MASTER EN SCIENCES MATHÉMATIQUES À FINALITÉ SPÉCIALISÉE EN PERSPECTIVES PROFESSIONNELLES DES MATHÉMATIQUES APPLIQUÉES

Erreurs de mesure dans la régression linéaire multivariée et fonctionnelle

MOUTON, Bertrand

Award date:
2022

Awarding institution:
Universite de Namur

UNIVERSITE DE NAMUR

Faculté des Sciences

ERREURS DE MESURE DANS LA RÉGRESSION LINÉAIRE MULTIVARIÉE ET FONCTIONNELLE

Promoteur : Germain VAN BEVER

Mémoire présenté pour l'obtention du grade académique de master en sciences mathématiques à finalité spécialisée en project engineering

Bertrand MOUTON

Août 2022

Remerciements

Je tiens à remercier avant toute chose toutes les personnes qui m'ont permis de réaliser de près ou de loin ce mémoire. Je tiens à saluer le département de mathématiques de l'Université de Namur, l'ensemble des professeurs et assistants qui m'ont accompagné tout au long de mon cursus universitaire. Aussi, je remercie plus particulièrement mon promoteur de mémoire, Germain Van Bever, dans son accompagnement.

J'ai également une pensée pour tous ceux dont j'ai croisé le chemin au cours de toutes ces années d'étude. Je remercie plus particulièrement Célia, Lara, Sarah, Marine, Margaux, Gaëtan, Laura pour leur soutien et les bons moments passés ensemble qui m'ont permis d'en arriver là où je suis. Merci aussi à Jordan pour son écoute et ses encouragements au cours de mes études ainsi qu'à ma famille qui m'a donné la possibilité d'atteindre mes objectifs.

Résumé

Dans le cadre de la régression linéaire multiple, où les modèles tentent d'établir l'interaction linéaire entre des régresseurs et une variable réponse, nous pouvons faire face à la présence d'erreurs de mesure. Ce cas de figure peut arriver lorsque certaines variables explicatives ne peuvent être directement observées. Leur substitution par d'autres données mesurées entraîne alors un biais dans l'estimation des coefficients de régression et met en péril l'identifiabilité du modèle. Ce mémoire a pour but, dans un premier temps, d'expliciter les effets de la présence d'erreurs de mesure dans un modèle de régression et de les illustrer via des données simulées. Deux méthodes de correction dans l'estimation du paramètre d'intérêt sont détaillées et illustrées par la suite. Dans un second temps, nous élargirons la présence des erreurs de mesure au cas fonctionnel de la régression. La théorie permettra d'acheminer le lecteur à un algorithme capable de retrouver là aussi un estimateur convergent du modèle de régression et à illustrer les résultats de cet algorithme sur des données simulées.

Mots-clefs : régression, SIMEX, variable instrumentale, régression fonctionnelle, erreurs de mesure

Abstract

In the context of multiple linear regression, where models try to establish the linear interaction between some regressors and a response variable, we can face to measurement errors. This case can happen when some explicative variables cannot be directly observed. Their substitution by other measured data implies a bias in the estimation of regression coefficients and jeopardizes the identifiability of the model. First, this work explains the effects of the presence of measurement errors in a model of regression and illustrate them with simulated data. Two different methods of correction in the estimation of the parameter of interest are also detailed and illustrated in the following. Second, we extend the presence of measurement errors to the functional case of the regression. The theory leads the reader to an algorithm able to find a consistent estimator of the regression model and to the illustration of results on simulated data.

Keywords : regression, SIMEX, instrumental variable, functional regression, measurement errors

Table des matières

Introduction

Le premier chapitre de ce mémoire se positionne dans le cadre de la régression multivariée où le but est d'établir l'interaction statistique entre des variables explicatives, dites régresseurs, et une variable réponse. Néanmoins, il est possible de faire face à des données contaminées par des erreurs de mesure. En effet, certaines variables à considérer dans le modèle de régression peuvent par exemple se retrouver inobservables pour différentes raisons. C'est pourquoi, pour une question de faisabilité des études statistiques, ces régresseurs sont remplacés par d'autres variables alors dites observables. Néanmoins, l'estimation du coefficient de régression associé à une de ces variables est sujette à un biais dont l'importance peut se montrer cruciale, notamment dans le test de significativité du régresseur. L'identification du bon modèle de régression se retrouve alors menacée.

Après une mise en contexte et une présentation de la problématique dans le premier chapitre, deux méthodes pour permettre de corriger les effets néfastes de la présence d'erreurs de mesure dans un modèle de régression seront proposées et détaillées au lecteur dans le deuxième chapitre. Les résultats de la mise en application d'un algorithme pour chacune des méthodes sur des données simulées seront également présentées.

Le troisième chapitre de ce mémoire porte sur la généralisation des erreurs de mesure à la régression fonctionnelle où les régresseurs sont des données fonctionnelles et où la variable réponse peut être aussi bien une réponse scalaire que fonctionnelle. Le paramètre à estimer devient alors lui aussi fonctionnel. L'objectif persiste dans l'élimination du biais dans l'estimation du paramètre d'intérêt fonctionnel. La théorie est détaillée dans le but d'acheminer le lecteur à la conception d'un algorithme amenant à remplir l'objectif du chaptire.

Le quatrième chapitre termine ce mémoire en appliquant sur des données simulées l'algorithme obtenu grâce à la théorie portant sur les erreurs de mesure au sein de la régression fonctionnelle et présenté dans le troisième chapitre. Les résultats présentés et commentés permettront une vérification de la qualité de la méthode décrite.

Chapitre 1

Erreurs de mesure dans la régression linéaire multivariée

Ce premier chapitre introduit le sujet des erreurs de mesure dans le cadre de la régression linéaire multivariée. Elle comprend d'abord une mise en contexte de la régression et une explication des notions nécessaires à la compréhension du mémoire ainsi qu'une définition des erreurs de mesure. Deux exemples concrets de la vie réelle viendront ensuite illustrer la présence d'erreurs de mesure dans un modèle de régression. Enfin, les problématiques de biais dans l'estimation du paramètre et de perte de puissance dans le test de significativité seront expliquées dans la suite de cette première partie. Des illustrations seront notamment utilisées pour aider à la compréhension de la partie théorique.

1.1 Mise en contexte

Cette section va nous permettre d'introduire le sujet en posant les bases de la régression et de définir la notion d'erreurs de mesure dans un modèle.

Tout d'abord, rappelons ce qu'est une régression. Le but de celle-ci est d'identifier et quantifier le lien pouvant exister entre une variable dépendante à expliquer, notée Y, et des variables explicatives, notées $X_1, \ldots, X_p$, appelées aussi régresseurs. Toutes ces variables sont supposées quantitatives. La régression s'exprime au moyen d'une équation mettant en relation linéaire la variable Y avec $X = (X_1, \ldots, X_p)$, où $p = 1$ pour une régression linéaire simple avec un seul régresseur et $p > 1$ pour une régression linéaire multiple avec plusieurs régresseurs. Cette relation se traduit par

$$Y = \beta_0 + \beta_x X + \varepsilon, \tag{1.1}$$

où Y est la variable dépendante à expliquer, $X = (X_1, \ldots, X_p)$ est le vecteur de régresseurs, ε est le terme d'erreur indépendant des autres variables et suit une loi $\mathcal{N}(0, \sigma_\varepsilon^2)$, β_0 est l'ordonnée à l'origine et $\beta_x = (\beta_1, \ldots, \beta_p)^t$ est le vecteur de coefficients associé à X.

L'enjeu de la régression est notamment l'estimation du paramètre β_x qui quantifie le lien du régresseur X avec la variable Y ainsi que l'identification des régresseurs qui possèdent une influence sur la variable Y. Dans le cadre d'une régression linéaire simple, le premier objectif se traduit par une estimation de la pente de la droite de régression qui tente d'approximer au mieux les données du modèle et qui est illustrée en rouge sur la FIGURE 1.1. Il s'agit donc de construire un estimateur, noté $\hat{\beta}_x$, du paramètre β_x et qui répond à la propriété d'être non-biaisé, c'est-à-dire $\mathrm{E}(\hat{\beta}_x) = \beta_x$. Nous verrons plus tard que cette propriété importante est mise à mal par la présence d'erreurs de mesure dans le modèle.

Ce mémoire se place dans le contexte où des erreurs de mesure entachent le régresseur X. Un raison de la présence de celles-ci est qu'il n'est pas possible de réaliser des observations directes de la variable explicative. Des exemples concrets de situations auxquelles nous pouvons faire face sont expliqués dans la suite de ce travail. Dans ce cas, le modèle ne contient plus la variable X qui est qualifiée d'inobservable mais une nouvelle variable, notée W, qui intègre la variable inobservable ainsi que des erreurs de mesure indépendantes. Cette variable observée est liée à X par la relation

$$W = X + U,$$

où U représente l'erreur de mesure telle que $\mathrm{E}(U|W) = 0$. Le modèle de régression (1.1) est alors remplacé par

$$Y = \beta_0 + \beta_w W + \varepsilon. \tag{1.2}$$

Cette régression observée (1.2) n'est donc plus la même que la régression réelle (1.1), ce qui entraîne un biais dans l'estimation du vrai paramètre β_x. En effet, nous verrons plus tard que $\mathrm{E}(\hat{\beta}_w) \neq \beta_x$, ce qui implique que nous ne pouvons plus nous contenter d'estimer naïvement le paramètre β_w pour obtenir une estimation du vrai paramètre β_x. Le lien entre le régresseur et la variable à expliquer est donc affecté par ce biais.

Il est possible, notamment dans une régression multiple, que les erreurs de mesure n'affectent pas tous les régresseurs. En effet, certains peuvent par exemple être observés directement sans erreur de mesure contrairement aux régresseurs W. Nous notons dès à présent ces variables Z. Dans la majeure partie de ce mémoire, sauf contre-indication, nous supposerons par simplicité que notre modèle ne possède pas de régresseur de ce type.

1.2 Exemples d'illustration

Un grand nombre d'exemples auxquels s'applique la problématique des erreurs de mesure dans le cadre de la régression sont présentés dans la référence [3]. Nous allons nous attarder plus particulièrement sur deux d'entre-eux pour comprendre plus facilement l'enjeu de la régression avec erreurs de mesure dans le monde réel et dans quelles situations nous pouvons y faire face.

1.2.1 Étude sur les maladies pulmonaires chez l'enfant

Tosteson, Stefanski, & Schafer (1989) ont décrit une étude permettant de déceler si l'exposition des enfants à des concentrations de dioxyde d'azote (NO_2) avait un lien avec la survenue d'une maladie pulmonaire. Dans ce modèle de régression, la variable dépendante Y prend comme valeurs

$$\begin{cases} Y = 1 & \text{si présence d'une maladie du poumon} \\ Y = 0 & \text{si absence d'une maladie du poumon} \end{cases}$$

et la variable explicative X représente l'exposition individuelle au NO_2. Or, il n'est pas possible de mesurer la dose exacte de dioxyde d'azote que chaque individu a reçue au cours de se vie. La variable X est donc considérée comme inobservable et doit être remplacée par une variable observable W soumise à des erreurs de mesure. Dans cette étude, c'est une variable bivariée mesurant la concentration de NO_2 dans la chambre et dans la cuisine de la maison de chaque enfant qui est utilisée. Il est évident que ces deux quantités ne peuvent représenter à elles seules toute l'exposition reçue par chaque enfant. En effet, d'autres sources, parfois non-négligeables, sont à prendre en compte comme par exemple la dose de dioxyde d'azote reçue à l'école et qui est donc à l'origine d'erreurs de mesure dans ce modèle. Nous verrons plus tard les graves conséquences que peuvent avoir ces erreurs de mesure sur le test de significativité des coefficients de régression et donc sur les conclusions de cette étude.

1.2.2 Étude sur le cancer du sein

Une étude de la NHANES (National Health and Nutrition Examination Survey) a interrogé des femmes sur leurs habitudes alimentaires en vue de déceler un cancer du sein [12]. La variable dépendante Y est une variable binaire comme dans l'exemple précédent, c'est-à-dire

$$\begin{cases} Y = 1 & \text{si présence d'un cancer du sein,} \\ Y = 0 & \text{si absence d'un cancer du sein.} \end{cases}$$

Une partie des régresseurs sont observés sans erreur de mesure, à savoir l'âge, l'indice de pauvreté, l'indice de masse corporelle (IMC), la consommation d'alcool, les antécédents familiaux, l'âge de la ménarche (premières règles) et le statut de la ménopause. Ces variables sont rassemblées dans le vecteur de régresseurs observés sans erreur de mesure Z. L'étude questionne également les individus sur leurs habitudes nutritionnelles mais celles-ci sont mesurées de manière imprécise car il est coûteux de suivre le régime alimentaire d'un grand nombre de personnes sur une période de temps importante. Au lieu de mesurer la variable X de la prise nutritionnelle de chaque patient, l'étude mesure le régime alimentaire des participants sur les 24 heures qui précèdent l'enquête et qui contient des erreurs de mesure. Ces observations correspondent à la variable W dans le modèle. Ces erreurs de mesure émanent principalement du fait que les repas sur 24 heures ne sont pas véritablement représentatifs du régime alimentaire d'une personne. En effet, il existe des variations journalières et saisonnières. Cette variable W ne correspond donc pas idéalement à la variable inobservable X du modèle et comporte donc un biais.

1.3 Biais dans l'estimation du paramètre

Pour rappel, notre objectif est d'estimer le paramètre β_x du vrai modèle sans erreur de mesure en ayant uniquement à disposition le modèle avec erreurs de mesure. La contamination du modèle par ces erreurs implique un biais dans notre estimateur. Pour mieux comprendre, illustrons le dans un premier temps avec un exemple de régression linéaire simple. Générons 1000 valeurs $X \sim \mathcal{N}(0,1)$ indépendantes et identiquement distribuées et considérons le vrai modèle de régression décrit par

$$Y = \beta_x X + \varepsilon,$$

où $\varepsilon \sim \mathcal{N}(0,1)$ sont des erreurs indépendantes et X correspond à la variable inobservable. Nous choisissons une valeur de coefficient $\beta_x = 3$. Considérons un deuxième modèle de régression qui sera notre modèle avec des erreurs de mesure U et décrit par

$$Y = \beta_w W + \varepsilon,$$

où $W = X + U$ correspond à la variable observée, avec $U \sim \mathcal{N}(0, \sigma_u^2)$. Dans ces deux modèles, nous calculons une régression linéaire avec le logiciel statistique R grâce à la commande *lm* dans le but d'estimer les paramètres β_x et β_w. Cette commande calcule l'estimateur des moindres carrés qui correspond au candidat qui minimise la somme des carrés des différences entre les observations et la prédiction réalisée par la régression.

Observons le résultat sur la FIGURE 1.1. Chaque cercle représente une valeur générée et la droite rouge correspond à la droite de régression. Nous pouvons remarquer que cette droite présente un coefficient de pente plus faible dans le modèle avec erreurs de mesure par rapport à celui sans erreur de mesure en raison d'une plus grande variabilité des valeurs. Ce phénomène est appelé phénomène d'atténuation, dans le sens où la valeur du coefficient de pente se rapproche de 0 par rapport à sa valeur dans le modèle sans erreur de mesure.

