HAL
open science

Sélection de variables dans les modèles additifs avec des estimateurs en plusieurs étapes

Anestis Antoniadis, Yannig Goude, Jean-Michel Poggi, Vincent Thouvenot

To cite this version:

Anestis Antoniadis, Yannig Goude, Jean-Michel Poggi, Vincent Thouvenot. Sélection de variables dans les modèles additifs avec des estimateurs en plusieurs étapes. [Rapport Technique] Université d'Orsay; EDF R&D; Université Joseph Fourier; Université Cap Town; Université Paris Descartes. 2015. hal-01116100

HAL Id: hal-01116100

https://hal.science/hal-01116100

Submitted on 12 Feb 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SELECTION DE VARIABLES DANS LES MODELES ADDITIFS AVEC DES ESTIMATEURS EN PLUSIEURS ETAPES

ANESTIS ANTONIADIS, Université Joseph Fourier,
YANNIG GOUDE, EDF R&D,
JEAN-MICHEL POGGI, Université Paris Descartes et Université Paris-Sud,
VINCENT THOUVENOT, EDF R&D et Université Paris-Sud

11 février 2015

Abstract

Dans ce document, nous présentons des méthodes d'estimation non paramétrique en plusieurs étapes de modèles additifs dont les composantes sont approchées par leurs développements dans des bases de B-splines. Nous nous plaçons dans un contexte asymptotique dans lequel le nombre d'observations tend vers l'infini et le nombre de covariables candidates pour expliquer le modèle peut éventuellement être plus élevé que le nombre d'observations disponibles, mais pour lequel on suppose qu'il y a moins de covariables "influentes" que d'observations. Pour notre travail, la notion d'effet significatif d'une variable ne se traduit pas, comme il est habituel dans ce contexte, par une norme de la variable bornée inférieurement par une constante strictement positive, car nous supposons que la norme de chaque composante significative est minorée par une suite décroissante dépendant du nombre d'observations et pouvant tendre vers 0 asymptotiquement. Nous étudions ainsi les problèmes de sélection et d'estimation de modèles additifs creux. Nous combinons les techniques des moindres carrés ordinaires (MCO) ou les P-Splines avec le Group LASSO. Nous discutons aussi du choix du critère de sélection de modèle (AIC, GCV ou BIC). Nous établissons la la consistance en sélection et en estimation d'un de nos estimateurs, puis illustrons le bien fondé des méthodes développées par des simulations.

Mots-clés : AIC, B-Spline, BIC, Consistance, Estimateurs en plusieurs étapes, GCV, Group LASSO, MCO, Modèle additif creux, P-Spline, Sélection de variables.

In this document, we present some multi-step nonparametric estimators used for additive models, whose components are approximated by their series developments in B-splines. We assume that the number of covariates can be larger than the number of observations, but that the number of influent covariates is less than the number of observations. In our work, the fact that a covariate has a significant effect does not mean that the norm of the corresponding component is bounded below by a constant positive bound as it is usually assumed in this context, since we only request that norms of significant components to be bounded below by a bound that may decrease to zero at an appropriate speed. We focus on selection and estimation of sparse additive models in this asymptotic context. Our multi-step estimators combine least squares or P-Splines estimators with Group LASSO. We discuss several model selection criteria (AIC, GCV or BIC) and we establish the proofs of selection and estimation consistency of one of our estimators. The behaviour of the resulting estimators is illustrated via simulations.

Keywords : AIC, B-Splines approximation, BIC, Consistency, GCV, Group LASSO, Multi-step estimator, OLS, P-Spline, Sparse additive model, Variable selection.

1 Introduction

Les modèles additifs, introduits par Hastie et Tibshirani [9], permettent de réaliser un bon compromis entre fléxibilité et complexité. Nous utilisons des bases de B-Splines (Stone, [18]) pour en approximer ces composantes. Le problème de sélection de composantes est un sujet classique en statistique. Pour y répondre dans le cas des modèles additifs, Jianqing Fan et Jiancheng Jiang [8] utilisent des tests de déviance. Wood [15] propose une version modifiée des P-Splines [7]. Pour avoir des méthodes efficaces à la fois en termes de sélection et d'estimation, de nombreux auteurs utilisent des estimateurs en plusieurs étapes. Par exemple, Antoniadis *et al.* [3] utilisent un estimateur Nonnegative Garrote ayant pour estimateur initial un estimateur P-Splines ou Huang *et al*. [10] utilisent un estimateur Group LASSO adaptatif, dont l'estimateur initial est un estimateur Group LASSO [22].

Nous nous intéressons (asymptotiquement) à la sélection et à l'estimation de composantes dans le modèle (1) :

$$Y_i = \beta_0^* + \sum_{j=1}^{d} f_j^*(X_{i,j}) + \epsilon_i, \tag{1}$$

avec $i = 1, \ldots, n$, $E(\epsilon_i|\underline{X}_i) = 0$, avec $\underline{X}_i := (X_{i,1}, \ldots, X_{i,d})^T$, ϵ_i i.i.d de variance σ^2 inconnue et des fonctions f_j^* suffisamment régulières en un sens qui sera précisé ci-dessous. Ces hypothèses forment un cadre classique pour la régression additive non paramétrique. Nous n'imposerons pratiquement pas de contraintes sévères sur la dimension d mais nous supposons qu'il n'existe qu'un certain nombre (inconnu) fini et borné de composantes f_j^* qui ont un effet significatif sur la régression. Le modèle (1) est donc un modèle additif creux. Nous supposons sans perte de généralité que le vecteur aléatoire $\underline{X} = (X_1, \ldots, X_d)^T$ est de loi à support dans $[0, 1]^d$.
Nous supposons que les fonctions f_j^* appartiennent à un espace de Sobolev :

$$H_2^2([0,1]) = \{f : [0,1] \to \mathbb{R} | f^{(1)} \text{absolument continue et } \int_0^1 \left(f^{(2)}(x)\right)^2 dx < +\infty\}.$$

De plus, pour des raisons d'identifiabilité, nous supposons que pour tout j, $E(f_j^*(X_j)) = 0$.
Nous notons

$$S^* = \{j \in \{1, \ldots, d\} | E(f_j^*(X_j)^2) \neq 0\},$$

l'ensemble des composantes non nulles du modèle et $s^* = card(S^*)$ le cardinal de S^*. Nous supposons de plus que $s^* << n$ et $s^* \leq D < d = d(n)$. Avec ces hypothèses, nous supposons qu'il y a un nombre restreint de variables explicatives significatives comparé au nombre d'observations. Par contre, nous autorisons le nombre de variables candidates (la dimension du modèle) à augmenter en fonction du nombre d'observations.

Notre objectif est d'obtenir un estimateur qui soit creux, c'est-à-dire qu'il ne doit pas contenir trop de composantes additives non informatives, et qui ait un bon comportement asymptotique, c'est-à-dire que son erreur quadratique moyenne tende vers 0.

Nous avons choisi d'approcher les composantes fonctionnelles du modèle additif (1) par leur développement tronqué dans des bases de B-Splines [18]. Avec cette approximation, nous pouvons approcher le modèle (1) par le modèle (2) :

$$E\left(Y|(X_1, \ldots, X_d) = (x_1, \ldots, x_d)\right) = \beta_0 + \sum_{j=1}^{d} \sum_{k=1}^{m_j} \beta_{j,k} B_{j,k}^{q_j}(x_j), \tag{2}$$

avec $\mathbf{B_j}(\cdot) = \left(B_{j,k}^{q_j}(\cdot) | k = 1, \ldots, K_j + q_j = m_j\right)$ la base de l'espace vectoriel engendré par les B-Splines de degré q_j et K_j noeuds, sur lequel est projetée la $j$ième composante additive $f_j^*(X_j)$. Soit $\mathbf{B_j} = (\mathbf{B_j}(x_{1j})^T, \ldots, \mathbf{B_j}(x_{nj})^T)^T \in \mathbb{R}^{n \times m_j}$, où $x_{i,j}$ constitue la $i$ème observation de la variable X_j. Les

paramètres à estimer du modèle sont alors β_0 et $\boldsymbol{\beta} = (\beta_{1,1}, \ldots, \beta_{d,m_d})^T$. Les variables explicatives sont les $(X_i)_{i\in[\![1,d]\!]}$. Elles peuvent être influentes, ou non, sur la réponse Y. Les composantes du modèle sont les f_j et les paramètres à estimer du modèle sont β_0 et $\boldsymbol{\beta} = (\beta_{1,1}, \ldots, \beta_{d,m_d})^T$. Nous supposons que les m_j, avec $j \in [\![1, d]\!]$ sont connus. Sauf mention contrainte, nous admettrons que $m_1 = \cdots = m_d$.

En projetant les composantes additives dans des bases de B-Splines, nous nous ramenons aux cas d'un modèle linéaire. Lorsque nous retenons une covariable, nous devons donc sélectionner tout le groupe de coefficients qui lui est associé dans la base correspondante de B-Splines. Nous avons donc besoin d'une méthode qui sélectionne groupe par groupe. Nous avons ainsi choisi d'utiliser le Group LASSO pour ajuster et sélectionner le nombre approprié de composantes additives du modèle.

Si le Group LASSO [22] permet de réaliser la sélection de composantes dans les modèles additifs, il a les mêmes inconvénients que le LASSO [19]. Du fait que la pénalisation sur les groupes de coefficients soit convexe (pénalité l_1), ce qui est équivalent à un seuillage doux, les coefficients estimés, même les plus forts, sont sur-rétrécis, amenant à un biais important sur l'estimation des composantes. Le fait que le degré de pénalisation soit constant quelque soit l'amplitude des coefficients explique en partie ce biais. Pour compenser cet effet, le Group LASSO a tendance à inclure des variables non pertinentes. Pour corriger ce biais, et pour mettre au point des méthodes simultanément optimales en sélection et en prédiction asymptotique, nous pouvons utiliser des estimateurs en plusieurs étapes. Nous avons trouvé peu de résultats théoriques démontrés concernant les estimateurs en plusieurs étapes. Antoniadis *et al.* [2] appliquent des variantes de nos procédures avec des bons résultats pratiques. Ceci motive notre étude.

Dans la suite, nous définissons les procédures d'estimateur en plusieurs étapes que nous avons étudié. $S = \{1, \ldots, d\}$ contient les numéros associés aux variables explicatives présentes dans le dictionnaire de covariables. Supposons qu'il existe $\lambda_{\max}$ tel que tous les coefficients estimés par l'estimateur Group LASSO associé à ce λ soient nuls et $\lambda_{\min}$ tel que $\lambda_{\min} << \lambda_{\max}$. Nous construisons une grille de paramètres de lissage uniforme Λ_{GrpL} entre $\lambda_{\min}$ et $\lambda_{\max}$.

Les procédures de type Post, que nous étudions plus en détail par la suite, sont de la forme suivante :

1. **Première étape : Construction de sous-plans d'expériences candidats :** Pour chaque $\lambda_i \in \Lambda_{GrpL}$
   - Notons S_{λ_i} le sous-ensemble des numéros des covariables sélectionnées par l'estimateur associé au Group LASSO de paramètre λ_i.
   - Résoudre $\tilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda_i}} = \arg\min \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{d} \sum_{k=1}^{m_j} \beta_{j,k} B_{j,k}^{q_j}(X_{i,j})\right)^2 + \lambda_i \sum_{j=1}^{d} \sqrt{m_j} \|\boldsymbol{\beta_j}\|_2$.
   - Pour chaque $j \in S$, si $\|\tilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda_i},\boldsymbol{j}}\|_2 \neq 0$, ajouter j dans S_{λ_i}, sinon ne pas considérer j dans S_{λ_i}.

   Pour faciliter les notations, nous supposons que tous les sous-plans d'expériences candidats sont différents.
2. **Seconde étape : Estimation des modèles candidats :** Pour chaque $S_{\lambda_s} \in \{S_{\lambda_{min}}, \ldots, S_{\lambda_{max}}\}$
   - Résoudre un problème du type $\hat{\boldsymbol{\beta}}_{S_{\lambda_s}} = \arg\min Q$ où Q est la fonction objective des moindres carrés ordinaires (MCO) ou des P-Splines [7].
   - Calcul du critère de sélection de modèle (CSM), typiquement le BIC [17], l'AIC [1] ou le GCV [20], de l'estimateur $\hat{\boldsymbol{\beta}}_{S_{\lambda_s}}$.
3. **Troisième étape : Sélection de l'estimateur final :** Sélectionner $\hat{\boldsymbol{\beta}}_{S_{\lambda_b}}$ qui minimise le CSM choisi

Les procédures de type Ante sont de la forme suivante :

1. **Première étape : Construction de sous-plans d'expériences candidats :** Pour chaque $\lambda_i \in \Lambda_{GrpL}$

- Notons S_{λ_i} le sous-ensemble des indices des covariables sélectionnées par l'estimateur associé au Group LASSO de paramètre λ_i.
- Résoudre $\tilde{\boldsymbol{\beta}}_{\lambda_i} = \arg\min \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{d} \sum_{k=1}^{m_j} \beta_{j,k} B_{j,k}^{q_j}(X_{i,j})\right)^2 + \lambda_i \sum_{j=1}^{d} \sqrt{m_j} \|\boldsymbol{\beta_j}\|_2$
- Pour chaque $j \in S$, si $\|\tilde{\boldsymbol{\beta}}_{\lambda_i,j}\|_2 \neq 0$, ajouter j dans S_{λ_i}, sinon ne pas considérer j dans S_{λ_i}.
- Calcul du critère de sélection de modèle (CSM), typiquement le BIC, l'AIC ou le GCV, de l'estimateur $\tilde{\boldsymbol{\beta}}_{\lambda_i}$.

2. **Deuxième étape : Sélection du plan d'expériences candidat final :** Sélectionner $\tilde{\boldsymbol{\beta}}_{\lambda_i}$ qui minimise le CSM choisi, notons S_{λ_s} le plan d'expériences associé à l'estimateur.
3. **Troisième étape : Estimation du modèle sélectionné :** Résoudre un problème du type $\hat{\boldsymbol{\beta}}_{S_{\lambda_s}} = \arg\min Q$ où Q est la fonction objective des MCO ou des P-Splines.

Dans le cadre usuel de la théorie de la statistique inférentielle, nous nous intéressons généralement aux propriétés asymptotiques des estimateurs sous les hypothèses d'un modèle donné a priori. Nous ne sommes pas dans ce cadre, puisque nos procédures comportent une étape de sélection de modèles fondée sur les données. Pour avoir des résultats à la fois en termes de sélection et d'estimation, il faut que les composantes non nulles soient uniformément plus grandes qu'une borne inférieure de rapport de signal sur bruit. Ceci n'est généralement pas vérifié en pratique. Si cette condition n'est pas vérifiée, une estimation consistante des composantes après sélection est généralement impossible. Leeb et Potscher [13] considèrent le problème qui consiste à estimer la distribution d'un estimateur conditionnellement au fait qu'on ait sélectionné au préalable un modèle. Ils proposent des procédures de type Ante (sélection d'un modèle, suivi d'une réestimation des paramètres) et montrent que l'estimateur ne peut généralement pas être consistant uniformément. L'hypothèse que les composantes non nulles soient plus fortes qu'un certain niveau dépendant du bruit est intuitive : si certaines composantes sont trop proches de 0, elles seront plus difficiles à sélectionner et à estimer. Néanmoins, pour démontrer nos résultats, nous supposons seulement que la norme Euclidienne des composantes non nulles ne tend pas trop vite vers 0 asymptotiquement. Nous nous affranchissons donc de l'hypothèse classique qui consiste à supposer que la norme Euclidienne de chaque composante non nulle est supérieure à une constante strictement positive et indépendante de n.

Par la suite, nous nous concentrons sur l'application du BIC, pour lequel nous montrons des résultats asymptotiques. Nous comparons ensuite sur un jeu de simulations les performances numériques obtenus lors de l'utilisation du BIC, du GCV et de l'AIC. Notre objectif est d'obtenir un estimateur qui soit creux (dans le sens qu'il ne contient pas trop de composantes additives non informatives) et qui ait un bon comportement asymptotique. Ce problème est plus difficile à résoudre dans le cas non paramétrique que dans le cas paramétrique, parce que la classe de fonctions dans laquelle nous cherchons à estimer la fonction de régression est plus compliquée. Il faut tenir compte dans l'étude théorique du fait que les coefficients que nous estimons permettent d'estimer une approximation des composantes des modèles additifs, et non les composantes elles-mêmes. Meier *et al.* [16] proposent, dans un article de référence, une méthode d'estimation par moindres carrés pénalisés en introduisant la pénalité (3) suivante :

$$\lambda_1 \sum_{j=1}^{d} \sqrt{\|f_j\|_{2n}^2 + \lambda_2 I(f_j)^2}, \tag{3}$$

où $\|f_j\|_{2n}^2 = 1/n \sum_{i=1}^{n} f_j(X_{i,j})^2$ et $I(f_j)^2 = \int_0^1 f_j^{(2)}(x)^2 dx$. Le premier terme permet de pénaliser l'entrée ou non de la covariable j et le second permet de pénaliser la non régularité de la composante f_j et permet donc de s'assurer du bon comportement asymptotique de l'estimateur et d'éviter le sur-apprentissage. Posons Ω_j tel que $\Omega_{j,kl} = \int B_{j,k}^{q_j(2)}(x) B_{j,l}^{q_j(2)}(x) dx$, $M_j = B_j^T B_j / n + \lambda_2 \Omega_j$. A l'aide d'une décomposition de Cholesky, il existe R_j telle que nous puissions écrire $M_j = R_j^T R_j$. Nous notons $\tilde{\boldsymbol{\beta}}_j = R_j \boldsymbol{\beta}_j$ et $\tilde{\mathbf{B}}_\mathbf{j} = \underline{\mathbf{B}}_\mathbf{j} R_j^{-1}$. Notons $\tilde{\mathbf{B}} = (\tilde{\mathbf{B}}_\mathbf{1} | \ldots | \tilde{\mathbf{B}}_\mathbf{d})$. Meier *et al.* [16] montrent que le

problème d'optimisation des moindres carrés pénalisés par la pénalisation (3) est alors équivalent à optimiser :

$$\arg\min \|\mathbf{Y} - \tilde{\mathbf{B}}\tilde{\boldsymbol{\beta}}\|_{2,n}^2 + \lambda_1 \sum_{j=1}^{d} \sqrt{m_j}\|\tilde{\boldsymbol{\beta}}_j\|_2,$$

et nous reconnaissons là un problème de type Group LASSO. Meier *et al.* [16] proposent de sélectionner λ_2 dans une grille de taille 15. Nous avons remarqué sur une simulation, non présentée dans ce document, qu'il n'y a pas de faux négatifs, mais qu'il y a des faux positifs avec cette méthode. En plus de la possibilité d'introduire des faux positifs, en pratique, cette double pénalisation conduit à des résultats peu satisfaisants en termes de prévision car l'estimateur résultant est moins performant que l'estimateur oracle, qui serait un estimateur construit lorsque S^* est supposé connu, et ceci vaut même si le vrai modèle est sélectionné. Cela semble justifier le fait de séparer la phase de sélection et la phase d'estimation.

Dans le cas des modèles de régression linéaire multiple, Belloni *et al.* [5] ont considéré une procédure en deux étapes : LASSO pour sélectionner les bonnes composantes puis MCO pour estimer et ont observé que cette procédure de type Ante est significativement meilleure que le LASSO car elle le débiaise. Ces résultats sont cohérents avec ce que nous avons aussi constaté sur nos simulations, où les procédures Ante permettent d'améliorer significativement les prédictions faites à partir d'un estimateur Group LASSO. Ce résultat de Belloni *et al.* [5] inspire notre travail dans ce document. Nous nous focalisons sur le Group LASSO et les B-Splines. Pour ce travail, nous avons besoin de tenir plus précisément compte du fait que :

- Chaque composante additive f_j du modèle est approchée par une combinaison linéaire de fonctions de bases (les B-Splines), leur nombre pouvant croître avec le nombre d'observations n pour que l'approximation soit de bonne qualité. Nous avons donc besoin de résultats nous assurant que cette approximation dans des bases de B-Splines n'est pas mauvaise. Stone [18] permet de justifier l'approximation.
- Nous appliquons le Group LASSO aux vecteurs des coefficients résultant de cette approximation puis nous ré-appliquons du Group LASSO ou des MCO ou des P-Splines pour améliorer la sélection (dans les cas des procédures de type Post) et l'estimation en débiaisant l'estimation par Group LASSO.

Nous avons constaté au cours des simulations que la première étape de nos procédures peut inclure des faux positifs ou des faux négatifs, quelque soit le critère de sélection de modèle utilisé (BIC, AIC ou GCV). Les simulations montrent qu'intégrer la seconde étape permet aux procédures de type Post d'améliorer les capacités de sélection de variables des procédures, et aux procédures de type Ante et Post d'améliorer fortement la capacité de prédiction des estimateurs par rapport à l'estimateur Group LASSO.

Dans la section 2, nous donnons les résultats qui prouvent la consistance en sélection et en estimation d'une de nos procédures. Dans la section 3, nous illustrons l'application de nos procédures sur des simulations. Les annexes A et B contiennent les démonstrations des résultats annoncés. L'annexe C contient les détails des résultats sur des simulations.

