^j)_{j=1}^p \in \mathcal{F}^p$

$$\frac{1}{np}\sum_{i,j}(Y_i^j - G^j(X_i))^2 + \qquad \sum_{j=1}^p \lambda^j \|G^j\|_{\mathcal{F}}^2$$

Généralisation de la pénalisation simple-tâche.

Équivalent à p problèmes simple-tâches indépendants.

$$M = \begin{pmatrix} \lambda^1 & & \\ & \ddots & \\ & & \lambda^p \end{pmatrix}$$

Minimiser sur $(G^j)_{j=1}^p \in \mathcal{F}^p$

$$\frac{1}{np}\sum_{i,j}(Y_i^j - G^j(X_i))^2 + \underbrace{\sum_{j=1}^p \lambda^j \|G^j\|_{\mathcal{F}}^2}_{\sum_{j,k} M_{j,k}\langle G^j, G^k\rangle := \|(G_1,\dots,G_p)\|_M^2}$$

Généralisation de la pénalisation simple-tâche.

Généralisation multi-tâches, les f^j ne sont plus estimées indépendamment.

$$M \in \mathcal{S}_p^{++}(\mathbb{R}) \text{ [EMP05]}^{5}$$

Minimiser sur $(G^j)_{j=1}^p \in \mathcal{F}^p$

$$\frac{1}{np} \sum_{i,j} (Y_i^j - G^j(X_i))^2 + \underbrace{\sum_{j,k} M_{j,k} \langle G^j, G^k \rangle}_{\|(G_1,\dots,G_p)\|_M^2}$$

On obtient un estimateur linéaire : $\widehat{f}_M = A_M y$.

---

5. T. Evgeniou, C. A. Micchelli et M. Pontil, Learning Multiple Tasks with Kernel Methods, 2005.

Quels types de similarité entre les tâches peut-on exprimer ?

Exemple de matrice M n° 1.

$$M_{\mathrm{ind}}(\lambda_1, \ldots, \lambda_p) = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

Critère obtenu :

$$\underbrace{\frac{1}{np} \sum_{i,j} (Y_i^j - G^j(X_i))^2}_{\text{Risque empirique}} + \sum_{j=1}^{p} \lambda_j \| G^j \|_{\mathcal{F}}^2$$

Traite indépendamment les p fonctions $\rightarrow$ simple-tâche.

Exemple de matrice M n° 2.

$$M_{\mathrm{SD}}(\lambda,\mu) = \begin{pmatrix} \lambda+(p-1)\mu & & -\mu \\ & \ddots & \\ -\mu & & \lambda+(p-1)\mu \end{pmatrix}$$

Critère obtenu :

$$\underbrace{\frac{1}{np}\sum_{i,j}(Y_i^j - G^j(X_i))^2}_{\text{Risque empirique}} + \lambda\sum_{j=1}^{p}\|G^j\|_{\mathcal{F}}^2 + \frac{\mu}{2}\sum_{j,k}\|G^j - G^k\|_{\mathcal{F}}^2 \ .$$

Exemple de matrice M n° 2.

Il vaut mieux utiliser

$$M_{\mathrm{AV}}(\lambda, \mu) = \frac{\lambda}{p} \underbrace{\frac{\mathbf{1}\mathbf{1}^\top}{p}}_{\text{Moyenne}} + \frac{\mu}{p} \underbrace{\left(I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p} \right)}_{\text{Variance}} .$$

Peut-on en partie s'adapter à la similarité entre les tâches ?

S'adapter au problème : sélectionner une matrice.

Problème restant : Sélectionner une matrice M dans un ensemble de matrices $\mathcal{M}$. On peut combiner les approches précédentes.

- Ensembles paramétrés (exemples n° 1 et 2).
- Ensembles discrets.
- Unions de différents ensembles de matrices.

Sélection de la matrice M dans $\mathcal{M}$.

$$\text{But : estimer } M^* \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \frac{1}{np} \mathbb{E}\left[\left\| \widehat{f}_M - f \right\|^2 \right] \right\}.$$

Sélection de la matrice M dans $\mathcal{M}$.

$$\text{But : estimer } M^* \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \frac{1}{np} \mathbb{E}\left[\left\| \widehat{f}_M - f \right\|^2 \right] \right\}.$$

$$\text{Minimiser : } \operatorname{crit}(M) = \frac{1}{np} \left\| y - \widehat{f}_M \right\|_2^2 + \operatorname{pen}(M).$$

Sélection de la matrice M dans $\mathcal{M}$.

