GETS modelling ou LASSO ?
Les différentes méthodes de sélection de variables avec des séries temporelles

Alizé Papp (*), Clément Rousset (**)

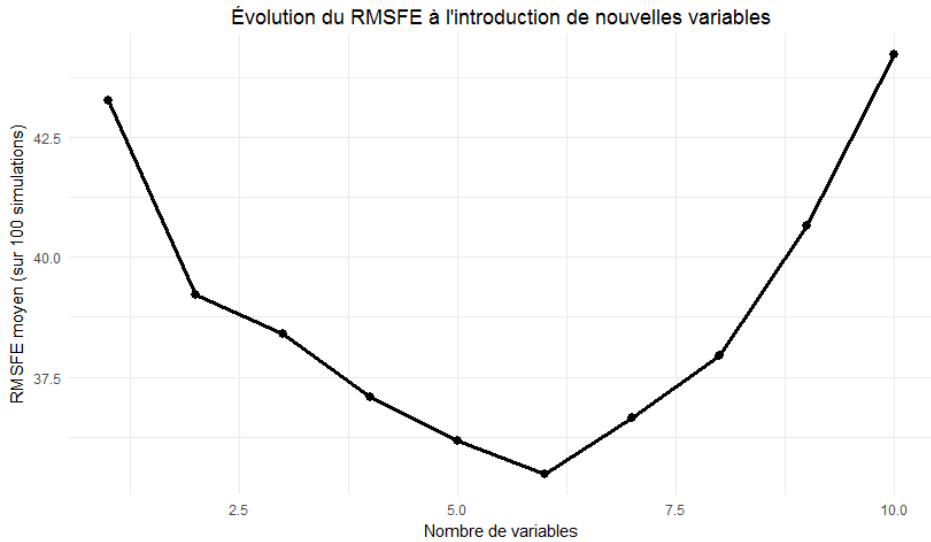
(*) ENSAE ParisTech
(**) INSEE, Division de la Synthèse conjoncturelle

13 juin 2018

Pourquoi sélectionner des variables ?

- Ajouter de nouvelles variables ✓
- Mais pourquoi ne pas toutes les ajouter ?

⚠**Risque d'overfitting**



Comment sélectionner des variables ?

De nombreuses séries pourraient améliorer la prévision mais...

- **Quelles variables tester ?**
- **Comment comparer** deux étalonnages ?
- Comment comparer en un **temps raisonnable** ?

1. Objectif : trouver le meilleur étalonnage
   - Cadre théorique du problème
   - Qu'est-ce qu'un bon étalonnage ?
   - La méthode intuitive de sélection et ses inconvénients
2. Les méthodes de sélection de variables
   - Le General-to-specific modelling
   - Le LASSO
3. GETS modelling ou Lasso : qu'en disent les simulations ?
   - Description de la méthode pour les simulations
   - TPR, FNR et RMSFE : critères d'évaluation
   - Résultats et premières conclusions
4. Utilisation pratique
   - Qu'est-ce qu'un étalonnage interprétable ?
   - Algorithme de recherche d'étalonnages interprétables
   - Exemple d'application

Cadre théorique

Y, l'agrégat à prévoir
$X_1, X_2, ..., X_p$: **p variables candidates**
Parmi elles, q réellement pertinentes :

$$Y_t = \sum_{i=1}^{q} \alpha_i X_{i,t} + \epsilon_t$$

Qu'est-ce qu'un bon étalonnage ?

Un étalonnage performant

Une première définition : un étalonnage avec une faible erreur de prévision (RMSFE).

$$RMSFE = \sqrt{E[(Y_{t+1} - \hat{Y}_{t+1|t})^2]}$$

où Y_{t+1} est la valeur de Y en t+1 et $\hat{Y}_{t+1|t}$ est la valeur de Y en t+1 prédite avec l'information disponible en t.
L'erreur de prévision à laquelle on peut s'attendre pour une prévision.
⚠**Computationnellement coûteux**

Comment sélectionner le meilleur étalonnage ?