2 Etude asymptotique des procédures

Notations

Par la suite, nous adoptons les notations suivantes :

$$\underline{B}_j(\cdot) = \{B_{j,k}^{q_j}(\cdot) | k = 1, \ldots, K_j + q_j = m_j\},$$

avec $m_j = m_j(n) \to +\infty$ tel que $m_j = o(n)$. Avec ces deux hypothèses, nous nous assurons que m_j croît lorsque le nombre d'observations augmente (nécessaire pour avoir une bonne approximation

des fonctions dans les bases de B-Splines) de manière lente. Nous poserons des hypothèses plus fortes sur m_j par la suite. Pour simplifier les notations, nous supposons que pour tout j, $j = 1, \ldots, d$, les $m_j = m_n$ et les $\underline{B}_j(\cdot)$ sont identiques. Nous notons

$$\mathcal{B}_j = \{f : [0,1] \to \mathbb{R} | f(\cdot) = \beta_0 + \sum_{k=1}^{m_n} \beta_k B_{j,k}^{q_j}(\cdot); \beta_k \in \mathbb{R}, \beta_0 \in \mathbb{R}\}.$$

Nous notons pour $1 \leq j \leq d$:

$$\underline{\tilde{\mathbf{x}}}_{\mathbf{i},\mathbf{B_j}} = \underline{B}_j^T(x_{ij}) - \frac{1}{n}\sum_{i=1}^{n} \underline{B}_j^T(x_{ij}) \in \mathbf{R}^{m_n},$$

et

$$\tilde{\mathbf{X}}_{\mathbf{i}} = [\underline{\tilde{\mathbf{x}}}_{\mathbf{i},\mathbf{B_1}}^T, \ldots, \underline{\tilde{\mathbf{x}}}_{\mathbf{i},\mathbf{B_d}}^T]^T \in \mathbf{R}^{d \times m_n},$$

et pour $1 \leq j \leq d$,

$$\hat{\Sigma}_j = \frac{1}{n}\sum_{i=1}^{n} \underline{\tilde{\mathbf{x}}}_{\mathbf{i},\mathbf{B_j}} \underline{\tilde{\mathbf{x}}}_{\mathbf{i},\mathbf{B_j}}^T,$$

de dimension $m_n \times m_n$. Nous notons la matrice de covariance empirique

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathbf{X}}_{\mathbf{i}} \tilde{\mathbf{X}}_{\mathbf{i}}^{\mathbf{T}},$$

de dimension $m_n d \times m_n d$. Enfin, nous notons le vecteur des coefficients inconnus

$$\underline{\boldsymbol{\beta}} = (\underline{\boldsymbol{\beta_1}}^T, \ldots, \underline{\boldsymbol{\beta_d}}^T)^T \in \mathbb{R}^{dm_n},$$

et

$$\underline{\boldsymbol{\beta}}_{\boldsymbol{B_j}} = \underline{\boldsymbol{\beta}}_{\boldsymbol{j}} = (\beta_{j1}, \ldots, \beta_{jm_n})^T.$$

La solution du Group LASSO standardisé consiste en :

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i \text{ et } \hat{f}_j(X_j) = \sum_{k=1}^{m_n} \hat{\beta}_{jk}(B_{j,k}^{q_j}(X_j) - \bar{B}_{j,k}^{q_j}),$$

où $\bar{B}_{j,k}^{q_j} = 1/n \sum_{k=1}^{m_n} B_{j,k}^{q_j}(X_{i,j})$ et

$$\hat{\underline{\boldsymbol{\beta}}} = \arg\min_{\underline{\boldsymbol{\beta}} \in \mathbf{R}^{dm_n}} + \frac{1}{2}\sum_{i=1}^{n}(Y_i - \tilde{X}_i^T \underline{\boldsymbol{\beta}})^2 + \sqrt{m_n}\lambda \sum_{j=1}^{d} \|\hat{\Sigma}^{1/2} \underline{\boldsymbol{\beta}}_{\boldsymbol{j}}\|_2, \tag{4}$$

où $\|\cdot\|_2$ désigne la norme euclidienne de $\mathbb{R}^{m_n}$.

Le Group LASSO est creux par groupe, ce qui signifie que si λ est suffisamment fort, il existe j telle que $\hat{\underline{\boldsymbol{\beta}}}_{\boldsymbol{j}} = \mathbf{0}$. Notons aussi que $\hat{\underline{\boldsymbol{\beta}}}_{\boldsymbol{j}} = \mathbf{0}$ si, et seulement si, $\hat{f}_j(.) = 0$ (sauf si $\hat{\Sigma}_j$ est singulière). Nous avons utilisé une base commune pour toutes les covariables, ce qui n'est pas nécessaire pour utiliser le Group LASSO. Nos procédures Ante et Post dépendant de méthodes déjà existantes, l'implémentation est simple. Nous nous sommes concentrés sur le Group LASSO pour la première étape des procédures, nous aurions pu utiliser une autre pénalisation, qui tiendrait par exemple compte de la corrélation inter-groupe, s'il y en a.

Nous utilisons le Group LASSO standardisé qui a des bonnes propriétés asymptotiques quand le paramètre λ est choisi de manière optimale selon le critère de sélection adopté.

Nous notons

$$\|\mathbf{f}^*(\mathbf{x})\| = \|\mathbf{f}^*(\mathbf{x})\|_2 = \sqrt{\sum_{i=1}^{n}\left(\sum_{j=1}^{d} f_j^*(x_{i,j})\right)^2}$$

et

$$\|\mathbf{f}^*\|_\infty = \max_{\mathbf{x_i}, i\in 1,\dots n} |\mathbf{f}^*(\mathbf{x_i})| = \max_{\mathbf{x_i}, i\in 1,\dots n} |\sum_{j=1}^{d} f_j^*(x_{i,j})|.$$

Nous noterons indifféremment $\|\cdot\|$ et $\|\cdot\|_2$ lorsque nous utilisons la norme euclidienne.

Par la suite, nous avons tout écrit avec $m_j = m = m(n)$ et la même base pour simplifier les notations mais les résultats obtenus par la suite sont généraux.

2.1 Hypothèses pour l'étude asymptotique

Meier *et al.* [16] donnent les conditions sous lesquelles une version modifiée de l'équation (3) permet de ne pas avoir de faux négatifs (voir le corollaire 1 de [16]). Pour que les conditions soient remplies, il faut que les composantes additives non nulles soient suffisamment éloignées de 0 au sens de la métrique que ces auteurs définissent. Cependant, ce résultat est insuffisant pour démontrer la consistance en sélection. De plus, dans les conditions du corollaire, il y a une condition de compatibilité, qui n'est pas vérifiable en pratique puisqu'elle nécessite la connaissance de S^*.

Nous énumérons et expliquons les différentes conditions qui sont nécessaires à la démonstration des résultats théoriques de nos estimateurs. Selon les résultats recherchés, il n'est pas nécessaire de supposer que toutes les conditions soient simultanément réalisées, mais il nous semble plus clair d'en énumérer la liste complète en premier.

Une référence importante dans le travail qui suit est l'article de Wang, Li et Leng [21]. Les résultats de Wang *et al.* [21] sont consacrés à l'analyse des modèles linéaires avec un nombre de covariables possibles tendant vers l'infini. Wang *et al.* [21] montrent que le BIC est un critère permettant d'identifier de manière consistante le vrai modèle lorsque le nombre de covariables potentielles tend vers l'infini pour les estimateurs, qu'ils soient pénalisés ou non. Nous nous sommes inspirés de leur travail pour l'appliquer au modèle additif décrit en introduction.

Nous nous plaçons dans le même cadre que Huang *et al.* [10] au niveau des dimensions des plans d'expériences, c'est-à-dire :

- Le nombre de covariables candidates d peut éventuellement tendre vers l'infini
- Le nombre de composantes non nulles s^* ne dépend pas de n
- $s^* \leq D < n$ avec D fixé indépendant de n

Notons, pour $j = 1, \dots, d$

$$\mathbf{U_j} = \begin{pmatrix} B_{j1}(x_{1j}) - \bar{B}_{j1} & \dots & B_{jm_n}(x_{1j}) - \bar{B}_{jm_n} \\ \vdots & & \vdots \\ B_{j1}(x_{nj}) - \bar{B}_{j1} & \dots & B_{jm_n}(x_{nj}) - \bar{B}_{jm_n} \end{pmatrix},$$

que nous pouvons réécrire :

$$\mathbf{U_j} = \left(\tilde{\mathbf{x}}_{\mathbf{1},\mathbf{B_j}} \dots \tilde{\mathbf{x}}_{\mathbf{n},\mathbf{B_j}}\right)^T,$$

et posons $\underline{\mathbf{Y}} = (Y_1, \dots Y_n)^T$. Enfin, notons $\mathbf{U} = [\mathbf{U_1} \dots \mathbf{U_d}]_{n\times dm_n}$.

Pour tout sous modèle $S \subseteq \{1, \dots, d\}$ nous notons $\mathbf{U_S}$ la matrice extraite de $\mathbf{U}$ constituée des colonnes indexées par S, c'est-à-dire que

$$\mathbf{U_S} = [\mathbf{U_j} | j \in S].$$

De même, nous définissons $\boldsymbol{\beta}_S = (\boldsymbol{\beta}_j)_{j\in S}$, $\hat{\boldsymbol{\beta}}_S = (\hat{\boldsymbol{\beta}}_j)_{j\in S}$ (estimés sur S), et plus généralement $\boldsymbol{\psi}_S = (\boldsymbol{\psi}_j)_{j\in S}$.

Nous notons $\mathbf{1_n}$ le vecteur colonne de taille n dont toutes les composantes sont égales à 1. Pour simplifier les notations, nous ajoutons $\mathbf{1_n}/\sqrt{m_n}$ comme première colonne de $\mathbf{U}$ et de $\mathbf{U_S}$.
Posons

$$\underline{\mathbf{b}} = \left(\sqrt{m_n}\beta_0, \boldsymbol{\beta}^T\right)^T \text{ et } \underline{\mathbf{b_S}} = \left(\sqrt{m_n}\beta_0, \boldsymbol{\beta_S}^T\right)^T.$$

Nous pouvons alors réécrire matriciellement le critère MCO associé au sous-modèle S par l'équation (5) :

$$Q_S^{MCO}(\mathbf{b_S}) = \|\underline{\mathbf{Y}} - \mathbf{U_S}\mathbf{b_S}\|^2. \tag{5}$$

Notons $\underline{\hat{\mathbf{b}}}_\mathbf{S}$ un vecteur minimisant (5), c'est-à-dire

$$\underline{\hat{\mathbf{b}}}_\mathbf{S} = \arg\min_{\mathbf{b_S}} Q_S^{MCO}(\mathbf{b_S}).$$

Notons que ce vecteur n'est pas unique si $n \leq card(S)m$.
Nous définissons, (voir Wang *et al.* [21]), le critère BIC par l'équation (6) :

$$BIC(S) = \log(\|\underline{\mathbf{Y}} - \mathbf{U_S}\underline{\hat{\mathbf{b}}}_\mathbf{S}\|^2) + card(S)m_n\frac{\log(nd)}{n}. \tag{6}$$

Pour être cohérent avec la définition usuelle du BIC, nous aurions dû utiliser $\log(\frac{\|\underline{\mathbf{Y}}-\mathbf{U_S}\underline{\hat{\mathbf{b}}}_\mathbf{S}\|^2}{n})$ à la place de $\log(\|\underline{\mathbf{Y}} - \mathbf{U_S}\underline{\hat{\mathbf{b}}}_\mathbf{S}\|^2)$, mais cela ne modifie pas le critère puisque nous cherchons S tel que $S = \arg\min BIC(S)$ et $\log\frac{(\|\underline{\mathbf{Y}}-\mathbf{U_S}\underline{\hat{\mathbf{b}}}_\mathbf{S}\|^2)}{n} = \log(\|\underline{\mathbf{Y}} - \mathbf{U_S}\underline{\hat{\mathbf{b}}}_\mathbf{S}\|^2) - \log(n)$.

Nous faisons les hypothèses suivantes :

1. **Hypothèse sur les vecteurs aléatoires** $(Y_i, \underline{\mathbf{X_i}})_{i=1,\dots,n}$ **:** $(Y_i, \underline{\mathbf{X_i}})_{i=1,\dots,n}$ est un n échantillon de vecteurs aléatoires $(Y_1, \underline{\mathbf{X_1}})$ de loi satisfaisant le modèle additif postulé dans le modèle (1).
2. **Hypothèse sur la loi des erreurs :** la loi conditionnelle de ϵ_1 sachant $\underline{\mathbf{X_1}}$ est une loi $N(0, \sigma^2(\underline{\mathbf{X_1}}))$ et $\sigma_1^2 \leq \sigma^2(\underline{\mathbf{X_1}}) \leq \sigma^2$ p.s. avec σ^2 et σ_1^2 des constantes indépendantes de n (ce cas recouvre le cas classique où les erreurs sont indépendantes de $\underline{\mathbf{X_1}}$).
3. **Hypothèse sur la loi de** $\underline{\mathbf{X_1}}$: le support de $\underline{\mathbf{X_1}}$ est compact. Pour alléger les notations, nous supposons que le support est contenu dans $[0, 1]^d$. De plus, les densités marginales de $\underline{\mathbf{X_1}}$ sont bornées inférieurement et supérieurement par deux constantes strictement positives fixes. La densité de $\underline{\mathbf{X_1}}$ est continue par rapport à la mesure de Lebesgue.
4. **Hypothèse sur la régularité des composantes additives** : pour tout $j \in S^*$, $f_j \in H_2^\nu([0, 1])$ avec ν entier supérieur ou égale à 2. L'inégalité de Morrey permet de dire que f_j est Holdérienne d'indice $> 1/2$.
5. **Hypothèses sur** $d, m_n,$ **et** s^* :
   (a) $m_n = An^{\frac{1}{2\nu+1}}$, avec A constante. Nous supposons donc que nous prenons des bases de splines de degré $q \geq 3$ (splines cubiques) pour approximer les composantes.
   (b) $\frac{\log(d)}{n^{\frac{2\nu}{2\nu+1}}} \to 0$
   (c) $s^* \leq D \leq \min(d, \frac{n}{m_n})$
   (d) $\frac{m_n \log(nd)}{n \min_{j\in S^*} \|f_j^*\|^2} \to 0$, c'est à dire que $\frac{m_n \log(nd)}{n} = o(\min_{j\in S^*} \|f_j^*\|^2)$.

L'hypothèse 1 est une hypothèse classique en régression. L'hypothèse 2 est une hypothèse suffisante pour s'assurer une propriété de concentration de loi normale (autour de sa moyenne et sachant $\underline{\mathbf{X_1}}$) d'une variable aléatoire de la forme $\sup_{\mathbf{t}\in T} \sum_{i=1}^n \epsilon_i t_i$, où T est un sous ensemble de $\mathbb{R}^n$ borné, dénombrable et dépendant de $\underline{\mathbf{X_1}}, \dots, \underline{\mathbf{X_n}}$. L'hypothèse 3 est classique. L'hypothèse 4 est une hypothèse sur la régularité des composantes additives. L'hypothèse 5.a est imposée par le choix de Stone pour l'approximation des composantes additives dans des bases de B-Splines permettant de réaliser un compromis optimal entre le biais et la variance. L'hypothèse 5.b assure que le logarithme de la dimension d de covariables candidates puisse croître mais de manière moins

rapide que $n^{\frac{2\nu}{2\nu+1}}$. Cette hypothèse nous permet donc d'affirmer que le nombre de covariables candidates croît mais n'explose pas. L'hypothèse 5.c permet de s'assurer de l'unicité de l'estimateur des MCO et est nécessaire pour des raisons d'identifiabilité. Enfin, l'hypothèse 5.d est suffisante pour nous permettre de démontrer des résultats de consistance en sélection et en estimation. Cette hypothèse est intéressante, parce qu'elle n'oblige pas la norme euclidienne de chaque composante à être strictement supérieure à une constante strictement positive indépendante de n.

2.2 Consistance asymptotique en sélection du BIC appliqué à l'estimateur des MCO

Dans cette sous-section, nous montrons que sous les conditions 1 à 5, le critère BIC défini par l'équation (6) est consistant en sélection de variables, c'est-à-dire que

$$P(\hat{S} = S^*) \to_{n\to\infty} 1.$$

Nous considérons deux cas. Dans le premier cas, nous supposons que nous ajustons un modèle S tel que $S^* \not\subseteq S$. Dans le second cas, nous supposons que nous ajustons un modèle S tel que $S^* \subseteq S$.

2.2.1 Cas où $S^* \not\subseteq S$

Considérons le cas où $S^* \not\subseteq S$. Notons $\underline{\hat{\mathbf{b}}}_{\mathbf{S}}$ et $\underline{\hat{\mathbf{b}}}_{\mathbf{S}^*}$ les estimateurs des MCO sous les modèles S et S^*, c'est-à-dire

$$\underline{\hat{\mathbf{b}}}_{\mathbf{S}} = \arg\min_{\underline{\mathbf{b}}} ||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\underline{\mathbf{b}}||^2 \text{ et } \underline{\hat{\mathbf{b}}}_{\mathbf{S}^*} = \arg\min_{\underline{\mathbf{b}}} ||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}||^2.$$

Nous avons supposé que $card(S) \leq D$. L'hypothèse 5.c nous assure que $m_n D < n$.
Notons alors

$$\bar{S} = S^* \cup S,$$

et notons aussi $\underline{\tilde{\mathbf{b}}}_{\mathbf{S}}$ le vecteur de dimension $card(\bar{S})m_n + 1$ coïncidant avec $\underline{\hat{\mathbf{b}}}_{\mathbf{S}}$ pour les coefficients appartenant à S et nul pour les autres composantes, c'est-à-dire

$$\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} = \left(\bigcup_{j\in\bar{S}} \underline{\tilde{\mathbf{b}}}_{\mathbf{S},\mathbf{j}}^T\right) \text{ avec } \underline{\tilde{\mathbf{b}}}_{\mathbf{S},\mathbf{j}} = \underline{\hat{\mathbf{b}}}_{\mathbf{S},\mathbf{j}} \text{ si } j \in S \text{ , } \mathbf{0} \text{ sinon.}$$

De même, nous notons

$$\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*} = \left(\bigcup_{j\in\bar{S}} \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*,\mathbf{j}}^T\right) \text{ avec } \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*,\mathbf{j}} = \underline{\hat{\mathbf{b}}}_{\mathbf{S}^*,\mathbf{j}} \text{ si } j \in S^* \text{ , } \mathbf{0} \text{ sinon,}$$

et

$$\underline{\mathbf{b}}_{\bar{\mathbf{S}}}^* = \left(\bigcup_{j\in\bar{S}} \underline{\mathbf{b}}_{\bar{\mathbf{S}},\mathbf{j}}^{*,T}\right) \text{ avec } \underline{\mathbf{b}}_{\bar{\mathbf{S}},\mathbf{j}}^* = \underline{\mathbf{b}}_{\mathbf{S}^*,\mathbf{j}}^* \text{ si } j \in S^* \text{ , } \mathbf{0} \text{ sinon.}$$

Avec ces notations, nous prolongeons par des 0 lorsque cela est nécessaire les divers vecteurs considérés pour que les produits matriciels aient un sens.
Nous remarquons que

$$||\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}}||^2 = ||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}}||^2,$$

et de plus,

$$BIC(S) - BIC(S^*) = \log\left(1 + \frac{||\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}}||^2/n - ||\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}||^2/n}{||\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}||^2/n}\right) + (card(S) - card(S^*))\, m_n \frac{\log(nd)}{n}. \tag{7}$$

Nous montrons que $BIC(S) - BIC(S^*) > 0$ avec probabilité tendant vers 1. Plus précisément,

Théorème 1 *Soit S tel que $S^* \not\subseteq S$. Alors,*

$$P(\min_{S^* \not\subseteq S \,|\, card(S) \leq M} BIC(S) - BIC(S^*) > 0) \to_{n \to +\infty} 1,$$

où M constante.

En particulier, une des conséquences est qu'il n'y a pas de faux négatifs avec probabilité tendant vers 1 lorsque le vrai plan d'expériences est candidat.

Pour démontrer le Théorème 1, nous étudions d'abord le terme suivant donné dans l'équation (8) :

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}} \underline{\tilde{\mathbf{b}}}_S\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}} \underline{\tilde{\mathbf{b}}}_{S^*}\|^2. \tag{8}$$

Nous pouvons montrer grâce à des calculs classiques de développement de la norme euclidienne et en utilisant le fait que $\underline{\mathbf{Y}} = \underline{\mathbf{f}}^*(\underline{\mathbf{X}}) + \underline{\boldsymbol{\epsilon}}$ l'identité (9) suivante :

$$\begin{aligned} &\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}} \underline{\tilde{\mathbf{b}}}_{\mathbf{S}}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}} \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}\|^2 \\ &\qquad = -2\underline{\boldsymbol{\epsilon}}^T \mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}) + 2\left(\mathbf{U}_{\bar{\mathbf{S}}} \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{f}}^*(\underline{\mathbf{X}})\right)^T \mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}) \\ &\qquad + \|\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})\|^2, \end{aligned} \tag{9}$$

où $[\underline{\mathbf{f}}^*(\underline{\mathbf{X}})]_{ij} = \beta_0^* + \sum_j f_j^*(X_{i,j})$.
Le premier terme du membre de droite de l'équation (9) correspond à la multiplication de l'erreur propre au modèle par la différence des prédictions faites lorsque nous utilisons les deux plans d'expériences différents. Le second terme de droite correspond au coût de l'approximation dans des bases de B-Splines tronquées et de l'utilisation de MCO pour estimer la vraie fonction de régression. On multiplie ce coût par la différence des prédictions selon le plan d'expériences utilisé. Le dernier terme de droite représente la norme L2 au carré de cette différence.

Stone [18] a montré que lorsque S^* est connu, nous avons :

$$\|\underline{\hat{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{b}}^*_{\mathbf{S}^*}\| = O\left(\frac{m_n}{\sqrt{n}} + m_n^{-\nu+1/2}\right),$$

et de plus, lorsque $f_j^* \in H_2^\nu$,

$$\|f_j^* - \sum_{k=1}^{m_n} b_{jk}^* \tilde{B}_{j,k}\|_\infty = O(m_n^{-\nu}).$$

La première égalité permet de contrôler l'erreur commise parce que nous utilisons des estimateurs MCO pour estimer les coefficients associés aux bases de B-Splines, alors que la seconde permet de contrôler uniformément le biais pour approcher les composantes additives par des projections dans des bases de B-Splines.

Le contrôle de la différence donnée dans le terme du membre de gauche de l'équation (9) est obtenu en étudiant séparément les trois termes du membre de droite de l'équation (9). Nous commençons par l'analyse de troisième terme. Le Lemme 1 en donne une minoration.

Lemme 1

$$\|\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})\|^2 \geq A_2 \frac{n}{m_n} v_n^2, \tag{10}$$

où A_2 une constante et

$$v_n = A_3\left(\sqrt{m_n} \min_{j \in S^*} \|f_j^*\| - \frac{m_n}{\sqrt{n}} - m_n^{-\nu+1/2}\right), \tag{11}$$

avec A_3 une constante.

Nous constatons que nous pouvons minorer le troisième terme par une suite indépendante du plan d'expériences S. De plus, nous remarquons que

$$\frac{m_n}{\sqrt{n}} + m_n^{-\nu+1/2} = 2An^{\frac{-\nu+1/2}{2\nu+1}} \to 0$$

Nous montrons également en annexe A le Lemme 2, qui établit que le second terme du membre de droite de l'équation (9) est négligeable par rapport au troisième terme.