But : estimer $M^* \in \underset{M \in \mathcal{M}}{\operatorname{argmin}} \left\{ \dfrac{1}{np} \mathbb{E}\left[\left\| \widehat{f}_M - f \right\|^2 \right] \right\}$.

Minimiser : $\operatorname{crit}(M) = \frac{1}{np} \left\| y - \widehat{f}_M \right\|_2^2 + \operatorname{pen}(M)$.

On prend : Pénalité idéale (en espérance)

$$\begin{aligned} \operatorname{pen}_{\mathrm{id}}(M) &:= \mathbb{E}\left[\frac{1}{np} \| \widehat{f}_M - f \|_2^2 \right] - \mathbb{E}\left[\frac{1}{np} \left\| y - \widehat{f}_M \right\|_2^2 \right] \\ &= \frac{2 \operatorname{tr}\left(A_M \cdot ({\color{red}\Sigma} \otimes I_n) \right)}{np} \end{aligned}$$

avec

$$A_M = (M^{-1} \otimes K) \left((M^{-1} \otimes K) + npI_{np} \right)^{-1} \quad .$$

Estimation de variance simple-tâche par les pénalités minimales.

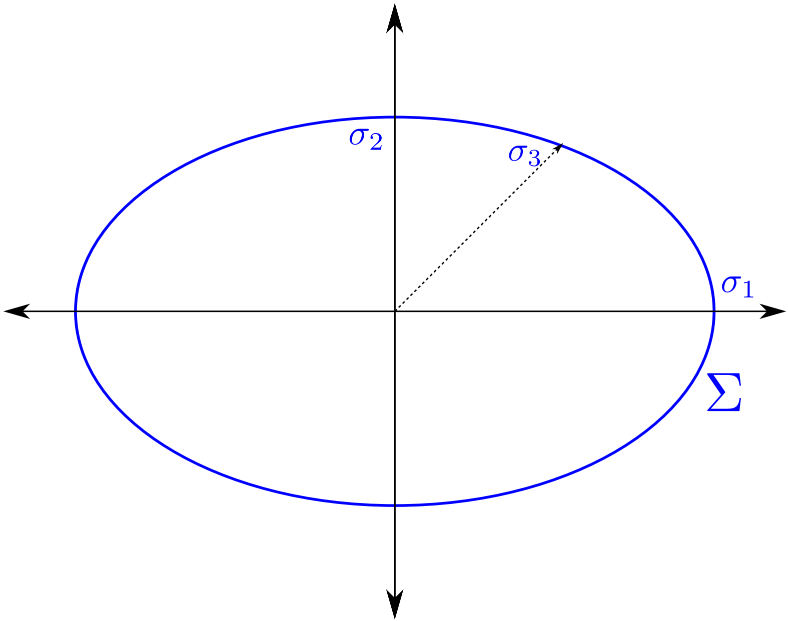
- Pout toute tâche $j \in \{1, \dots, p\}$, estimer la variance du bruit ε^j.
- On peut utiliser des pénalités minimales, [BM07][6] et [AM09][7].
- Pour tout $C > 0$, calculer

$$\widehat{\lambda}_0(C) \in \underset{\lambda \in \mathbb{R}_+}{\operatorname{argmin}} \left\{ \frac{1}{n} \left\| \widehat{f}_\lambda^j - y^j \right\|^2 + C \frac{2\,\mathrm{tr}(A_\lambda) - \mathrm{tr}(A_\lambda^\top A_\lambda)}{n} \right\}$$