L'explication du biais provient du calcul de l'estimateur des moindres carrés du paramètre β_w. En effet, nous calculons

$$\begin{aligned}
\hat{\beta}_w &\overset{\text{def}}{=} \frac{\widehat{\text{Cov}}(W,Y)}{\widehat{\sigma}_w^2} \\
&= \frac{\widehat{\text{Cov}}(X,Y) + \widehat{\text{Cov}}(U,Y)}{\widehat{\sigma}_x^2 + \widehat{\sigma}_u^2} \\
&= \frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_x^2 + \widehat{\sigma}_u^2}\ \beta_x \\
&:= \lambda \beta_x,
\end{aligned}$$

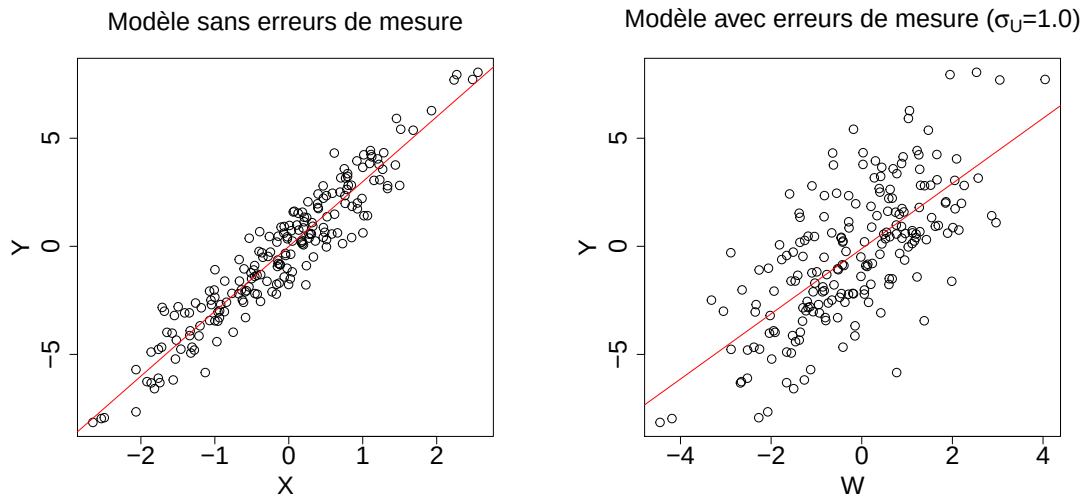

FIGURE 1.1 – Exemple d'une régression linéaire simple sans (à gauche) et avec erreurs de mesure (à droite), où $\sigma_x^2 = 1$ et $\sigma_u^2 = 1$. La taille de l'échantillon vaut $n = 1000$ dans les deux cas et la valeur du paramètre sans erreur de mesure est $\beta_x = 3$. Par (1.3), le paramètre du modèle avec erreurs de mesure vaut $\beta_w = \beta_x/2 = 1.5$.

où $\widehat{\sigma}_{uy} = 0$ ($U \perp\!\!\!\perp Y$) et

$$\begin{aligned}\widehat{\mathrm{Cov}}(X, Y) &= \widehat{\mathrm{Cov}}(X, \beta_0 + \beta_x X + \varepsilon) \\ &= \beta_x \widehat{\mathrm{Cov}}(X, X) \\ &= \beta_x \widehat{\sigma}_x^2.\end{aligned}$$

Ainsi, pour montrer que l'estimateur $\hat{\beta}_w$ est un estimateur biaisé de β_x, nous calculons

$$\mathrm{E}(\hat{\beta}_w) = \lambda \beta_x,$$

où nous appelons facteur d'atténuation

$$\lambda := \frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_x^2 + \widehat{\sigma}_u^2} \leq 1. \tag{1.3}$$

Comme cette valeur est inférieure à 1, notre estimateur est biaisé sur une valeur inférieure à β_x, d'où l'appellation de ce facteur λ. La régression dans le modèle avec erreurs de mesure présente donc une moindre qualité dans le sens où la variabilité du modèle est bien plus grande puisque les observations sont plus dispersées et donc la prédiction est moins précise. Notons que le cas où

$$\begin{aligned}\lambda = 1 &\iff \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} = 1 \\ &\iff \sigma_u^2 = 0\end{aligned}$$

correspond au modèle où les erreurs de mesure sont nulles, c'est-à-dire à une régression classique sans erreur de mesure.

Pour mieux nous rendre compte de ce phénomène d'atténuation et donc de biais, nous allons l'observer sur un grand nombre de simulations. Répétons 1000 fois l'opération réalisée pour obtenir la FIGURE 1.1 et calculons $\hat{\beta}_{w,k}$, l'estimateur du paramètre β_w pour la k^e réalisation, $k = 1, \ldots, 1000$. Traçons ensuite l'histogramme des valeurs obtenues. Cette opération est réalisée pour plusieurs valeurs de σ_u^2 et les résultats sont présentés dans la FIGURE 1.2.

L'histogramme supérieur gauche de la FIGURE 1.2 représente les valeurs de $\hat{\beta}_w$ pour le modèle de régression sans erreur de mesure ($\sigma_u = 0$). Nous pouvons remarquer qu'elles sont centrées en la vraie valeur du paramètre $\beta_x = 3$ que nous avons choisi. Les trois autres histogrammes représentent quant à eux les valeurs de $\hat{\beta}_w$, c'est-à-dire l'estimateur du paramètre du modèle de régression avec erreurs de mesure, pour différentes valeurs de σ_u. Dans ces histogrammes, les valeurs semblent être centrées en une valeur différente de la vraie valeur du paramètre $\beta_x = 3$. Calculons la valeur de $\mathrm{E}(\hat{\beta}_w)$ obtenue via la formule (1.3) et assurons nous que les histogrammes sont bien centrés en ces valeurs.

Dans notre modèle de régression, nous avons choisi que la variable inobservable X suivait une loi $\mathcal{N}(0, 1)$. Sa variance vaut donc $\sigma_x^2 = 1$. Dans le premier histogramme de la FIGURE 1.2, le modèle ne présente pas d'erreurs de mesure. Il s'agit du cas où $\sigma_u = 0$ et donc que la variable observée W est égale à la vraie variable inobservable X. Dans ce cas, $\lambda = 1$ et l'estimateur $\hat{\beta}_w$ est donc non-biaisé et centré 3 comme décrit dans la TABLE 1.1. En vérifiant la FIGURE 1.2, nous pouvons remarquer que les valeurs sont effectivement centrées en 3 qui est la vraie valeur du paramètre β_x.

Dans l'histogramme supérieur droit de la FIGURE 1.2, le modèle présente des erreurs de mesure de variance $\sigma_u^2 = 0.1^2 = 0.01$. Dans ce cas, $\lambda = \frac{100}{101}$ et donc l'estimateur $\hat{\beta}_w$ présente un biais et est centré en 2.97 comme indiqué dans la TABLE 1.1 ou comme nous pouvons l'observer sur l'histogramme.

σ_u	λ	$\mathrm{E}(\hat{\beta}_w) = \lambda\beta_x$
0	1	3
0.1	100/101	2.97
0.5	4/5	2.4
1	1/2	1.5

TABLE 1.1 – Valeurs de $\mathrm{E}(\hat{\beta}_w)$ pour un modèle avec erreurs de mesure en fonction de la valeur de σ_u et où $\sigma_x^2 = 1$.

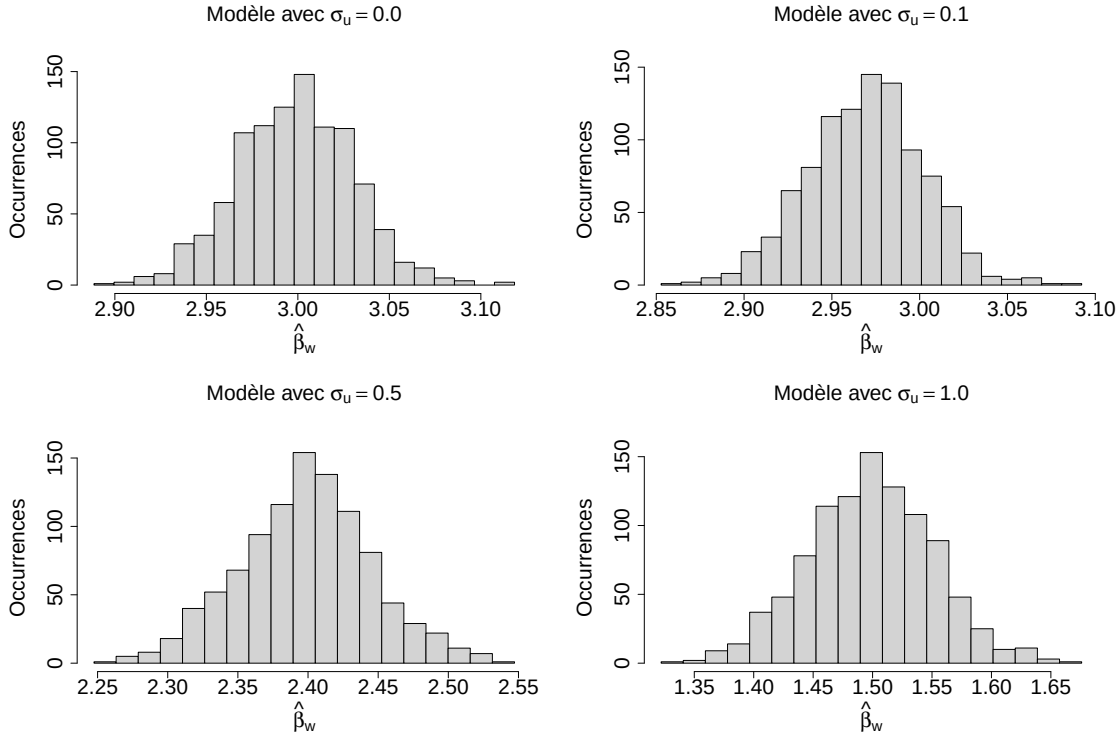

FIGURE 1.2 – Histogrammes des valeurs de $\hat{\beta}_w$ d'une régression linéaire simple avec erreurs de mesure pour différentes valeurs de σ_u. La taille de l'échantillon vaut $n = 1000$ et le nombre d'itérations est de 1000. La valeur du paramètre sans erreur de mesure est $\beta_x = 3$ et $\sigma_x^2 = 1$.

Dans l'histogramme inférieur gauche, le modèle présente des erreurs de mesure dont la variance vaut $\sigma_u^2 = 0.5^2 = 0.25$. Dans ce cas, $\lambda = \frac{4}{5}$ et donc l'estimateur $\hat{\beta}_w$ est biaisé et est centré en 2.4 comme indiqué dans la TABLE 1.1 ou comme nous pouvons le voir sur le troisième histogramme de la FIGURE 1.2.

Enfin, dans l'histogramme inférieur droit, la variance des erreurs de mesure vaut $\sigma_u^2 = 1$. Dans ce cas, $\lambda = \frac{1}{2}$ et donc l'estimateur $\hat{\beta}_w$ est biaisé et est centré en la valeur de 1.5 comme indiqué dans la TABLE 1.1. Il s'agit également de la valeur en laquelle sont centrées les valeurs dans le dernier histogramme de la FIGURE 1.2.

1.4 Perte de puissance du test de significativité

Le biais dans l'estimation du coefficient des régresseurs n'est pas le seul problème qu'apporte la présence d'erreurs de mesure dans un modèle. Cela affecte également la puissance du test de significativité des coefficients. Rappelons brièvement ce qu'est la puissance d'un test d'hypothèses. Un test d'hypothèses est un test statistique opposant deux hypothèses antinomiques. Celles-ci sont l'hypothèse nulle H_0 et l'hypothèse alternative H_1. Le résultat d'un test d'hypothèses rejette ou non H_0 sur base d'une

statistique construite sur un échantillon de la population étudiée. La puissance de ce test est alors la probabilité de rejeter l'hypothèse nulle en sachant que l'hypothèse nulle est fausse. Plus cette valeur est élevée, plus le test est susceptible de nous fournir un bon résultat car il évite les erreurs de type II, c'est-à-dire ne pas rejeter l'hypothèse nulle sachant que celle-ci est fausse. Dans ce cas-ci, l'erreur de type II surviendra lorsque la p-valeur du test de significativité associée à $\hat{\beta}_w$ est supérieure au niveau choisi $\alpha = 0.05$ alors que la p-valeur du test associée à $\hat{\beta}_x$ est inférieure à $\alpha = 0.05$.

Considérons le modèle de régression linéaire simple (1.1) défini précédemment, avec $p = 1$. Considérons aussi le test d'hypothèses

$$\begin{cases} H_0 : & \beta_x = 0 \\ H_1 : & \beta_x \neq 0 \end{cases}.$$

Nous allons réaliser des simulations permettant d'estimer la puissance de ce test en fonction de la valeur de σ_u. Pour chacune de ces valeurs, nous allons simuler n observations $X \sim \mathcal{N}(0, 1)$ et déterminer l'estimateur $\hat{\beta}_x$ du paramètre β_x dans le modèle (1.1). Nous conservons la valeur $\beta_x = 3$. De plus, nous calculons les observations $W = X + U$, où $U \sim \mathcal{N}(0, \sigma_u^2)$ avec des valeurs de σ_u variant de 0 à 5 par pas de 0.5, et déterminons l'estimateur $\hat{\beta}_w$ du paramètre β_w dans le modèle (1.2). Pour chaque valeur de σ_u, nous pouvons estimer la puissance du test par

$$P = 1 - E_{II},$$

où E_{II} est le rapport du nombre de modèles réalisant une erreur de type II sur le nombre total de modèles générés (au nombre de 1000 pour chaque valeur de σ_u).

La FIGURE 1.3 présente les résultats de l'algorithme pour quatre valeurs de n. Nous remarquons que plus la valeur de σ_u est élevée, plus l'estimation de la puissance du test est faible, c'est-à-dire plus la proportion de tests réalisant une erreur de type II est élevée. Dans le premier graphe, où $n = 10$, la puissance du test est estimée à 92.6% lorsque $\sigma_u = 0.5$ mais chute à 22.3% lorsque $\sigma_u = 2$. La taille de l'échantillon influence aussi la puissance du test. En effet, plus n est grand, plus la puissance du test est importante.

La raison de cette perte de puissance est la suivante. Lorsque nous sommes face à un modèle avec des erreurs de mesure U, le paramètre β_x est atténué comme nous l'avons vu précédemment par le facteur $\lambda \leq 1$ et se rapproche ainsi de 0 au plus la valeur de σ_u est élevée. Dans ce cas, la probabilité de rejeter l'hypothèse nulle si celle-ci est fausse faiblit. Cela augmente de ce fait les erreurs de type II dans ce test statistique, ce qui entraîne un plus grand risque de ne pas rejeter l'hypothèse nulle alors que celle-ci est fausse. Le test de significativité aura donc moins tendance à rejeter la nullité des coefficients associés aux régresseurs alors qu'il le devrait puisque la vraie valeur du paramètre $\beta_x = 3$ est ici clairement non nul.

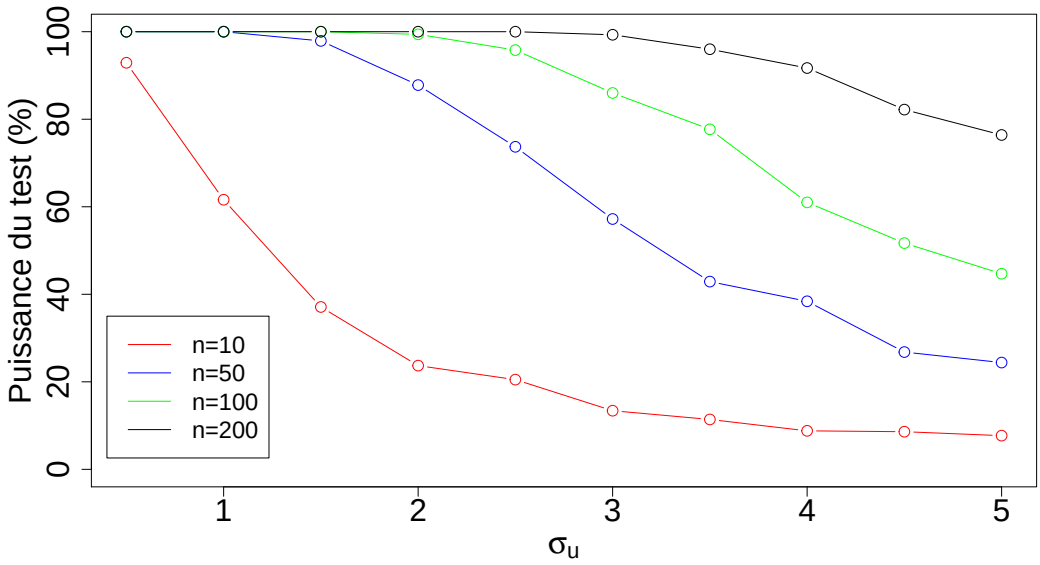

FIGURE 1.3 – Estimation de la puissance du test de significativité du coefficient associé au régresseur X dans un modèle de régression linéaire simple avec erreurs de mesure. L'échantillon est de taille n et les variables du modèle $X \sim \mathcal{N}(0,1)$ et $\varepsilon \sim \mathcal{N}(0,1)$. Le vrai paramètre $\beta_x = 3$ et le nombre de modèles utilisés pour estimer la puissance est de 1000 à chaque étape.

La conséquence de cette perte de puissance du test de significativité peut mettre en péril la bonne identification du modèle de régression. Cela peut menacer grandement les résultats d'une étude statistique. Revenons à l'exemple d'illustration détaillé dans la section 1.2.1 qui, pour rappel, faisait référence à une étude permettant d'établir si l'exposition au NO_2 était liée à une maladie du poumon chez les enfants. En cas de perte de puissance du test de significativité dans ce cas particulier, cela aurait pour conséquence une probabilité plus importante d'un non-rejet de l'hypothèse nulle alors que celle-ci est fausse. En d'autres mots, nous pourrions conclure que la concentration de NO_2 est un régresseur statistiquement non-significatif (au niveau de confiance choisi) pour une maladie du poumon chez les enfants alors qu'en réalité, dans le modèle de régression sans erreur de mesure, l'hypothèse nulle est fausse, c'est-à-dire que le régresseur est significatif et donc qu'il a bel et bien un impact sur la variable réponse.

1.5 Conclusion

Dans ce premier chapitre, nous avons rappelé l'objectif d'une régression. Nous avons introduit la notion d'erreurs de mesure et les effets néfastes de leur présence dans un modèle sur l'estimation des coefficients des régresseurs. Deux problèmes surviennent

en effet. Le premier est l'introduction d'un biais dans l'estimateur du coefficient de régression qui ne correspond plus au paramètre du vrai modèle mais bien à une valeur réduite de celui-ci. Le second est une augmentation de la probabilité de réaliser une erreur de type II dans le test de significativité des coefficients de régression. Cela peut mettre à mal la bonne identification du modèle de régression.

L'objectif du chapitre 2 est donc, en ayant à disposition le modèle de régression avec erreurs de mesure, d'expliciter des méthodes permettant de trouver un estimateur non-biaisé de la valeur du paramètre dans le modèle sans erreur de mesure. Deux approches vont être exposées, la première via le SIMEX et la seconde via la méthode de la variable instrumentale. Après une explication théorique, un algorithme sera proposé pour chacune des méthodes et des résultats sur des données simulées seront présentés pour les illustrer.