Lemme 2

$$|2\left(\mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{f}}^*(\underline{\mathbf{X}})\right)^T \mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})| \leq C_4 \sqrt{nm_n^{-2\nu}} \sqrt{\frac{n}{m_n}}|v_n|,$$

avec $C_4 > 0$ constante. L'ordre de ce terme est négligeable par rapport au troisième terme de l'équation (9).

Il reste à étudier le premier terme du membre de droite de l'équation (9), que nous ne pouvons pas minorer par un terme strictement positif. Cependant, nous montrons que nous pouvons minorer la somme du premier terme et du troisième terme de droite par une suite strictement positive. Pour prouver ce résultat, nous établissons des résultats à propos du coût d'approximation des composantes du modèle additif dans des bases de B-Splines (Proposition 1 de l'annexe A), ainsi que sur l'expression $\underline{\boldsymbol{\epsilon}}^T \Pi_{\bar{S}} \underline{\boldsymbol{\epsilon}}$ (Proposition 2 de l'annexe A). Nous établissons aussi la Proposition 3 de l'annexe A, qui est centrale pour la suite de la démonstration et montre que $\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\hat{\mathbf{b}}}_{\bar{\mathbf{S}}}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\|^2 = O_p(nm_n^{-2\nu}) + o_p(m_n \log(nd))$ lorsque $S^* \subseteq \bar{S}$.

Puisque $S^* \subseteq \bar{S}$, nous pouvons écrire

$$\begin{aligned}\underline{\mathbf{Y}} &= \underline{\mathbf{f}}^*(\underline{\mathbf{X}}) + \underline{\boldsymbol{\epsilon}} \\ &= \underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}} + \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}} + \underline{\boldsymbol{\epsilon}} \\ &= B + \underline{\boldsymbol{\epsilon}} + \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}},\end{aligned}$$

avec $B = \underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}$.

De plus, nous notons $\Pi_{\bar{S}}$ le projecteur orthogonal sur l'espace engendré par les colonnes de $\mathbf{U}_{\bar{\mathbf{S}}}{}^T$:

$$\Pi_{\bar{S}} = \mathbf{U}_{\bar{\mathbf{S}}}(\mathbf{U}_{\bar{\mathbf{S}}}{}^T\mathbf{U}_{\bar{\mathbf{S}}})^{-1}\mathbf{U}_{\bar{\mathbf{S}}}{}^T.$$

Le lemme suivant permet d'affirmer que la somme du premier et du troisième terme du membre de droite de l'équation (9) est minorée par une suite strictement positive.

Lemme 3 *Sous nos hypothèses,*

$$-2\underline{\boldsymbol{\epsilon}}^T\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}) \geq -4\underline{\boldsymbol{\epsilon}}^T\Pi_{\bar{S}}\Pi_{\bar{S}}\underline{\boldsymbol{\epsilon}} - \frac{1}{4}\|\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})\|^2. \tag{12}$$

Nous avons donc

$$-2\underline{\boldsymbol{\epsilon}}^T\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}) + \|\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})\|^2 \geq \frac{3}{4}\|\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})\|^2 + o_P(m_n \log(nd)) \geq R_1\frac{n}{m_n}v_n^2,$$

avec R_1 constante.

Grâce aux trois lemmes précédents, nous obtenons que

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}\|^2 \geq N_1 \frac{n}{m_n} v_n^2 + o_p(\frac{n}{m_n} v_n^2), \tag{13}$$

avec N_1 constante.

Les hypothèses sur les normes des composantes additives impliquent en particulier que :

$$\frac{v_n^2}{m_n} \geq \min \|f_j^*\|^2 + o(\min \|f_j\|^2).$$

Ceci permet d'établir le Lemme 4 :

Lemme 4 *Les suites*

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\mathbf{b}_{\mathbf{S}^*}\|^2/n$$

et

$$\inf_{S^* \subseteq S | card(S) \leq D} \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\mathbf{b}}_{\mathbf{S}}\|^2/n$$

sont des suites bornées inférieurement par une constante strictement positive avec une probabilité tendant vers 1.

D'après ce lemme, $\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\hat{\mathbf{b}}_{\mathbf{S}^*}\|/n$ est bornée inférieurement. L'équation (13) et les résultats précédents montrent qu'il existe une constante N_2 telle que, avec une probabilité tendant vers 1,

$$\frac{\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}}\|^2/n - \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}\|^2/n}{\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}\|^2/n} \geq N_2 \frac{v_n^2}{m_n} > \min \|f_j^*\|^2 + o(\min \|f_j^*\|^2) > 0.$$

Comme il est montré en annexe A, nous pouvons conclure la démonstration du Théorème 1 en utilisant $\log(x) \leq \min(x/2, \log(2))$.

2.2.2 Cas où $S^* \subseteq S$

Nous montrons ensuite le résultat du Théorème 2, qui traite des faux positifs.

Théorème 2

$$P(\min_{S^* \subset S | card(S) \leq M} BIC(S^*) - BIC(S) < 0) \to_{n \to +\infty} 1.$$

Notons, de la même manière que dans le paragraphe précédent, par $\underline{\mathbf{b}}_{\mathbf{S}}^*$ le vecteur $\underline{\mathbf{b}}_{\mathbf{S}^*}^*$ prolongé par des 0 pour les j tels que $j \in S$ et $j \notin S^*$.

Nous étudions l'expression :

$$BIC(S^*) - BIC(S) = \log\left(\frac{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\hat{\mathbf{b}}}_{\mathbf{S}^*}\|^2}{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}}\|^2}\right) + (card(S^*) - card(S))\, m_n \frac{\log(nd)}{n}.$$

Par définition de l'estimateur MCO,

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\hat{\mathbf{b}}}_{\mathbf{S}^*}\|^2 \leq \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}_{\mathbf{S}^*}^*\|^2.$$

Donc, nous avons

$$BIC(S^*) - BIC(S) \leq \log\left(1 + \frac{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}_{\mathbf{S}^*}^*\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}}\|^2}{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}}\|^2}\right) + (card(S^*) - card(S))\, m_n \frac{\log(nd)}{n}.$$

Nous montrons en annexe A le lemme suivant :

Lemme 5

$$\frac{||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*}||^2 - ||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}||^2}{||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}||^2} \in [0, 1], \tag{14}$$

avec une probabilité tendant vers 1.

Dans la démonstration du Lemme 5, nous montrons que

$$||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*}||^2/n - ||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}||^2/n = ||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\underline{\mathbf{b}}^*_{\mathbf{S}}||^2/n - ||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}||^2/n \to^P 0,$$

ce qui montre que les erreurs quadratiques moyennes obtenues lorsque les “vrais" coefficients sont utilisés tendent à prendre les mêmes valeurs que les erreurs quadratiques moyennes des estimateurs MCO (qui sont minimales).

En annexe A, nous utilisons le Lemme 5 et le développement limité en 0 du logarithme pour démontrer le Théorème 2.

Le Théorème 1 nous permet d'affirmer que si S^* est candidat, tous les ensembles S de cardinal fini ne contenant pas S^* ont un BIC supérieur avec une probabilité tendant vers 1. Ce théorème permet ainsi de se prémunir des faux négatifs. Ce résultat et le Théorème 2 permettent d'affirmer qu'il n'y a pas de faux positifs avec une probabilité tendant vers 1.

2.2.3 Théorème de consistance du BIC

Nous pouvons alors énoncer le Théorème 3 qui est analogue au Théorème 3 de Wang *et al.* [21] et qui permet d'affirmer la consistance du BIC.

Théorème 3 *Sous les conditions 1 à 5, le critère BIC défini par (6) est consistant en sélection de variables, c'est-à-dire*

$$P(\hat{S} = S^*) \to_{n\to\infty} 1.$$

Le BIC défini par l'équation (6) est donc consistant en termes de sélection de variables. Ceci n'est pas suffisant pour prouver la consistance de notre procédure d'estimation. En effet, nous venons de démontrer que dans le cas où nous estimons à l'aide de MCO un modèle additif dont les composantes sont approchées dans une base de B-Splines, si le vrai modèle est parmi les modèles candidats, alors il est sélectionné avec probabilité tendant vers 1 lorsque le BIC est utilisé comme critère de sélection.

Toutefois, comme nous venons de le voir, le critère BIC que nous avons étudié n'est pas le critère BIC de l'estimateur que nous avons défini dans l'équation (4), car il ne tient pas compte de la pénalisation.

2.3 Etude de la procédure Post1

2.3.1 Consistance en sélection de la procédure Post1

Le Théorème 3 permet d'affirmer la consistance en sélection du BIC lorsqu'un estimateur MCO est utilisé. Nous montrons ensuite que le BIC appliqué à l'estimateur défini en (4) est consistant.

Sous nos hypothèses, Kato [11] établit la Proposition 4 (voir annexe B), qui permet de simplifier l'équation (4) et d'affirmer que nous pouvons remplacer $\hat{\Sigma}_j$ par la matrice identité dans l'expression (4). De plus, le centrage des B-Splines n'a pas d'impact sur les résultats asymptotiques. Nous pouvons ainsi réécrire l'équation (4) sous la forme de l'équation (15) :

$$\tilde{\underline{\mathbf{b}}}_\lambda = \arg\min_{\underline{b}\in\mathbf{R}^{dm_n+1}} ||\underline{\mathbf{Y}} - \mathbf{U}^{\mathbf{T}}\underline{\mathbf{b}}||^2 + \sqrt{m_n}\lambda \sum_{j=1}^{d} ||\underline{\mathbf{b}}_{\mathbf{j}}||_2. \tag{15}$$

Il se pose la question du choix du paramètre λ. Le paramètre de lissage λ est choisi sur une grille de valeurs positives en sélectionnant celui qui minimise le BIC pénalisé.

Comme Huang *et al.* [10] l'ont constaté et utilisé pour le LASSO adaptatif, ce choix de λ permet de réduire le nombre de composantes additives et d'estimer un nombre $card(\hat{S})$ de composantes avec une très faible probabilité de faux négatifs.

Par la suite, nous précisons quelques notations. Nous notons

$$S_\lambda = \{j; \|\underline{\tilde{\mathbf{b}}}_{\lambda,\mathbf{j}}\| \neq 0\},$$

De même, nous notons

$$BIC(\lambda) = \log(\|\underline{\mathbf{Y}} - \mathbf{U}\underline{\tilde{\mathbf{b}}}_\lambda\|^2) + card(S_\lambda) m_n \frac{\log(nd)}{n}.$$

De plus, nous notons

$$\underline{\hat{\mathbf{b}}}_\lambda = (\underline{\tilde{\mathbf{b}}}_{\lambda,\mathbf{j}})_{\mathbf{j}\in\mathbf{S}_\lambda}.$$

A λ fixé, nous justifions dans la Proposition 5 (voir l'annexe B) qu'il est équivalent de résoudre le problème d'optimisation en considérant $\mathbf{U}$ ou $\mathbf{U}_{\mathbf{S}_\lambda}$ (à la dimension des vecteurs près). Ce résultat est important, puisqu'il permet de travailler sur le plan d'expériences $\mathbf{U}_{\mathbf{S}_\lambda}$, qui est de dimension finie par hypothèse, et non sur $\mathbf{U}$.

Nous avons donc

$$\begin{aligned} BIC(\lambda) &= \log(\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}_\lambda}\underline{\hat{\mathbf{b}}}_\lambda\|^2) + card(S_\lambda) m_n \frac{\log(nd)}{n} \\ &= \log(\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\underline{\tilde{\mathbf{b}}}_\lambda\|^2) + card(S_\lambda) m_n \frac{\log(nd)}{n}. \end{aligned}$$

Nous énonçons le théorème affirmant la consistance du BIC pénalisé.

Théorème 4 *Sous nos hypothèses,*
Soit $\lambda_n \sim \sqrt{n\log(m_n^2 d)/m_n}$,
Alors

$$P(S_{\lambda_n} = S^*) \to 1,$$

et pour tout λ *tel que* $S_\lambda \neq S^*$ *et* $card(S_\lambda) \leq M$

$$P(BIC(\lambda) - BIC(\lambda_n) > 0) \to 1.$$

Le Théorème 4 permet d'affirmer la consistance en sélection de l'estimateur issu de la procédure Post1. Pour démontrer ce théorème, nous montrons qu'à un terme négligeable près, il est équivalent de minimiser le BIC pénalisé et le BIC étudié dans la section précédente. La première étape est de donner une solution au programme d'optimisation associé au Group LASSO.

Soit λ_n tel que $S_{\lambda_n} = S^*$. Alors nous pouvons facilement montrer que

$$\underline{\hat{\mathbf{b}}}_{\lambda_\mathbf{n}} = ({\mathbf{U}_{\mathbf{S}^*}}^T \mathbf{U}_{\mathbf{S}^*})^{-1}({\mathbf{U}_{\mathbf{S}^*}}^T \underline{\mathbf{Y}} + \underline{\mathbf{r}}),$$

avec

$$\begin{aligned} \underline{\mathbf{r}} &= -\lambda_n \sqrt{m_n} \frac{\partial}{\partial \underline{\mathbf{b}}} \sum_{j\in S^*} \|\underline{\mathbf{b}}_\mathbf{j}\|_{|\underline{\mathbf{b}}=\underline{\hat{\mathbf{b}}}_{\lambda_\mathbf{n}}} \\ &= -\lambda_n \sqrt{m_n} \left(0 \ldots (\frac{\underline{\hat{\mathbf{b}}}^{\mathbf{T}}_{\lambda_\mathbf{n},\mathbf{j}}}{\|\underline{\hat{\mathbf{b}}}_{\lambda_\mathbf{n},\mathbf{j}}\|})_{j\in S^*} \right)^T \end{aligned}$$

Le terme $\underline{\mathbf{r}}$ est inconnu, puisqu'il dépend des dérivées partielles des coefficients. Nous pouvons facilement montrer que

$$\|\underline{\mathbf{r}}\|^2 = O(\lambda_n^2 m_n).$$

$\underline{\mathbf{r}}$ est le terme qui différencie l'estimateur Group LASSO de l'estimateur MCO sur le plan d'expériences $\mathbf{U_{S^*}}$ et de coefficient de pénalisation λ_n. En effet,

$$\hat{\mathbf{b}}_{\lambda_\mathbf{n}} = \hat{\mathbf{b}}_{\mathbf{S^*}} + ({\mathbf{U_{S^*}}}^T \mathbf{U_{S^*}})^{-1} \underline{\mathbf{r}}.$$

Nous énonçons ensuite les résultats qui montrent que si $\lambda_n \sim \sqrt{n \log(m_n^2 d)/m_n}$, alors

$$P(S_{\lambda_n} = S^*) \to 1.$$

La Proposition 20 (voir annexe B) permet d'affirmer que $S_{\lambda_n} = S^*$ si

$$\begin{cases} \|\underline{\mathbf{b}}_\mathbf{j}^*\| - \|\tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}}\| < \|\underline{\mathbf{b}}_\mathbf{j}^*\| \text{ si } j \in S^* \\ \|{\mathbf{U_j}}^T(\underline{\mathbf{Y}} - \mathbf{U_{S^*}} \tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{S^*}})\| \leq \lambda_n \sqrt{m_n} \text{ si } j \notin S_{\lambda_n} \end{cases}$$

Nous voulons montrer qu'avec $\lambda_n \sim \sqrt{n \log(m_n^2 d)/m_n}$, $P(S_{\lambda_n} = S^*) \to 1$. Nous montrons plutôt que $P(S_{\lambda_n} \neq S^*) \to 0$. D'après la Proposition 6, il est facile de montrer que

$$P(S_{\lambda_n} \neq S^*) \leq P(\text{il existe } j \in S^*, \|\underline{\mathbf{b}}_\mathbf{j}^* - \tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}}\| \geq \|\underline{\mathbf{b}}_\mathbf{j}^*\|) + P(\text{il existe } j \notin S^*, \|{\mathbf{U_j}}^T(\underline{\mathbf{Y}} - \mathbf{U_{S^*}} \tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{S^*}})\| > \lambda_n \sqrt{m_n}).$$

Par la suite, nous montrons que si $\lambda_n \sim \sqrt{n \log(m_n^2 d)/m_n}$, alors

$$P(\text{il existe } j \in S^*, \|\underline{\mathbf{b}}_\mathbf{j}^* - \tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}}\| \geq \|\underline{\mathbf{b}}_\mathbf{j}^*\|) \to 0 \text{ et } P(\text{il existe } j \notin S^*, \|{\mathbf{U_j}}^T(\underline{\mathbf{Y}} - \mathbf{U_{S^*}} \tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{S^*}})\| > \lambda_n \sqrt{m_n}) \to 0.$$

Lemme 6

$$P(\textit{il existe } j \in S^*, \|\underline{\mathbf{b}}_\mathbf{j}^* - \tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}}\| \geq \|\underline{\mathbf{b}}_\mathbf{j}^*\|) \to 0.$$

Ce lemme, démontré en annexe B, permet d'affirmer qu'avec une probabilité tendant vers 1, il n'y a pas de faux négatifs. Les résultats de Stone sont nécessaires pour montrer ce résultat.

Lorsque $\lambda_n \sim \sqrt{n \log(m_n^2 d)/m_n}$, il n'y a pas de faux négatifs avec une probabilité tendant vers 1 pour le Group LASSO. Nous travaillons ensuite sur l'absence de faux positifs. Pour montrer ce résultat, nous supposons que pour tout $j \neq j'$, $\|{\mathbf{U_j}}^T \mathbf{U_{j'}}\| = o_P(\frac{n}{m_n})$. Les hypothèses de ce type sont classiques en sélection de variables. Elles signifient que la matrice de plan d'expériences satisfait une hypothèse d'isométrie restreinte. Cette propriété requiert essentiellement que chaque couple ${\mathbf{U_j}}^T \mathbf{U_{j'}}$ soit semblable à un système orthonormal. Nous pouvons facilement trouver des exemples où cette hypothèse est vérifiée (par exemple, dans le cas de covariables indépendantes et suivant une loi uniforme).

Le Lemme 7 montre qu'il ne peut exister de faux positifs sous nos hypothèses avec une probabilité tendant vers 1 lorsque $\lambda_n \sim \sqrt{n \log(m_n^2 d)/m_n}$.

Lemme 7

$$P(\textit{il existe } j \notin S^*, \|{\mathbf{U_j}}^T(\underline{\mathbf{Y}} - \mathbf{U_{S^*}} \tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{S^*}})\| > \lambda_n \sqrt{m_n}) \to 0.$$

La première étape de la démonstration consiste à se ramener au cas des EMCO, ce qui introduit une première source d'erreur, que l'hypothèse qui suppose que pour tout $j \neq j'$, $\|{\mathbf{U_j}}^T \mathbf{U_{j'}}\| = o_P(\frac{n}{m_n})$ permet de contrôler par un terme négligeable par rapport à $\lambda_n \sqrt{m_n}$. Ensuite, nous cherchons à majorer par trois suites négligeables par rapport à $\lambda_n \sqrt{m_n}$ les trois sources d'erreurs suivantes :

- celles dues aux résidus du modèle ;
- celles dues à l'approximation du modèle dans des bases de B-Splines ;

– celles dues aux erreurs commises par l'EMCO pour estimer les coefficients approchant le modèle.

Nous avons exhibé un λ_n tel que, sous nos hypothèses, le Group LASSO soit consistant en sélection.

En annexe B, nous montrons la Proposition 7, qui permet de montrer $\|\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}\|$ est bornée par $O(\sqrt{m_n \log(m_n^2 d)})$. La démonstration se fait principalement grâce au Lemme 3 de [10]. La Proposition 8 montrée en annexe B permet de donner les expressions de $\tilde{\mathbf{b}}_{\lambda_\mathbf{n}} - \underline{\mathbf{b}}^*_{\mathbf{S}^*}$ et $\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\tilde{\mathbf{b}}_{\lambda_\mathbf{n}}$ en fonction de $\underline{\mathbf{r}}$, de $\underline{\boldsymbol{\epsilon}}$, du plan d'expériences et du biais d'approximation dans des bases de B-Splines.

La suite de la démonstration consiste à montrer que son BIC pénalisé est minimal avec une probabilité tendant vers 1. Nous montrons que la différence des erreurs quadratiques de l'estimateur associé au Group LASSO et de l'estimateur par MCO lorsque le vrai plan d'expériences a été sélectionné peut être majoré. Nous montrons d'abord, grâce à des développements classiques de la norme euclidienne, que cette différence est dépendante de $\underline{\mathbf{r}}$ lorsque le vrai plan d'expériences est sélectionné.

$$\begin{aligned}&\|\underline{\mathbf{Y}} - \mathbf{U}\underline{\tilde{\mathbf{b}}}_{\lambda_\mathbf{n}}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}_{\lambda_\mathbf{n}}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}_{\lambda_\mathbf{n}}}\|^2\\ &= \|\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}\|^2 - 2(\underline{\mathbf{Y}} - \Pi_{S^*}\underline{\mathbf{Y}})^T(\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}})\end{aligned}$$

Le Lemme 8 permet de montrer que cette différence ne croit pas trop vite.

Lemme 8

$$\|\underline{\mathbf{Y}} - \mathbf{U}\underline{\hat{\mathbf{b}}}_{\lambda_\mathbf{n}}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}_{\lambda_\mathbf{n}}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}_{\lambda_\mathbf{n}}}\|^2 = o_P(m_n\sqrt{\log(m_n^2 d)\log(nd)})$$

Maintenant que nous avons pu majorer la différence entre l'erreur de l'estimateur Group LASSO et de l'EMCO lorsque le vrai plan d'expériences est sélectionné, nous pouvons majorer la différence entre les deux BIC associés à ces estimateurs.

Nous avons $card(S_{\lambda_n}) = card(S^*)$ et donc

$$BIC(\lambda_n) - BIC(S^*) = \log(\|\underline{\mathbf{Y}} - \mathbf{U}\underline{\hat{\mathbf{b}}}_{\lambda_\mathbf{n}}\|^2) - \log(\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}_{\lambda_\mathbf{n}}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}_{\lambda_\mathbf{n}}}\|^2).$$

Le Lemme 9 permet de majorer $BIC(\lambda_n) - BIC(S^*)$.

Lemme 9

$$BIC(\lambda_n) - BIC(S^*) = o_P(m_n\frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n}) + o(o_P(m_n\frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n})),$$

et $m_n\frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n}$ *négligeable par rapport à* $\min(\log(2), v_n^2/m_n)$.