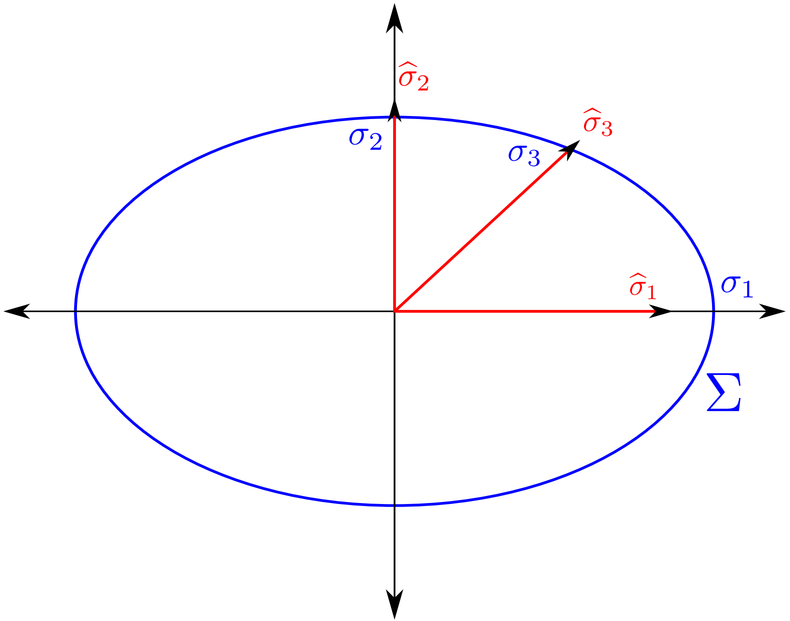
  - Si $C > \sigma^2$, $\mathrm{df}(\widehat{\lambda}_0(C)) = tr(A_{\widehat{\lambda}_0(C)})$ est petit.
  - Si $C < \sigma^2$, $\mathrm{df}(\widehat{\lambda}_0(C))$ est grand.
  - On estime σ^2 en repérant un grand saut dans la courbe de $\mathrm{df}(\widehat{\lambda}_0(C))$.

6. L. Birgé et P. Massart, Minimal Penalties for Gaussian Model Selection, 2007.
7. S. Arlot et P. Massart, Data-driven calibration of penalties for least-squares regression, 2009.

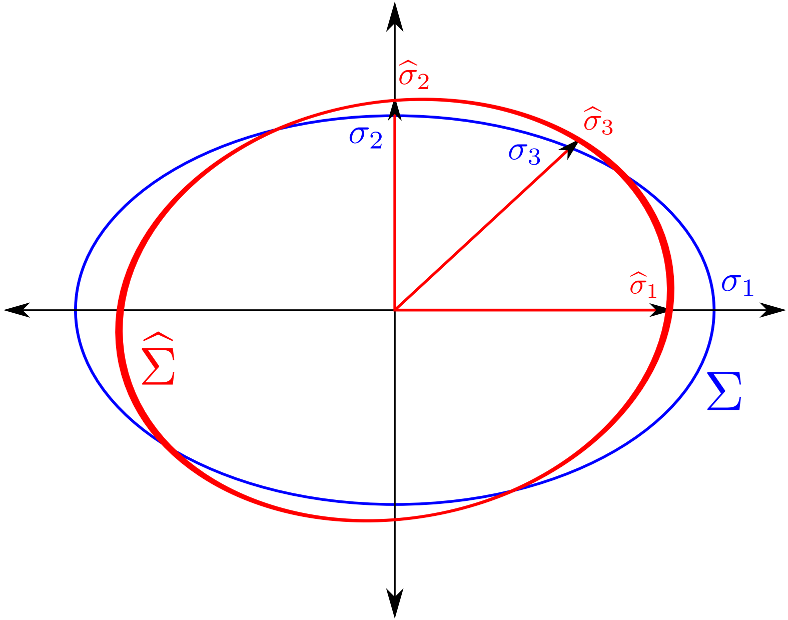
Estimation de Σ.



Estimation de Σ.



Estimation de Σ.



Hypothèse sur le biais.

$$\left.\begin{aligned} &\forall j \in \{1, \ldots, p\}\,,\ \exists \lambda_{0,j} \in (0, +\infty)\,, \\ &\qquad \mathrm{df}(\lambda_{0,j}) \leq \sqrt{n} \quad \text{et} \quad \frac{1}{n} \left\| (A_{\lambda_{0,j}} - I_n) F_{e_j} \right\|_2^2 \leq \Sigma_{j,j} \sqrt{\frac{\ln n}{n}} \end{aligned} \right\} \ \textbf{(Hdf)}$$

avec $\mathrm{df}(\lambda) = \mathrm{tr}(A_\lambda)$ et $A_\lambda = K(K + \lambda I_n)^{-1}$.

Estimation consistante de Σ.

$\widehat{\Sigma}$: matrice obtenue en sortie de l'algorithme précédent.

