First TASSILI workshop on Shrinkage Estimation and Data Science

Organized by LAMOPS at ENSSEA, Tipaza (Algeria) and LITIS at INSA Rouen Normandie, Madrillet, Salle A-RC-02 at the Bougainville building, December the 21th[1]

Cadre général du workshop

Le but du workshop est de faire le point sur les avancée récentes dans le domaine de l'amélioration du SURE matriciel en réunissant des spécialistes du domaine. En effet, les modèles additifs de la forme

$$\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{\mathcal{E}}, \tag{1}$$

où $\boldsymbol{Y}$ est une matrice $n \times m$ observée, $\boldsymbol{M}$ est une matrice inconnue que l'on souhaite estimer et $\boldsymbol{\mathcal{E}}$ est un bruit aléatoire, ont connu un accroissement d'intérêt ces récentes années. Typiquement, dans la littérature, il est supposé que $m < n$ et que la loi de $\boldsymbol{\mathcal{E}}$ est gaussienne, soit $\boldsymbol{\mathcal{E}} \sim \mathcal{N}(0, I_n \otimes \Sigma)$, où Σ est une matrice de covariance $m \times m$ définie positive et I_n est la matrice identité d'ordre n. Le plus souvent, $\Sigma = \sigma^2 I_m$ ce qui correspond au cas i.i.d. et qui constitue un exemple typique de la problématique abordée que nous développons ci -après. Notons que le cadre du modèle (1) est précisé pour être adapté aux situations souvent rencontrées en pratique où les colonnes de M peuvent être très liées entre elles : il est donc supposé la matrice $\boldsymbol{M}$ est de rang petit ou que celle-ci peut être approchée par une matrice de rang $p < m$.

Dans ce contexte, il s'agit d'envisager des estimateurs alternatifs à l'estimateur naïf qu'est l'observation $\boldsymbol{Y}$. Une approche naturelle consiste à tronquer la décomposition en valeurs singulières de la matrice $\boldsymbol{Y}$ (par hard thresholding ou soft thresholding), ce qui revient à effectuer une opération de "shrinkage" initiée par Stein. Dans cette optique, de nombreux estimateurs $\widehat{\boldsymbol{M}} = \widehat{\boldsymbol{M}}(\boldsymbol{Y})$ ont été proposés dans la littérature. En présence de tant d'alternatives, il est nécessaire de les évaluer. À cette fin, en théorie de la décision statistique, cette évaluation est faite au travers d'une fonction de coût (souvent choisie quadratique) et de sa fonction de risque associée (qui est l'espérance de ce coût). Comme, pour tout estimateur $\widehat{\boldsymbol{M}}$, contrôler le niveau de "shrinkage" importe (trop de "shrinkage" donne de grand biais et pas assez aboutit à une grande variance), le SURE (Stein's Unbiased Risk Estimate) est alors un instrument privilégié (en moyenne, il donne le risque de $\widehat{\boldsymbol{M}}$).

La théorie rappelée ci-dessus est gaussienne, et cette hypothèse distributionnelle est loin d'être toujours satisfaite. Récemment, le modèle

$$\boldsymbol{Y} = \boldsymbol{M} + \boldsymbol{\mathcal{E}}, \quad \boldsymbol{\mathcal{E}} \sim \mathcal{E}(\mathbf{0}_{nm}, I_n \otimes \Sigma), \tag{2}$$

[1] `http://www.insa-rouen.fr/institution/Identite/acces/plan-acces-campus-insa-rouen`

a été considéré, où $\boldsymbol{\mathcal{E}}$ est un bruit elliptiquement distribué de matrice de covariance proportionnelle à $I_n \otimes \Sigma$, où Σ est une matrice de dispersion (ou d'échelle) $m \times m$ définie positive et I_n est la matrice identité d'ordre n. Le modèle (2) est une extension du modèle gaussien où $\mathcal{E} = \mathcal{N}(0, I_n \otimes \Sigma)$. Ces auteurs ont développé des estimateurs de type SURE relatifs à un coût quadratique invariant qui correspondent à différentes situations suivant que Σ est connu ou inconu, que la loi du bruit est normale ou non. Ils ont mis en évidence le fait que ces estimateurs SURE sont efficaces comme sélecteurs de modèles et robustes par rapport à la classe des lois à symétrie elliptique considérée.