Nous suivons le même genre de raisonnement que dans la sous-section précédente pour démontrer le Lemme 9.

Nous avons maintenant tous les résultats nécessaires pour démontrer le théorème 4. Pour cela, nous utilisons un raisonnement similaire à Wang *et al.* [21] pour montrer que si $\lambda_n \sim \sqrt{n\log(m_n^2 d)/m_n}$, alors

$$P(BIC(\lambda) - BIC(\lambda_n) > 0) \to 1.$$

Nous venons ainsi d'énoncer les résultats permettant de montrer que le modèle choisi par le BIC pénalisé est consistant en sélection de variables. Il reste à démontrer l'optimalité en estimation de la procédure.

2.3.2 Consistance en estimation

Nous rappelons le Théorème 4.2 de Kato [12] utilisé pour démontrer l'optimalité de la procédure Post1Bic.

Théorème 5 *Sous les hypothèses 1 à 4 et 5.b, et avec des restrictions sur les bases utilisées à la première étape, si de plus* $(s^*)^2 m_n \log(\max(n,d))/n \to 0$*, si* $\max_{card(S)\leq s, \boldsymbol{\alpha}\in\mathbb{S}_T^{dm_n-1}} \|\hat{\boldsymbol{\Sigma}}^{\mathbf{1/2}}\mathbf{b}\| \leq O(1)$ *tel que* $s/s^* \to +\infty$*, avec* $\mathbb{S}^{dm_n-1}$ *la sphère unité sur* $\mathbb{R}^{dm_n}$ *et* $\mathbb{S}_T^{dm_n-1} = \mathbb{S}^{dm_n-1} \cap \{\boldsymbol{\alpha} \in \mathbb{R}^{dm_n} | \boldsymbol{\alpha}_{\boldsymbol{T}^c} = \mathbf{0}\}$*. Et si* $\max_{\boldsymbol{\alpha}\in\mathbb{S}^{dm_n-1}\cap\mathbb{C}} \|\hat{\boldsymbol{\Sigma}}^{\mathbf{1/2}}\mathbf{b}\| \geq O(1)$*, avec* $\mathbb{C} = \{\boldsymbol{\alpha} \in \mathbb{R}^{dm_n} | \sum_{j\in(S^{*,c})} \|\boldsymbol{\alpha}_j\| \leq 21 \sum_{j\in(S^*)} \|\boldsymbol{\alpha}_j\|\}$*. Si de plus* $\|g^*\| \leq O(1)s^*$*. Alors il existe une constante* A_f *telle que*

$$1/n\|\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \hat{\mathbf{f}}(\mathbf{X})\|^2 \leq_P \frac{s^* m_n}{n^2} A_f n(1 + \sqrt{\log(d/m_n)})^2,$$

où $\leq_P$ *signifie inférieur ou égale avec une probabilité 1.*

Les bases de B-Splines vérifient bien les conditions posées par Kato [12]. Or

$$\begin{aligned}
\frac{s^* m_n}{n^2} A_f n(1 + \sqrt{\log(d/m_n)})^2 &= A_s \frac{m_n}{n}(1 + \sqrt{\log(d/m_n)})^2 \text{ (avec } A_s \text{ constante)} \\
&= A_{s_2} \frac{(1 + \sqrt{\log(d/m_n)})^2}{n^{2\nu/(2\nu+1)}} \text{ (avec } A_{s_2} \text{ constante)} \\
&\to 0 \text{ (d'après l'hypothèse 5.b)}
\end{aligned}$$

L'optimalité de la procédure Post1Bic découle donc du Théorème 5 de Kato [12] qui montre que le modèle B-Splines pénalisé avec une pénalisation conduit à une estimation consistante de la fonction de régression avec une erreur empirique quadratique moyenne de l'ordre

$$O(1)\frac{s^* m_n}{n}(1 + \sqrt{\log(d/m_n)})^2 \to 0.$$

Cela justifie entièrement la procédure Post1Bic.

3 Simulations

L'objectif des simulations est d'illustrer les trois propriétés suivantes de nos procédures :

- la capacité à sélectionner les composantes ;
- la capacité à reproduire la forme des effets ;
- la capacité à prévoir.

Les questions subsidiaires à ces trois propriétés sont :

- L'intensité du bruit influence-t-elle la qualité de sélection et d'estimation des modèles ? Si oui, comment ?
- Comment le nombre d'observations influence-t-il la qualité de sélection et d'estimation des modèles ?
- Quel impact a la corrélation entre les variables explicatives ?

Les critères d'évaluation des modèles usuels sont présentés dans le tableau 1 :

Critère	Définition
MAPE	Mean Absolute Percentage Error ; $MAPE(f, \hat{f}) = 1/n \sum_{i=1}^{n} \lvert(f(x_i) - \hat{f}(x_i))/f(x_i)\rvert$
RMSE	Root Mean Square Error ; $RMSE(f, \hat{f}) = \sqrt{1/n \sum_{i=1}^{n}((f(x_i) - \hat{f}(x_i))^2)}$
MS	Médiane du nombre de composantes sélectionnées
MTZ	Médiane du nombre de composantes nulles parmi les vraies composantes nulles (vrais zéros)
MFZ	Médiane du nombre de composantes nulles parmi les vraies composantes non nulles (faux zéros)
MTP	Médiane du nombre de composantes non nulles parmi les vraies composantes non nulles (vrais positifs)
MFP	Médiane du nombre de composantes non nulles parmi les vraies composantes nulles (faux négatifs)
Median Curve 1	Courbe estimée correspondant à de la simulation associée au RMSE médian sur toutes les simulations pour une méthode donnée

Tableau 1 – Critère d'évaluation des modèles

MTZ permet de juger les méthodes sur leur capacité à bien seuiller à 0 les composantes nulles. MTP permet d'évaluer leur capacité à ne pas seuiller à 0 des composantes non nulles. MFP permet de quantifier le nombre de composantes que la méthode à conserver alors qu'elles étaient non influentes. Enfin, MFZ permet de quantifier le nombre de composantes que la méthode a seuillées à 0 alors qu'elles étaient influentes. Ceci peut être résumé par le tableau 2.

	Composantes réellement influentes	Composantes réellement nulles
Composantes conservées par la méthode	MTP	MFP
Composantes seuillées à 0 par la méthode	MFZ	MTZ

Tableau 2 – Critère d'évaluation des modèles

Le rapport signal sur bruit (SNR) permet de mesurer l'intensité du signal par rapport au bruit. Nous avons choisi d'utiliser un SNR empirique, qui dépend donc du plan d'expériences utilisé. Nous considérons le modèle suivant :

$$Y_i = \sum f_j(X_{ji}) + \epsilon_i, \; i = 1, \dots n,$$

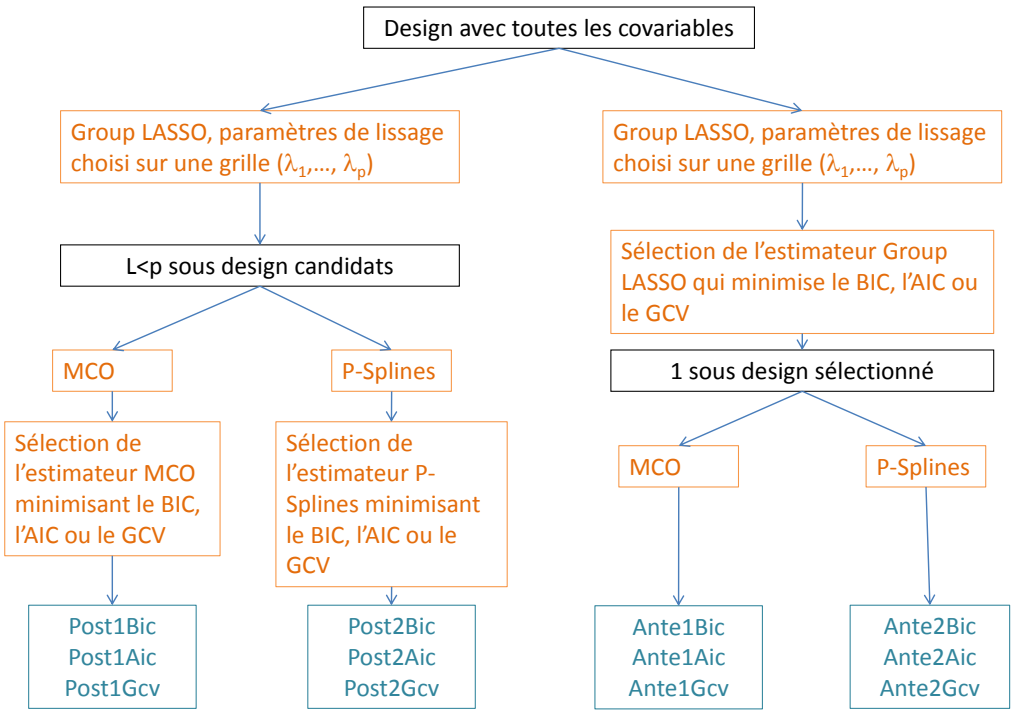
où ϵ est un bruit centré de variance σ^2.

Alors le SNR est défini par :

$$SNR = \sqrt{\frac{1/(n-1) \sum_{j=1}^{d} \sum_{i=1}^{n} \left(f_j(x_{ji}) - 1/n \sum_{i=1}^{n} f_j(x_{ji})\right)^2}{\sigma^2}} \tag{16}$$

Le SNR permet de comparer le signal non bruité avec le bruit. Plus le SNR est fort, moins le signal est bruité, et donc plus il est facile à estimer. Nous considérons qu'un SNR compris entre 2 et 3 correspond à un signal fortement bruité et un SNR supérieur à 4 correspond à un signal faiblement bruité.

Le graphique 1 rappelle les procédures définies en introduction.



Graphique 1 – Illustration des procédures type Post (gauche) et type Ante (droite)

Nous nous intéressons aussi au Group LASSO seul. Les méthodes de références sont des MCO et des P-Splines en utilisant l'ensemble des variables (influentes ou non) comme design. De plus, nous avons ajouté les résultats que nous obtenons lorsque nous utilisons la fonction *gam* du package *mgcv* avec les options select=T (GAMSelect) et avec "cs" comme lisseur (GAMShrinkage) (voir Wood [15]), qui consiste à modifier la pénalisation P-Splines. Elles permettent de réduire les coefficients associés à certaines variables non informatives, mais aucun coefficient n'est strictement nul. Nous avons alors décidé qu'une composante est nulle si la norme 2 de ces coefficients associés est inférieure à un certain seuil. Dans les faits, les normes 2 sont soit extrêmement faibles (de l'ordre de 10^{-15}), soit proche de 10^{-2}.

Le tableau 3 comporte le dictionnaire des méthodes utilisées. Nous donnons deux exemples pour expliquer au lecteur comment le tableau doit être lu.
Lorsque Ante2Bic est utilisé, *BIC(Group LASSO)+P-Spline* signifie que le plan d'expériences est sélectionné après l'étape Group LASSO grâce au BIC, le modèle étant ensuite réestimé à l'aide de P-Splines (méthode de type Ante).
Lorsque Post2Bic est utilisé, *BIC(Group LASSO+P-Spline)* signifie tous les plans d'expériences issus du Group LASSO sont estimés à l'aide de P-Splines et l'estimateur qui minimise le BIC est sélectionné (méthode de type Post).

Modèle	Méthode d'estimation
Ante1Bic	BIC(Group LASSO)+LASSO
Ante1Gcv	GCV(Group LASSO)+LASSO
Ante2Bic	BIC(Group LASSO)+P-Splines
Ante2Gcv	GCV(Group LASSO)+P-Splines
Post1Bic	BIC(Group LASSO+MCO)
Post1Aic	AIC(Group LASSO+MCO)
Post1Gcv	GCV(Group LASSO+MCO)
Post2Bic	BIC(Group LASSO+P-Spline)
Post1Aic	AIC(Group LASSO+P-Spline)
Post2Gcv	GCV(Group LASSO+P-Spline)
GrpLASSOBic	BIC(Group LASSO)
GrpLASSOGcv	GCV(Group LASSO)
MCO	MCO
PSP	P-Splines
GAMSelect	P-Splines modifiées 1
GAMShrinkage	P-Splines modifiées 2

Tableau 3 – Dictionnaire des méthodes utilisées

3.1 Description de la simulation

Cette simulation est inspirée de l'exemple de simulation de Lin et Zhang [14], et qui a aussi été utilisée par Cantoni et *al* [6] et Antoniadis *et al*. [3]. Il y a 10 variables explicatives dont 4 sont informatives. Les composantes de la simulation sont les suivantes :

$$
\begin{aligned}
f_1 &: \ x \mapsto 5x \\
f_2 &: \ x \mapsto 3(2x-1)^2 \\
f_3 &: \ x \mapsto \frac{4\sin(2\pi x)}{2-\sin(2\pi x)} \\
f_4 &: \ x \mapsto 6(0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^4(2\pi x)) \\
f_5 &: \ x \mapsto 0
\end{aligned}
$$

Les cinq fonctions sont $C^\infty[0,1]$.

Nous simulons différentes covariables dont nous faisons varier les corrélations. La simulation des covariables se fait selon le principe :

1. Générer $W_{i1}, \dots W_{i10}, U_i$ indépendamment d'une loi uniforme sur $[0,1]$, avec $i \in [\![1,n]\!]$
2. $X_{ij} = \frac{W_{ij}+tU_i}{1+t}$ avec $j \in [\![1,10]\!]$

Notons que t permet de contrôler la corrélation entre les variables. En effet, $Cor(X_{ij}, X_{il}) = \frac{t^2}{1+t^2}$ pour $j \neq l$.

Le modèle simulé est :

$$Y_i = f_1(X_{i1}) + f_2(X_{i2}) + f_3(X_{i3}) + f_4(X_{i4}) + \sum_{j=5}^{10} f_5(X_{ij}) + \epsilon_i,$$

où ϵ_i est un bruit gaussien centré.

Les variables explicatives sont projetées dans des bases de B-Splines de degré 3 et qui ont 10 noeuds placés au niveau des déciles des variables explicatives.

Nous simulons neuf variantes du modèle lorsque le nombre d'observations $n = 3000$. Nous prenons un nombre élevé d'observations parce que les applications concrètes que nous étudions

ont beaucoup d'observations. Nous considérons trois valeurs de variance du bruit (0.5, 1.5 et 3.5) et trois valeurs de corrélation ($t = 0$, 1 et 2, *i.e.* conduisant à une corrélation entre les variables explicatives soit 0 , 1/2 et 4/5 respectivement). Pour chaque fonction, nous définissons un $SNRf_i = \sqrt{\frac{1/(n-1)\sum_{i=1}^{n}\left(f_i(x_{ji})-1/n\sum_{i=1}^{n} f_i(x_{ji})\right)^2}{\sigma^2}}$, qui permet de mesurer l'influence de la composante considérée.

Simulation	SNR	$SNRf_1$	$SNRf_2$	$SNRf_3$	$SNRf_4$
Variance 0.5, t=0	5.29	2.04	1.25	2.59	3.94
Variance 1.5, t=0	3.09	1.17	0.73	1.49	2.32
Variance 3.5, t=0	2.02	0.78	0.48	0.97	1.51
Variance 0.5, t=1	5.42	1.44	0.85	2.55	4.49
Variance 1.5, t=1	3.12	0.84	0.48	1.49	2.56
Variance 3.5, t=1	2.05	0.55	0.32	0.97	1.69
Variance 0.5, t=2	5.38	1.54	0.83	2.67	4.33
Variance 1.5, t=2	3.10	0.88	0.48	1.53	2.50
Variance 3.5, t=2	2.04	0.58	0.32	1.01	1.64

Tableau 4 – SNR des différentes simulations

Le tableau 4 montre que les quatre composantes n'ont pas la même influence sur la variable à expliquer. La composante 4 est la plus influente, alors que la composante 2 est la moins influente. Cependant, toutes les composantes influentes semblent respecter l'hypothèse 5.d. L'intérêt d'introduire des variables ayant des influences d'intensités différentes est de vérifier que les méthodes parviennent à sélectionner toutes les composantes, qu'elles soient fortement comme faiblement influentes. Ajouter de la corrélation entre les variables explicatives introduit un changement dans la distribution de celles-ci par rapport à la loi uniforme. Les variables explicatives étant distribuées différemment, les $SNRf_i$ et donc indirectement les SNR évoluent aussi différemment selon les corrélations. Nous considérons que le signal est fortement bruité lorsque le SNR est compris entre 2 et 3. Le signal est moyennement bruité lorsque le SNR est compris entre 3 et 4. Il est faiblement bruité si le SNR est supérieur à 4.

3.2 Capacité de sélection de composantes

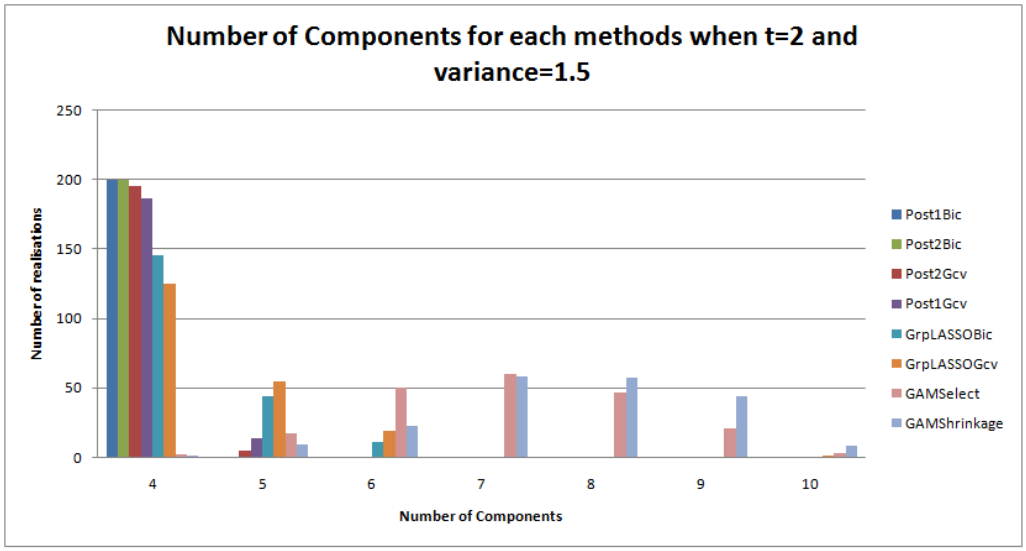
Le tableau 5 récapitule les critères d'évaluation des capacités de sélection dans le cas où la variance du bruit est de 1.5 et où t=2. Des tableaux semblables pour les huit autres scénarios sont donnés dans l'annexe C. Seul un des scénarios est présenté ici. Il s'agit du scénario le plus clivant si nous considérons à la fois les capacités de sélection et de prédiction. Si nous avions choisi un scénario avec une variance plus forte, les clivages auraient été plus forts en termes de prédiction, mais plus faibles en termes de capacité de sélection de variables.
Post1Aic et Post2AIC n'apparaissent pas dans le tableau 5 par soucis de lisibilité. Elles sont légèrement moins bonnes en termes de sélection de variable que respectivement Post1Gcv et Post2Gcv. De même, GAMSelect et GAMShrinkage n'apparaissent pas dans le tableau, car les deux procédures sont moins performantes.

Critère	Post1Bic	Post1Gcv	Post2Bic	Post2Gcv	GrpLASSOBic	GrpLASSOGcv	Idéal
MS	4(0)	4(0.26)	4(0)	4(0.16)	4(0.58)	4(0.77)	4
MTZ	6(0)	6(0.26)	6(0)	6(0.16)	6(0.58)	6(0.77)	6
MFZ	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0
MTP	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4
MFP	0(0)	0(0.26)	0(0)	0(0.16)	0(0.58)	0(0.77)	0

Tableau 5 – Capacité à sélectionner les composantes selon le critère utilisé avec une variance du bruit de 1.5 lorsque $t = 2$ (SNR=3.1)

Avec 3000 observations, il n'y a pas de faux négatifs quelque soit le SNR ou la corrélation entre les covariables. Le vrai modèle est à chaque fois sélectionné par les méthodes de type "Post" combiné au BIC, ce qui n'est pas le cas avec les autres procédures. La méthode "GrpLASSOGcv", et donc les méthodes de type "Ante" utilisant le GCV, est la plus sensible aux variations de SNR (à corrélation constante, plus celui-ci est faible, plus la méthode est proche de sélectionner à chaque fois le vrai modèle). Les méthodes de type "Ante" sont plus impactées par les variations de la corrélation entre les variables explicatives, que les méthodes de type "Post". Les tableaux 6 à 13 de l'annexe C montrent que lorsque le BIC est utilisé, les méthodes type "Post" sont plus performantes en termes de sélection de composantes que les méthodes type "Ante".

Dans tous les cas, nous avons constaté que l'ensemble des composantes influentes a été sélectionné. Sachant cela, le nombre de composantes sélectionnées est entièrement représentatif de la performance en termes de sélection de composantes. Le graphique 2 présente les nombres de composantes pour les 200 simulations des différentes méthodes lorsque la variance est de 1.5 et $t = 2$. Sur ces graphiques, les méthodes sont classées selon le nombre de fois où quatre composantes ont été sélectionnées, du plus grand au plus petit (et donc du plus efficace au moins efficace).



Graphique 2 – Nombre de composantes lorsque la variance du bruit est 1.5 et $t = 2$ (SNR=3.1) pour chaque méthode

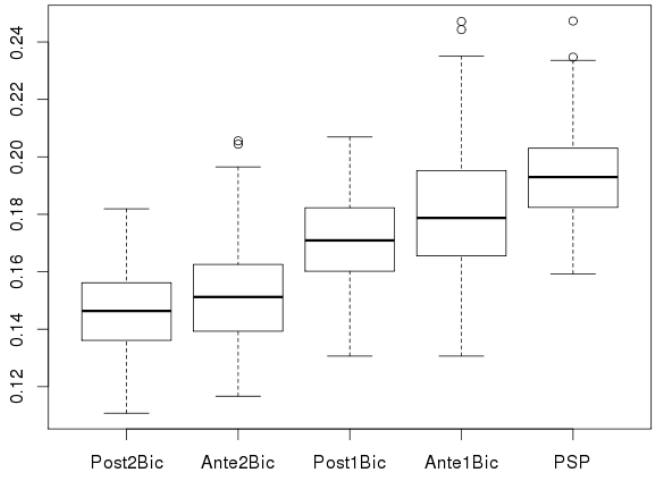
GAMSelect et GAMShrinkage sont les moins régulières dans le nombre de composantes sélectionnées. Sur les 200 simulations, Post1Bic et Post2Bic ont à chaque fois sélectionné le nombre idoine de composantes. Ce graphique illustre les différences de performances en termes de sélection de variables entre les méthodes de type "Post", et les méthodes de type "Ante", issues du Groupe LASSO. Nous notons que le GroupLASSOGcv a sélectionné une fois les dix covariables.