Chapitre 2

Méthodes de correction de l'estimateur

2.1 Méthode du SIMEX

Comme nous l'avons vu dans la première partie de ce travail, le présence d'erreurs de mesure dans un modèle de régression est problématique d'une part parce qu'elle cause un biais dans l'estimation du coefficient des régresseurs et d'autre part parce qu'elle implique une perte de puissance dans le test de significativité. Dans cette section, nous allons aborder une méthode permettant de corriger l'estimateur du paramètre. Cette méthode, appelée SIMEX, va nous amener à approcher la vraie valeur du paramètre dans notre modèle à partir d'un algorithme d'extrapolation. Il sera détaillé en fonction de différentes situations qui peuvent survenir et il sera illustré via un exemple de simulations numériques dont nous tirerons des graphiques.

2.1.1 Objectifs

Le SIMEX est une abréviation pour "Simulation Extrapolation" et est une méthode permettant d'estimer le paramètre dans le modèle de régression sans erreur de mesure. Pour rappel, le vrai modèle de régression sans erreur de mesure est décrit par (1.1) avec β_x le paramètre de ce modèle. Le modèle avec erreurs de mesure est quant à lui décrit par (1.2) avec β_w le paramètre de ce modèle. L'estimateur de ce dernier est pour rappel un estimateur biaisé de β_x puisque

$$\hat{\beta}_w = \lambda\beta_x = \frac{\widehat{\sigma}_x^2}{(\widehat{\sigma}_x^2 + \widehat{\sigma}_u^2)}\ \beta_x.$$

L'idée du SIMEX est d'intégrer des erreurs de mesure de plus en plus grandes dans notre modèle en générant celles-ci avec une variance σ_u^2 croissante puis de procéder à une extrapolation des valeurs obtenues jusqu'au modèle sans erreur de mesure, c'est-à-dire lorsque $\sigma_u^2 = 0$.

2.1.2 Description de la méthode

Dans cette section, nous allons décrire brièvement le fonctionnement général du SIMEX. Définissons tout d'abord la fonction

$$\mathcal{G}(\xi) := (1+\xi)\sigma_u^2, \tag{2.1}$$

où σ_u^2 est la variance de l'erreur de mesure U et ξ est un paramètre réel. Cette fonction va représenter la variance totale des erreurs de mesure dans le modèle. Le paramètre ξ va quantifier la grandeur de ces erreurs de mesure. En effet, la variance totale des erreurs de mesure dans le modèle équivaut à la variance des erreurs de mesure dans le modèle naïf, c'est-à-dire σ_u^2, multiplié par le facteur $(1+\xi)$. Dans notre modèle naïf, la variance des erreurs de mesure vaut bien la fonction (2.1) évaluée en $\xi = 0$, c'est-à-dire

$$\mathcal{G}(0) = \sigma_u^2.$$

L'idée du SIMEX est de considérer des modèles successifs avec des erreurs de mesure dont les variances sont de plus en plus grandes, ce qui se traduit avec l'augmentation de la valeur ξ dans la fonction (2.1). En effet, celle-ci peut se réécrire sous la forme

$$\mathcal{G}(\xi) = \sigma_u^2 + \xi\sigma_u^2,$$

ce qui implique qu'en augmentant la valeur de ξ, on augmente la variance des erreurs de mesure d'une quantité $\xi\sigma_u^2$.

Soit l'ensemble $\{\xi_1, \ldots, \xi_M\}$, où $0 = \xi_1 < \cdots < \xi_M$, l'ensemble des valeurs que nous choisissons pour le paramètre ξ. Nous générons des modèles dont les erreurs de mesure ont une variance de

$$\mathcal{G}(\xi_m) := (1+\xi_m)\sigma_u^2$$

et nous calculons l'estimateur du coefficient de régression pour ξ_m, avec $m = 1, \ldots, M$, en procédant à une régression des moindres carrés. Nous répétons cette opération K fois pour chaque ξ_m et nous estimons $\hat{\beta}_{w,m}$, l'estimateur de β_w pour ξ_m, par la moyenne empirique

$$\frac{1}{K}\sum_{k=1}^{K} \hat{\beta}_{w,m,k}.$$

Cette méthode qui permet de réaliser une estimation par la moyenne empirique s'appelle méthode de Monte-Carlo et est expliquée plus en détail dans la section suivante. Remarquons que $\hat{\beta}_{w,1}$ correspond à l'estimateur dans le modèle de régression naïf puisqu'il s'agit du cas où $\xi_1 = 0$.

Une fois les valeurs de $\hat{\beta}_{w,m}$ obtenues pour chaque $m = 1, \ldots, M$, nous les disposons sur un graphe dont l'axe des abscisses correspond aux valeurs de $\{\xi_1, \ldots, \xi_M\}$. L'idée du SIMEX pour retrouver l'estimateur du paramètre β_x dans le vrai modèle sans erreur de mesure est de réaliser une extrapolation des valeurs jusqu'en $\xi = -1$.

En effet, il ne nous est pas permis de retirer de la variance dans les données, c'est pourquoi nous allons ajouter de la variance étape par étape pour finalement réaliser une extrapolation jusqu'au cas où nous n'aurions plus de variance σ_u^2. Il s'agit en effet de la valeur du paramètre ξ qui décrit le modèle en l'absence d'erreur de mesure, c'est-à-dire $\xi = -1$ puisque

$$\mathcal{G}(-1) = (1-1)\sigma_u^2 = 0.$$

La FIGURE 2.1 illustre le principe du SIMEX avec le même modèle de régression que celui considéré dans l'exemple de la FIGURE 1.1 où $\beta_x = 3$. L'ensemble des valeurs du paramètre ξ choisi est $\{0, 0.5, 1, 1.5, 2\}$. Pour chacune des valeurs de cet ensemble, le nombre d'estimateurs que nous calculons pour chaque ξ_m est $K = 1000$, qui est donc le nombre total de modèles générés pour chaque ξ_m. Nous calculons ensuite pour $m = 1, \ldots, M$,

$$\hat{\beta}_{w,m} = \frac{1}{1000} \sum_{k=1}^{1000} \hat{\beta}_{w,m,k}.$$

Ces valeurs sont ensuite tracées en noir sur la FIGURE 2.1. Nous pouvons notamment observer la diminution de la valeur de l'estimateur lorsque la variance totale de l'erreur augmente, ce qui montre bien le phénomène d'atténuation par le facteur $\lambda \leq 1$ expliqué dans la première partie de ce mémoire. La droite rouge pointillée décrit quant à elle l'extrapolation que nous devrions idéalement obtenir grâce à l'algorithme du SIMEX. Elle devrait donc atteindre la valeur de $\beta_x = 3$ en $\xi = -1$. Attention, il ne s'agit pas de la vraie valeur de l'extrapolation obtenue par un algorithme. Il s'agit d'un graphe purement illustratif.

2.1.3 Algorithme du SIMEX

Dans cette section, nous allons passer en revue l'algorithme du SIMEX dans plusieurs cas. Ceux-ci se différencient par la variance σ_u^2 des erreurs de mesure. En effet, celle-ci peut être connue ou inconnue et nous pouvons également nous retrouver dans le cas homoscédastique (variance identique) ou hétéroscédastique (variances différentes). La théorie détaillée dans cette section provient essentiellement de la lecture de la référence [3]. Des résultats découlant de simulations numériques illustreront aussi l'algorithme du SIMEX. Ces simulations sont basées sur la lecture des références [6] et [13]. La première aborde également le cas particulier du SIMEX dans le cas discret et la seconde dans le cadre d'une régression logistique. Nous ne nous pencherons pas sur ces cas dans ce mémoire.

Variance des erreurs connue

Cas homoscédastique : Supposons que la variance des erreurs de mesure U est connue et vaut σ_u^2. Nous nous plaçons dans le cas homoscédastique où l'ensemble des erreurs de mesures $\{U_i\}_{i=1}^n$ possède la même variance. Notons θ le paramètre d'intérêt à estimer. Tout d'abord, nous devons réaliser les simulations qui augmentent

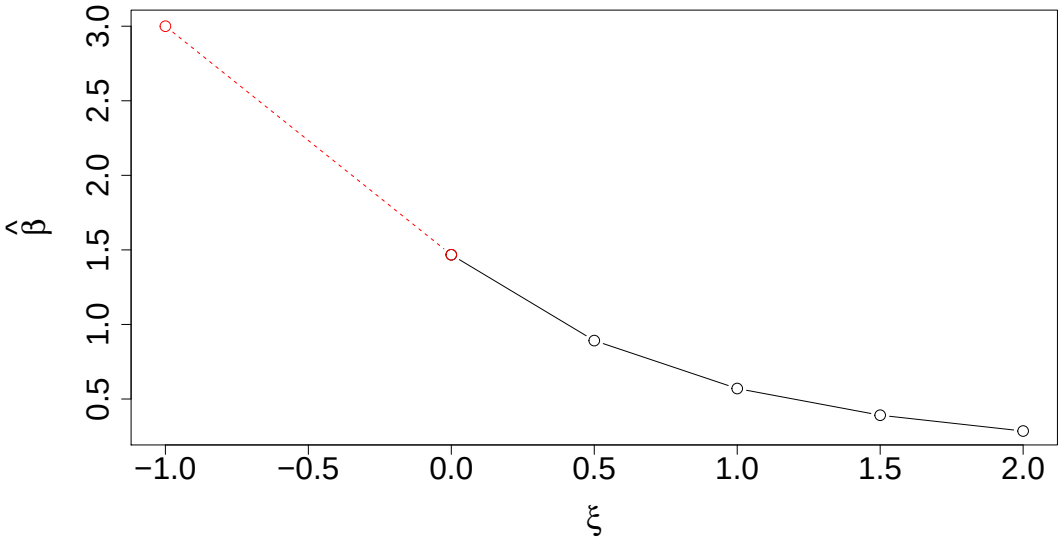

FIGURE 2.1 – Illustration de la méthode du SIMEX sur le modèle de la FIGURE 1.1. En noir, les valeurs obtenues pour $\hat{\beta}_x$ en fonction des valeurs de ξ et en rouge, l'extrapolation que devrait donner idéalement l'algorithme du SIMEX en $\xi = -1$, c'est-à-dire $\beta_x = 3$.

étape par étape la variance des erreurs de mesure en faisant varier le paramètre ξ. La variance totale des erreurs dans le modèle est pour rappel calculée par la fonction

$$\mathcal{G}(\xi) = (1 + \xi)\sigma_u^2.$$

Le processus SIMEX commence avec ce que nous appelons le pas de simulation qui va générer de nouvelles données à partir des observations W_i en notre possession. Leurs valeurs sont calculées, $\forall \xi \geq 0$, par

$$W_{i,k}(\xi) = W_i + \sqrt{\xi}\, U_{i,k},$$

avec $i = 1, \ldots, n$ et $k = 1, \ldots, K$, où n est la taille de notre échantillon et $\{U_{i,k}\}_{i=1}^{n}$ sont mutuellement indépendantes, indépendantes des observations et normalement distribuées de moyenne 0 et de variance σ_u^2. Cette formule, qui ajoute les erreurs de mesure d'un facteur $\sqrt{\xi}$, augmente bien la variance des observations W de la variance des erreurs de mesure σ_u^2 mulitpliée par ξ (propriété de la variance). Cela nous ramène ainsi bel et bien à la formule de $\mathcal{G}(\xi)$. Pour chaque valeur du paramètre ξ, tout l'échantillon est regénéré K fois via cette formule. Nous avons donc créé K ensembles de n observations au total pour chaque valeur de ξ.

La propriété fondamentale de ces mesures est que leur erreur quadratique moyenne sachant X_i converge vers 0 lorsque $\xi \longrightarrow -1$. En effet,

$$\begin{aligned}
\mathrm{MSE}(W_{i,k}(\xi)|X_i) &= \mathrm{E}\left(\left[W_{i,k}(\xi) - X_i\right]^2 |X_i\right) \\
&= \mathrm{E}\left(W_{i,k}(\xi)^2|X_i\right) + \mathrm{E}\left(X_i^2|X_i\right) - 2\ \mathrm{E}(W_{i,k}(\xi)X_i|X_i) \\
&= \mathrm{E}\left(W_{i,k}(\xi)^2|X_i\right) + X_i^2 - 2X_i\ \mathrm{E}(W_{i,k}(\xi)|X_i) \\
&= (1+\xi)\sigma_u^2 + X_i^2 + X_i^2 - 2X_i^2 \qquad (2.2) \\
&= (1+\xi)\sigma_u^2 \xrightarrow[\xi\to -1]{} 0.
\end{aligned}$$

L'étape (2.2) s'explique par les résultats

$$\begin{aligned}
\mathrm{E}(W_{i,k}(\xi)|X_i) &= \mathrm{E}(W_i + \sqrt{\xi}\ U_{i,k}|X_i) \\
&= \mathrm{E}(W_i|X_i) + \sqrt{\xi}\ \mathrm{E}(U_{i,k}|X_i) \\
&= \mathrm{E}(W_i|X_i) + 0 \\
&= X_i,
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left(W_{i,k}(\xi)^2|X_i\right) &= \mathrm{Var}(W_{i,k}(\xi)|X_i) + \left(\mathrm{E}(W_{i,k}(\xi)|X_i)\right)^2 \\
&= \mathrm{Var}(W_i + \sqrt{\xi}\ U_{i,k}|X_i) + X_i^2 \\
&= \mathrm{Var}(W_i|X_i) + \xi\ \mathrm{Var}(U_{i,k}|X_i) + X_i^2 \\
&= \sigma_u^2 + \xi\sigma_u^2 + X_i^2 \\
&= (1+\xi)\sigma_u^2 + X_i^2.
\end{aligned}$$

Définissons $\hat{\theta}_{m,k}(\xi)$ l'estimateur obtenu du paramètre d'intérêt θ pour le k^e ensemble de données générées par le pas de simulation pour ξ_m. L'estimateur du paramètre θ pour ξ_m est alors la moyenne empirique de l'ensemble $\{\hat{\theta}_{m,k}\}_{k=1}^K$, c'est-à-dire

$$\hat{\theta}_m(\xi) = \frac{1}{K}\sum_{k=1}^{K} \hat{\theta}_{m,k}(\xi).$$

Il est important de réaliser l'estimation du paramètre sur plusieurs échantillons de même taille. En effet, cela permet de réaliser des simulations de Monte-Carlo, c'est-à-dire des simulations utilisant de l'aléatoire et qui, lorsqu'elles sont répétées un grand nombre de fois, permettent d'approcher au mieux la valeur réelle du paramètre par la moyenne empirique. Cette dernière converge en effet presque sûrement vers l'espérance par la loi forte des grands nombres. Cela permet alors de prendre en compte le biais intégré dans le modèle par les erreurs de mesure sans y ajouter de la variabilité dû à l'aléatoire. Une méthode de Monte-Carlo connue est celle permettant de calculer l'aire d'un disque en générant un grand nombre de fois des échantillons aléatoires uniformément distribués et en réalisant la moyenne empirique des rapports entre les valeurs dans et en dehors du disque pour chaque échantillon. Cet exemple permet aussi d'avoir une valeur approchée de π. Plus d'informations à propos de la

méthode sont disponibles dans la référence [11].

Remarquons que dans la régression linéaire multiple, où nous tentons d'estimer un paramètre à plusieurs composantes, l'extrapolation réalisée par le SIMEX se fait composante par composante.

Cas hétéroscédastique : Si nous nous trouvons dans le cas hétéroscédastique, c'est-à-dire que les variances des erreurs de mesure ne sont pas toutes égales, mais toujours avec des variances connues, seul le pas de simulation des nouvelles données est affecté. Celle-ci se traduit alors, $\forall \xi \geq 0$, par la formule

$$W_{i,k}(\xi) = W_i + \sqrt{\xi}\, U_{i,k},$$

avec $i = 1, \ldots, n$ et $k = 1, \ldots, K$, où n est la taille de notre échantillon et K est le nombre d'ensembles de données générées par l'algorithme pour chaque ξ_m, avec $\{U_{i,k}\}_{i=1}^{n}$ mutuellement indépendantes, indépendantes des observations et normalement distribuées de moyenne 0 et de variance $\sigma_{u,i}^2$.