Méthode intuitive : tester toutes les combinaisons possibles.
Problèmes :

1. Coût exponentiel (2^p).
2. Risque de suraprentissage

1 Objectif : trouver le meilleur étalonnage
- Cadre théorique du problème
- Qu'est-ce qu'un bon étalonnage ?
- La méthode intuitive de sélection et ses inconvénients

2 Les méthodes de sélection de variables
- Le General-to-specific modelling
- Le LASSO

3 GETS modelling ou Lasso : qu'en disent les simulations ?
- Description de la méthode pour les simulations
- TPR, FNR et RMSFE : critères d'évaluation
- Résultats et premières conclusions

4 Utilisation pratique
- Qu'est-ce qu'un étalonnage interprétable ?
- Algorithme de recherche d'étalonnages interprétables
- Exemple d'application

GETS modelling

Intuition fondamentale

- Contexte :
  - Les ancêtres : le *forward* et *backward selection*
  - Contexte : l'école d'économétrie de la LSE.
  - Hoover et Perez (1999) puis Sucarrat et Genaro (2009)
- **Objectif : trouver un chemin plus rapide jusqu'au meilleur modèle.**
  Contraintes : les combinaisons linéaires prédictives et le risque de dépendance au sentier
- Idée :
  - ***General***... : partir des p variables => tient compte des combinaisons linéaires,
  - ***to specific*** : trouver les q justes (en retirant petit à petit les variables non significatives tant que le choix de variables passe des tests d'hétéroscédasticité des résidus).

GETS modelling

Algorithme

1. Estimation du modèle à p variables, u **variables ne sont pas significatives** $(u \leqslant p)$
2. **Chaque u définit un chemin.**
   1. Estimation du modèle sans la u-ième variable la moins significative
   2. Tant que le modèle passe les tests, on retire la variable la moins significative.
3. Chaque chemin fournit au maximum un modèle, on **choisit celui avec le meilleur BIC**.

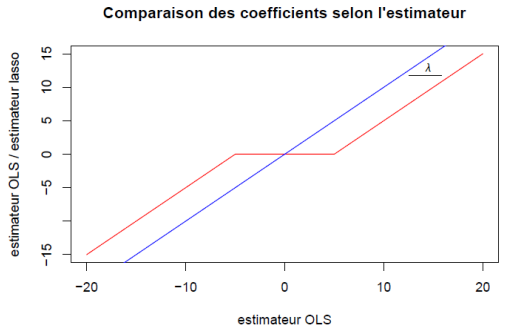
Coût : $o(p^2)$

$$Y_t = \alpha_1 X_{1,t} + \alpha_2 X_{2,t} + \cdots + \alpha_p X_{p,t} + \epsilon_t$$

LASSO

Least absolute shrinkage and selection operator, Robert Tibshirani, 1996

Une méthode de régularisation.
Intuition : à la régression linéaire standard, on rajoute un **terme de pénalité** sur la taille des coefficients => restreint les coefficients vers 0.



Coefficient par OLS vs par Lasso

$$\arg \min_{\beta_0,\beta_1,\ldots,\beta_p} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Coût constant.

1 Objectif : trouver le meilleur étalonnage
- Cadre théorique du problème
- Qu'est-ce qu'un bon étalonnage ?
- La méthode intuitive de sélection et ses inconvénients