3.3 Capacité d'estimation et de prédiction

Les formes des courbes des effets sont relativement bien reproduites : par exemple, la parabole est estimée par une parabole. Cependant, l'échelle est moins bien respectée lorsque le Group LASSO est utilisé seul, à cause du biais introduit par le Group LASSO. Les étapes suivantes sont donc justifiées.

Le graphique 3 récapitule les boîtes à moustaches des RMSE en prédiction lorsque la variance du bruit est 1.5 et $t = 2$. On trouvera les mêmes graphiques sont représentés pour les huit autres scénarios dans l'annexe C. Dans cette annexe, il y a aussi un graphique de boîte à moustaches rassemblant toutes les méthodes pour le cas où la variance est 1.5 et $t = 2$. Nous n'avons représenté qu'une méthode de référence, qui est généralement le meilleur des modèles de références,

c'est-à-dire PSP. Nous avons quatre familles de procédures (Post1, Post2, Ante1 et Ante2). Pour chacune de ces procédures, nous avons gardé celle utilisant le BIC, afin d'avoir un représentant par famille et aussi parce que le BIC est dans la majorité des cas le critère amenant aux meilleures performances en termes de sélections de composantes à l'intérieur d'une famille de méthodes.



Graphique 3 – Boîtes à moustaches des RMSE lorsque la variance du bruit est 1.5 et $t = 2$ (SNR=3.1)

Le graphique 3 illustre l'importance de la sélection de variables pour la phase de prédiction. En effet, Post2Bic, utilisant la même méthode d'estimation qu'Ante2Bic et PSP, mais plus performante en termes de sélection de composantes, a des meilleures performances en termes de prédiction. De même, Post1Bic est plus performante en termes de prédiction qu'Ante1Bic. Sur ce jeu de simulation, la régularisation est importante, comme l'illustre le fait que Post2Bic et Ante2Bic soient plus performantes que Post1Bic et Ante1Bic.

Lorsque les variables ne sont pas corrélées, quelle que soit la variance, lorsque la médiane des RMSE est considérée, quatre groupes de méthodes ayant des médianes de RMSE proche se dégagent. Par la suite, nous décrivons ces groupes par performances décroissantes. Le premier groupe est composé des méthodes ayant une phase de régularisation avec de la sélection de variables, le deuxième groupe des méthodes sans régularisation avec sélection de variables. Le troisième groupe est composé de la méthode de régularisation sans sélection de variables. Enfin, le quatrième groupe est composé des MCO sans sélection de variables et des méthodes utilisant du Group LASSO sans étape de correction de biais. Enfin, en cas de variance faible des bruits, les méthodes GAMSelect et GAMShrinkage ont une médiane des RMSE forte (à un niveau équivalent du MCO sans sélection de variables). Par contre, lorsque la variance des bruits est forte, les méthodes GAMSelect et GAMShrinkage sont moins dégradées en prédiction que les autres procédures, tout en restant moins efficaces que les méthodes utilisant de la régularisation avec sélection de variables. Il y a une augmentation attendue des RMSE lorsque la variance des bruits augmente. Les méthodes avec régularisation et avec sélection de variables sont les plus performantes en termes de prédiction.

Lorsque les covariables sont corrélées, les mêmes quatre groupes se dégagent en termes de RMSE. La méthode Post2Bic devient à chaque fois meilleure que la méthode Ante2Bic, exceptée lorsque la variance des bruits et la corrélation entre les covariables sont faibles. Par rapport aux autres méthodes, les performances de GAMSelect et GAMShrinkage sont plus dégradées lorsque les covariables sont corrélées. Il peut arriver qu'utiliser un critère de type GCV échoue en termes

de RMSE, à tel point que Ante2Gcv est moins bon en termes de RMSE que Ante1Bic et Post1Bic lorsque $t = 2$ et que la variance des bruits est de 0.5. Avec une corrélation des covariables plus forte ou une variance des bruits plus élevée, les différences entre les différentes méthodes augmentent. Sur 9 variantes de la simulation, la méthode Post2Bic conduit dans 7 cas à une médiane des RMSE la plus faible parmi toutes les méthodes étudiées.

4 Conclusion

En guise de conclusion, nous donnons ici une synthèse de la contribution.

L'intérêt des estimateurs en plusieurs étapes est illustré par à une simulation de Meier [16] qui montre que l'estimateur Group LASSO oracle (oracle dans le sens que l'estimateur Group LASSO est estimé sur le plan d'expériences comportant uniquement les variables influentes) a des meilleurs performances prédictives que l'estimateur Group LASSO estimé sur un plan d'expériences comportant des faux positifs, et ce même si le plan d'expériences sélectionné par l'estimateur Group LASSO ne comporte que les variables réellement influentes. Belloni *et al*. [5] partent du même constat et combinent, dans le cas des modèles linéaires, le LASSO avec les MCO dans une procédure de type Ante. Comme nous avions pu le constater sur des simulations, Belloni *et al*. [5] constatent une amélioration des performances en prédiction, qui s'explique par le fait que les MCO corrigent le biais introduit par le LASSO.

En résumé, ce rapport est consacré à la démonstration de la consistance en sélection et en estimation de la procédure Post1Bic. Une des difficultés de la démonstration provient du fait que nous travaillons sur une approximation du modèle additif. En effet, nous utilisons des bases de B-Splines pour approcher les composantes additives du modèle et cette approximation introduit un terme d'erreur qu'il faut contrôler.

Après avoir posé quelques notations, nous donnons les hypothèses nécessaires à la démonstration de la consistance de Post1Bic. Certaines hypothèses sont très classiques dans le cadre de la régression et concernent le lissage des composantes, les observations et les erreurs. Nous posons aussi une hypothèse sur le nombre de noeuds, qui ne peut pas croître plus vite que le nombre des observations. Cette hypothèse est faite pour que l'erreur due à l'approximation du modèle dans des bases de B-Splines puissent être contrôlée. Nous autorisons le nombre de variables explicatives candidates à croître en fonction du nombre d'observations. Cependant, le nombre de variables explicatives candidates ne peut croître qu'à une vitesse moindre qu'une vitesse exponentielle. Cette hypothèse est nécessaire pour que la pénalisation associé au BIC ne soit pas trop forte. Le nombre de covariables influentes est constant, par souci d'identifiabilité. La norme 2 des composantes influentes est minorée, ce qui est nécessaire pour la consistance en sélection. Contrairement à Huang *et al*. [10], nous ne supposons pas que la norme 2 des composantes influentes est strictement supérieure à une constante strictement positive indépendante du nombre d'observations.

La démonstration de la consistance en sélection se fait en deux étapes. Tout d'abord, nous prouvons que, sous nos hypothèses, le BIC est consistant en sélection lorsque des estimateurs MCO sont utilisés pour estimer le modèle additif. Une difficulté de cette partie de la démonstration résulte du fait qu'il faut contrôler le coût du à l'approximation du modèle additif dans des bases de B-Splines. La seconde étape de la démonstration de la consistance est de montrer que lorsque le coefficient de lissage du Group LASSO a une bonne forme, que nous donnons, le BIC pénalisé est suffisamment proche du BIC associé à l'EMCO, ce qui nous permet, en utilisant le définition de l'estimateur MCO de conclure quand à la consistance en sélection lorsque le BIC est utilisé.

A Démonstration pour la consistance asymptotique en sélection du BIC appliqué à l'EMCO

A.1 Cas où $S^* \not\subseteq S$

Démonstration Lemme 1 *La démonstration du lemme est en deux parties. D'abord, nous minorons* $\|\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})\|$*, puis nous utilisons un résultat de Huang* et al. *([10]).*

Rappelons que nous sommes dans le cas où $S^* \not\subseteq S$ *et que* $\bar{S} = S^* \cup S$*. Nécessairement, au moins une composante non nulle de* $\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}$ *est estimée comme nulle dans* $\underline{\tilde{\mathbf{b}}}_{\mathbf{S}}$*.*
Nous avons donc nécessairement

$$\|\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\| \geq \min_{j \in S^*} \|\underline{\mathbf{b}}_{\mathbf{S}^*,\mathbf{j}}{}^*\|.$$

De plus, $\sup_{x\in[0,1]} |\sum_{k=1}^{m_n} \underline{\mathbf{B}}^2_{j,k}(x)|^{1/2} = O(m_n^{1/2})$.
Pour tout $j \in S^*$*,*

$$\|f_j^*\|_2 - \|\sum_{k=1}^{m_n} b^*_{jk}\tilde{B}_{j,k}\|_2 \leq \|f_j^* - \sum_{k=1}^{m_n} b^*_{jk}\tilde{B}_{j,k}\|_2$$
$$\leq A_4 m_n^{-\nu} \text{ (grâce à Stone [18] et parce que nous travaillons sur un compact)}$$

Nous avons donc

$$\|\sum_{k=1}^{m_n} b^*_{jk}\tilde{B}_{j,k}\|_2 \geq \|f_j^*\|_2 - A_4 m_n^{-\nu}.$$

Or

$$\|\mathbf{b}^*_{\mathbf{j}}\|_2 \sim \sqrt{m_n}\|\sum_{k=1}^{m_n} b^*_{jk}\tilde{B}_{j,k}\|_2,$$

car les B-Splines sont uniformément bornées et évaluées en m_n *noeuds.*
D'où, en utilisant les propriétés d'approximation des B-Splines, nous pouvons conclure que :

$$\min_{j\in S^*} \|f_j^*\|_2 \leq A_5\left(\frac{1}{\sqrt{m_n}} \min_{j\in S^*} \|\underline{\mathbf{b}}^*_{\mathbf{j}}\|_2 + m_n^{-\nu+1/2}\right).$$

D'où

$$\min_{j\in S^*} \|\underline{\mathbf{b}}^*_{\mathbf{j}}\|_2 \geq A_6\left(\sqrt{m_n} \min_{j\in S^*} \|f_j^*\|_2 - m_n^{-\nu+1/2}\right).$$

Donc uniformément, pour tout modèle $S \not\supseteq S^*$*, nous avons*

$$\|\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}\| \geq \|\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\| - \|\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}\|$$
$$\geq A_8\left(\sqrt{m_n} \min_{j\in S^*} \|f_j^*\| - \frac{m_n}{\sqrt{n}} - m_n^{-\nu+1/2}\right)$$

Nous utilisons le lemme 3 de Huang et al. *([10]).*
Ici, $m_n = O(n^{\frac{1}{2\nu+1}})$*, avec* $\nu \geq 2$*, d'où* $0 < \frac{1}{2\nu+1} < 0.5$*. De plus, le cardinal de* $\bar{S}$ *est inférieur ou égale à 2D, qui est une constante indépendante de n et d. De plus, avec nos hypothèses 1 et 3, nous respectons les conditions (A3) et (A4) de l'article de Huang* et al. *([10]).*
Alors, avec une probabilité tendant vers 1, il existe une constante c_1 *telle que*

$$\frac{nc_1}{m_n} \leq \rho_{\min}(\mathbf{U}_{\bar{\mathbf{S}}}{}^T\mathbf{U}_{\bar{\mathbf{S}}}),$$

où $\rho_{\min}(\mathbf{U}_{\bar{\mathbf{S}}}{}^T\mathbf{U}_{\bar{\mathbf{S}}})$ *est la valeur propre minimale de* $(\mathbf{U}_{\bar{\mathbf{S}}}{}^T\mathbf{U}_{\bar{\mathbf{S}}})$*.*
Nous avons donc

$$\|\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})\|^2 = (\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})^T\mathbf{U}_{\bar{\mathbf{S}}}{}^T\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}) \geq A_2\frac{n}{m_n}v_n^2.$$

□

Démonstration Lemme 2 *Pour démontrer le Lemme 2, nous utilisons d'abord l'inégalité de Cauchy Schwarz, puis la formule de l'erreur d'estimation de la régression additive et le Lemme 3 de Huang* et al. *([10]).*

D'après l'inégalité de Cauchy Schwarz, nous avons :

$$|2\left(\mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{f}}^*(\underline{\mathbf{X}})\right)^T \mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})| \leq 2\|\left(\mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{f}}^*(\underline{\mathbf{X}})\right)^T\|\|\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})\|.$$

Etudions chacun des deux termes :

– *Etude de* $\|\left(\mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{f}}^*(\underline{\mathbf{X}})\right)^T\|$ *:*

$$\begin{aligned}\|\left(\mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{f}}^*(\underline{\mathbf{X}})\right)^T\|^2 &\leq \|\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{b}}^*_{\mathbf{S}^*})\|^2 + \sum_{j\in S^*}\|f_j^* - \sum_{k=1}^{m_n} b^*_{jk}\tilde{B}_{j,k}\|^2 \\ &\leq C\left(m_n + nm_n^{-2\nu} + 2\sqrt{n}m_n^{-\nu+1/2} + m_n^{-2\nu}\right)\end{aligned}$$

d'où

$$\|\left(\mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{f}}^*(\underline{\mathbf{X}})\right)^T\|^2 \leq C(m_n + nm_n^{-2\nu} + o(m_n)),$$

et donc

$$\|\left(\mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{f}}^*(\underline{\mathbf{X}})\right)^T\| = O_P\left((m_n + nm_n^{-2\nu})^{1/2}\right).$$

– *Etude de* $\|\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})\|$ *:*
D'après le Lemme 3 de [10], il existe une constante c_2 *telle que*

$$\frac{nc_2}{m_n} \geq \rho_{\max}(\mathbf{U}_{\bar{\mathbf{S}}}{}^T\mathbf{U}_{\bar{\mathbf{S}}}),$$

où $\rho_{\max}(\mathbf{U}_{\bar{\mathbf{S}}}{}^T\mathbf{U}_{\bar{\mathbf{S}}})$ *est la valeur propre maximale de* $(\mathbf{U}_{\bar{\mathbf{S}}}{}^T\mathbf{U}_{\bar{\mathbf{S}}})$.
Nous avons donc

$$\|\mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})\| \leq \sqrt{\frac{nc_2}{m_n}}|v_n|.$$

Nous avons donc

$$|2\left(\mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{f}}^*(\underline{\mathbf{X}})\right)^T \mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*})| \leq C_4\sqrt{nm_n^{-2\nu}}\sqrt{\frac{n}{m_n}}|v_n|,$$

avec $C_4 > 0$ *constante.*
Nous pouvons en conclure que

$$-C_4\sqrt{nm_n^{-2\nu}}\sqrt{\frac{n}{m_n}}|v_n| \leq 2\left(\mathbf{U}_{\bar{\mathbf{S}}}\underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{f}}^*(\underline{\mathbf{X}})\right)^T \mathbf{U}_{\bar{\mathbf{S}}}(\underline{\tilde{\mathbf{b}}}_{\mathbf{S}} - \underline{\tilde{\mathbf{b}}}_{\mathbf{S}^*}) \leq C_4\sqrt{nm_n^{-2\nu}}\sqrt{\frac{n}{m_n}}|v_n|.$$

Il nous reste à montrer que ce terme est bien d'un ordre inférieur au troisième terme du membre de droite de l'équation (9).

$$\begin{aligned}\frac{\sqrt{nm_n^{-2\nu}}\sqrt{\frac{n}{m_n}}|v_n|}{\frac{n}{m_n}v_n^2} &= \frac{m_n^{-\nu+1-1/2}}{|v_n|} \\ &= C_5\frac{n^{\frac{1/2-\nu}{2\nu+1/2}}}{|v_n|},\end{aligned}$$

avec $C_5 > 0$ *constante.*

Calculons $n^{-\frac{1/2-\nu}{2\nu+1/2}}|v_n|$, *et en notant* $C_{10} > 0$ *une constante :*

$$\begin{aligned} n^{-\frac{1/2-\nu}{2\nu+1/2}}|v_n| &\geq C_{10}\left(n^{-\frac{1/2-\nu}{2\nu+1}} n^{1/2\frac{1}{2\nu+1}} \min_{j\in S^*} \|f_j^*\|_2 - 2n^{(-\nu+1/2)/(2\nu+1)} n^{-\frac{1/2-\nu}{2\nu+1}}\right) \\ &>> C_{10}\left(\sqrt{\log(nd)} - 2n^{\frac{-\nu}{2\nu+1}}\right) \end{aligned}$$

Or $\sqrt{\log(nd)} \to_{n\to+\infty} \infty$ *et* $n^{\frac{-\nu}{2\nu+1}} \to_{n\to+\infty} 0$. *Donc*

$$n^{-\frac{1/2-\nu}{2\nu+1/2}}|v_n| >> C_{10}\left(\sqrt{\log(nd)} - 2n^{\frac{-\nu}{2\nu+1}}\right) \to_{n\to+\infty} +\infty,$$

et donc

$$\frac{\sqrt{nm_n^{-2\nu}}\sqrt{\frac{n}{m_n}}|v_n|}{\frac{n}{m_n}v_n^2} \to_{n\to+\infty} 0.$$

D'où asymptotiquement, le second terme de l'équation (9) est négligeable devant le troisième terme de (9).

□

Proposition 1 *Pour tout* $\bar{S}$ *tel que* $S^* \subseteq \bar{S}$, *nous avons*

$$O_p(\|\Pi_{\bar{S}} B\|^2) = O_p(\|B\|^2) = O_p(nm_n^{-2\nu}). \tag{17}$$

Proposition 2 *Soit* M *constante telle que* $M < n$ *et* $\bar{S}$ *tel que* $S^* \subseteq \bar{S}$, *alors*

$$\sup_{\bar{S}|card(\bar{S})<M} \underline{\boldsymbol{\epsilon}}^T \Pi_{\bar{S}} \underline{\boldsymbol{\epsilon}} = o_P\left(m_n \log(nd)\right). \tag{18}$$

De plus, $m_n \log(nd)$ *est d'un ordre plus petit que* $\frac{n}{m_n}v_n^2$.

Les deux propositions précédentes permettent de montrer que, lorsque le vrai plan d'expériences S^* est inclus dans $\bar{S}$, alors $\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\hat{\mathbf{b}}}_{\bar{\mathbf{S}}}\|^2/n - \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\|^2/n$ tend vers 0 avec une probabilité tendant vers 1, c'est-à-dire :

Proposition 3 *Soit* $\bar{S}$ *tel que* $S^* \subseteq \bar{S}$,
alors,

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\hat{\mathbf{b}}}_{\bar{\mathbf{S}}}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\|^2 = O_p(nm_n^{-2\nu}) + o_p(m_n \log(nd)). \tag{19}$$

Montrons les trois propositions dans l'ordre.

Démonstration Proposition 1 *Nous avons* $\Pi_{\bar{S}}\mathbf{U}_{\bar{\mathbf{S}}} = \mathbf{U}_{\bar{\mathbf{S}}}, \Pi_{\bar{S}}^T\Pi_{\bar{S}} = \Pi_{\bar{S}}^T, (\Pi_{\bar{S}}^T\Pi_{\bar{S}})^T = \Pi_{\bar{S}}^T\Pi_{\bar{S}}$ *et* $(\Pi_{\bar{S}}^T\Pi_{\bar{S}})^T = (\Pi_{\bar{S}}^T)^T = \Pi_{\bar{S}}$. *Nous avons donc* $\Pi_{\bar{S}} = (\Pi_{\bar{S}}^T\Pi_{\bar{S}})^T = \Pi_{\bar{S}}^T\Pi_{\bar{S}} = \Pi_{\bar{S}}^T$. *Enfin,* $(I - \Pi_{\bar{S}})$ *est un projecteur.*

$$\begin{aligned} \|\Pi_{\bar{S}} B\|^2 &= B^T\Pi_{\bar{S}}^T\Pi_{\bar{S}} B \\ &= (\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}})^T\Pi_{\bar{S}}(\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}) \\ &= \underline{\mathbf{f}}^*(\underline{\mathbf{X}})^T(\Pi_{\bar{S}} - I)\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) + \underline{\mathbf{f}}^*(\underline{\mathbf{X}})^T\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \underline{\mathbf{b}}^{*T}_{\bar{\mathbf{S}}}\mathbf{U}_{\bar{\mathbf{S}}}^T\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \underline{\mathbf{f}}^*(\underline{\mathbf{X}})^T\mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}} + \underline{\mathbf{b}}^{*T}_{\bar{\mathbf{S}}}\mathbf{U}_{\bar{\mathbf{S}}}^T\mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}} \\ &= -\underline{\mathbf{f}}^*(\underline{\mathbf{X}})^T(I - \Pi_{\bar{S}})\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) + \|B\|^2 \\ &= -\underline{\mathbf{f}}^*(\underline{\mathbf{X}})^T(I - \Pi_{\bar{S}})^T(I - \Pi_{\bar{S}})\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) + \|B\|^2 \\ &= \|B\|^2 - \|(I - \Pi_{\bar{S}})\underline{\mathbf{f}}^*(\underline{\mathbf{X}})\|^2 \\ &\leq \|B\|^2 \end{aligned}$$

D'où

$$\|\Pi_{\bar{S}} B\|^2 = O_p(\|B\|^2).$$

De plus,

$$\begin{aligned}
\|\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\|^2 &= \|\sum_{j\in\bar{S}} \underline{\mathbf{f}}^*_{\mathbf{j}}(\underline{\mathbf{X}}^{\mathbf{j}}) - \sum_{j\in\bar{S}}\sum_{k=1}^{m_n} b_{j,k}(B_{jk}(x_{ij}) - \bar{B}_{jk})\|^2 \\
&\leq \sum_{j\in\bar{S}} \|\underline{\mathbf{f}_{\mathbf{j}}}^*(\underline{\mathbf{X}}^{\mathbf{j}}) - \sum_{k=1}^{m_n} b_{j,k}(B_{jk} - \bar{B}_{jk})\|^2 \\
&\leq \sum_{j\in\bar{S}} n\|f_j^* - \sum_{k=1}^{m_n} b_{j,k}(B_{jk} - \bar{B}_{jk})\|^2_{+\infty} \\
&\leq \sum_{j\in\bar{S}} n m_n^{-2\nu} \\
&\leq D_1 n m_n^{-2\nu} \text{ avec } D_1 \text{ constante car } s^* \text{ supposé constant}
\end{aligned}$$