Le reste du processus reste similaire au cas homoscédastique décrit précédemment. Le cas hétéroscédastique survient lorsque nous réalisons plusieurs mesures pour chaque observation W_i. La j^e mesure pour l'observation i est notée $W_{i,j}$, où $j = 1, \ldots, k_i$, avec $k_i \geq 1$ le nombre de mesures effectuées pour la i^e observation de l'échantillon. Notons $\overline{W}_{i,\cdot}$ la moyenne de ces mesures, c'est-à-dire pour $i = 1, \ldots, n$,

$$\overline{W}_{i,\cdot} = \frac{1}{k_i} \sum_{j=1}^{k_i} W_{i,j}.$$

Comme la variance vaut σ_u^2 pour chacune des mesures $W_{i,j}$,

$$\begin{aligned}
\mathrm{Var}(\overline{W}_i, \cdot) &= \mathrm{Var}\left(\frac{1}{k_i} \sum_{j=1}^{k_i} W_{i,j} \right) \\
&= \frac{1}{k_i^2} \sum_{j=1}^{k_i} \mathrm{Var}(W_{i,j}) \\
&= \frac{1}{k_i^2}\ k_i\ \mathrm{Var}(W_{i,1}) \\
&= \frac{\sigma_u^2}{k_i} := \sigma_{u,i}^2.
\end{aligned}$$

En suivant les étapes de la même manière que le cas précédent, nous obtenons pour l'erreur quadratique moyenne la même conclusion

$$\mathrm{MSE}(W_{i,k}(\xi)|X_i) = (1+\xi)\sigma_{u,i}^2 \xrightarrow[\xi \to -1]{} 0.$$

Variance des erreurs inconnue

Nous nous plaçons maintenant dans le cas d'un modèle où les erreurs sont hétéroscédastiques et dont les variances sont inconnues. Pour identifier ces variances $\sigma^2_{u,i}$, il est nécessaire de réaliser $k_i \geq 2$ mesures de la i^e observation de l'échantillon. La j^e mesure de l'observation i est alors décrite par

$$W_{i,j} = X_i + U_{i,j},$$

avec $j = 1, \ldots, k_i$, où $U_{i,j}$ sont les erreurs de mesure distribuées suivant une loi normale de moyenne 0 et de variance $\sigma^2_{u,i}$, mutuellement indépendantes et indépendantes des observations.

Soit $c_{i,m} = ((c_{i,m})_1, \ldots, (c_{i,m})_{k_i})$ un vecteur normalisé de taille k_i, appelé vecteur de contraste, tel que

$$\sum_{j=1}^{k_i} (c_{i,m})_j = 0 \quad \text{et} \quad \sum_{j=1}^{k_i} (c_{i,m})_j^2 = 1. \tag{2.3}$$

Alors les nouvelles mesures nous sont données, $\forall \xi > 0$, par la formule

$$W_{k,m}(\xi) = \overline{W}_i, . + \sqrt{\frac{\xi}{k_i}} \sum_{j=1}^{k_i} (c_{i,m})_j \; W_{i,j},$$

avec $i = 1, \ldots, n$ et $k = 1, \ldots, K$, où n est la taille de l'échantillon et K est le nombre d'ensembles de données utilisés dans l'algorithme pour ξ_m. Le vecteur de contraste $c_{i,m}$ permet de faire apparaître le facteur $(1+\xi)$ dans la variance de l'erreur de mesure comme dans les cas précédents. Il effectue une combinaison linéaire des nouvelles données. Réalisons le même calcul sur l'erreur quadratique moyenne comme dans les cas précédents.

$$\begin{aligned}
\mathrm{MSE}(W_{i,m}(\xi)|X_i) &= \mathrm{E}\left([W_{i,m}(\xi) - X_i]^2 \, |X_i \right) \\
&= \mathrm{E}\left(W_{i,m}(\xi)^2 | X_i \right) + \mathrm{E}\left(X_i^2 | X_i \right) - 2\mathrm{E}(W_{i,m}(\xi) X_i | X_i) \\
&= \mathrm{E}\left(W_{i,m}(\xi)^2 | X_i \right) + X_i^2 - 2X_i \; \mathrm{E}(W_{i,m}(\xi) | X_i) \\
&= (1+\xi)\frac{\sigma^2_{u,i}}{k_i} + X_i^2 + X_i^2 - 2X_i^2 \\
&= (1+\xi)\frac{\sigma^2_{u,i}}{k_i} \xrightarrow[\xi \to -1]{} 0.
\end{aligned} \tag{2.4}$$

L'étape (2.4) s'explique par les résultats suivants.

$$\begin{aligned}
\mathrm{E}(W_{i,m}(\xi)|X_i) &= \mathrm{E}\Big(\overline{W}_{i,\cdot} + \sqrt{\frac{\xi}{k_i}}\sum_{j=1}^{k_i}(c_{i,m})_j\ W_{i,j}|X_i\Big) \\
&= \mathrm{E}(\overline{W}_{i,\cdot}|X_i) + \sqrt{\frac{\xi}{k_i}}\sum_{j=1}^{k_i}(c_{i,m})_j\ \mathrm{E}(W_{i,j}|X_i) \\
&= \frac{1}{k_i}\sum_{j=1}^{k_i}\mathrm{E}(W_{i,j}|X_i) + \sqrt{\frac{\xi}{k_i}}\sum_{j=1}^{k_i}(c_{i,m})_j\ \mathrm{E}(W_{i,j}|X_i) \\
&= \frac{1}{k_i}\sum_{j=1}^{k_i}X_i + \sqrt{\frac{\xi}{k_i}}X_i\sum_{j=1}^{k_i}(c_{i,m})_j \\
&= X_i.
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}\left(W_{i,m}(\xi)^2|X_i\right) &= \mathrm{Var}(W_{i,m}(\xi)|X_i) + \left(\mathrm{E}(W_{i,m}(\xi)|X_i)\right)^2 \\
&= \mathrm{Var}\Big(\overline{W}_{i,\cdot} + \sqrt{\frac{\xi}{k_i}}\sum_{j=1}^{k_i}(c_{i,m})_j\ W_{i,j}|X_i\Big) + X_i^2 \\
&= \frac{1}{k_i^2}\sum_{j=1}^{k_i}\mathrm{Var}(W_{i,j}|X_i) + \frac{\xi}{k_i}\sum_{j=1}^{k_i}(c_{i,m})_j^2\ \mathrm{Var}(W_{i,j}|X_i) + X_i^2 \\
&= \frac{1}{k_i^2}\sum_{j=1}^{k_i}\sigma_{u,i}^2 + \frac{\xi}{k_i}\sigma_{u,i}^2 + X_i^2 \\
&= (1+\xi)\frac{\sigma_{u,i}^2}{k_i} + X_i^2.
\end{aligned}$$

Le processus de simulation des nouvelles données nécessite de générer des vecteurs de contraste aléatoirement en générant des valeurs $Z_{i,m,1}, ..., Z_{i,m,k_i}$ suivant une loi $\mathcal{N}(0,1)$. En définissant les composantes du vecteur de contraste comme

$$(c_{i,m})_j := \frac{Z_{i,m,j} - \overline{Z}_{i,m,\cdot}}{\sqrt{\sum_{j=1}^{k_i}\left(Z_{i,m,j} - \overline{Z}_{i,m,\cdot}\right)^2}},$$

les propriétés (2.3) du vecteur sont bien vérifiées et la distribution de ce vecteur est uniforme sur l'ensemble des vecteurs de contraste de dimension k_i. [3]

2.1.4 Simulations numériques

Observons le résultat de l'application du SIMEX sur un exemple réalisé grâce au logiciel *R* par la fonction *simex*. Le modèle auquel nous appliquons cette commande

ξ	-1	0	0.5	1	1.5	2
Linéaire	1.786	1.491	1.194	0.994	0.851	0.749
Quadratique	2.235	1.491	1.194	0.994	0.851	0.749
Non-linéaire	3.023	1.491	1.194	0.994	0.851	0.749

TABLE 2.1 – Valeurs de $\hat{\beta}_{w,m}$ avec $\xi_m = 0, 0.5, 1, 1.5, 2$, pour chaque type d'extrapolation. La valeur de $\hat{\beta}_x$ extrapolée par le SIMEX est reprise dans la colonne pour $\xi = -1$.

est le même que celui utilisé dans la FIGURE 1.1. Nous nous plaçons donc dans le cas homoscédastique avec une variance des erreurs σ_u^2 connue. La méthode d'extrapolation utilisée par le SIMEX est laissé au choix de l'utilisateur. Il s'agit des méthodes d'extrapolation linéaire, quadratique et non linéaire.

La FIGURE 2.2 illustre le résultat obtenu pour les 3 types d'extrapolation. L'ensemble utilisé pour les valeurs de ξ est $\{0, 0.5, 1, 1.5, 2\}$. Pour chacune de ces valeurs est associé une valeur de l'estimateur du paramètre. Il s'agit des courbes noires sur les graphiques. Les courbes rouges pointillées sont les résultats du type d'extrapolation associé à chaque graphe. Nous pouvons observer que l'extrapolation non linéaire fournit la meilleure valeur de l'estimateur en $\xi = -1$ sachant que la vraie valeur du paramètre dans le modèle sans erreur de mesure est $\beta_x = 3$. Nous pouvons l'observer dans la TABLE 2.1 où la valeur de l'estimateur $\hat{\beta}_x$ nous est fournie en rouge dans la première colonne. Nous remarquons que l'extrapolation non linéaire nous donne une valeur de 3.023, ce qui est très proche de la vraie valeur. La mauvaise performance de l'extrapolation linéaire est dû au fait que la courbe du SIMEX obtenue dans la FIGURE 2.2 ne suit pas une allure linéaire à cause du facteur d'atténuation non linéaire en σ_u. En effet, la valeur de $\hat{\beta}_x$ n'atteint que 1.786, bien éloignée de la vraie valeur $\beta_x = 3$.

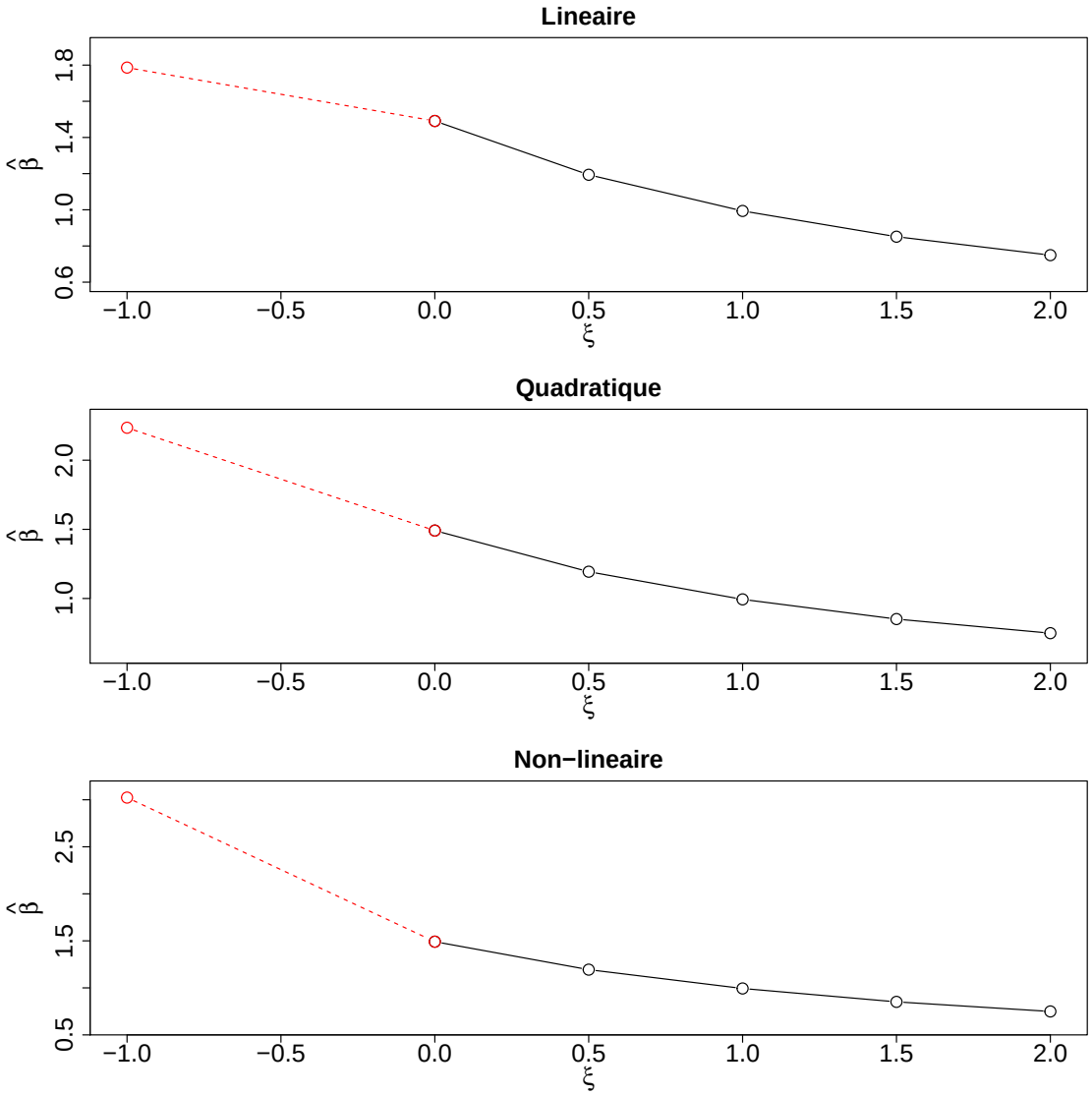

FIGURE 2.2 – Illustration des résultats obtenus par le SIMEX sur le modèle associé à la FIGURE 1.1. Le vrai paramètre $\beta_x = 3$, la variance des erreurs de mesure $\sigma_u^2 = 1$ et la taille de l'échantillon $n = 1000$. Les différents types d'extrapolation sont linéaire, quadratique et non-linéaire.

2.2 Méthode de la variable instrumentale

La méthode des variables instrumentales est une méthode d'estimation du paramètre d'intérêt β_x grâce à la présence d'une variable autre que les régresseurs et appelée *variable instrumentale*. Nous notons dès à présent ce type de variable T.

2.2.1 Concept et définition d'une variable instrumentale

Dans cette section, nous allons passer à travers la construction de la méthode des variables instrumentales qui nous permettra d'estimer le paramètre du modèle de régression.

Soient

$$Y = f(X) + \varepsilon \quad \text{et} \quad W = X + U.$$

Nous pouvons en utilisant les dérivées partielles écrire

$$\frac{\partial f}{\partial X} = \frac{\partial Y}{\partial X} - \frac{\partial \varepsilon}{\partial X} \quad \text{et} \quad \frac{\partial W}{\partial T} = \frac{\partial X}{\partial T} + \frac{\partial U}{\partial T}.$$

Il en découle ainsi directement que

$$\begin{aligned}\frac{\partial f}{\partial X}\frac{\partial W}{\partial T} &= \left(\frac{\partial Y}{\partial X} - \frac{\partial \varepsilon}{\partial X}\right)\left(\frac{\partial X}{\partial T} + \frac{\partial U}{\partial T}\right) \\ &= \frac{\partial Y}{\partial X}\frac{\partial X}{\partial T} - \frac{\partial \varepsilon}{\partial X}\frac{\partial X}{\partial T} + \frac{\partial Y}{\partial X}\frac{\partial U}{\partial T} - \frac{\partial \varepsilon}{\partial X}\frac{\partial U}{\partial T} \\ &= \frac{\partial Y}{\partial T} - \frac{\partial \varepsilon}{\partial T} + \frac{\partial Y}{\partial X}\frac{\partial U}{\partial T} - \frac{\partial \varepsilon}{\partial X}\frac{\partial U}{\partial T}. \end{aligned} \tag{2.5}$$

En posant les trois conditions suivantes

$$\frac{\partial U}{\partial T} = 0, \quad \frac{\partial \varepsilon}{\partial T} = 0, \quad \frac{\partial W}{\partial T} \neq 0, \tag{2.6}$$

l'expression (2.5) devient

$$\begin{aligned}&\frac{\partial f}{\partial X}\frac{\partial W}{\partial T} = \frac{\partial Y}{\partial T} \\ \iff &\frac{\partial f}{\partial X} = \frac{\partial Y/\partial T}{\partial W/\partial T}. \end{aligned} \tag{2.7}$$

Cette dernière égalité peut être traduite en français de la manière suivante. Nous pouvons déterminer la variation de la fonction f décrivant la régression par rapport à la variable X en calculant le rapport entre la variation de la réponse Y par rapport à la variable T et la variation de la variable observée W par rapport à la variable T. Toute la complexité réside donc ici dans l'identification d'une telle variable T qui vérifie les trois conditions en (2.6). Cette variable est alors nommée "variable instrumentale" (IV).

En posant $f(X) = \beta_0 + \beta_x X + \varepsilon$,

$$\frac{\partial f}{\partial X} = \beta_x.$$

Ainsi, la relation (2.7) permet de trouver un estimateur du paramètre β_x du modèle sans erreur de mesure.

Les trois conditions décrites en (2.6) définissent la variable instrumentale T et peuvent être traduites comme suit : la variable instrumentale T n'a pas de relation avec la variable d'erreur U ni du terme d'erreur de la régression ε mais est liée à la variable explicative observée W (équivalent à dire que T est liée à la variable X).

Définition 1. *Une variable T est dite variable instrumentale (IV) si et seulement si*

1. *T n'est pas indépendante de X,*
2. *T est indépendante de U,*
3. *T est indépendante de $Y - \mathrm{E}(Y|Z, X)$.*

Remarquons que la dernière condition est équivalente à l'indépendance entre la variable T et le terme d'erreur ε. En effet, puisque $\mathrm{E}(\varepsilon|Z, X) = 0$,

$$\begin{aligned} Y &= \beta_0 + \beta_x X + \beta_z Z + \varepsilon \\ \mathrm{E}(Y|Z, X) &= \beta_0 + \beta_x X + \beta_z Z + \mathrm{E}(\varepsilon|Z, X) \\ Y - \mathrm{E}(Y|Z, X) &= \varepsilon - \mathrm{E}(\varepsilon|Z, X) = \varepsilon. \end{aligned}$$

Un exemple possible de variable instrumentale dans un modèle est une variable dont la mesure a été répétée via une autre méthode.