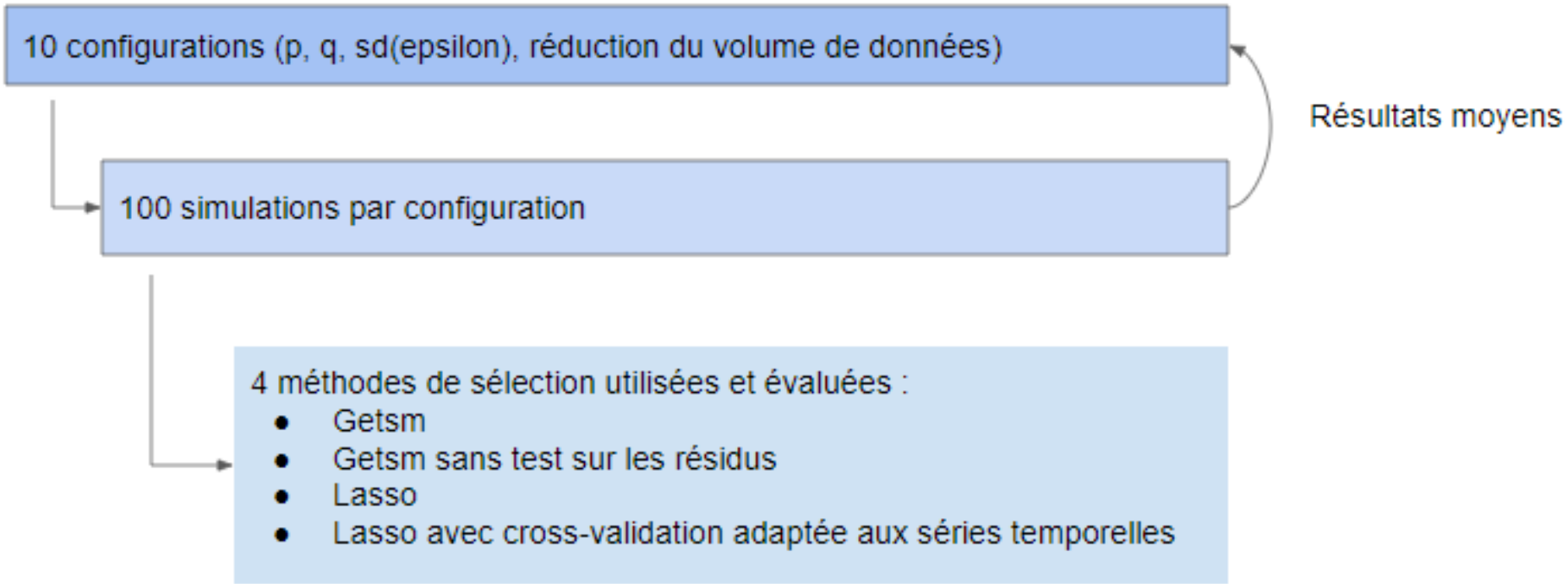
2 Les méthodes de sélection de variables
- Le General-to-specific modelling
- Le LASSO

3 GETS modelling ou Lasso : qu'en disent les simulations ?
- Description de la méthode pour les simulations
- TPR, FNR et RMSFE : critères d'évaluation
- Résultats et premières conclusions

4 Utilisation pratique
- Qu'est-ce qu'un étalonnage interprétable ?
- Algorithme de recherche d'étalonnages interprétables
- Exemple d'application

Méthode de simulation

Idée : générer des Y en tirant au hasard des X, des coefficients et des bruits.



Les méthodes de sélection comparées

- **Getsm avec variantes sur les tests Portmanteau :**
  - Getsm *avec* test Portmanteau de Ljung-Box sur les résidus.
  - Getms *sans* test Portmanteau (que des tests de significativité)
- **Lasso avec variantes sur la cross-validation :**
  - Cross-validation standard,
  - Cross-validation adaptée aux séries temporelles (par fenêtres glissantes)
- **Et deux OLS de repère :**
  - Sur le modèle avec les p variables candidates,
  - Sur le modèle avec les q vraies variables.

TPR, FNR et RMSFE

Critères d'évaluation

Soient P l'ensemble des variables candidates, S celui des variables sélectionnées et Q celui des variables pertinentes.

$$TPR = \frac{card(Q \cap S)}{card(S)}$$ proportion de sélectionnées à raison

$$FNR = \frac{card(Q \cap \bar{S})}{card(Q)}$$ proportion de rejetées à tort parmi les pertinentes

$$RMSFE = \sqrt{E[(Y_{t+1} - \hat{Y}_{t+1|t})^2]}$$

où Y_{t+1} est la valeur de Y en t+1 et $\hat{Y}_{t+1|t}$ est la valeur de Y en t+1 prédite avec l'information disponible en t.