Théorème ([SAB12] [8])

Avec (**Hdf**) *et pour* n *assez grand, avec probabilité supérieure à* $1 - p(p+1)/2 \times n^{-\delta}$, *on a*

$$(1-\eta)\Sigma \preceq \widehat{\Sigma} \preceq (1+\eta)\Sigma \ ,$$

avec

$$\eta = L_1(2+\delta)p\sqrt{\frac{\ln n}{n}}c(\Sigma)^2 \ .$$

8. M. Solnon, S. Arlot et F. Bach, Multi-task Regression using Minimal Penalties, 2012.

Hypothèses sur la structure multi-tâches de $\mathcal{M}$.

$\mathcal{M}$ est discret.

$$\exists (C, \alpha_{\mathcal{M}}) \in]0, +\infty[^2, \ \operatorname{card}(\mathcal{M}) < C n^{\alpha_{\mathcal{M}}} \ . \qquad (\mathcal{M}\textbf{discret})$$

Les éléments de $\mathcal{M}$ sont co-diagonalisables dans une même base, comme dans les exemples n° 1 et 2.

$$\left. \begin{array}{l} \exists P \in O_p(\mathbb{R}) \ , \\ \mathcal{M} \subseteq \left\{ P^\top \operatorname{Diag}(d_1, \ldots, d_p) P \ , \ (d_j)_{j=1}^p \in]0, +\infty[^p \right\} \end{array} \right\} \ (\mathcal{M}\textbf{codiag})$$

Inégalité oracle, cas discret.

On sélectionne

$$\widehat{M} \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - y \right\|_2^2 + \frac{2}{np} \operatorname{tr}\left(A_M \cdot (\widehat{\Sigma} \otimes I_n) \right) \right\} \quad .$$

Théorème ([SAB12])

Avec (**Hdf**) *et* ($\mathcal{M}$**discret**) *et pour n assez grand, avec probabilité supérieure à* $1 - \kappa' p(p + C) n^{-\delta}$, *on a*

$$\frac{1}{np} \left\| \widehat{f}_{\widehat{M}} - f \right\|_2^2 \leq \left(1 + \frac{1}{\ln n} \right)^2 \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} + \theta_n \quad ,$$

avec

$$\theta_n = L_2 c(\Sigma)^4 \sigma_{\max} (\alpha + \delta)^2 \frac{p^4 \ln(n)^3}{np} \quad .$$

Inégalité oracle, cas continu et co-diagonalisable.

On a un algorithme simplifié qui donne une matrice $\widehat{\Sigma}_{\mathrm{HM}}$, on sélectionne

$$\widehat{M} \in \operatorname*{argmin}_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - y \right\|_2^2 + \frac{2}{np} \operatorname{tr}\left(A_M \cdot (\widehat{\Sigma}_{\mathrm{HM}} \otimes I_n) \right) \right\} \quad .$$

Théorème ([SAB12])

Avec (**Hdf**) *et* ($\mathcal{M}$**codiag**) *et pour* n *assez grand, avec probabilité supérieure à* $1 - \kappa'' p n^{-\delta}$*, on a*

$$\frac{1}{np} \left\| \widehat{f}_{\widehat{M}} - f \right\|_2^2 \leq \left(1 + \frac{1}{\ln n} \right)^2 \inf_{M \in \mathcal{M}} \left\{ \frac{1}{np} \left\| \widehat{f}_M - f \right\|_2^2 \right\} + \theta_n \quad ,$$