Dans la suite de ce chapitre et pour plus de facilité, nous allons introduire des notations permettant d'identifier sans ambiguïté les différents coefficients se rapportant aux différentes variables X, Z, T ou encore le terme indépendant β_0. L'équation de régression à plusieurs dimensions avec laquelle nous travaillons est alors notée

$$(\beta_0, \beta_z, \beta_x) \xrightarrow{\text{notation}} \left(\beta_{Y|\underline{1}ZX}, \beta_{Y|1\underline{Z}X}, \beta_{Y|1Z\underline{X}}\right).$$

Nous lirons donc le coefficient $\beta_{Y|1\underline{Z}X}$ comme le coefficient de la variable Z dans le modèle de régression mettant la réponse Y en relation avec les régresseurs Z et X. La lecture des autres coefficients se fait de manière analogue. Le vecteur contenant ces trois éléments peut tout simplement s'exprimer sous la forme réduite $\beta_{Y|\underline{1Z}T}$. Il est tout a fait possible avec cette notation de ne considérer par exemple que les deux premières composantes de ce vecteurs en écrivant $\beta_{Y|\underline{1}\underline{Z}T}$.

Enfin, introduisons également ces dernières notations

$$\begin{aligned}\widetilde{X} &:= (1, Z^t, X^t)^t \\ \widetilde{T} &:= (1, Z^t, T^t)^t \\ \widetilde{W} &:= (1, Z^t, W^t)^t \\ \widetilde{U} &:= \widetilde{W} - \widetilde{X}.\end{aligned}$$

2.2.2 Description de la méthode des IV

Rappelons que notre vrai modèle de régression linéaire simple est décrit par

$$Y = \beta_{Y|\underline{1}X} + \beta_{Y|1\underline{X}} X + \varepsilon,$$

où Y est la variable de réponse, X est la variable explicative inobservable et ε est le terme d'erreur. Rappelons également que notre variable explicative observable est

$$W = X + U,$$

où U est la variable d'erreur de mesure d'espérance nulle. Toutes ces variables aléatoires possèdent des variances finies.

Notons

$$\begin{aligned}\sigma_{ty} &:= \mathrm{Cov}(T, Y), \\ \sigma_{tw} &:= \mathrm{Cov}(T, W), \\ \sigma_{tx} &:= \mathrm{Cov}(T, X).\end{aligned}$$

En calculant

$$\begin{aligned}\sigma_{ty} &= \mathrm{Cov}(T, Y) \\ &= \mathrm{Cov}\left(T, \beta_{Y|\underline{1}X} + \beta_{Y|1\underline{X}} X + \varepsilon\right) \\ &= 0 + \beta_{Y|1\underline{X}} \mathrm{Cov}(T, X) + \mathrm{Cov}(T, \varepsilon) \\ &:= \beta_{Y|1\underline{X}} \sigma_{tx} + \sigma_{t\varepsilon},\end{aligned}$$

$$\begin{aligned}\sigma_{tw} &= \mathrm{Cov}(T, W) \\ &= \mathrm{Cov}(T, X + U) \\ &:= \sigma_{tx} + \sigma_{tu},\end{aligned}$$

nous obtenons

$$\frac{\sigma_{ty}}{\sigma_{tw}} = \frac{\beta_{Y|1\underline{X}} \sigma_{tx} + \sigma_{t\varepsilon}}{\sigma_{tx} + \sigma_{tu}}. \tag{2.8}$$

En imposant les conditions

$$\sigma_{t\varepsilon} = 0 \;\; ; \;\; \sigma_{tu} = 0 \;\; ; \;\; \sigma_{tx} \neq 0, \tag{2.9}$$

la formule (2.8) devient

$$\frac{\sigma_{ty}}{\sigma_{tw}} = \beta_{Y|1\underline{X}}.$$

Ainsi, l'estimateur IV du paramètre d'intérêt $\beta_{Y|1\underline{X}}$ est défini par

$$\hat{\beta}^{IV}_{Y|1\underline{X}} = \frac{\widehat{\sigma}_{ty}}{\widehat{\sigma}_{tw}}.$$

Notons cependant qu'il convient de mettre l'accent sur les conséquences sérieuses d'une mauvaises identification de la variable instrumentale dans le modèle. Il est nécessaire de s'assurer du respect des trois conditions (2.9) par la variable instrumentale T et qui figurent d'ailleurs explicitement dans la définition de celle-ci. En effet, prenons l'exemple où nous supposons que la variable T faillit à la première condition, c'est-à-dire $\sigma_{t\varepsilon} \neq 0$, alors

$$\hat{\beta}^{IV}_{Y|1\underline{X}} = \frac{\widehat{\sigma}_{ty}}{\widehat{\sigma}_{tw}} = \beta_{Y|1\underline{X}} + \frac{\widehat{\sigma}_{t\varepsilon}}{\widehat{\sigma}_{tx}}. \tag{2.10}$$

Il apparaît donc clairement qu'un biais intervient directement dans l'estimation du paramètre $\beta_{Y|1\underline{X}}$. Ce biais est égal au rapport entre la covariance de T et ε et la covariance de T et X. De plus, au plus cette dernière covariance est proche de 0, au plus le biais est élevé et donc au plus l'estimation correcte du paramètre est clairement menacée. Cette conséquence nous oblige ainsi à être certain de soi dans la désignation d'une variable comme variable instrumentale.

Nous ferons également attention à ce que σ_{tw}, présent au dénominateur de l'estimateur, soit différent de 0 pour garantir son existence. Ainsi, nous pourrons réaliser un test d'hypothèse où l'hypothèse nulle H_0 considère que $\sigma_{tw} = 0$.

2.2.3 Cas avec plusieurs variables instrumentales

Dans cette section, nous nous intéressons au cas où nous sommes en présence de plusieurs variables instrumentales. Nous considérons que nous avons autant de variables instrumentales que de régresseurs inobservables, c'est-à-dire de composantes dans X. Il est tout à fait possible de considérer la présence d'un plus grand nombre de variables instrumentales que de composantes dans X mais nous ne considérerons pas ce cas-là dans ce mémoire. La théorie s'y rapportant peut être consultée dans le chapitre 6 de la référence [3].

Supposons que

$$Y = \widetilde{X}^t \beta_{Y|\underline{\widetilde{X}}} + \varepsilon \quad \text{et} \quad \widetilde{W} = \widetilde{X} + \widetilde{U},$$

où ε et U sont de moyennes nulles et où toutes les variables ont des moments finis d'ordre 2.

L'objectif est donc de déterminer l'estimateur de $\beta_{Y|\underline{\widetilde{X}}}$. Pour cela, nous allons calculer de manière analogue à la section précédente les expressions de la covariance entre $\widetilde{T}$ et

Y ainsi que de la covariance entre $\widetilde{T}$ et $\widetilde{W}$, où $\widetilde{T}$ et $\widetilde{W}$ sont maintenant des matrices contenant des 1 en première ligne, $Z_1, \ldots, Z_{d_z}$ en deuxième ligne et $T_1, \ldots, T_{d_t}$ ou $W_1, \ldots, W_{d_w}$ en dernière ligne respectivement. Notons que d_z, d_t et d_w correspondent au nombre de composantes pour les variables Z, T et W respectivement.

Calculons les deux quantités suivantes

$$\begin{aligned}\Sigma_{\widetilde{T}\widetilde{X}} &:= \mathrm{Cov}(\widetilde{T}, \widetilde{X}) \\ &= \mathrm{E}(\widetilde{T}\widetilde{X}^t) \\ &= \mathrm{E}(\widetilde{T}\widetilde{W}^t) - \mathrm{E}(\widetilde{T}\widetilde{U}^t) \\ &= \Sigma_{\widetilde{T}\widetilde{W}} - \Sigma_{\widetilde{T}\widetilde{U}},\end{aligned}$$

$$\begin{aligned}\Sigma_{\widetilde{T}Y} &:= \mathrm{Cov}(\widetilde{T}, Y) \\ &= \mathrm{E}(\widetilde{T}Y) \\ &= \mathrm{E}\left(\widetilde{T}\left(\widetilde{X}^t\beta_{Y|\underline{\widetilde{X}}} + \varepsilon\right)\right) \\ &= \beta_{Y|\underline{\widetilde{X}}}\mathrm{E}(\widetilde{T}\widetilde{X}^t) + \mathrm{E}(\widetilde{T}\varepsilon) \\ &:= \beta_{Y|\underline{\widetilde{X}}}\Sigma_{\widetilde{T}\widetilde{X}} + \Sigma_{\widetilde{T}\varepsilon}.\end{aligned}$$

Remarquons que $\mathrm{E}(\widetilde{T})\mathrm{E}(\widetilde{X}) = 0$ car $\mathrm{E}(\widetilde{T}) = 0$. En effet, $\widetilde{T}$ possède une colonne remplie de 1, ce qui nous permet de considérer, sans nuire à la généralité, une espérance de $\widetilde{T}$ nulle (voir, par exemple, [3]). Si tel n'était pas le cas, nous pourrions modifier la colonne de constantes pour y arriver.

Grâce à ces deux formules, nous obtenons que

$$\begin{aligned}\left(\Sigma^t_{\widetilde{T}\widetilde{W}}\Sigma_{\widetilde{T}\widetilde{W}}\right)^{-1}\Sigma^t_{\widetilde{T}\widetilde{W}}\Sigma_{\widetilde{T}Y} &= \left(\Sigma^t_{\widetilde{T}\widetilde{W}}\Sigma_{\widetilde{T}\widetilde{W}}\right)^{-1}\Sigma^t_{\widetilde{T}\widetilde{W}}\left(\Sigma_{\widetilde{T}\widetilde{X}}\beta_{Y|\underline{\widetilde{X}}} + \Sigma_{\widetilde{T}\varepsilon}\right) \\ &= \left(\Sigma^t_{\widetilde{T}\widetilde{W}}\Sigma_{\widetilde{T}\widetilde{W}}\right)^{-1}\Sigma^t_{\widetilde{T}\widetilde{W}}\left(\left(\Sigma_{\widetilde{T}\widetilde{W}} - \Sigma_{\widetilde{T}\widetilde{U}}\right)\beta_{Y|\underline{\widetilde{X}}} + \Sigma_{\widetilde{T}\varepsilon}\right). \quad (2.11)\end{aligned}$$

Enfin, en imposant les conditions

$$\Sigma_{\widetilde{T}\varepsilon} = 0 \;\; ; \;\; \Sigma_{\widetilde{T}\widetilde{U}} = 0 \;\; ; \;\; \mathrm{rang}(\Sigma_{\widetilde{T}\widetilde{X}}) = \dim(\widetilde{X}) \quad (2.12)$$

et comme

$$\mathrm{rang}(\Sigma_{\widetilde{T}\widetilde{X}}) = \mathrm{rang}(\Sigma_{\widetilde{T}\widetilde{W}}),$$

la formule (2.11) devient

$$\left(\Sigma^t_{\widetilde{T}\widetilde{W}}\Sigma_{\widetilde{T}\widetilde{W}}\right)^{-1}\Sigma^t_{\widetilde{T}\widetilde{W}}\Sigma_{\widetilde{T}Y} = \beta_{Y|\underline{\widetilde{X}}}.$$

Ainsi, l'estimateur IV du paramètre d'intérêt $\beta_{Y|\underline{\widetilde{X}}}$ est défini par

$$\hat{\beta}^{IV}_{Y|\underline{\widetilde{X}}} = \left(\widehat{\Sigma}^t_{\widetilde{T}\widetilde{W}}\widehat{\Sigma}_{\widetilde{T}\widetilde{W}}\right)^{-1}\widehat{\Sigma}^t_{\widetilde{T}\widetilde{W}}\widehat{\Sigma}_{\widetilde{T}Y},$$

où $\widehat{\Sigma}_{\widetilde{T}\widetilde{W}} := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\widetilde{T}_i\widetilde{W}_i^t$ est un estimateur de $\Sigma_{\widetilde{T}\widetilde{W}}$.

2.2.4 Construction de l'algorithme

Dans cette section, nous allons détailler les étapes de l'algorithme permettant de retrouver l'estimateur du paramètre d'intérêt $\hat{\beta}_{Y|\widetilde{X}}$, c'est-à-dire les coefficients de la régression du modèle duquel nous avons enlevé les effets des erreurs de mesure. Il s'agit d'un algorithme adapté au cas où $\dim(T) = \dim(X)$. Comme évoqué précédemment, le cas où le nombre de variables instrumentales est plus grand que le nombre de variables inobservables n'est pas abordé et nous renvoyons le lecteur vers la théorie qui s'y rapporte dans [3].

Notons tout d'abord d_z, d_x et d_t la dimension de Z, X et T respectivement.

1. Calculons la matrice $\hat{\beta}^t_{\widetilde{W}|\widetilde{T}}$, de dimension $(1 + d_z + d_x) \times (1 + d_z + d_t)$. Elle s'exprime comme

$$\hat{\beta}^t_{\widetilde{W}|\widetilde{T}} = \begin{pmatrix} \hat{\beta}_{1|1} & \hat{\beta}_{1|Z_1} & \cdots & \hat{\beta}_{1|Z_{d_z}} & \hat{\beta}_{1|T_1} & \cdots & \hat{\beta}_{1|T_{d_t}} \\ \hat{\beta}_{Z_1|1} & \hat{\beta}_{Z_1|Z_1} & & & & & \\ \vdots & & \ddots & & & & \\ \hat{\beta}_{Z_{d_z}|1} & \cdots & & \hat{\beta}_{Z_{d_z}|Z_{d_z}} & & & \\ \hat{\beta}_{W_1|1} & \hat{\beta}_{W_1|Z_1} & \cdots & \hat{\beta}_{W_1|Z_{d_z}} & \hat{\beta}_{W_1|T_1} & \cdots & \hat{\beta}_{W_1|T_{d_t}} \\ \vdots & & & & & & \\ \hat{\beta}_{W_{d_x}|1} & \hat{\beta}_{W_{d_x}|Z_1} & \cdots & \hat{\beta}_{W_{d_x}|Z_{d_z}} & \hat{\beta}_{W_{d_x}|T_1} & \cdots & \hat{\beta}_{W_{d_x}|T_{d_t}} \end{pmatrix}.$$

   Chaque ligne de cette matrice correspond aux coefficients des régressions expliquant chacune des composantes de Z (partie supérieure de la matrice) et de W (partie inférieure de la matrice) par les composantes de Z et de T. Il

apparaît que les $(1 + d_z)$ premières lignes de cette matrice sont en fait des 0 et des 1 puisque les composantes Z_i, pour $i = 1, \ldots, d_z$, peuvent être expliquées entièrement par elles-mêmes. Pour mieux comprendre, prenons par exemple la régression expliquant l'élément Z_1 par (Z, T). Elle s'exprime comme

$$Z_1 = \hat{\beta}_{Z_1|1} + \hat{\beta}_{Z_1|Z_1} Z_1 + \ldots + \hat{\beta}_{Z_1|Z_{d_z}} Z_{d_z} + \hat{\beta}_{Z_1|T_1} T_1 + \ldots + \hat{\beta}_{Z_1|T_{d_t}} T_{d_t}.$$

Ainsi, seul le coefficient $\hat{\beta}_{Z_1|Z_1}$ est non-nul et vaut 1. Le résultat est similaire pour $Z_2, \ldots, Z_{d_z}$ et donc seuls les coefficients

$$\left(\hat{\beta}_{1|1}, \hat{\beta}_{Z_1|Z_1}, \ldots, \hat{\beta}_{Z_{d_z}|Z_{d_z}}\right)$$

sont non-nuls sur leurs lignes et valent tous 1. En d'autres termes, la matrice peut se réexprimer comme

$$\hat{\beta}^t_{\widetilde{W}|\widetilde{T}} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ & & & \\ 0 & \cdots & 0 \; 1 \; 0 & \cdots \; 0 \\ \hat{\beta}_{W_1|1} & \hat{\beta}_{W_1|Z_1} & \cdots & \hat{\beta}_{W_1|Z_{d_z}} \; \hat{\beta}_{W_1|T_1} \; \cdots \; \hat{\beta}_{W_1|T_{d_t}} \\ & & & \\ \hat{\beta}_{W_{d_x}|1} & \hat{\beta}_{W_{d_x}|Z_1} & \cdots & \hat{\beta}_{W_{d_x}|Z_{d_z}} \hat{\beta}_{W_{d_x}|T_1} \; \cdots \; \hat{\beta}_{W_{d_x}|T_{d_t}} \end{pmatrix}.$$

2. Calculons les valeurs de $\hat{\beta}^t_{\widetilde{W}|\widetilde{T}} \widetilde{T}$ et réaliser une régression entre la réponse Y et ces nouvelles valeurs. Les coefficients alors obtenus via cette régression correspondent à $\hat{\beta}_{Y|\widetilde{X}}$.

En d'autres termes, cet algorithme revient à réaliser ce qui est appelé un « two-stage least squares (2SLS) regression ». Il est donc composé comme son nom l'indique de deux étapes de régression des moindres carrés. La première correspond alors à réaliser une régression expliquant W par la variable instrumentale T et la seconde à intégrer ce résultat dans la régression expliquant la réponse Y par la variable X.