Le RMSFE et ses cousins

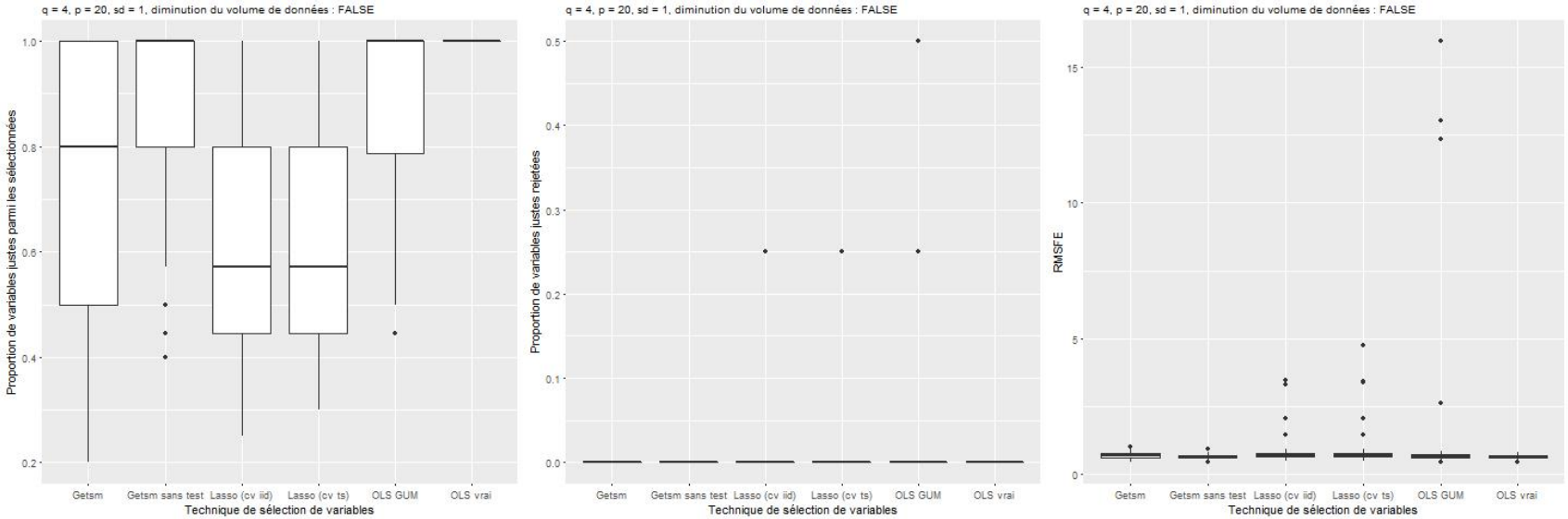
Calcul des erreurs de prévision.
Plusieurs façons de les agréger ensuite :

1. **RMSFE,** la racine carrée de la somme des erreurs de prévision au carré
2. **max FE,** l'erreur de prévision maximale
3. **WRMSFE,** le RMSFE pondéré pour pénaliser davantage les erreurs sur les derniers points.

Un large succès du GETS modelling

Cas où p = 20, q = 4, sd = 1

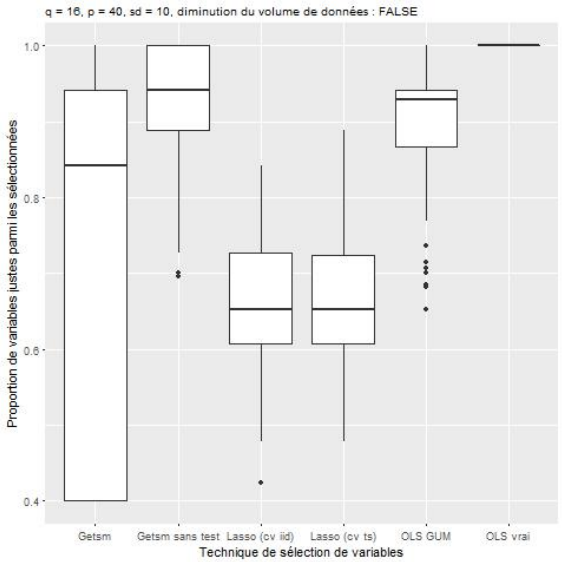
Boîte de Tuckey des résultats des simulations pour chaque configuration (100 simulations par configuration)