Objectifs scientifiques

La plupart des résultats d'estimation d'une matrice de paramètres ont été développés dans le contexte de la loi normale multivariée. Dans la mesure où l'hypothèse de normalité est souvent mise en défaut, un plus large contexte distributionnel est maintenant considéré. Bien que de récents résultats aient été obtenu dans le cadre les lois à symétrie elliptique qui contient le cas gaussien, dans ce contexte distributionnel, de multiples questions se posent encore. En particulier, il est nécessaire d'obtenir des résultats robustes quant aux propriétés des estimateurs au sens où ceux-ci n'exigeraient pas de spécifier la loi elliptique du bruit mais seraient valides pour une grande classe de telles lois.

Problématique

Problématique fondamentale

Les méthodes modernes d'estimation (Lasso, SCAD, MCP...) dépendant d'un paramètre de réglage conduisent à des familles d'estimateurs. Un des problèmes qui nous intéresse dans ce workshop est de déterminer un critère précis et robuste permettant de sélectionner au sein de la famille considérée le meilleur estimateur pour l'application à traiter. Dans un contexte gaussien i.i.d., le SURE fournit un tel critère. Le but de ce workshop est de faire le point sur les techniques permettant d'améliorer celui-ci en termes de performance et de robustesse. En effet on sait que le SURE est inadmissible, et donc qu'il existe des estimateurs de risque le dominant, et ce dans le cas dépendant et non gaussien. Notre projet vise à déterminer de tels estimateurs améliorés.

Domaines d'application

Le workshop est aussi l'occasion de faire le point sur les applications de ces méthodes dans les domaines de :

1. la factorisation de faible rang pour le traitement d'image dans le cadre de la reconstruction à partir de données incomplètes et à partir de données bruitées ou altérées
2. la complétion de matrices
    - pour la recommandation, par exemple, dans le cas de la compétition Netflix
    - pour le traitement d'image (connue sous la dénomination *in painting*)
3. l'apprentissage de dictionnaire
    - pour le traitement d'image
    - pour l'apprentissage multi tâche

First TASSILI workshop on Shrinkage Estimation and Data Science

Organized by LAMOPS at ENSSEA, Tipaza (Algeria) and LITIS at INSA Rouen Normandie, Madrillet, Salle A-RC-02 at the Bougainville building, December the 21th[2]

Provisional program

9:30 to 10:00 Welcome coffee break

10:00 to 10:15 Stéphane Canu: *Openning address*

10:15 to 11:15 Fatiha Mezoued: *Improved estimation of discriminant coefficients*

11:05 to 11:30 *Coffee break*

11:30 to 12:30 Dominique Fourdrinier: *Shrinkage estimation of a location parameter for a multivariate skew elliptic distribution*

12:30 to 2:00 Lunch break

2:00 to 2:30 Mohamed Anis Haddouche: *Improved estimation of a covariance matrix*

3:00 to 3:30 Ismaila Seck: *Deep learning and adversarial examples*

2:30 to 3:00 Djamila Boukehil: *Improved estimation of a precision matrix*

3:30 to 4:00 Nihad Nouri: *SURE and image processing*

4:00 to 4:30 Coffe break and open discussion about the continuation of the collaboration

Contact: Fatiha Mezoued <famezoued@yahoo.fr>,
Dominique Fourdriner <dominique.fourdrinier@univ-rouen.fr>
Stéphane Canu <stephane.canu@insa-rouen.fr>

[2] `http://www.insa-rouen.fr/institution/Identite/acces/plan-acces-campus-insa-rouen`