	Lecture	Mathématiques	Sciences
FRIENDS	-0.733	-0.340	-2.983 (**)
PISA_{2009}	0.779 (***)	0.882 (***)	0.842 (***)
ESCS	2.274 (**)	2.440 (**)	1.443
COUNT	0.186	0.230	-1.170
VOC	-14.69 (***)	-2.025	-9.460 (**)
BVOC	-39.53 (***)	-30.22 (***)	-34.16 (***)
FEMALE	5.942 (***)	-0.821	-0.312
N	4951	4951	4951
R_a^2	0.793	0.833	0.777

TABLE 2.2 – Résultats de la régression des moindres carrés du modèle décrit en (2.13) provenant de l'article [1] où la variable dépendante est PISA_{2010}.

(*) : $p < 0.1$	(**) : $p < 0.05$	(***) : $p < 0.01$

2.2.5 Exemple de modèle avec une variable instrumentale

Dans cette section, nous allons illustrer l'utilisation de variables instrumentales dans un exemple concret de modèle de régression avec erreurs de mesure. Cet exemple est tiré de l'article [1] et tous les résultats présentés dans cette sous-section proviennent de cette référence.

Le sujet de cet exemple porte sur la performance des étudiants au test PISA (Programme International pour le Suivi des Acquis des Élèves) en lecture, en mathématiques et en sciences. L'objectif est de répondre à la question suivante : comment la présence d'amis dans la classe influence les résultats PISA des élèves ? Nous considérons un modèle de régression qui tente d'expliquer le score PISA de l'année 2010 par différents facteurs. Le modèle de régression s'exprime comme

$$\begin{aligned}\text{PISA}_{2010} = \beta_0 + \beta_1\ \text{FRIENDS} + \beta_2\ \text{PISA}_{2009} + \beta_3\ \text{ESCS} + \beta_4\ \text{COUNT} \\ + \beta_5\ \text{VOC} + \beta_6\ \text{BVOC} + \beta_7\ \text{FEMALE} + \varepsilon,\end{aligned} \tag{2.13}$$

où FRIENDS est le nombre d'amis d'un individu dans sa classe, PISA_{2009} est le score PISA de l'individu lors de l'année précédente (en 2009), ESCS est un indice concernant le statut economique, social et culturel de la famille de l'individu, COUNT est la variable décrivant si l'individu vit dans la campagne ou non (1 ou 0 respectivement), VOC/BVOC sont des variables décrivant le type de l'établissement scolaire de l'individu, FEMALE est le genre féminin ou masculin de l'individu (1 ou 0 respectivement) et ε est le terme d'erreur.

Dans la TABLE 2.2 présentant les résultats de la régression naïve du modèle (2.13), nous remarquons que le régresseur FRIENDS possède des valeurs de coefficients négatives pour les trois scores PISA. Une des interprétations que nous pourrions donner est qu'avoir un ami dans la même classe que soi pourrait inciter à moins travailler

	Lecture	Mathématiques	Sciences
DISTANCE	-0.0027 (***)	-0.0027 (***)	-0.0028 (***)
PISA_{2009}	-0.0005 (***)	-0.0004 (**)	-0.0004 (**)
ESCS	-0.0240 (**)	-0.0240 (**)	-0.0249 (**)
COUNT	0.0047	0.0043	0.0062
VOC	-0.0421 (*)	-0.0415 (*)	-0.0383
BVOC	0.0421	0.0543	0.0564 (*)
FEMALE	-0.0119	-0.0338 (*)	-0.0378
N	4951	4951	4951
Test F	41.7518	39.9005	41.821
R_a^2	0.0272	0.0266	0.0266

TABLE 2.3 – Résultats de la régression des moindres carrés du modèle décrit en (2.14) provenant de l'article [1] où la variable dépendante est FRIENDS.

(*) $: p < 0.1$	(**) $: p < 0.05$	(***) $: p < 0.01$

pour l'école et passer plus de temps à faire autre chose qu'étudier. Néanmoins, ces résultats viennent à l'encontre d'études citées par l'article [1] et décrites dans [4] et [5] sur la présence d'amis dans la classe d'élèves et les résultats scolaires. Il semble donc qu'un biais s'est glissé dans le modèle.

Pour remédier à cela, nous allons introduire dans le modèle la variable instrumentale DISTANCE qui correspond à la distance en minutes entre le lieu de domicile de l'individu et l'établissement scolaire. Cette variable peut être considérée comme une variable instrumentale puisqu'elle est liée à la variable observée FRIENDS (nous sommes amenés à penser qu'au plus la valeur de DISTANCE est faible, au plus la probabilité de trouver des amis dans sa classe est importante). Pour confirmer cette hypothèse, nous observons le modèle qui met en relation la variable FRIENDS et la variable instrumentale DISTANCE comme

$$\begin{aligned}\text{FRIENDS}^* = \alpha_0 + \alpha_1\ \text{DISTANCE} + \alpha_2\ \text{PISA}_{2009} + \alpha_3\ \text{ESCS} + \alpha_4\ \text{COUNT} \\ + \alpha_5\ \text{VOC} + \alpha_6\ \text{BVOC} + \alpha_7\ \text{FEMALE} + e.\end{aligned} \tag{2.14}$$

Il s'agit ici de l'étape 1 de la régression 2SLS comme décrit dans la section 2.2.4. En lisant la TABLE 2.3 nous montrant les résultats de cette régression, nous remarquons que la variable DISTANCE est bien significative dans les trois modèles de régressions dont la formule est décrite en (2.14) avec une p-valeur associée au test de significativité des coefficients inférieure au seuil α pris classiquement à 0.05. De plus, le test F renvoie de grandes valeurs pour les trois modèles. Nous pouvons considérer que la variable instrumentale possède un lien suffisamment fort avec DISTANCE si ce test donne des valeurs supérieures à 10, ce qui est bien le cas dans la TABLE 2.3 (voir [14] et [9], cités dans [1]).

	Lecture	Mathématiques	Sciences
FRIENDS*	6.885	17.24	6.171
PISA_{2009}	0.782 (***)	0.889 (***)	0.845 (***)
ESCS	2.466 (**)	2.878 (***)	1.682
COUNT	0.317	0.533	-1.026
VOC	-14.26 (***)	-1.088	-9.010 (**)
BVOC	-39.74 (***)	-30.86 (***)	-34.54 (***)
FEMALE	6.074 (***)	-0.133	-0.0123
N	4951	4951	4951
R_a^2	0.791	0.825	0.775

TABLE 2.4 – Résultats de la régression des moindres carrés du modèle décrit en (2.15) provenant de l'article [1] où la variable dépendante est PISA_{2010}.

(*) : $p < 0.1$	(**) : $p < 0.05$	(***) : $p < 0.01$

Elle remplit également la condition d'indépendance par rapport à la variable réponse. En effet, l'article [1] dont est tiré cet exemple affirme qu'il n'existe pas de lien entre la distance domicile-école et les résultats scolaires. Nous pouvons le supposer si les valeurs de DISTANCE ne sont pas trop importantes, ce qui est le cas dans cette étude. Cela ne serait vraissemblablement pas le cas dans le sens inverse où une trop grande distance domicile-école engendrerait alors une fatigue chez les élèves impactant ainsi leur apprentissage.

Notre modèle final considère à présent la deuxième étape de la régression 2SLS, c'est-à-dire le modèle

$$\begin{aligned}\text{PISA}_{2010} = \beta_0 + \beta_1\ \text{FRIENDS}^* + \beta_2\ \text{PISA}_{2009} + \beta_3\ \text{ESCS} + \beta_4\ \text{COUNT} \\ + \beta_5\ \text{VOC} + \beta_6\ \text{BVOC} + \beta_7\ \text{FEMALE} + \varepsilon,\end{aligned} \tag{2.15}$$

où la variable FRIENDS* est la variable réponse de la première étape de la régression 2SLS (2.14). Les résultats associés à (2.15) sont présentés dans la TABLE 2.4 où nous pouvons observer que les coefficients associés à la variable FRIENDS* sont devenus positifs, ce qui va dans le sens des études qui montraient que la présence d'amis dans la classe d'un individu favorisait ses résultats scolaires (voir [4] et [5]). Néanmoins, nous observons que le régresseur est devenu non-significatif dans la TABLE 2.4. En effet, l'inconvénient de la méthode des IV est qu'elle augmente la variance de la variable explicative et donc dans de rares cas peut réduire la significativité des régresseurs.

2.2.6 Simulations numériques

Dans cette section, nous allons observer les résultats de l'algorithme expliqué dans la section précédente sur un jeu de données.

Prenons un modèle de régression linéaire simple

$$Y = \beta_0 + \beta_x X + \varepsilon,$$

où $X \sim N(0,1)$, $\varepsilon \sim N(0,1)$ et la taille de l'échantillon vaut 200. Supposons que le modèle est pollué par des erreurs de mesure $U \sim N(0,1)$ telles que $W = X + U$ est la variable réellement observée. Nous calculons également une variable instrumentale T. Dans cet exemple, nous la générons telle que

$$T = 2X + \nu,$$

où $\nu \sim N(0,1)$. Il s'agit bien d'une variable instrumentale par définition puisqu'elle ne dépend que de la variable X.

L'algorithme calcule tout d'abord la matrice $\hat{\beta}^t_{\widetilde{W}|\widetilde{T}}$ comme décrit à l'étape 1. Une fois cette étape réalisée, nous calculons les valeurs des coefficients de la régression entre Y et $\hat{\beta}^t_{\widetilde{W}|\widetilde{T}}\widetilde{T}$ pour obtenir les coefficients de la régression de Y expliquée par X.

Nous répétons cette opération 1000 fois et affichons les valeurs de l'estimateur du paramètre d'intérêt dans un histogramme pour le modèle naïf et pour le modèle prenant en compte les effets des erreurs de mesure grâce à la méthode de la variable instrumentale. Comme nous le voyons à la FIGURE 2.3, l'histogramme du modèle naïf possède des valeurs centrées autour de 1.5, ce qui correspond bien a une atténuation de la vraie valeur du paramètre $\beta_x = 3$. Le facteur d'atténuation vaut en effet ici

$$\frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} = \frac{1^2}{1^2 + 1^2} = \frac{1}{2}.$$

Dans le second histogramme, les valeurs sont bel et bien centrées autour de la vraie valeur du paramètre $\beta_x = 3$ grâce à l'algorithme utilisant la méthode de la variable instrumentale. Nous pouvons néanmois remarquer que la variance de l'estimateur pour le modèle utilisant la variable instrumentale est plus grande que pour le modèle naïf, ce qui corrobore à l'explication donnée dans le perte de significativité du régresseur dans l'exemple de la section précédente.

En plus de biaiser la valeur de l'estimateur du paramètre d'intérêt, la présence d'erreurs de mesure affecte également la puissance du test de significativité comme expliqué dans le section 1.4. En observant la FIGURE 2.4, nous remarquons que la proportion de tests de significativité ne réalisant pas d'erreurs de type II diminue au plus la valeur de σ_u augmente pour la régression naïve avec erreurs de mesure (courbes rouges). Plus la taille de l'échantillon n est faible, plus la puissance du test diminue évidemment. Ce que nous remarquons à présent, c'est que pour le modèle de régression utilisant la méthode de la variable instrumentale pour la correction de l'estimateur (courbes bleues), la puissance du test reste proche de 100% peu importe la valeur de σ_u et même pour une valeur de n peu élevée. Cette méthode de correction de l'estimateur permet donc aussi de supprimer la perte de puissance.

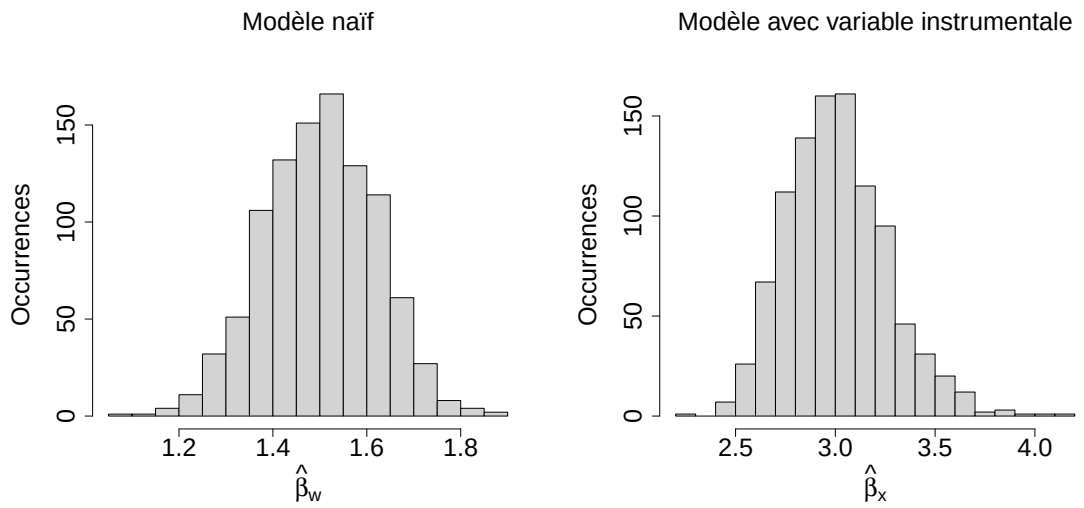

FIGURE 2.3 – Histogrammes des valeurs de $\hat{\beta}_w$ d'une régression linéaire simple avec erreurs de mesure. Le résultat pour le modèle naïf est représenté à gauche et le résultat avec la correction via la méthode de la variable instrumentale à droite. La vraie valeur du paramètre $\beta_x = 3$.

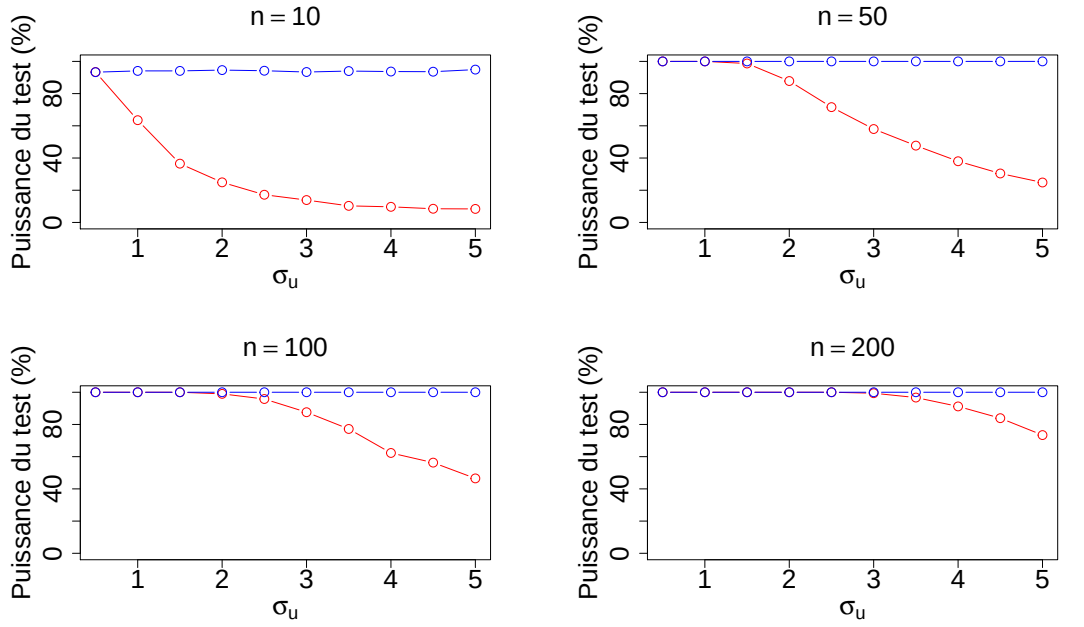

FIGURE 2.4 – Estimation de la puissance du test de significativité dans un modèle de régression avec erreurs de mesure pour la courbe en rouge et avec la correction apportée à l'estimateur par la méthode de la variable instrumentale pour la courbe en bleu, où n est la taille de l'échantillon. Le nombre de modèles utilisés pour estimer la puissance est de 1000 à chaque étape.

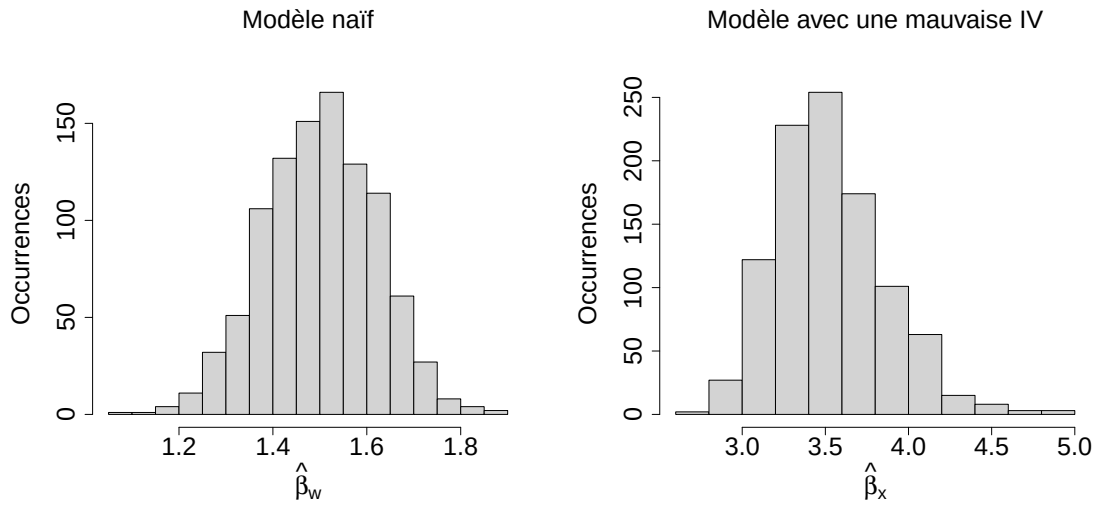

FIGURE 2.5 – Histogramme des valeurs de $\hat{\beta}_w$ d'une régression linéaire simple avec erreurs de mesure où la correction via la méthode de la variable instrumentale se fait avec une variable ne respectant pas le critère $\sigma_{t\varepsilon} = 0$. La vraie valeur du paramètre $\beta_x = 3$.