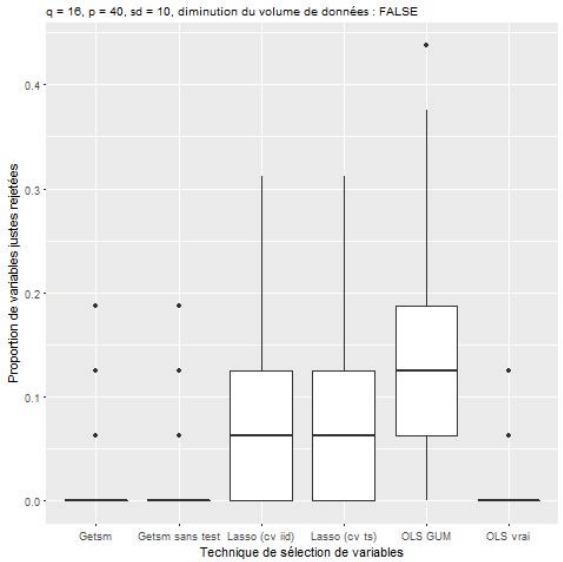


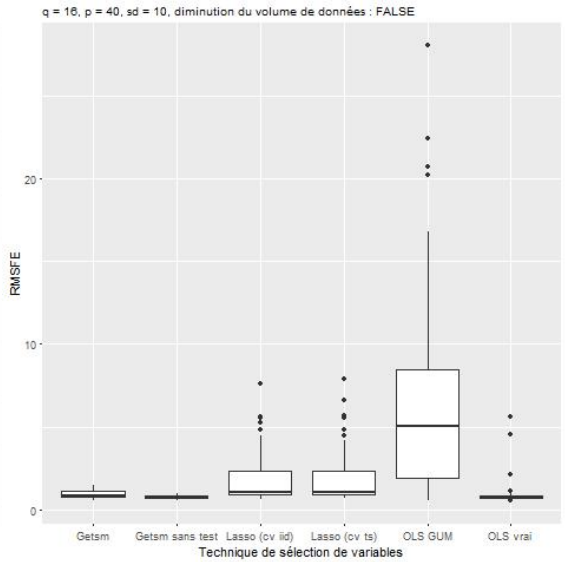
- TPR (gauche) : seul critère distinctif, Getsm sans test meilleur,
- FNR (milieu) : très bonnes performances pour tous,
- RMSFE (droite) : très bonnes qualités prédictives pour tous.

Un large succès du GETS modelling

Cas où p = 40, q = 16, sd = 10







- TPR (gauche) : getsm meilleur pour la moitié des cas que le lasso, mais bien moins stable,
- FNR (milieu) : le lasso a davantage tendance à rejeter des variables à tort (plus économe),
- RMSFE (droite) : qualités comparables pour getsm, getsm sans test et dans une moindre mesure le lasso.

Synthèse et limites

Avec quelques limites :

- Getsm évalué dans son cadre le plus propice,
- Lasso computationnellement plus économe,
- Lasso est la seule possibilité lorsqu'il y a plus de variables que d'observations.

Comparaison pratique du Lasso et du Gets modelling

	Getsm	Lasso
Simulations	✓	×
Temps de calcul	×	✓
Vrai modèle pas nécessairement inclus	×	✓
Variables corrélées	✓	×

Dans la pratique, les trois méthodes sont complémentaires et invitent à un choix humain éclairé.