□

Démonstration Proposition 2 *Nous avons supposé que le vecteur des erreurs était gaussien centré, indépendant et avec des composantes de variance bornée par σ^2 p.s. (voir l'hypothèse 2). $\|\Pi_{\bar{S}}\underline{\boldsymbol{\epsilon}}\|^2 = \underline{\boldsymbol{\epsilon}}^T\Pi_{\bar{S}}\underline{\boldsymbol{\epsilon}}$ est donc une forme quadratique d'un vecteur gaussien. La dimension de l'espace propre de $\Pi_{\bar{S}}$ est au plus M avec M fini. Cette forme quadratique peut donc s'exprimer comme une somme de variables du Chi-deux indépendantes.*
Calculons l'espérance de $\underline{\boldsymbol{\epsilon}}^T\Pi_{\bar{S}}\underline{\boldsymbol{\epsilon}}$.

$$\begin{aligned}
E(\underline{\boldsymbol{\epsilon}}^T\Pi_{\bar{S}}\underline{\boldsymbol{\epsilon}}) &= E\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(\Pi_{\bar{S}})_{ij}\epsilon_j\epsilon_i\right) \\
&\leq M\sigma^2
\end{aligned}$$

Grâce à l'inégalité de Markov, nous pouvons en conclure que

$$\begin{aligned}
P\left(\sup_{\bar{S}|card(\bar{S})\leq M} \underline{\boldsymbol{\epsilon}}^T\Pi_{\bar{S}}\underline{\boldsymbol{\epsilon}} \geq m_n\log(nd)\right) &\leq \frac{M\sigma^2}{m_n\log(nd)} \\
&\to_{n\to+\infty} 0
\end{aligned}$$

Montrons que $m_n\log(nd)$ est d'un ordre plus petit ou égal que $\frac{n}{m_n}v_n^2$

$$\frac{m_n\log(nd)}{\frac{n}{m_n}\left(\sqrt{m_n}\min\|f_j^*\|_2 - 2m_n^{-\nu+1/2}\right)^2} = \frac{m_n}{n}\frac{\log(nd)}{(\min\|f_j^*\|_2)^2}\frac{1}{\left(1 - 2\frac{m_n^{-\nu}}{\min\|f_j^*\|_2}\right)^2}$$

or

$$\frac{m_n^{-\nu}}{\min\|f_j^*\|_2} < G_1\frac{m_n^{-\nu}}{\sqrt{m_n^{-2\nu}\log(nd)}} \to_{n\to+\infty} 0 \text{ (d'après l'hypothèse 5 et } G_1 > 0 \text{ constante).}$$

D'où

$$\frac{1}{\left(1 - 2\frac{m_n^{-\nu}}{\min\|f_j^*\|_2}\right)^2} \to_{n\to+\infty} 1.$$

De plus, $\frac{\log(nd)}{(\min\|f_j^*\|_2)^2} < \frac{\log(nd)}{m_n^{-2\nu}\log(nd)} < m_n^{2\nu}$
Finalement,

$$\frac{m_n\log(nd)}{\frac{n}{m_n}\left(\sqrt{m_n}\min\|f_j^*\|_2 - 2m_n^{-\nu+1/2}\right)^2} < \frac{m_n^{2\nu+1}}{n} = A.$$

D'où $m_n\log(nd) = O(\frac{n}{m_n}v_n^2)$
D'où $\sup_{\bar{S}|card(\bar{S})}\underline{\boldsymbol{\epsilon}}^T\Pi_{\bar{S}}\underline{\boldsymbol{\epsilon}} = o_P(\frac{n}{m_n}v_n^2)$.
□

Démonstration Proposition 3

$$\begin{aligned}\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\hat{\underline{\mathbf{b}}}_{\bar{\mathbf{S}}}\|^2 &= \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}(\mathbf{U}_{\bar{\mathbf{S}}}{}^T\mathbf{U}_{\bar{\mathbf{S}}})^{-1}\mathbf{U}_{\bar{\mathbf{S}}}{}^T\underline{\mathbf{Y}}\|^2\\ &= \|(I-\Pi_{\bar{S}})\underline{\mathbf{Y}}\|^2\end{aligned}$$

et

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\|^2 = \|B + \underline{\boldsymbol{\epsilon}}\|^2,$$

d'où

$$\begin{aligned}\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\hat{\underline{\mathbf{b}}}_{\bar{\mathbf{S}}}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\|^2 &= \|(I-\Pi_{\bar{S}})\underline{\mathbf{Y}}\|^2 - \|B+\underline{\boldsymbol{\epsilon}}\|^2\\ &= \|(I-\Pi_{\bar{S}})(\mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}} + B + \underline{\boldsymbol{\epsilon}})\|^2 - \|B+\underline{\boldsymbol{\epsilon}}\|^2\\ &= \|(I-\Pi_{\bar{S}})(B+\underline{\boldsymbol{\epsilon}})\|^2 - \|B+\underline{\boldsymbol{\epsilon}}\|^2\\ &= (B+\underline{\boldsymbol{\epsilon}})^T(I-\Pi_{\bar{S}})^T(I-\Pi_{\bar{S}})(B+\underline{\boldsymbol{\epsilon}}) - (B+\underline{\boldsymbol{\epsilon}})^T(B+\underline{\boldsymbol{\epsilon}})\\ &= -(B+\underline{\boldsymbol{\epsilon}})^T\Pi_{\bar{S}}(B+\underline{\boldsymbol{\epsilon}})(car\ \Pi_{\bar{S}}^T\Pi_{\bar{S}} = \Pi_{\bar{S}}\ et\ \Pi_{\bar{S}}^T = \Pi_{\bar{S}})\\ &= -\|\Pi_{\bar{S}}(B+\underline{\boldsymbol{\epsilon}})\|^2\end{aligned}$$

Comme $\|X+Y\|^2 \le 2\|X\|^2 + 2\|Y\|^2$ *et* $\|X+Y\|^2 \ge 2\|X\|^2 - 2\|Y\|^2$ *, nous avons donc*

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\hat{\underline{\mathbf{b}}}_{\bar{\mathbf{S}}}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\|^2 \ge -2\left(\|\Pi_{\bar{S}}B\|^2 + \|\Pi_{\bar{S}}\underline{\boldsymbol{\epsilon}}\|^2\right),$$

et

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\hat{\underline{\mathbf{b}}}_{\bar{\mathbf{S}}}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\|^2 \le 2\left(\|\Pi_{\bar{S}}B\|^2 - \|\Pi_{\bar{S}}\underline{\boldsymbol{\epsilon}}\|^2\right).$$

Comme $S^* \subseteq \bar{S}$*, d'après les Propositions 1 et 2, nous avons*

$$O_p(\|\Pi_{\bar{S}}B\|^2) = O_p(\|B\|^2) = O_p(nm_n^{-2\nu})\ et \sup_{\bar{S}|card(\bar{S})<M}\underline{\boldsymbol{\epsilon}}^T\Pi_{\bar{S}}\underline{\boldsymbol{\epsilon}} = o_P\left(m_n\log(nd)\right).$$

□

Démonstration Lemme 3 *Puisque pour tout couple de réels* (a,b) *,* $-4a^2 - \frac{1}{4}b^2 \le -2ab$*, nous avons,*

$$-2\underline{\boldsymbol{\epsilon}}^T\mathbf{U}_{\bar{\mathbf{S}}}(\tilde{\underline{\mathbf{b}}}_{\mathbf{S}} - \tilde{\underline{\mathbf{b}}}_{\mathbf{S}^*}) \ge -4\underline{\boldsymbol{\epsilon}}^T\Pi_{\bar{S}}\Pi_{\bar{S}}\underline{\boldsymbol{\epsilon}} - \frac{1}{4}\|\mathbf{U}_{\bar{\mathbf{S}}}(\tilde{\underline{\mathbf{b}}}_{\mathbf{S}} - \tilde{\underline{\mathbf{b}}}_{\mathbf{S}^*})\|^2.$$

Donc

$$\begin{aligned}-2\underline{\boldsymbol{\epsilon}}^T\mathbf{U}_{\bar{\mathbf{S}}}(\tilde{\underline{\mathbf{b}}}_{\mathbf{S}} - \tilde{\underline{\mathbf{b}}}_{\mathbf{S}^*}) + \|\mathbf{U}_{\bar{\mathbf{S}}}(\tilde{\underline{\mathbf{b}}}_{\mathbf{S}} - \tilde{\underline{\mathbf{b}}}_{\mathbf{S}^*})\|^2 &\ge -4\underline{\boldsymbol{\epsilon}}^T\Pi_{\bar{S}}\Pi_{\bar{S}}\underline{\boldsymbol{\epsilon}} + \frac{3}{4}\|\mathbf{U}_{\bar{\mathbf{S}}}(\tilde{\underline{\mathbf{b}}}_{\mathbf{S}} - \tilde{\underline{\mathbf{b}}}_{\mathbf{S}^*})\|^2\\ &\ge \frac{3}{4}\frac{n}{m_n}v_n^2 + o_p(\frac{n}{m_n}v_n^2) \quad d'après\ le\ Lemme\ 2\end{aligned}$$

Ceci démontre le Lemme 3.
□

Démonstration Lemme 4 *Nous commençons par montrer le résultat pour*

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\mathbf{b}_{\mathbf{S}^*}\|^2/n.$$

Nous avons

$$\begin{aligned}\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\mathbf{b}_{\mathbf{S}^*}\|^2/n &= \|\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) + \underline{\boldsymbol{\epsilon}} - \mathbf{U}_{\mathbf{S}^*}\mathbf{b}_{\mathbf{S}^*}\|^2/n \quad \textit{par définition de } \underline{\mathbf{Y}} \\ &\geq \frac{\|\underline{\boldsymbol{\epsilon}}\|^2}{n} - \|\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U}_{\mathbf{S}^*}\mathbf{b}_{\mathbf{S}^*}\|^2/n \\ &\geq \frac{\|\underline{\boldsymbol{\epsilon}}\|^2}{n} + O_P(m_n^{-2\nu})\end{aligned}$$

Nous allons montrer que $\frac{\|\underline{\boldsymbol{\epsilon}}\|^2}{n}$ *est strictement supérieur à 0 avec une probabilité tendant vers 1. Pour cela, nous allons utiliser l'inégalité de Bienaymé-Tchebychev :*
Soit X une variable aléatoire de variance finie, alors pour tout $\alpha > 0$,

$$P(|X - E(X)| \geq \alpha) \leq \frac{V(X)}{\alpha}.$$

Nous considérons la variable aléatoire $\frac{1}{n}\sum \epsilon_i^2$. *D'après l'hypothèse 2,*

$$E(\frac{1}{n}\sum \epsilon_i^2) = \frac{1}{n}\sum \sigma^2(\underline{\mathbf{X_i}}),$$

et, par indépendance et comme ϵ_i *suit une loi normale centrée de variance* $\sigma^2(\underline{\mathbf{X_i}})$, *le moment d'ordre 4 de* ϵ_i *est* $3\sigma^4(X_i)$.
D'où

$$\begin{aligned}V(\frac{1}{n}\sum \epsilon_i^2) &= \frac{3}{n^2}\sum \sigma^4(X_i) \\ &\leq \frac{3}{n}\sigma^4 \quad \textit{d'après l'hypothèse 2}\end{aligned}$$

Finalement, d'après Bienaymé-Tchebychev, pour tout $\alpha > 0$,

$$P(|\frac{1}{n}\sum \epsilon_i^2 - \frac{1}{n}\sum \sigma^2(\underline{\mathbf{X_i}})| \geq \alpha) \leq \frac{3}{n}\frac{\sigma^4}{\alpha} \to 0,$$

d'où

$$\frac{1}{n}\sum \epsilon_i^2 \to^P \frac{1}{n}\sum \sigma^2(\underline{\mathbf{X_i}}) \geq \frac{n}{n}\sigma_1^2 \geq \sigma_1^2 > 0 \textit{ (d'après l'hypothèse 2).}$$

Donc, avec une probabilité tendant vers 1, $\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\mathbf{b}_{\mathbf{S}^*}\|^2/n$ *est bornée inférieurement par une constante strictement positive.*

Nous passons à l'étude de la suite $\inf_{S^* \subseteq S | card(S) \leq D} \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\mathbf{b}}_{\mathbf{S}}\|^2/n$
$S^* \subseteq S$, *d'où, d'après le Lemme 3,*

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\mathbf{b}}_{\mathbf{S}}\|^2/n = \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\|^2/n + O_p(m_n^{-2\nu}) + o_p(\frac{m_n}{n}\log(nd)) = \|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\|^2/n + o_P(1).$$

Or, par construction,

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\bar{\mathbf{S}}}\underline{\mathbf{b}}^*_{\bar{\mathbf{S}}}\|^2/n = \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}_{\mathbf{S}^*}\|^2/n \geq \sigma_1^2 + O_P(m_n^{-2\nu}),$$

d'où

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\mathbf{b}}_{\mathbf{S}}\|^2/n \geq \sigma_1^2 + o_P(1) > 0,$$

avec une probabilité tendant vers 1 dès que $S^* \subseteq S$.
D'où la conclusion du lemme.

□

Démonstration Théorème 1 *Nous avons*

$$BIC(S) - BIC(S^*) \geq \min(\frac{L_1}{m_n} v_n^2, \log(2)) + (card(S) - card(S^*))\frac{m_n \log(nd)}{n},$$

où $L_1 > 0$ *constante.*
Or, lorsque $n \to +\infty, L_1 m_n v_n^2 > 0$ *et*

$$\frac{m_n \log(nd)}{n} = A\frac{n^{1/(2\nu+1)}}{n}\log(nd) = An^{-2\nu/(2\nu+1)}(\log(n) + \log(d)).$$

Or, d'après les hypothèses 5, $n^{-2\nu/(2\nu+1)}\log(d) \to_{n\to+\infty} 0$
et il est évident que $n^{-2\nu/(2\nu+1)}\log(n) \to_{n\to+\infty} 0$.

De plus, $|card(S) - card(S^*)| \leq |M - D|$, *or M et D indépendantes de n et finies, d'où*

$$|card(S) - card(S^*)|\frac{m_n \log(nd)}{n} \to_{n\to+\infty} 0.$$

Et de plus,

$$|card(S) - card(S^*)|\frac{m_n \log(nd)}{n} = o(\min \|f_j^*\|^2).$$

D'où, avec probabilité 1, lorsque $n \to +\infty$, *pour tout* S *tel que* $S^* \nsubseteq S$,

$$BIC(S) - BIC(S^*) > 0.$$

D'où la conclusion du théorème.

□

A.2 Cas où $S^* \subseteq S$

Démonstration Lemme 5 *Commençons par l'étude de*

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2.$$

Tout d'abord, nous constatons que $\mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*} = \mathbf{U}_{\mathbf{S}}\underline{\mathbf{b}}^*_{\mathbf{S}}$, *et donc*

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*}\|^2 = \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\underline{\mathbf{b}}^*_{\mathbf{S}}\|^2.$$

De toute évidence, nous avons bien

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2 = \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\underline{\mathbf{b}}^*_{\mathbf{S}}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2 \geq 0,$$

par définition de l'estimateur MCO, et dans nos conditions, le programme des MCO est identifiable, nous avons égalité si, et seulement si, $\hat{\underline{\mathbf{b}}}_{\mathbf{S}} = \underline{\mathbf{b}}^*_{\mathbf{S}}$.

Or $S^* \subseteq S$, *donc d'après la Proposition 3,*

$$\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\underline{\mathbf{b}}^*_{\mathbf{S}}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2 = O_P(nm_n^{-2\nu}) + o_P(m_n \log(nd)).$$

Comme $S^* \subseteq S$, *nous pouvons appliquer le Lemme 4. Avec une probabilité tendant vers 1,* $\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2/n$ *tend vers une constante strictement supérieure à 0. Or*

$$O_P(m^{-2\nu}) + o_P(m_n \log(nd)/n) \to^P 0.$$

Donc, avec une probabilité tendant vers 1,

$$0 \leq \frac{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2}{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2} \to^P 0.$$

□

Démonstration Théorème 2 *D'après le Lemme 5, avec une probabilité tendant vers 1,*

$$\frac{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2}{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2} \to^P 0,$$

et

$$\frac{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2}{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2} = O_P(m^{-2\nu}) + o_P(m_n \log(nd)/n).$$

Or, si $x_n \to 0, \log(1 + x_n) = x_n + o(x_n)$ *pour n suffisamment grand.*

Comme

$$BIC(S^*) - BIC(S) \leq \log\left(1 + \frac{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2}{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}}\|^2}\right) + (card(S^*) - card(S))\, m_n \frac{\log(nd)}{n},$$

nous avons

$$\begin{aligned} BIC(S^*) - BIC(S) \leq O_P(m_n^{-2\nu}) + o_P(\frac{m_n}{n}\log(nd)) \\ + o(O_P(m_n^{-2\nu}) + o_P(m_n \log(nd)/n)) \\ + (card(S^*) - card(S))\, m_n \frac{\log(nd)}{n}. \end{aligned}$$

Or si $S^* \subsetneq S$*, alors*

$$(card(S^*) - card(S)) < 0,$$

car les cardinaux sont finis par hypothèse.

Montrons que $O_P(m_n^{-2\nu})$ *est négligeable par rapport à* $m_n \frac{\log(nd)}{n}$.

$$\begin{aligned} \frac{O_P(m_n^{-2\nu})}{m_n \frac{\log(nd)}{n}} &= \frac{O_P(n^{-2\nu/(2\nu+1)})}{n^{1/(2\nu+1)} \frac{\log(nd)}{n}} \\ &= \frac{O_P(1)}{\log(nd)} \\ &\underset{n\to+\infty}{\overset{P}{\to}} 0 \end{aligned}$$

donc

$$BIC(S^*) - BIC(S) < 0,$$

avec une probabilité tendant vers 1.

□

Démonstration Théorème 3 *D'après les Théorèmes 1 et 2, le BIC est minimum lorsque l'estimateur MCO est calculé sur* S^* *avec une probabilité tendant vers 1, lorsque le cardinal maximal des plans d'expériences candidats est inférieur à une constante indépendante de n.*

□

B Démonstration de la consistance en sélection de la procédure Post1

Proposition 4 *Notons*

$$\Omega_0 = \{\|\hat{\Sigma}_j^{1/2} - I_m\| \leq 0.5, 1 \leq j \leq d\}.$$

Supposons que

$$\frac{m_n \log(d)}{n} \to 0.$$

Alors $P(\Omega_0) \to 1$.
Et de plus,

$$\max_{1\le j\le d} \|\hat{\Sigma}_j^{1/2} - I_m\| \to^P 0,$$

et

$$\max_{1\le j\le d} \|\bar{\mathbf{B}}_{\mathbf{j}}\| \to^P 0.$$

Démonstration Proposition 4 *Le résultat découle de Kato 2011 [11] et 2012 [12] (Lemme B.1, page 40).*

□

Proposition 5 *Si*

$$\underline{\tilde{\mathbf{b}}}_\lambda = \arg \min_{\underline{b}\in\mathbf{R}^{dm_n+1}} \|\underline{\mathbf{Y}} - \mathbf{U}^{\mathbf{T}}\underline{\mathbf{b}}\|^2 + \sqrt{m_n}\lambda \sum_{j=1}^{d} \|\underline{\mathbf{b}}_{\mathbf{j}}\|_2,$$

et $card(S_\lambda) \le M$,
Alors

$$\underline{\hat{\mathbf{b}}}_\lambda = \arg \min_{\underline{b}\in\mathbf{R}^{card(S_\lambda)m_n+1}} \|\underline{\mathbf{Y}} - \mathbf{U}^{\mathbf{T}}_{\mathbf{S}_\lambda}\underline{\mathbf{b}}\|^2 + \sqrt{m_n}\lambda \sum_{j\in S_\lambda} \|\underline{\mathbf{b}}_{\mathbf{j}}\|_2,$$

Démonstration Proposition 5 *Pour démontrer ce résultat, nous montrons qu'il n'existe pas de* $\underline{\tilde{\mathbf{a}}}_\lambda$ *tel que*

$$\|\underline{\mathbf{Y}} - \mathbf{U}^{\mathbf{T}}_{\mathbf{S}_\lambda}\underline{\hat{\mathbf{b}}}_\lambda\|^2 + \sqrt{m_n}\lambda \sum_{j\in S_\lambda} \|\underline{\hat{\mathbf{b}}}_{\lambda,\mathbf{j}}\|_2 > \|\underline{\mathbf{Y}} - \mathbf{U}^{\mathbf{T}}_{\mathbf{S}_\lambda}\underline{\tilde{\mathbf{a}}}_\lambda\|^2 + \sqrt{m_n}\lambda \sum_{j\in S_\lambda} \|\underline{\tilde{\mathbf{a}}}_{\lambda,\mathbf{j}}\|_2.$$

Nous faisons un raisonnement par l'absurde. Supposons qu'il existe $\underline{\tilde{\mathbf{a}}}_\lambda$ *tel que*

$$\|\underline{\mathbf{Y}} - \mathbf{U}^{\mathbf{T}}_{\mathbf{S}_\lambda}\underline{\hat{\mathbf{b}}}_\lambda\|^2 + \sqrt{m_n}\lambda \sum_{j\in S_\lambda} \|\underline{\hat{\mathbf{b}}}_{\lambda,\mathbf{j}}\|_2 > \|\underline{\mathbf{Y}} - \mathbf{U}^{\mathbf{T}}_{\mathbf{S}_\lambda}\underline{\tilde{\mathbf{a}}}_\lambda\|^2 + \sqrt{m_n}\lambda \sum_{j\in S_\lambda} \|\underline{\tilde{\mathbf{a}}}_{\lambda,\mathbf{j}}\|_2.$$