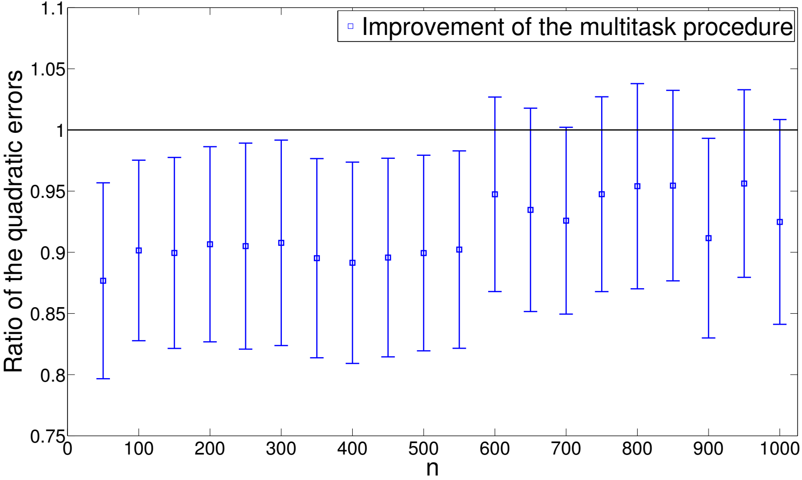
avec

$$\theta_n = L_2 \sigma_{\max} (2 + \delta)^2 \frac{p \ln(n)^3}{np} \quad .$$

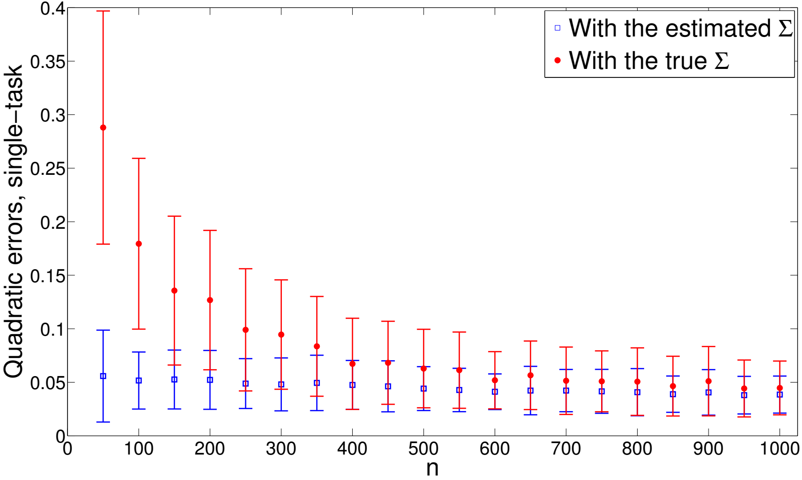
Comparaison entre les deux cas.

	Hypothèse ($\mathcal{M}$**discret**)	Hypothèse ($\mathcal{M}$**codiag**)
Coût de $\widehat{\Sigma}$	$p(p+1)/2$	p
Probabilité	$1-\kappa' p(p+C)n^{-\delta}$	$1-\kappa'' p n^{-\delta}$
Erreur θ_n	$L_2 c(\Sigma)^4 \sigma_{\max}(\alpha+\delta)^2 \frac{p^4 \ln(n)^3}{np}$	$L_2 \sigma_{\max}(2+\delta)^2 \frac{p \ln(n)^3}{np}$

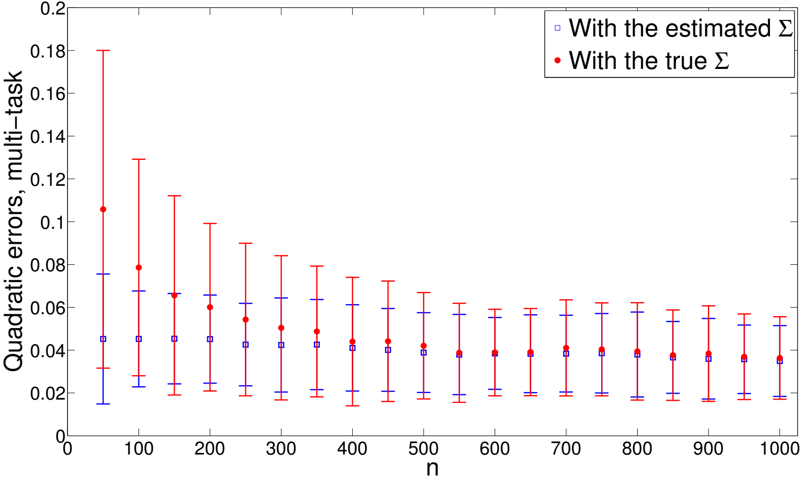
Simulations : comparaison multi-tâches - simple-tâche.



Simulations : risque simple-tâche.



Simulations : risque multi-tâches.



Le multi-tâches fonctionne-t-il ?

Le multi-tâches fonctionne-t-il ?

Petites hypothèses

- Pour simplifier : $\Sigma = \sigma^2 I_p$.
- Paramètres de régularisation :
  $\mathcal{M}_{\mathrm{AV}} = \left\{ M_{\mathrm{AV}}(\lambda, \mu),\ (\lambda, \mu) \in]0, +\infty[^2 \right\}$.