1 Objectif : trouver le meilleur étalonnage
- Cadre théorique du problème
- Qu'est-ce qu'un bon étalonnage ?
- La méthode intuitive de sélection et ses inconvénients

2 Les méthodes de sélection de variables
- Le General-to-specific modelling
- Le LASSO

3 GETS modelling ou Lasso : qu'en disent les simulations ?
- Description de la méthode pour les simulations
- TPR, FNR et RMSFE : critères d'évaluation
- Résultats et premières conclusions

4 Utilisation pratique
- Qu'est-ce qu'un étalonnage interprétable ?
- Algorithme de recherche d'étalonnages interprétables
- Exemple d'application

Notion d'étalonnage interprétable

Un étalonnage est interprétable pour un conjoncturiste s'il peut résumer son message en une phrase simple :

$$\Delta log(PIB) = \alpha + \beta Confiance_{\text{dernier mois}},$$

avec $\beta > 0 \rightarrow$ "le haut niveau de confiance des entreprises donnerait une croissance en hausse"
Pour

$$\Delta log(PIB) = \alpha + \beta Confiance_{\text{dernier mois}} + \gamma Confiance_{\text{dernier mois-1}},$$

avec $\beta > 0$ et $\gamma < 0$, on est coincé.

Ce qu'on retient comme "interprétable"

Un indicateur trimestriel, X_t pour lequel on attend des contributions positifs peut apparaître sous les formes (avec $\beta > 0$ et $gamma > 0$ et significatifs)

- βX_t "le haut niveau de ... impliquerait"
- βX_{t-1} "le haut niveau passé de ... impliquerait"
- $\beta \Delta X_t$ "la hausse de ... impliquerait"
- $\beta \Delta X_t / X_t$ "la variation positive de ... impliquerait"
- $\beta X_t + \gamma \Delta X_t$ "la haut niveau combiné à la hausse de ... impliquerait"

Idem pour les données mensuelles avec quelques subtilités en plus

Algorithme de recherche d'étalonnages interprétables

Pour différents candidats $X_{1t}, X_{2t}, X_{3t}, X_{4t}$.... L'utilisateur les répartit entre indicateurs devant avoir une contribution positive (exp : confiance sur PIB), négative (exp : confiance emploi sur le chômage) ou non déterminée (TUC sur investissement)

- Générer toutes les formules interprétables avec un indicateur seul ou en combinant deux indicateurs
- Estimer chaque formule
- Exclure les formules si un candidat est non significatif et si les signes ne sont pas ceux attendus
- Classer les formules restantes par R^2 ou RMFSE, etc.

Exemple de l'indice des prix en Espagne

Résultat du Lasso

$$\begin{aligned}\Delta log(ipch_{sj}) = \quad & prixattendconst_{mois2} + lag(prixattendconst_{mois2})^{**} + \\ & prixattendconso^{**}_{mois2} + lag(prixattendconst_{mois1}) + \\ & prixattendservice_{mois1} - lag(prixconstconso_{mois2})^{***} + \\ & acquis_{mois1}(ipch_sj)^{***},\end{aligned}$$

$R^2 = 0.937$ et $RMFSE = 0.0013214$

Meilleure formule interprétable

$$\Delta log(ipch_{sj}) = \quad prixattendconst^{***}_{mois2} + acquis_{mois1}(ipch_sj)^{***},$$

$R^2 = 0.916$ et $RMFSE = 0.00115334$

Conclusion

Deux définitions complémentaires d'un bon étalonnage :

1. **Une erreur de prévision faible**
   - Les simulations : Gets modelling l'emporte en moyenne et en stabilité,
   - La pratique : les méthodes ont chacune leurs points forts et un jugement éclairé est nécessaire pour trancher.
   - Autres méthodes possibles :
     - Régularisation : Ridge, Elastic Net,
     - Heuristiques : recuit simulé, cross-entropy,
     - Facteurs dynamiques.
2. **Un étalonnage interprétable**