Pour rappel, il est crucial de s'assurer que les trois critères qui définissent une variable instrumentale soient bien remplis. Illustrons le résultat de la violation d'une des conditions, comme celle évoquée en (2.10) par exemple où $\sigma_{t\varepsilon} \neq 0$. Choisissons, au lieu de la relation $T = 2X + \nu$, une variable instrumentale définie comme

$$T = 2X + \nu + \epsilon.$$

Dès à présent, la variable T dépend du régresseur X mais est également liée à la variable réponse (violation de la première condition dans la définition 1). Il s'agit donc d'un mauvais choix pour une variable instrumentale. Vérifions la conséquence portée sur l'estimation du coefficient de régression grâce à la FIGURE 2.5. Nous observons en effet que les valeurs de l'histogramme se retrouvent centrées non plus en $\beta_x = 3$ mais en 3.5. Cette valeur correspond effectivement à la valeur du paramètre additionnée du biais présent dans la formule (2.10) et qui vaut $\sigma_{t\varepsilon}/\sigma_{tx}$, où $\sigma_{t\varepsilon} = 1$ et $\sigma_{tx} = 2$ dans notre cas.

2.3 Conclusion

Dans ce deuxième chapitre, nous avons explicité deux méthodes dont l'objectif est la correction de l'estimateur du coefficient de régression.

La première méthode, le SIMEX, est une méthode qui se base sur l'extrapolation pour remplir l'objectif. Le désavantage de cette méthode est qu'elle nécessite de connaître ou d'identifier la variance des erreurs de mesure avant de l'appliquer. Aussi, les extrapolations ne donnent pas nécessairement de bons résultats. Nous l'avons

constaté de manière assez flagrante avec l'extrapolation linéaire dans nos simulations numériques.

La seconde méthode, la variable instrumentale, est une méthode qui se base sur l'observation d'une autre variable liée au régresseur mais indépendante des erreurs de mesure et de la variable réponse. Dans nos simulations numériques, nous avons remarqué que l'estimateur corrigé était proche de la vraie valeur du paramètre même s'il présentait une variance plus grande. Le désavantage de cette méthode réside dans une correcte identification d'une variable instrumentale qui peut s'avérer être fatale si telle n'était pas le cas.

Chapitre 3

Erreurs de mesure dans la régression linéaire fonctionnelle

3.1 Introduction

Dans ce troisième chapitre, nous nous intéressons au cas fonctionnel de la régression où les régresseurs fonctionnels ont été contaminés par des erreurs de mesure également fonctionnelles. Dans cette section, nous allons introduire les données fonctionnelles et le concept de régression fonctionnelle. Dans la section suivante, nous réaliserons une mise en contexte des erreurs de mesure dans le cadre d'une régression linéaire fonctionnelle. Nous passerons notamment en revue les hypothèses faites sur le modèle et sur la structure des erreurs de mesure. Ensuite, nous décrirons la théorie permettant d'obtenir une méthode retrouvant un estimateur du paramètre fonctionnel dans le modèle sans erreur de mesure sur base de l'observation du modèle contaminé, objectif similaire au chapitre précédent dans le cas multivarié. Enfin, ce mémoire enchaînera sur le dernier chapitre mettant en pratique l'algorithme sur deux modèles différents tirés de l'article [7].

Jusqu'à présent, dans les précédents chapitres, nous considérions un modèle de régression linéaire multivariée où les variables étaient toutes des variables aléatoires scalaires. A partir de maintenant, le régresseur X et le paramètre à estimer β_X deviennent des fonctions, que nous considérons dépendante du temps t. La variable réponse peut soit être scalaire soit fonctionnelle.

Notre modèle devient donc un modèle de régression linéaire fonctionnelle qui peut s'écrire dans le cas d'une réponse Y scalaire comme

$$Y = \beta_0 + \mathcal{B}_X X + \varepsilon,$$

où $\mathcal{B}_X : L^2[0,1] \longrightarrow \mathbb{R}$, ou d'une autre manière

$$Y = \beta_0 + \int_0^1 \beta_X(t)X(t)\ dt + \varepsilon, \tag{3.1}$$

où β_0 est le terme indépendant (scalaire), $X(t)$ est le régresseur fonctionnel, $\beta_X(t)$ est le paramètre fonctionnel à estimer et ε est le terme d'erreur (scalaire). Ainsi, l'objectif de cette régression est de trouver un estimateur $\hat{\beta}_X(t)$ convergent de $\beta_X(t)$. Pour plus de facilité, nous considérerons dans la suite un modèle de régression fonctionnel sans terme indépendant β_0.

Dans le cas d'une réponse $Y = Y(t)$ fonctionnelle, le modèle de régression est défini par

$$Y(t) = \beta_0(t) + \mathcal{B}_X X + \varepsilon(t),$$

où $\mathcal{B}_X : L^2[0,1] \longrightarrow L^2[0,1]$, ou d'une autre manière

$$Y(t) = \beta_0(t) + \int_0^1 \beta_X(s,t)X(s)\ ds + \varepsilon(t), \tag{3.2}$$

où $\beta_0(t)$ est le terme indépendant, $X(t)$ est le régresseur, $\beta_X(t)$ est le paramètre à estimer et ε est le terme d'erreur. Tous dans ce cas-ci sont fonctionnels. Ainsi, l'objectif de cette régression est de trouver un estimateur $\hat{\beta}_X(s,t)$ convergent de $\beta_X(s,t)$. Pour plus de facilité, nous considérerons dans la suite un modèle de régression fonctionnel sans terme indépendant $\beta_0(t)$.

Néanmoins, il nous est généralement impossible de mesurer de manière continue ces variables sur leur domaine, généralement pris $[0,1]$. C'est pourquoi il nous faut faire une distinction entre la fonction régresseur continue $X = X(t)$, où $t \in [0,1]$, et le régresseur sous forme vectorielle $(X(t_1), \ldots, X(t_L))$, où $0 \leq t_1 \leq \cdots \leq t_L \leq 1$, avec L la taille de la discrétisation du domaine de définition des variables. La même distinction peut être faite sur le paramètre β_X et la variable réponse Y. Nous supposons que les variables sont définies sur un espace de Hilbert H, c'est-à-dire un espace complet pour la norme $||\cdot|| = \langle \cdot, \cdot \rangle^{1/2}$.

3.2 Mise en contexte

Dans cette section, nous allons nous intéresser au cas fonctionnel de la régression comme décrit précédemment mais où des erreurs de mesure se sont glissées parmi les observations du régresseur X. La lecture des articles [7] et [8] ont été nécessaires pour la réalisation de ce chapitre. Nous reprendrons des notations similaires à celles utilisées dans les chapitres précédents. Nous notons $X = X(t)$ le régresseur fonctionnel inobservable sans erreur de mesure, $U = U(t)$ les erreurs de mesures fonctionnelles et $W = W(t)$ la variable fonctionnelle observée entachée des erreurs de mesure telle que $W = X + U$. Comme dans la section précédente, toutes ces variables aléatoires fonctionnelles sont définies sur l'intervalle $[0,1]$ et L est la taille de la grille sur laquelle sont mesurées de manière discrète les variables fonctionnelles et donc la taille des vecteurs associés.

Ainsi, l'observation de la variable fonctionnelle W dans le modèle au lieu de la variable fonctionnelle inobservable X définit la modèle observé comme

$$Y = \beta_0 + \int_0^1 \beta_W(t)W(t)\ dt + \varepsilon, \tag{3.3}$$

pour une réponse Y scalaire et

$$Y(t) = \beta_0(t) + \int_0^1 \beta_W(s,t)W(s)\ ds + \varepsilon(t), \tag{3.4}$$

pour une réponse Y fonctionnelle, où $\beta_W(t)$ et $\beta_W(s,t)$ sont maintenant les paramètres du modèle observé associés à la variable contaminée par les erreurs de mesure U. Le but est alors de retrouver un estimateur $\hat{\beta}_X(t)$ ou $\hat{\beta}_X(s,t)$ qui converge vers $\beta_X(t)$ ou $\beta_X(s,t)$ pour le modèle de régression inobservable sans erreur de mesure (3.1) ou (3.2) en ayant uniquement à disposition le modèle de régression avec erreurs de mesure (3.3) ou (3.4).

Au cours de ce chapitre, nous allons régulièrement utiliser la notion de covariance dans le cas fonctionnel. Nous définissons l'opérateur de covariance associé à la variable fonctionnelle X, noté $\mathcal{K}_X$, comme

$$\begin{aligned} \mathcal{K}_X \ : \ L^2[0,1] &\longrightarrow L^2[0,1] \\ f &\longmapsto \mathcal{K}_X(f) := \mathrm{E}\left(\langle X - \mathrm{E}(X), f\rangle (X - \mathrm{E}(X))\right), \end{aligned}$$

avec $\mathrm{E}(X) \overset{\text{déf}}{=} \displaystyle\int_0^1 X(t)\ dt$. Nous pouvons aussi exprimer cet opérateur de covariance grâce à son noyau, c'est-à-dire sous la forme

$$\mathcal{K}_X(f)(t) := \int_0^1 k_X(s,t)f(s)\ ds,$$

$\forall f \in L^2[0,1]$ et $\forall t \in [0,1]$, où $k_X : (s,t) \in [0,1]^2 \longmapsto k_X(s,t) := \mathrm{Cov}(X(s), X(t))$ est le noyau de l'opérateur de covariance $\mathcal{K}_X$. Nous supposons que le noyau k_X est réel analytique, c'est-à-dire qu'il admet une décomposition de Karhunen-Loève finie (voir [7]), et donc que nous pouvons écrire l'opérateur sous la forme

$$\mathcal{K}_X = \sum_{j=1}^{r} \lambda_j(\eta_j \otimes \eta_j),$$

où $\forall e \in H$, $(\eta_j \otimes \eta_j)(e) = \langle \eta_j, e\rangle \eta_j$, r est le rang fini de $\mathcal{K}_X$, λ_j sont les valeurs propres ordonnées et η_j les fonctions propres orthonormées réelles analytiques de $\mathcal{K}_X$. Pour rappel, une fonction réelle analytique f est une fonction telle que, pour tout point x_0 de son domaine, la série

$$\sum_{n=0}^{\infty} a_n(x - x_0)^n,$$

où $(a_n)_{n\in\mathbb{N}}$ sont des nombres réels, converge vers $f(x)$ où x est dans un voisinage de x_0. Cela permet notamment d'écrire la fonction f sous forme d'une série de Taylor autour de x_0. [15]

Toutes ces définitions peuvent être peuvent également être exprimées pour les opérateurs de covariance des autres variables fonctionnelles U et W, notés $\mathcal{K}_U$ et $\mathcal{K}_W$.

Nous allons faire l'hypothèse suivante sur l'opérateur de covariance de l'erreur de mesure U (voir [7]). L'opérateur $\mathcal{K}_U$ possède une structure bande, c'est-à-dire que $\forall t_i, t_j \in [0,1]$,

$$Cov(U(t_i), U(t_j)) = 0 \quad \text{si} \quad |t_i - t_j| > \delta, \tag{3.5}$$

où le paramètre $\delta \in [0,1]$ est appelé largeur de bande. Ainsi, nous supposons que ces erreurs U sont indépendantes au-delà d'une bande de temps δ.

3.3 Description théorique de la méthode

Pour obtenir un estimateur $\hat{\beta}_X$ du paramètre du modèle sans erreur de mesure, nous calculons l'estimateur des moindres carrés. Pour rappel, cet estimateur dans le cas de la régression linéaire simple s'obtient par la formule

$$\frac{\widehat{Cov}(W,Y)}{\widehat{\mathrm{Var}}(X)},$$

où les variables W, Y et X sont univariées. Ainsi, dans le cadre fonctionnel, l'estimateur des moindres carrés devient

$$\widehat{\mathcal{K}}_X^{-1}\, \widehat{\mathcal{C}}_{W,Y},$$

avec $\widehat{\mathcal{K}}_X^{-1}$ l'inverse de l'opérateur de covariance de la variable fonctionnelle X et où $\widehat{\mathcal{C}}_{W,Y}$ est l'estimateur de covariance entre la variable fonctionnelle W et la réponse Y. Ainsi, pour retrouver notre estimateur $\hat{\beta}_X$, il nous faut estimer l'opérateur $\mathcal{K}_X^{-1}$, ce qui n'est pas trivial à cause de l'inobservabilité de X dans notre modèle avec erreurs de mesure. Cela peut se faire en estimant le noyau de cet opérateur.

Le problème réside donc ici dans l'estimation de $\mathcal{K}_X^{-1}$ en ayant uniquement en notre possession la variable observée W. De plus, trouver un estimateur convergent $\hat{\mathcal{J}}$ de $\mathcal{K}_X$ ne nous garantit pas forcément que son inverse $\hat{\mathcal{J}}^{-1}$ est un estimateur convergent de $\mathcal{K}_X^{-1}$. Introduisons le résultat suivant que nous considérons comme admis (voir [7]).

Proposition 1 (Admis)**.** *La fonction $\mathcal{J} \longrightarrow \mathcal{J}^{-1}$ est continue sur l'espace des opérateurs auto-adjoints de rangs finis si et seulement si $\forall$ suite $(\mathcal{J}_n)$ telle que $\mathcal{J}_n \longrightarrow \mathcal{J}$, $\mathrm{rang}(\mathcal{J}_n) = \mathrm{rang}(\mathcal{J})$, $\forall n$ suffisamment grand.*

Il apparaît donc avec la Proposition 1 que la convergence de l'estimateur vers l'opérateur de covariance $\mathcal{K}_X$ ne suffit plus. Nous devons également vérifier que le rang de cet estimateur coïncide avec le rang de $\mathcal{K}_X$.

Comme il n'est pas possible de mesurer des données fonctionnelles de manière continue dans le temps, nous devons passer par le cas discret. Ainsi, notre estimateur des moindres carrés sera calculé par

$$\widehat{K}_X^{-1}\ \widehat{\mathrm{Cov}}(W,Y),$$

où $\widehat{K}_X^{-1}$ est l'inverse de l'estimateur de covariance de $(X(t_1),\ldots,X(t_L))$, c'est-à-dire $\widehat{K}_X(i,j)=\mathrm{Cov}(X(t_i),X(t_j))$, pour $i,j=1,\ldots,L$, et où $\widehat{\mathrm{Cov}}(W,Y)$ est l'estimateur de covariance entre $(W(t_1),\ldots,W(t_L))$ et Y ou $(Y(t_1),\ldots,Y(t_L))$ en cas de réponse scalaire ou fonctionnelle.

Dans la section 3.3.3, nous allons passer en revue les différents théorèmes qui construisent la théorie qui gravite autour de l'algorithme permettant trouver un estimateur convergent de l'inverse de l'opérateur de covariance associé à la variable inobservable X. Cette section est précédée de la section 3.3.1 qui présente quelques définitions importantes et de la section 3.3.2 qui cite quelques résultats préliminaires utiles dans les démonstrations des résultats principaux.

3.3.1 Rappel de quelques définitions

Définition 2. *Une matrice carrée A de dimension n est appelée matrice bande si et seulement si, pour $k \geq 0$,*

$$a_{ij}=0 \quad si \quad |i-j|>k,$$

$i,j=1,\ldots,n$. La constante k est alors appelée largeur de bande.

Définition 3. *Soient A et B deux matrices de dimension $n \times m$, le produit de Hadamard entre la matrice A et B est défini par*

$$(A\circ B)_{ij} \overset{déf}{=} A_{ij}B_{ij},$$

$i=1,\ldots,n$ et $j=1,\ldots,m$.

Définition 4. *Soit A une matrice de dimension $n \times m$, la norme de Frobenius de A est définie par*

$$||A||_F \overset{déf}{=} tr(A^*A)^{1/2}=tr(AA^*)^{1/2}=\sqrt{\sum_{i,j=1}^{n,m}|A_{ij}|^2},$$

où A^* est la matrice adjointe (ou transposée conjuguée) de la matrice A.

Définition 5. *Soit (e_n) une base dans un espace de Hilbert H, la norme de Hilbert-Schmidt associée à un opérateur de Hilbert-Schmidt $\mathcal{A}$ est définie par*

$$|||\mathcal{A}|||_{HS} = \left(\sum_{n=0} ||\mathcal{A}e_n||^2 \right)^{1/2}.$$

Un opérateur est dit de Hilbert-Schmidt si sa norme de Hilbert-Schmidt est finie.