Décomposition biais-variance : le risque dépend

- des valeurs propres de la matrice de noyau K : $(\gamma_i)_{i=1}^n$;
- des coefficients de la moyenne des signaux f^j sur la base qui a servi à diagonaliser K : $\left(\dfrac{\mu_i}{p}\right)_{i=1}^n$;
- des coefficients de la « variance » des signaux f^j sur la base qui a servi à diagonaliser K : $(\varsigma_i^2)_{i=1}^n$.

Hypothèses.

Régularité du noyau :

$$\forall i \in \{1, \dots, n\},\ \gamma_i = n i^{-2\beta}\ . \qquad \textbf{(RegNoyau)}$$

Régularité du signal :

$$\forall i \in \{1, \dots, n\},\ \begin{cases} \frac{\mu_i^2}{p} & = & C_1 n i^{-2\delta} \\ \varsigma_i^2 & = & C_2 n i^{-2\delta} \end{cases} . \qquad \textbf{(RegSignal)}$$

Notations

- Risque oracle multi-tâches : $\mathfrak{R}^\star_{\mathrm{MT}}$.
- Risque oracle simple-tâche : $\mathfrak{R}^\star_{\mathrm{ST}}$.

Contrôle du risque oracle multi-tâches.

Théorème ([Sol13][9])

Soit n, p, C_1, C_2, σ^2, β et δ tels que (**RegNoyau**) *et* (**RegSignal**) *sont vérifiés et tels que $1 < 2\delta < 4\beta + 1$, on a*

$$\mathfrak{R}^\star_{\mathrm{MT}} \leq 2^{1/(2\delta)} \left(\frac{np}{\sigma^2}\right)^{1/(2\delta)-1} \kappa(\beta, \delta) \left[C_1^{1/(2\delta)} + (p-1)^{1-(1/2\delta)} C_2^{1/2\delta} \right] \ .$$

De plus, il existe des constantes N et $\alpha \in (0,1)$ telles, si $n \geq N$, $p/\sigma^2 \leq n$ et si $2 < 2\delta < 4\beta$, on a

$$\mathfrak{R}^\star_{\mathrm{MT}} \geq \alpha \left(\frac{np}{\sigma^2}\right)^{1/(2\delta)-1} \kappa(\beta, \delta) \left[C_1^{1/2\delta} + (p-1)^{1-(1/2\delta)} C_2^{1/2\delta} \right] \ .$$

9. M. Solnon, Comparison bewteen multi-task and single-task oracle risks in kernel ridge regression, 2013.

Hypothèses pour calculer le risque oracle simple-tâche.

2 points : p est pair et

$$f^1 = \cdots = f^{p/2} \quad \text{et} \quad f^{p/2+1} = \cdots = f^p \ . \qquad \textbf{(2Points)}$$

1 *outlier* :

$$f^1 = \cdots = f^{p-1} \ . \qquad \textbf{(1Out)}$$

Notation

- $r = C_2 / C_1$

Comparaison multi-tâches - simple-tâche.

Corollaire ([Sol13])

Soit n, p, C_1, C_2, σ^2, β *et* δ *tels que* $2 < 2\delta < 4\beta$ *et* $n\sigma^2 > p$, *si* (**RegSignal**) *et* (**RegNoyau**) *sont vérifiés, alors*

avec (**2Points**) :

$$\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}} \asymp \frac{p^{1/(2\delta)-1} + (\frac{p-1}{p})^{1-(1/2\delta)} r^{1/2\delta}}{\left(1+\sqrt{r}\right)^{1/\delta} + \left|1-\sqrt{r}\right|^{1/\delta}} \quad ;$$

avec (**1Out**) :

$$\frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}} \asymp \frac{p^{1/(2\delta)-1} + \left(\frac{p-1}{p}\right)^{1-(1/2\delta)} r^{1/2\delta}}{\frac{p-1}{p}\left(1+\sqrt{\frac{r}{p-1}}\right)^{1/\delta} + \frac{1}{p}\left|1-\sqrt{r(p-1)}\right|^{1/\delta}} \quad .$$