Notons $\underline{\bar{\mathbf{a}}}_\lambda$ *tel que* $\underline{\bar{\mathbf{a}}}_{\lambda,\mathbf{j}} = \underline{\tilde{\mathbf{a}}}_{\lambda,\mathbf{j}}$ *si* $j \in S_\lambda$, $= \mathbf{0}$ *sinon.*
Alors il est clair que

$$\sqrt{m_n}\lambda \sum_{j\in S_\lambda} \|\underline{\tilde{\mathbf{a}}}_{\lambda,\mathbf{j}}\|_2 = \sqrt{m_n}\lambda \sum_{j=1}^{d} \|\underline{\bar{\mathbf{a}}}_{\lambda,\mathbf{j}}\|_2,$$

par définition de $\underline{\bar{\mathbf{a}}}_\lambda$.
De même, puisque $\underline{\bar{\mathbf{a}}}_{\lambda,\mathbf{j}} = \underline{\tilde{\mathbf{a}}}_{\lambda,\mathbf{j}}$ *si* $j \in S_\lambda$, *nous avons*

$$\|\underline{\mathbf{Y}} - \mathbf{U}\underline{\bar{\mathbf{a}}}_\lambda\|^2 = \|\underline{\mathbf{Y}} - \mathbf{U}^{\mathbf{T}}_{\mathbf{S}_\lambda}\underline{\tilde{\mathbf{a}}}_\lambda\|^2.$$

De même,

$$\|\underline{\mathbf{Y}} - \mathbf{U}^{\mathbf{T}}_{\mathbf{S}_\lambda}\underline{\hat{\mathbf{b}}}_\lambda\|^2 + \sqrt{m_n}\lambda \sum_{j\in S_\lambda} \|\underline{\hat{\mathbf{b}}}_{\lambda,\mathbf{j}}\|_2 = \|\underline{\mathbf{Y}} - \mathbf{U}\underline{\tilde{\mathbf{b}}}_\lambda\|^2 + \sqrt{m_n}\lambda \sum_{j=1}^{d} \|\underline{\tilde{\mathbf{b}}}_{\lambda,\mathbf{j}}\|_2,$$

puisque

$$\|\underline{\mathbf{Y}} - \mathbf{U}^{\mathbf{T}}_{\mathbf{S}_\lambda}\underline{\hat{\mathbf{b}}}_\lambda\|^2 + \sqrt{m_n}\lambda \sum_{j\in S_\lambda} \|\underline{\hat{\mathbf{b}}}_{\lambda,\mathbf{j}}\|_2 > \|\underline{\mathbf{Y}} - \mathbf{U}^{\mathbf{T}}_{\mathbf{S}_\lambda}\underline{\tilde{\mathbf{a}}}_\lambda\|^2 + \sqrt{m_n}\lambda \sum_{j\in S_\lambda} \|\underline{\tilde{\mathbf{a}}}_{\lambda,\mathbf{j}}\|_2,$$

nous avons donc

$$\|\underline{\mathbf{Y}} - \mathbf{U}\underline{\tilde{\mathbf{b}}}_\lambda\|^2 + \sqrt{m_n}\lambda \sum_{j=1}^{d} \|\underline{\tilde{\mathbf{b}}}_{\lambda,\mathbf{j}}\|_2 > \|\underline{\mathbf{Y}} - \mathbf{U}\underline{\bar{\mathbf{a}}}_\lambda\|^2 + \sqrt{m_n}\lambda \sum_{j=1}^{d} \|\underline{\bar{\mathbf{a}}}_{\lambda,\mathbf{j}}\|_2,$$

ce qui est contradictoire avec

$$\underline{\tilde{\mathbf{b}}}_\lambda = \arg \min_{\underline{b} \in \mathbf{R}^{dm_n+1}} \|\underline{\mathbf{Y}} - \mathbf{U}^{\mathbf{T}}\underline{\mathbf{b}}\|^2 + \sqrt{m_n}\lambda \sum_{j=1}^{d} \|\underline{\mathbf{b}}_{\mathbf{j}}\|_2.$$

Ceci permet de conclure la proposition.

□

Proposition 6 $\tilde{\mathbf{b}}_{\lambda_\mathbf{n}}$ *est solution de (15) si*

$$\begin{cases} \mathbf{U_j}^T(\underline{\mathbf{Y}} - \mathbf{U}\tilde{\mathbf{b}}_{\lambda_\mathbf{n}}) = \lambda_n \sqrt{m_n} \frac{\tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}}}{\|\tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}}\|} \; si \; j \in S_{\lambda_n} \\ \|\mathbf{U_j}^T(\underline{\mathbf{Y}} - \mathbf{U}\tilde{\mathbf{b}}_{\lambda_\mathbf{n}})\| \leq \lambda_n \sqrt{m_n} \; si \; j \notin S_{\lambda_n} \end{cases} \tag{20}$$

Soit $w_n = (\tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}}/\|\tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}}\|, j \in S^*)$. *Notons*

$$\hat{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{S}^*} = (\mathbf{U_{S^*}}^T\mathbf{U_{S^*}})^{-1}(\mathbf{U_{S^*}}^T\underline{\mathbf{Y}} - \lambda_n \sqrt{m_n} w_n).$$

Si $S_{\lambda_n} = S^*$ *, alors d'après le Lemme 5, si* S^* *est composé des* s^* *premières covariables,* $\tilde{\mathbf{b}}_{\lambda_\mathbf{n}} = (\hat{\mathbf{b}}^T_{\lambda_\mathbf{n},\mathbf{S}^*}, \mathbf{0}^T)^T$.
Comme $\mathbf{U}\tilde{\mathbf{b}}_{\lambda_\mathbf{n}} = \mathbf{U_{S^*}}\hat{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{S}^*}$ *et* $\{\mathbf{U_j}, j \in S^*\}$ *sont linéairement indépendant,*
$S_{\lambda_n} = S^*$ *si*

$$\begin{cases} \|\underline{\mathbf{b}}^*_{\mathbf{j}}\| - \|\tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}}\| < \|\underline{\mathbf{b}}^*_{\mathbf{j}}\| \; si \; j \in S^* \\ \|\mathbf{U_j}^T(\underline{\mathbf{Y}} - \mathbf{U_{S^*}}\tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{S}^*})\| \leq \lambda_n \sqrt{m_n} \; si \; j \notin S_{\lambda_n} \end{cases}$$

Démonstration Proposition 6 *La première partie du lemme est classique (voir la Proposition 8 de Bach [4] par exemple) et consiste en l'écriture des équations normales du Group LASSO.*
$S_{\lambda_n} = S^*$ *si*

$$\begin{cases} \|\underline{\mathbf{b}}^*_{\mathbf{j}}\| - \|\tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}}\| < \|\underline{\mathbf{b}}^*_{\mathbf{j}}\| \; si \; j \in S^* \\ \|\mathbf{U_j}^T(\underline{\mathbf{Y}} - \mathbf{U_{S^*}}\tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{S}^*})\| \leq \lambda_n \sqrt{m_n} \; si \; j \notin S_{\lambda_n} \end{cases}$$

ce qui est démontré au début de la démonstration par Huang et al. *[10].*

□

Démonstration Lemme 6 *Soit* $j \in S^*$.
Notons

$$T_j = (\mathbf{0_{m_n \times m_n}}, \ldots, \mathbf{0_{m_n \times m_n}}, \mathbf{I_{m_n \times m_n}}, \mathbf{0_{m_n \times m_n}}, \ldots, \mathbf{0_{m_n \times m_n}}),$$

avec $\mathbf{0_{m_n \times m_n}}$ *matrice nulle de taille* $m_n \times m_n$ *et* $\mathbf{I_{m_n \times m_n}}$ *matrice identité de taille* $m_n \times m_n$. *Le jeme bloc de* T_j *est* $\mathbf{I_{m_n \times m_n}}$.

$$\begin{aligned} \tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}} &= T_j\tilde{\mathbf{b}}_{\lambda_\mathbf{n}} \\ &= T_j(\underline{\hat{\mathbf{b}}}_{\mathbf{S}^*} - (\mathbf{U_{S^*}}^T\mathbf{U_{S^*}})^{-1}\underline{\mathbf{r}}) \end{aligned}$$

Notons $C_{S^*} = \frac{1}{n}\mathbf{U_{S^*}}^T\mathbf{U_{S^*}}$, *alors* $\|\tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}} - \underline{\mathbf{b}}^*_{\mathbf{S}^*,\mathbf{j}}\| = \|T_j(\underline{\hat{\mathbf{b}}}_{\mathbf{S}^*} - \frac{1}{n}C^{-1}_{S^*}\underline{\mathbf{r}} - \underline{\mathbf{b}}^*_{\mathbf{S}})\| \leq \|T_j(\underline{\hat{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{b}}^*_{\mathbf{S}})\| + \frac{1}{n}\|T_jC^{-1}_{S^*}\underline{\mathbf{r}}\|$, *or* $\|\underline{\hat{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{b}}^*_{\mathbf{S}}\| = O(\frac{m_n}{\sqrt{n}})$, *d'après Stone [18], d'où*

$$\max_j \|T_j(\underline{\hat{\mathbf{b}}}_{\mathbf{S}^*} - \underline{\mathbf{b}}^*_{\mathbf{S}})\| \leq \frac{m_n}{\sqrt{n}}.$$

De plus,

$$\begin{aligned}\frac{1}{n}\|T_j C_{S^*}^{-1}\underline{\mathbf{r}}\| &\leq \frac{1}{n}\|T_j\|\|C_{S^*}^{-1}\|\|\underline{\mathbf{r}}\| \\ &\leq O(1)\frac{1}{n}m_n\lambda_n\sqrt{m_n} \\ &\leq O(1)\frac{m_n}{\sqrt{n}}\sqrt{\log(m_n^2 d)} \\ &\leq O(1)\frac{m_n}{\sqrt{n}}\sqrt{\log(nd)} \\ &\leq o(\sqrt{m_n}\min\|f_j^*\|)\end{aligned}$$

d'où

$$\|\tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{j}} - \underline{\mathbf{b}}^*_{\mathbf{S}^*,\mathbf{j}}\| \leq \frac{m_n}{\sqrt{n}} + o(\sqrt{m_n}\min\|f_j^*\|).$$

De plus,

$$\begin{aligned}\min_{j\in S^*}\|\underline{\mathbf{b}}^*_\mathbf{j}\| &\geq A_6\left(\sqrt{m_n}\min_{j\in S^*}\|f_j^*\| - O(1)\frac{m_n}{\sqrt{n}}\right) \\ &\geq O(1)\left(\sqrt{m_n}\sqrt{\frac{m_n}{n}\log(nd)} + O(\frac{m_n}{\sqrt{n}})\right) \\ &\geq O(1)\left(\frac{m_n}{\sqrt{n}}\sqrt{\log(nd)} + O(\frac{m_n}{\sqrt{n}})\right)\end{aligned}$$

Or, il est clair que

$$\frac{m_n}{\sqrt{n}} = o(\frac{m_n}{\sqrt{n}}\sqrt{\log(nd)}).$$

et donc

$$P(il\ existe\, j \in S^*, \|\underline{\mathbf{b}}^*_\mathbf{j} - \tilde{\mathbf{b}_{\lambda_\mathbf{n},\mathbf{j}}}\| \geq \|\underline{\mathbf{b}}^*_\mathbf{j}\|) \to 0.$$

□

Pour démontrer le résultat suivant, nous supposons que pour tout $j \neq j'$, $\|\mathbf{U_j}^T\mathbf{U'_j}\| = o_P(\frac{n}{m_n})$.

Démonstration Lemme 7 *Soit $j \notin S^*$.*

$$\begin{aligned}\|\mathbf{U_j}^T(\underline{\mathbf{Y}} - \mathbf{U_{S^*}}\tilde{\mathbf{b}}_{\lambda_\mathbf{n}})\| &\leq \|\mathbf{U_j}^T(\underline{\mathbf{Y}} - \mathbf{U_{S^*}}\hat{\mathbf{b}}_{\mathbf{S}_{\lambda_\mathbf{n}}} - \mathbf{U_{S^*}}(\mathbf{U_{S^*}}^T\mathbf{U_{S^*}})^{-1}\underline{\mathbf{r}})\| \\ &\leq \|\mathbf{U_j}^T(\underline{\mathbf{Y}} - \mathbf{U_{S^*}}\hat{\mathbf{b}}_{\mathbf{S}_{\lambda_\mathbf{n}}})\| + \|\mathbf{U_j}^T\mathbf{U_{S^*}}(\mathbf{U_{S^*}}^T\mathbf{U_{S^*}})^{-1}\underline{\mathbf{r}}\| \\ &\leq \|\mathbf{U_j}^T(\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U_{S^*}}\underline{\mathbf{b}}^*_{\mathbf{S}^*} + \underline{\boldsymbol{\epsilon}} + \mathbf{U_{S^*}}\underline{\mathbf{b}}^*_{\mathbf{S}^*} - \mathbf{U_{S^*}}\hat{\mathbf{b}}_{\mathbf{S}_{\lambda_\mathbf{n}}})\| + \|\mathbf{U_j}^T\mathbf{U_{S^*}}(\mathbf{U_{S^*}}^T\mathbf{U_{S^*}})^{-1}\underline{\mathbf{r}}\| \\ &\leq \|\mathbf{U_j}^T(\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U_{S^*}}\underline{\mathbf{b}}^*_{\mathbf{S}^*})\| + \|\mathbf{U_j}^T\Pi_j\underline{\boldsymbol{\epsilon}}\| + \|\mathbf{U_j}^T(\mathbf{U_{S^*}}\underline{\mathbf{b}}^*_{\mathbf{S}^*} - \mathbf{U_{S^*}}\hat{\mathbf{b}}_{\mathbf{S}_{\lambda_\mathbf{n}}})\| + \|\mathbf{U_j}^T\mathbf{U_{S^*}}(\mathbf{U_{S^*}}^T\mathbf{U_{S^*}})^{-1}\underline{\mathbf{r}}\|\end{aligned}$$

Etudions chacun des 4 termes.

– *Etude de $\|\mathbf{U_j}^T(\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U_{S^*}}\underline{\mathbf{b}}^*_{\mathbf{S}^*})\|$:*

$$\begin{aligned}\|\mathbf{U_j}^T(\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U_{S^*}}\underline{\mathbf{b}}^*_{\mathbf{S}^*})\| &\leq \max_{j\notin S^*}\|\mathbf{U_j}^T\mathbf{U_j}\|^{1/2}\|(\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U_{S^*}}\underline{\mathbf{b}}^*_{\mathbf{S}^*})\| \\ &\leq O(1)\sqrt{\frac{n}{m_n}}m_n^{-\nu} \\ &\leq o(\lambda_n\sqrt{m_n})\end{aligned}$$

– *Etude de* $\|\mathbf{U_j}^T \Pi_j \underline{\boldsymbol{\epsilon}}\|$ *:*

$$\begin{aligned}
\|\mathbf{U_j}^T \Pi_j \underline{\boldsymbol{\epsilon}}\| &\leq \max_{j \notin S^*} \|\mathbf{U_j}^T \mathbf{U_j}\|^{1/2} \|\|\Pi_j \underline{\boldsymbol{\epsilon}}\| \\
&\leq \sqrt{\frac{n}{m_n}} o_P(\sqrt{m_n \log(nd)}) \\
&\leq o_P(\lambda_n \sqrt{m_n})
\end{aligned}$$

– *Etude de* $\|\mathbf{U_j}^T (\mathbf{U_{S^*}} \underline{\mathbf{b}}^*_{\mathbf{S^*}} - \mathbf{U_{S^*}} \hat{\mathbf{b}}_{\mathbf{S}_{\lambda_\mathbf{n}}})\|$ *:*

$$\begin{aligned}
\|\mathbf{U_j}^T (\mathbf{U_{S^*}} \underline{\mathbf{b}}^*_{\mathbf{S^*}} - \mathbf{U_{S^*}} \hat{\mathbf{b}}_{\mathbf{S}_{\lambda_\mathbf{n}}})\| &\leq \max_{j \notin S^*} \|\mathbf{U_j}^T \mathbf{U_j}\|^{1/2} \|\mathbf{U_{S^*}}^T \mathbf{U_{S^*}}\|^{1/2} \|\underline{\mathbf{b}}^*_{\mathbf{S^*}} - \hat{\mathbf{b}}_{\mathbf{S}_{\lambda_\mathbf{n}}}\| \\
&\leq O(1) \sqrt{\frac{n}{m_n}} \sqrt{\frac{n}{m_n}} \frac{m_n}{\sqrt{n}} \\
&\leq O(1) \sqrt{n} \\
&\leq o(\lambda_n \sqrt{m_n})
\end{aligned}$$

– *Etude de* $\|\mathbf{U_j}^T \mathbf{U_{S^*}} (\mathbf{U_{S^*}}^T \mathbf{U_{S^*}})^{-1} \underline{\mathbf{r}}\|$ *:*
$\|\mathbf{U_j}^T \mathbf{U_{j'}}\| = o_P(n/m_n)$,

$$\begin{aligned}
\|\mathbf{U_j}^T \mathbf{U_{S^*}} (\mathbf{U_{S^*}}^T \mathbf{U_{S^*}})^{-1} \underline{\mathbf{r}}\| &= \|\mathbf{U_j}^T \mathbf{U_{S^*}} \sum_{i=1}^{s^*} T_i^T T_i (\mathbf{U_{S^*}}^T \mathbf{U_{S^*}})^{-1} \underline{\mathbf{r}}\| \\
&\leq \sum_{i=1}^{s^*} \|\mathbf{U_j}^T \mathbf{U_{S^*}} T_i^T T_i (\mathbf{U_{S^*}}^T \mathbf{U_{S^*}})^{-1} \underline{\mathbf{r}}\| \\
&\leq \sum_{i=1}^{s^*} \|\mathbf{U_j}^T \mathbf{U_{S^*}} T_i^T\| \|\|T_i^T T_i\|^{1/2} \|(\mathbf{U_{S^*}}^T \mathbf{U_{S^*}})^{-1}\| \|\|\underline{\mathbf{r}}\| \\
&\leq \sum_{i \in S^*} \|\mathbf{U_j}^T \mathbf{U_i}\| \|\|(\mathbf{U_{S^*}}^T \mathbf{U_{S^*}})^{-1}\| \|\|\underline{\mathbf{r}}\| \\
&\leq o_P(n/m_n) \frac{m_n}{n} \lambda_n \sqrt{m_n} \\
&\leq o_P(\lambda_n \sqrt{m_n})
\end{aligned}$$

Ceci permet de conclure que

$$\|\mathbf{U_j}^T (\underline{\mathbf{Y}} - \mathbf{U_{S^*}} \tilde{\mathbf{b}}_{\lambda_\mathbf{n}})\| \leq o_P(\lambda_n \sqrt{m_n}).$$

□

Proposition 7

$$\|\mathbf{U_{S^*}} (\mathbf{U_{S^*}}^T \mathbf{U_{S^*}})^{-1} \underline{\mathbf{r}}\| = O(1) \lambda_n \frac{m_n}{\sqrt{n}} = O(1) \sqrt{m_n \log(d m_n^2)}.$$

Démonstration Proposition 7

$$\begin{aligned}
\|\mathbf{U_{S^*}} (\mathbf{U_{S^*}}^T \mathbf{U_{S^*}})^{-1} \underline{\mathbf{r}}\| &\leq \|\mathbf{U_{S^*}}^T \mathbf{U_{S^*}}\|^{1/2} \|(\mathbf{U_{S^*}}^T \mathbf{U_{S^*}})^{-1}\| \|\|\underline{\mathbf{r}}\| \\
&\leq O(1) \lambda_n \sqrt{m_n} \sqrt{\frac{n}{m_n}} \frac{m_n}{n} \\
&\leq O(1) \lambda_n \frac{m_n}{\sqrt{n}} \\
&\leq O(1) \sqrt{m_n \log(d m_n^2)}
\end{aligned}$$

□

Proposition 8

$$\tilde{\mathbf{b}}_{\lambda_\mathbf{n}} - \underline{\mathbf{b}}^*_{\mathbf{S}^*} = (\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}(\mathbf{U}_{\mathbf{S}^*}{}^T(\underline{\boldsymbol{\epsilon}} + B) - \underline{\mathbf{r}}),$$

et

$$\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\tilde{\mathbf{b}}_{\lambda_\mathbf{n}} = (I - \Pi_{S^*})(\underline{\boldsymbol{\epsilon}} + B) + \mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}.$$

Démonstration Proposition 8 *Nous étudions les deux expressions de la Proposition 8.*

– *Etude de* $\tilde{\mathbf{b}}_{\lambda_\mathbf{n}} - \underline{\mathbf{b}}^*_{\mathbf{S}^*}$ *:*

$$\begin{aligned}
\tilde{\mathbf{b}}_{\lambda_\mathbf{n}} - \underline{\mathbf{b}}^*_{\mathbf{S}^*} &= (\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}(\mathbf{U}_{\mathbf{S}^*}{}^T\underline{\mathbf{Y}} - \underline{\mathbf{r}}) - \underline{\mathbf{b}}^*_{\mathbf{S}^*} \\
&= (\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}(\mathbf{U}_{\mathbf{S}^*}{}^T(\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*}) - \underline{\mathbf{r}}) \\
&= (\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}(\mathbf{U}_{\mathbf{S}^*}{}^T(\underline{\boldsymbol{\epsilon}} + B) - \underline{\mathbf{r}})
\end{aligned}$$

– *Etude de* $\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\tilde{\mathbf{b}}_{\lambda_\mathbf{n}}$ *:*

$$\begin{aligned}
\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\tilde{\mathbf{b}}_{\lambda_\mathbf{n},\mathbf{S}^*} &= \underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*} + \mathbf{U}_{\mathbf{S}^*}\underline{\mathbf{b}}^*_{\mathbf{S}^*} - \mathbf{U}_{\mathbf{S}^*}\tilde{\mathbf{b}}_{\lambda_\mathbf{n}} \\
&= B + \underline{\boldsymbol{\epsilon}} - \mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}(\mathbf{U}_{\mathbf{S}^*}{}^T(\underline{\boldsymbol{\epsilon}} + B) - \underline{\mathbf{r}}) \\
&= (I - \Pi_{S^*})(\underline{\boldsymbol{\epsilon}} + B) + \mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}
\end{aligned}$$