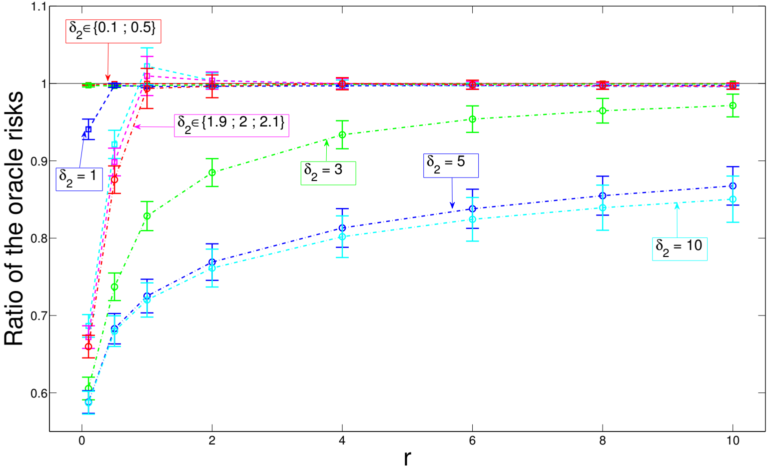
Situations positives et négatives pour l'oracle multi-tâches.

Soit

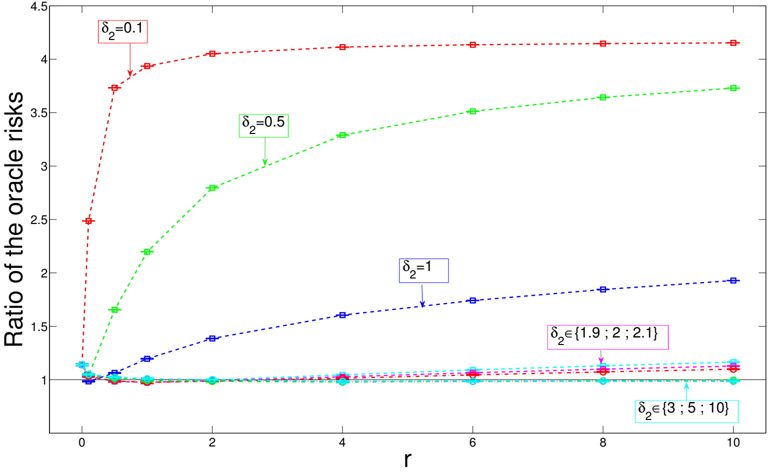
$$\rho = \frac{\mathfrak{R}^\star_{\mathrm{MT}}}{\mathfrak{R}^\star_{\mathrm{ST}}} \quad .$$

- Si r tend vers 0, alors ρ tend vers $Cst \times p^{1/2\delta-1}$.
- Quand r est grand :
  - Avec (**2Points**), ρ est borné ;
  - Avec (**1Out**), ρ tend vers $+\infty$ quand r tend vers ∞.

Simulations : un groupe de tâches.



Simulations : un groupe de tâches et un *outlier*.



Conclusion

Références bibliographiques. I

- Sylvain Arlot et Francis Bach: *Data-driven calibration of linear estimators with minimal penalties*, 2011.
  arXiv :0909.1884v2.
- Sylvain Arlot et Pascal Massart: *Data-driven calibration of penalties for least-squares regression*.
  Journal of Machine Learning Research, 10 :245–279 (electronic), 2009.
- Lucien Birgé et Pascal Massart: *Minimal Penalties for Gaussian Model Selection*.
  Probability Theory and Related Fields, 138 :33–73, 2007.
- Theodoros Evgeniou, Charles A. Micchelli et Massimiliano Pontil: *Learning Multiple Tasks with Kernel Methods*.
  Journal of Machine Learning Research, 6 :615–637, 2005.
- Chong Gu: *Smoothing spline ANOVA models*.
  Springer, 2002.

Références bibliographiques. II

Matthieu Solnon, Sylvain Arlot et Francis Bach: *Multi-task Regression using Minimal Penalties*.
Journal of Machine Learning Research, 13 :2773–2812, septembre 2012.

Matthieu Solnon: *Comparison between multi-task and single-task oracle risks in kernel ridge regression*.
2013.

Grace Wahba: *Spline Models for Observational Data*, tome 59 de *CBMS-NSF Regional Conference Series in Applied Mathematics*.
Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

The end.

Merci !

Et maintenant, avez-vous des questions ?

À suivre : pot en salle club !