□

Démonstration Lemme 8

$$\begin{aligned}
||\underline{\mathbf{Y}} - \mathbf{U} &= ||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\hat{\underline{\mathbf{b}}}_{\lambda_\mathbf{n}}||^2 - ||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}_{\lambda_\mathbf{n}}}\hat{\underline{\mathbf{b}}}_{\mathbf{S}_{\lambda_\mathbf{n}}}||^2 \\
&= ||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}(\mathbf{U}_{\mathbf{S}^*}{}^T\underline{\mathbf{Y}} + \underline{\mathbf{r}})||^2 - ||\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\hat{\underline{\mathbf{b}}}_{\mathbf{S}^*}||^2 \\
&= (\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}(\mathbf{U}_{\mathbf{S}^*}{}^T\underline{\mathbf{Y}} + \underline{\mathbf{r}}))^T(\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}(\mathbf{U}_{\mathbf{S}^*}{}^T\underline{\mathbf{Y}} + \underline{\mathbf{r}})) - (\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\hat{\underline{\mathbf{b}}}_{\mathbf{S}^*})^T(\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}^*}\hat{\underline{\mathbf{b}}}_{\mathbf{S}^*}) \\
&= ||\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}||^2 - 2(\underline{\mathbf{Y}} - \Pi_{S^*}\underline{\mathbf{Y}})^T(\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}})
\end{aligned}$$

□

Démonstration Lemme 9 *D'après le Lemme 7*

$$||\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}|| = O(m_n\sqrt{n\log(m_n^2 d)/m_n}/\sqrt{n}) = O(\sqrt{m_n\log(m_n^2 d)}).$$

De plus,

$$\begin{aligned}
(\underline{\mathbf{Y}} - \Pi_{S^*}\underline{\mathbf{Y}})^T(\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}) &= (\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) + \underline{\boldsymbol{\epsilon}} - \mathbf{U}_{\mathbf{S}^*}\hat{\underline{\mathbf{b}}}_{\mathbf{S}^*})^T(\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}) \\
&= (\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U}_{\mathbf{S}^*}\hat{\underline{\mathbf{b}}}_{\mathbf{S}^*})^T(\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}) + \underline{\boldsymbol{\epsilon}}^T(\Pi_{S^*}\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}) \\
&\textit{(car } \Pi_{S^*}\mathbf{U}_{\mathbf{S}^*} = \mathbf{U}_{\mathbf{S}^*}\textit{)}
\end{aligned}$$

D'après Cauchy-Schwarz,

$$|(\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U}_{\mathbf{S}^*}\hat{\underline{\mathbf{b}}}_{\mathbf{S}^*})^T(\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}})| \leq ||\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U}_{\mathbf{S}^*}\hat{\underline{\mathbf{b}}}_{\mathbf{S}^*}||||\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}||.$$

Nous avons

$$||\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}|| = O(\sqrt{m_n\log(m_n^2 d)}),$$

et

$$\|\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U}_{\mathbf{S}^*}\underline{\hat{\mathbf{b}}}_{\mathbf{S}^*}\| \leq O_P(\sqrt{m_n}) \quad \textit{d'après le Lemme 17},$$

d'où

$$|(\underline{\mathbf{f}}^*(\underline{\mathbf{X}}) - \mathbf{U}_{\mathbf{S}^*}\underline{\hat{\mathbf{b}}}_{\mathbf{S}^*})^T(\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}})| \leq O_P(m_n\sqrt{\log(m_n^2 d)}).$$

De même,

$$|\underline{\boldsymbol{\epsilon}}^T\Pi_{S^*}(\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}})| \leq \|\Pi_{S^*}\underline{\boldsymbol{\epsilon}}\|\|\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}}\|.$$

D'après le Lemme 3,

$$\|\Pi_{S^*}\underline{\boldsymbol{\epsilon}}\| \leq o_P(\sqrt{m_n\log(nd)}),$$

d'où

$$|\underline{\boldsymbol{\epsilon}}^T\Pi_{S^*}(\mathbf{U}_{\mathbf{S}^*}(\mathbf{U}_{\mathbf{S}^*}{}^T\mathbf{U}_{\mathbf{S}^*})^{-1}\underline{\mathbf{r}})| \leq o_P(m_n\sqrt{\log(nd)\log(m_n^2 d)}).$$

□

Démonstration Lemme 10

$$\begin{aligned}
BIC(\lambda_n) - BIC(S^*) &= \log\left(\frac{\|\underline{\mathbf{Y}} - \mathbf{U}\underline{\tilde{\mathbf{b}}}_{\lambda_\mathbf{n}}\|^2}{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}_{\lambda_\mathbf{n}}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}_{\lambda_\mathbf{n}}}\|^2}\right) \\
&= \log\left(1 + \frac{\|\underline{\mathbf{Y}} - \mathbf{U}\underline{\tilde{\mathbf{b}}}_{\lambda_\mathbf{n}}\|^2 - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}_{\lambda_\mathbf{n}}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}_{\lambda_\mathbf{n}}}\|^2}{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}_{\lambda_\mathbf{n}}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}_{\lambda_\mathbf{n}}}\|^2}\right) \\
&= \log\left(1 + \frac{\|\underline{\mathbf{Y}} - \mathbf{U}\underline{\tilde{\mathbf{b}}}_{\lambda_\mathbf{n}}\|^2/n - \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}_{\lambda_\mathbf{n}}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}_{\lambda_\mathbf{n}}}\|^2/n}{\|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}_{\lambda_\mathbf{n}}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}_{\lambda_\mathbf{n}}}\|^2/n}\right) \\
&= \log\left(1 + o_P\left(m_n\frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n}\right)\right) \textit{ car } \|\underline{\mathbf{Y}} - \mathbf{U}_{\mathbf{S}_{\lambda_\mathbf{n}}}\underline{\hat{\mathbf{b}}}_{\mathbf{S}_{\lambda_\mathbf{n}}}\|^2/n \textit{ borné d'après le Lemme 4} \\
&= o_P\left(m_n\frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n}\right) + o\left(o_P\left(m_n\frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n}\right)\right)
\end{aligned}$$

En effet,

$$m_n\frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n} \leq m_n\frac{\log(nd)}{n} = m_n\frac{\log(d)}{n} + m_n\frac{\log(n)}{n} \to 0.$$

Ensuite, nous montrons que $m_n\frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n}$ *est négligeable par rapport à* $\min(\log(2), v_n^2/m_n)$.
Comme $m_n\frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n} \leq \frac{m_n\log(nd)}{n}$, *nous montrons plutôt que* $\frac{m_n\log(nd)}{n}$ *est négligeable par rapport à* $\min(\log(2), v_n^2/m_n)$.
$\frac{m_n\log(nd)}{n} \to 0$, *donc* $\frac{m_n\log(nd)}{n} = o(\log(2))$, *donc si* $\min(\log(2), v_n^2/m_n) = \log(2)$, *le résultat est immédiat. Montrons que* $\frac{m_n\log(nd)}{n} = o(v_n^2/m_n)$.
D'après l'hypothèse 5.d,

$$\frac{m_n\log(nd)}{n} = o(\min_{j\in S^*}\|f_j\|^2)$$

et

$$\frac{v_n^2}{m_n} = \min_{j\in S^*}\|f_j\|^2 + o(\min_{j\in S^*}\|f_j\|^2),$$

donc

$$\frac{m_n\log(nd)}{n} = o\left(\frac{v_n^2}{m_n}\right),$$

et donc $m_n \frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n}$ *est négligeable par rapport à* $\min(\log(2), v_n^2/m_n)$.

□

Démonstration Théorème 4 *Soit* λ *tel que* $card(S_\lambda) \leq M$. *Alors, par définition des MCO,*

$$BIC(\lambda) \geq BIC(S_\lambda).$$

Supposons que $S^* \not\subseteq S_\lambda$, *alors*

$$\begin{aligned} BIC(\lambda) - BIC(\lambda_n) &\geq BIC(S_\lambda) - BIC(\lambda_n) \\ &\geq BIC(S_\lambda) - BIC(S^*) + o_P(m_n \frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n}) \quad \textit{d'après le Lemme 9} \end{aligned}$$

Or $BIC(S_\lambda) - BIC(S^*) \geq O(1)(\min(log(2), \frac{v_n^2}{m_n}) + o(\frac{v_n^2}{m_n}))$,
et $o_P(m_n \frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n})$ *négligeable par rapport à* $\min(1, \frac{v_n^2}{m_n})$.
D'où, d'après le Théorème 1,

$$P(\min_{\lambda | card(S_\lambda) \leq M \text{ et } S^* \not\subseteq S_\lambda} BIC(\lambda) - BIC(\lambda_n) > 0) \to_{n \to +\infty} 1.$$

Supposons maintenant que $S^* \subseteq S_\lambda$,
alors

$$-BIC(\lambda) \leq -BIC(S_\lambda),$$

c'est-à-dire que

$$BIC(\lambda_n) - BIC(\lambda) \leq BIC(\lambda_n) - BIC(S_\lambda) \leq BIC(S^*) + o_P(m_n \frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n}) - BIC(S_\lambda)$$

Or $o_p(m_n \frac{\sqrt{\log(m_n^2 d)\log(nd)}}{n})$ *négligeable par rapport à* $\frac{m_n \log(nd)}{n}$.
D'où, d'après le Théorème 2,

$$P(\min_{\lambda | S^* \subset S_\lambda \text{ et } card(S_\lambda) \leq M} BIC(\lambda_n) - BIC(\lambda) < 0) \to_{n \to +\infty} 1.$$

Ceci démontre le théorème.

□

C Résultats détaillés de la simulation

Critère	Post1Bic	Post1Gcv	Post2Bic	Post2Gcv	GrpLASSOBic	GrpLASSOGcv	GAMSelect	GAMShrinkage	Idéal
MS	4(0)	4(0.18)	4(0)	4(0.29)	4(0.10)	4(0.25)	7 (1.26)	8(1.16)	4
MTZ	6(0)	6(0.18)	6(0)	6(0.29)	6(0.10)	6(0.25)	3 (1.26)	2(1.16)	6
MFZ	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0
MTP	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4
MFP	0(0)	0(0.18)	0(0)	0(0.29)	0(0.10)	0(0.25)	3 (1.26)	4(1.16)	0

Tableau 6 – Capacité à sélectionner les composantes selon le critère utilisé avec une variance du bruit de 0.5 lorsque $t = 0$ (SNR=5.3)

Critère	Post1Bic	Post1Gcv	Post2Bic	Post2Gcv	GrpLASSOBic	GrpLASSOGcv	GAMSelect	GAMShrinkage	Idéal
MS	4(0)	4(0.36)	4(0)	4(0.20)	4(0.10)	4(0.12)	7 (1.23)	8 (1.12)	4
MTZ	6(0)	6(0.36)	6(0)	6(0.20)	6(0.10)	6(0.12)	3 (1.23)	2 (1.12)	6
MFZ	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0
MTP	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4
MFP	0(0)	0(0.36)	0(0)	0(0.20)	0(0.10)	0(0.12)	3 (1.23)	4 (1.12)	0

Tableau 7 – Capacité à sélectionner les composantes selon le critère utilisé avec une variance du bruit de 1.5 lorsque $t = 0$ (SNR=3.1)

Critère	Post1Bic	Post1Gcv	Post2Bic	Post2Gcv	GrpLASSOBic	GrpLASSOGcv	GAMSelect	GAMShrinkage	Idéal
MS	4(0)	4(0.33)	4(0)	4(0.20)	4(0)	4(0)	7.5(1.28)	8 (1.22)	4
MTZ	6(0)	6(0.33)	6(0)	6(0.20)	6(0)	6(0)	2.5(1.28)	2 (1.22)	6
MFZ	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0
MTP	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4
MFP	0(0)	0(0.33)	0(0)	0(0.20)	0(0)	0(0)	3.5(1.28)	4 (1.22)	0

Tableau 8 – Capacité à sélectionner les composantes selon le critère utilisé avec une variance du bruit de 3.5 lorsque $t = 0$ (SNR=2.0)

Lorsque les variables explicatives ne sont pas corrélées, les performances en termes de sélection de composantes des différentes méthodes issus du Group LASSO sont équivalentes, mis à part Post1Gcv, légèrement moins efficace.

Critère	Post1Bic	Post1Gcv	Post2Bic	Post2Gcv	GrpLASSOBic	GrpLASSOGcv	GAMSelect	GAMShrinkage	Idéal
MS	4(0)	4(0.35)	4(0)	4(0.24)	4(0.20)	4(0.52)	8(1.18)	8 (1.14)	4
MTZ	6(0)	6(0.35)	6(0)	6(0.24)	6(0.20)	6(0.52)	2(1.18)	2 (1.14)	6
MFZ	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0
MTP	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4
MFP	0(0)	0(0.35)	0(0)	0(0.24)	0(0.20)	0(0.52)	4(1.18)	4 (1.14)	0

Tableau 9 – Capacité à sélectionner les composantes selon le critère utilisé avec une variance du bruit de 0.5 lorsque $t = 1$ (SNR=5.42)

Critère	Post1Bic	Post1Gcv	Post2Bic	Post2Gcv	GrpLASSOBic	GrpLASSOGcv	GAMSelect	GAMShrinkage	Idéal
MS	4(0)	4(0.38)	4(0)	4(0.20)	4(0.10)	4(0.22)	7(1.31)	8 (1.24)	4
MTZ	6(0)	6(0.38)	6(0)	6(0.20)	6(0.10)	6(0.22)	3(1.31)	2 (1.24)	6
MFZ	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0
MTP	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4
MFP	0(0)	0(0.38)	0(0)	0(0.20)	0(0.10)	0(0.22)	3(1.31)	4 (1.24)	0

Tableau 10 – Capacité à sélectionner les composantes selon le critère utilisé avec une variance du bruit de 1.5 lorsque $t = 1$ (SNR=3.1)

Critère	Post1Bic	Post1Gcv	Post2Bic	Post2Gcv	GrpLASSOBic	GrpLASSOGcv	GAMSelect	GAMShrinkage	Idéal
MS	4(0)	4(0.27)	4(0)	4(0.12)	4(0.18)	4(0.14)	7(1.14)	8(1.24)	4
MTZ	6(0)	6(0.27)	6(0)	6(0.12)	6(0.18)	6(0.14)	3(1.14)	2(1.24)	6
MFZ	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0
MTP	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4
MFP	0(0)	0(0.27)	0(0)	0(0.12)	0(0.18)	0(0.14)	3(1.14)	4(1.24)	0

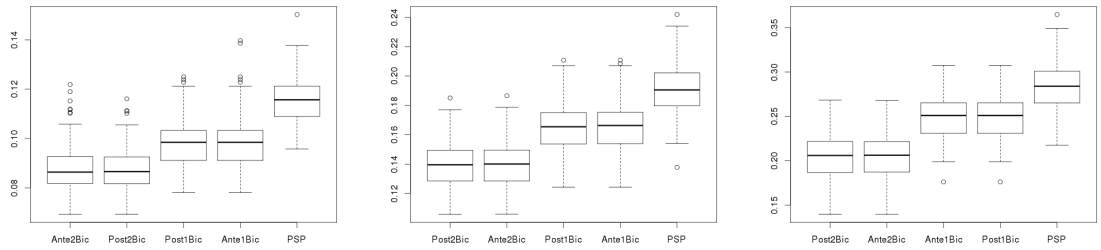
Tableau 11 – Capacité à sélectionner les composantes selon le critère utilisé avec une variance du bruit de 3.5 lorsque $t = 1$ (SNR=2.05)

Critère	Post1Bic	Post1Gcv	Post2Bic	Post2Gcv	GrpLASSOBic	GrpLASSOGcv	GAMSelect	GAMShrinkage	Idéal
MS	4(0)	4(0.34)	4(0)	4(0.20)	4(0.66)	4(2.67)	7(1.13)	8(1.08)	4
MTZ	6(0)	6(0.34)	6(0)	6(0.20)	6(0.66)	6(2.67)	3(1.13)	2(1.08)	6
MFZ	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0
MTP	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	0(0)	0(0)	4
MFP	0(0)	0(0.34)	0(0)	0(0.20)	0(0.66)	0(2.67)	3(1.13)	4(1.08)	0

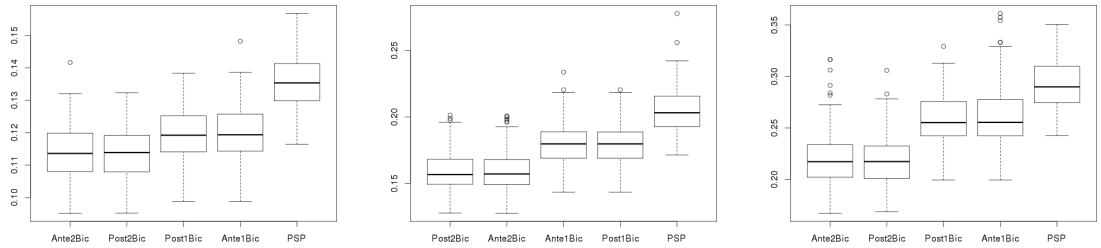
Tableau 12 – Capacité à sélectionner les composantes selon le critère utilisé avec une variance du bruit de 0.5 lorsque $t = 2$ (SNR=5.38)

Critère	Post1Bic	Post1Gcv	Post2Bic	Post2Gcv	GrpLASSOBic	GrpLASSOGcv	GAMSelect	GAMShrinkage	Idéal
MS	4(0)	4(0.25)	4(0)	4(0.14)	4(0.44)	4(0.39)	7(1.23)	8(1.23)	4
MTZ	6(0)	6(0.25)	6(0)	6(0.14)	6(0.44)	6(0.39)	3(1.23)	2(1.23)	6
MFZ	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0
MTP	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4(0)	4
MFP	0(0)	0(0.25)	0(0)	0(0.14)	0(0.44)	0(0.39)	3(1.23)	4(1.23)	0

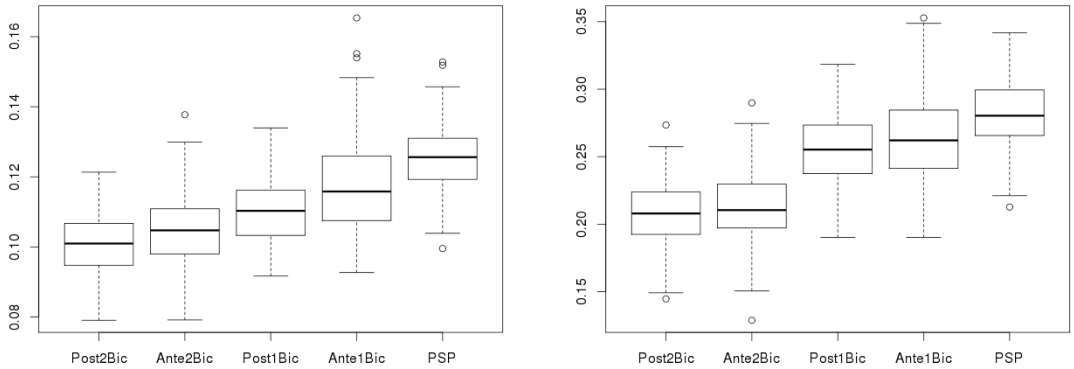
Tableau 13 – Capacité à sélectionner les composantes selon le critère utilisé avec une variance du bruit de 3.5 lorsque $t = 2$ (SNR=2.0)



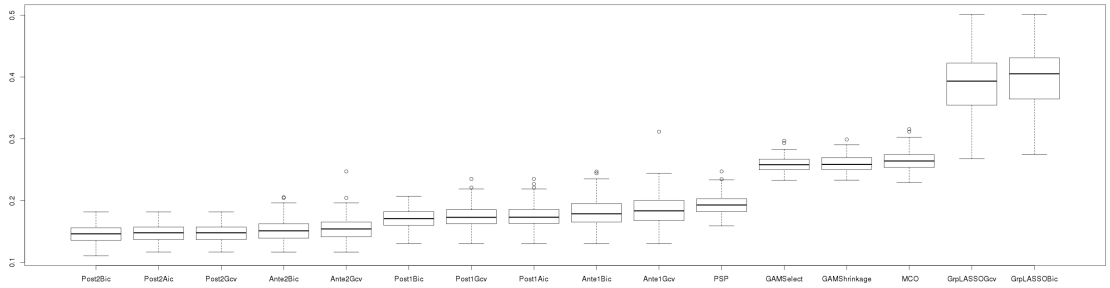
Graphique 4 – Boites à moustaches des RMSE lorsque la variance du bruit est respectivement 0.5, 1.5 et 3.5 et $t = 0$



Graphique 5 – Boites à moustaches des RMSE lorsque la variance du bruit est respectivement 0.5, 1.5 et 3.5 et $t = 1$



Graphique 6 – Boites à moustaches des RMSE lorsque la variance du bruit est respectivement 0.5 et 3.5 et $t = 2$



Graphique 7 – Boites à moustaches des RMSE lorsque la variance du bruit vaut 1.5 et $t = 2$ (SNR=3.1)

Références

[1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6) :716–723, Dec 1974.

[2] A. Antoniadis, I. Gijbels, and S. Lambert-Lacroix. Penalized estimation in additive varying coefficient models using grouped regularization. *Statistical Papers*, 55(3) :727–750, 2014.

[3] A. Antoniadis, I. Gijbels, and A. Verhasselt. Variable selection in additive models using p-splines. *Technometrics*, 54(4) :425–438, 2012.

[4] F.R. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9 :1179–1225, June 2008.

[5] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2) :521–547, 05 2013.

[6] E. Cantoni, J. Mills Flemming, and E. Ronchetti. Variable selection in additive models by nonnegative garrote. *Statistical Modelling*, 11(3) :237–252, March 2006.

[7] P.H.C. Eilers and B.D. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2) :89–121, 05 1996.

[8] J. Fan and J. Jiang. Generalized likelihood ratio tests for additive models. *Journal American Statistical Association*, 2005.

[9] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. London : Chapman & Hall, 1990.

[10] J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *The Annals of Statistics*, 38(4) :2282–2313, 08 2010.

[11] K. Kato. Group Lasso for high dimensional sparse quantile regression models. *ArXiv e-prints*, March 2011.

[12] K. Kato. Two-step estimation of high dimensional additive models. *ArXiv e-prints*, July 2012.

[13] H. Leeb and B. M. Potscher. Can one estimate the conditional distribution of post-model-selection estimators ? *The Annals of Statistics*, 34(5) :2554–2591, 10 2006.

[14] Y. Lin and H.H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5) :2272–2297, 10 2006.

[15] G. Marra and S. Wood. Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7) :2372 – 2387, 2011.

[16] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37(6B) :3779–3821, 12 2009.

[17] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2) :461–464, 1978.

[18] C.J. Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2) :689–705, 06 1985.

[19] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58 :267–288, 1994.

[20] G. Wahba and P. Craven. Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31 :377–404, 1978/79.

[21] H. Wang, B. Li, and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 71(3) :671–683, 2009.

[22] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68 :49–67, 2006.