3.3.2 Résultats préliminaires

Avant d'en venir aux résultats théoriques permettant de construire l'algorithme amenant à l'estimation du paramètre du modèle de régression sans erreur de mesure, il semble important de citer quelques résultats préliminaires qui nous seront très utiles dans les démonstrations des théorèmes principaux en section 3.3.3. Ainsi, dans la suite, le lemme 1, les propositions 3 et 4 sont tirés de [8] tandis que les propositions 5 et 6 proviennent de [10].

Lemme 1. *Soient $0 \leq \delta \leq 1$ et $b(s,t)$ un noyau continu sur $[0,1]^2$ tel que*

$$b(s,t) = 0 \text{ si } |s-t| > \delta$$

et soit $(t_1, \ldots, t_L) \in \mathcal{T}_L$, alors $B^L = \{b(t_i, t_j)\}_{i,j=1}^L$ est une matrice bande avec une largeur de bande $2[\delta L] + 1$, où $[.]$ désigne la partie entière.

Démonstration. Soient $i, j \in \{1, \ldots, L\}$,

$$B^L(i,j) = b(t_i, t_j) = 0 \quad \text{si} \quad |t_i - t_j| > \delta.$$

Comme $\forall u \in \{1, \ldots, L\}$, nous pouvons écrire $t_u = \dfrac{u}{L}$. De plus, comme $\delta \geq 0$ et $L > 0$, alors

$$|t_i - t_j| > \delta \Longleftrightarrow \left| \frac{i-j}{L} \right| > \delta \Longleftrightarrow |i-j| > \delta L \Longleftrightarrow |i-j| \geq [\delta L] + 1, \tag{3.6}$$

où [.] correspond à la partie entière.

Plaçons-nous dans la partie triangulaire supérieure de la matrice B^L, c'est-à-dire lorsque $i < j$. Alors $i - j < 0$ et la condition (3.6) devient $j - i \geq [\delta L] + 1$. Les éléments de la matrice B^L sont donc nuls sur les colonnes $i + [\delta L] + 1, \ldots, L$ de la matrice.

Plaçons-nous maintenant dans la partie triangulaire inférieure de la matrice B^L, c'est-à-dire lorsque $i > j$. Alors $i - j > 0$ et la condition (3.6) devient $i - j \geq [\delta L] + 1$. Les éléments de la matrice B^L sont donc nuls sur les lignes $j + [\delta L] + 1, \ldots, L$ de la matrice.

La matrice peut alors s'écrire sous la forme

$$B^L = \begin{pmatrix} b(t_1,t_1) & \cdots & b(t_1,t_{1+[\delta L]}) & & 0 & \cdots & 0 \\ \vdots & \ddots & & \ddots & & \ddots & \vdots \\ b(t_{1+[\delta L]},t_1) & \ddots & & & & \ddots & 0 \\ & \ddots & & & \ddots & & b(t_{L-[\delta L]},t_L) \\ & & \ddots & & & & \vdots \\ 0 & & & & \ddots & & b(t_L,t_L) \\ \vdots & \ddots & & \ddots & & & \\ 0 & \cdots & 0 & & b(t_L,t_{L-[\delta L]}) & \cdots & b(t_L,t_L) \end{pmatrix}$$

Ainsi, la largeur de bande de la matrice B^L, c'est-à-dire le nombre de diagonales comportant des éléments non-nuls, est égale à $2[\delta L]$ auquel nous devons ajouter la diagonale principale de la matrice, c'est-à-dire $2[\delta L] + 1$. □

Proposition 2 (Admis). *Soit une matrice A de dimension $k \times k$, la matrice A est de rang $r < k$ si et seulement si*

1. *il existe un mineur d'ordre r non-nul et*
2. *tous les mineurs d'ordre $r + 1$ sont nuls.*

Proposition 3 (Admis). *Soit $\mathcal{A}$ un opérateur avec un noyau*

$$a(s,t) = \sum_{i=1}^{r} \lambda_i \eta_i(t) \eta_i(s),$$

où $r < \infty$ et les fonctions propres $\{\eta_1, ..., \eta_r\}$ sont réelles analytiques.

Si $L > r$, alors tous les mineurs d'ordre r de la matrice $A^L := \{a(t_i, t_j)\}_{i,j=1}^{L}$ sont non-nuls presque partout sur $\mathcal{T}_L$.

Proposition 4 (Admis). *Supposons que* $E(||X||^4) < \infty$, $\delta < \frac{1}{4}$ *et* $\tau \xrightarrow{n\to\infty} 0$ *(voir théorème 2 pour la définition de* τ*) et définissons la taille de la grille* $L \geq 4(r+1)$, *alors*

$$n\ \tau\ L^{-2} \left| \text{rang}(\widehat{\mathcal{K}}_X) - r \right| = \mathcal{O}_P(1),$$

où r *est le rang de* $\mathcal{K}_X$.

Proposition 5 (Admis). *Soient deux opérateurs de covariance* $\mathcal{A}$ *et* $\mathcal{B}$ *tels que*

$$\mathcal{A} = \sum_{j\geq 1} \gamma_j (x_j \otimes x_j) \quad \text{et} \quad \mathcal{B} = \sum_{j\geq 1} \gamma'_j (x'_j \otimes x'_j),$$

où γ_j *et* γ'_j *sont les valeurs propres de* $\mathcal{A}$ *et* $\mathcal{B}$ *respectivement,* x_j *et* x'_j *sont les fonctions propres de* $\mathcal{A}$ *et* $\mathcal{B}$ *respectivement,*

1. $\forall j \geq 1$, $|\gamma_i - \gamma'_j| \leq ||\mathcal{A} - \mathcal{B}||_{\mathcal{L}}$;
2. *Si* $\mathcal{A}$ *et* $\mathcal{B}$ *sont symétriques, définis positifs et Hilbert-Schmidt, alors*

$$||x_j - x'_j|| \leq \frac{2\sqrt{2}}{\min_{k\neq j} |\gamma_j - \gamma_k|} ||\mathcal{A} - \mathcal{B}||_{\mathcal{L}}.$$

Proposition 6 (Admis). *Soit* $\{X_k\}$ *indépendants et identiquement distribués à valeurs dans un espace séparable de Hilbert* H *tels que* $\mathrm{E}(||X_1||) < \infty$, *alors*

1. $$\left|\left| \frac{1}{n} \sum_{k=1}^{n} X_k - \mathrm{E}(X_1) \right|\right| \longrightarrow 0 \quad p.s.$$
2. *Si* $\mathrm{E}(||X||^2) < \infty$, *alors*

$$n^{-1/2} \sum_{k=1}^{n} (X_k - \mathrm{E}(X_k)) \xrightarrow{\mathcal{D}} \mathcal{N}(0, C),$$

avec C *l'opérateur de covariance.*

3.3.3 Résultats principaux

Le théorème 1 est un résultat important qui permet de prouver le théorème 2 qui fonde la méthode de récupération de l'estimateur $\widehat{\mathcal{K}}_X$ (voir [8] pour ces deux théorèmes). En effet, le théorème 1 montre que sous certaines hypothèses, nous pouvons décomposer une matrice issue du noyau d'un opérateur de covariance en une somme de deux matrices de rangs finis A^L et B^L. La matrice B^L doit être une matrice bande

de largeur de bande inférieure à 1/2 et cette décomposition est unique.

Le théorème 2 nous montre que sous certaines hypothèses, la matrice A^L issue de la décomposition par le théorème 1 est l'unique matrice de rang minimal telle que

$$||P^L \circ (R^L - A^L)||_F^2 = 0,$$

où $R^L := A^L + B^L$. En effet, comme le théorème 1 nous impose une somme de matrices unique, cette formule nous permet de nous assurer l'égalité de R^L avec A^L en dehors d'une bande de largeur $[L/4]$. Dans le cas qui nous intéresse, cela nous assure que $\mathcal{K}_X$ et $\mathcal{K}_W$ sont égaux en dehors de cette bande. De plus, comme $\mathcal{K}_U$ peut être déterminée de manière unique, il est possible de reconstituer les éléments de la bande de $\mathcal{K}_X$.

Enfin, le théorème 3 prouve la convergence de l'estimateur du rang de $\widehat{\mathcal{K}}_X$, la vitesse de convergence de $\widehat{\mathcal{K}}_X$ et $\widehat{\mathcal{K}}_X^{-1}$ ainsi que la vitesse de convergence de l'estimateur $\hat{\beta}_X$ du paramètre β_X (voir [7]).

Théorème 1. *Soient $\mathcal{A}_1$ et $\mathcal{A}_2$ deux opérateurs de covariance avec des rangs finis r_1 et r_2 et de noyau a_1 et a_2 respectivement tels que, sans perdre de généralité, $r_1 \geq r_2$. Soient $\mathcal{B}_1$ et $\mathcal{B}_2$ deux opérateurs de covariance avec une largeur de bande $\delta_1 < \frac{1}{2}$ et $\delta_2 < \frac{1}{2}$ et de noyau b_1 et b_2 respectivement.*

Soit $(t_1, \ldots, t_L) \in \mathcal{T}_L$, définissons $\left(A_1^L, B_1^L, A_2^L, B_2^L\right) \in \mathbb{R}^{L \times L}$ tel que, pour $m = 1, 2$,

$$A_m^L(i,j) = a_m(t_i, t_j) \quad \text{et} \quad B_m^L(i,j) = b_m(t_i, t_j),$$

avec $i, j = 1, \ldots, L$.

Si les fonctions propres de $\mathcal{A}_1$ et $\mathcal{A}_2$ sont réelles analytiques et que

$$L \geq L^* := \max\left(\frac{2r_1 + 2}{1 - 2\delta_1}, \frac{2r_1 + 2}{1 - 2\delta_2}\right),$$

alors

$$A_1^L + B_1^L = A_2^L + B_2^L \Longleftrightarrow A_1^L = A_2^L \quad et \quad B_1^L = B_2^L,$$

presque partout sur $\mathcal{T}_L$.

Démonstration. Nous pouvons appliquer le Lemme 1 aux matrices B_1^L et B_2^L puisque par définition et pour $m = 1, 2$,

$$B_m^L(i,j) = b_m(t_i, t_j) = 0 \quad \text{si} \quad |t_i - t_j| > \delta_m.$$

Il en résulte que les matrices B_1^L et B_2^L sont des matrices bandes avec des largeurs de bandes égales à $2[\delta_1 L] + 1$ et $2[\delta_2 L] + 1$ respectivement.

Notons $\delta = \max\{\delta_1, \delta_2\}$ et Ω l'ensemble des indices des éléments simultanément nuls sur les matrices B_1^L et B_2^L. Comme la largeur de bande maximale de ces deux matrices vaut $2[\delta L] + 1$ par le Lemme 1, les éléments sont donc simultanément nuls en dehors de cette bande, c'est-à-dire sur les indices appartenant à l'ensemble

$$\Omega := \left\{(i,j) \in \{1,\dots,L\}^2 \ : \ |i-j| > [\delta L]\right\}.$$

Pour démontrer l'équivalence du théorème, nous allons uniquement démontrer que

$$A_1^L + B_1^L = A_2^L + B_2^L \Longrightarrow A_1^L = A_2^L \ \text{ et } \ B_1^L = B_2^L.$$

L'autre implication est en effet triviale. Supposons que $A_1^L + B_1^L = A_2^L + B_2^L$, nous savons que les matrices B_1^L et B_2^L sont toutes les deux nulles sur Ω. Ainsi, nous obtenons que

$$\{A_1^L\}_{ij} = \{A_2^L\}_{ij}, \ \forall (i,j) \in \Omega. \tag{3.7}$$

Définissions Ω_S l'ensemble des indices d'une sous-matrice S formée des r_1 premières lignes et r_1 dernières colonnes d'une matrice de dimension $L \times L$. Par l'hypothèse $L \geq L^* := \max\left(\dfrac{2r_1+2}{1-2\delta_1}, \dfrac{2r_1+2}{1-2\delta_2}\right)$, nous avons que

$$L \geq \frac{2r_1+2}{1-2\max(\delta_{1,2})} = \frac{2r_1+2}{1-2\delta}.$$

Par cette inégalité, nous obtenons que

$$\begin{aligned} L \geq \frac{2r_1+2}{1-2\delta} &\Longleftrightarrow L - 2\delta L \geq 2r_1 + 2 > 2r_1 \\ &\Longleftrightarrow L - 2r_1 > 2\delta L > \delta L \geq [\delta L]. \end{aligned} \tag{3.8}$$

De plus, dans la sous-matrice S, l'élément ayant la différence entre l'indice de ligne et l'indice de colonne la plus petite est l'élément le plus proche de la diagonale principale de la matrice initiale de taille $L \times L$. Il s'agit donc de l'élément inférieur gauche de la sous-matrice S. Ainsi, $\forall (i,j) \in \Omega_S$,

$$\begin{aligned} \min(|i-j|) &= |r_1 - (L - r_1 + 1)| \\ &= (L - r_1 + 1) - r_1 \\ &= L + 1 - 2r_1. \end{aligned} \tag{3.9}$$

L'égalité (3.9) est justifiée par le fait que

$$L \geq \frac{2r_1+2}{1-2\delta} > 2r_1 + 2 > 2r_1 - 1$$

car $0 < \delta < \dfrac{1}{2}$ par hypothèse et donc $1 - 2\delta \in]0,1[$.

En d'autres termes, $\forall (i,j) \in \Omega_S$,

$$|i-j| \geq L - 2r_1 + 1 > L - 2r_1. \tag{3.10}$$

En combinant (3.8) et (3.10), nous concluons que $\forall (i,j) \in \Omega_S$,

$$|i-j| > [\delta L]$$

et donc que $\Omega_S \subset \Omega$. Comme les matrices A_1^L et A_2^L sont égales sur Ω par (3.7), la sous-matrice S de dimension $r_1 \times r_1$ est commune à ces deux matrices.

Nous allons appliquer la Proposition 3 à la matrice A_1^L. En effet, nous savons par hypothèse que l'opérateur $\mathcal{A}_1$ qui lui est associé est de rang fini r_1 et que ses fonctions propres sont réelles analytiques. Par conséquent, comme $0 < \delta < \frac{1}{2}$ implique

$$L \geq \frac{2r_1+2}{1-2\delta} > 2r_1 + 2 > r_1,$$

nous avons que tous les mineurs d'ordre r_1 de la matrice A_1^L sont non-nuls presque partout sur $\mathcal{T}_L$. Comme le déterminant de la sous-matrice S est en fait un mineur d'ordre r_1 de la matrice A_1^L, il est donc non-nul. Ainsi, comme S est contenue dans la matrice A_2^L aussi, le rang de cette dernière est d'au moins r_1 par la Proposition 2. De plus, nous avons considéré par hypothèse que $rang(A_2^L) = r_2 \leq r_1$. En combinant avec le résultat précédent, nous obtenons finalement que

$$rang(A_2^L) = r_1.$$

Les matrices A_1^L et A_2^L sont donc deux matrices de rang r_1 égales sur Ω.

Soit $A^* \in \mathbb{R}^{L \times L}$ une matrice égale à A_1^L sur Ω mais dont ne sait rien de ses éléments sur Ω^c. Nous allons montrer qu'il existe une seule et unique matrice de complétion de A^* et qui est de rang r_1. Il nous faut dorénavant déterminer les éléments de A^* se trouvant sur Ω^c, c'est-à-dire les éléments tels que $|i-j| \leq [\delta L]$. Par l'hypothèse $L \geq L^*$, nous pouvons trouver une sous-matrice de A^* de dimension $(r_1+1) \times (r_1+1)$ avec un seul élément, noté x^*, dans Ω^c dont la valeur n'est donc pas connue. Comme cette sous-matrice de dimension $(r_1+1) \times (r_1+1)$, que nous notons à présent $\widetilde{A}$, est de rang incomplet, la colonne comprenant l'élément x^* est donc une combinaison linéaire des autres colonnes et peut donc être déterminé par calculs.

Par la Proposition 2, comme A^* est de rang r_1, tous les mineurs d'ordre r_1+1 de A^* sont nuls, ce qui implique que le

$$\det(\widetilde{A}) = 0.$$

En détails, pour $j \in \{1, \ldots, r_1+1\}$,

$$\det(\widetilde{A}) = \sum_{i=1}^{r_1+1} \widetilde{A}_{ij}(-1)^{i+j} \det(\widetilde{A}_{-(i,j)}),$$

où $\widetilde{A}_{-i,-j}$ est la matrice $\widetilde{A}$ à laquelle la ligne i et la colonne j ont été supprimées. Nous obtenons

$$ax^* + b = 0,$$

où $a = (-1)^{r_1+2}\det(\widetilde{A}_{-x^*})$ et b est égal aux autres termes du déterminant de $\widetilde{A}$. Remarquons que $a \neq 0$ car nous avons supposé précédemment que tous les mineurs d'ordre r_1 de la matrice A_1^L et donc de A^* également sont non-nuls. Ainsi, le déterminant de toute sous-matrice de dimension r_1 de A^* est non-nul. Par conséquent, cette équation possède une solution unique et peut être résolue.

En raisonnant de manière itérative, nous pouvons déterminer toute la matrice A^* en choisissant une sous-matrice $\widetilde{A}$ différente à chaque itération de manière à déterminer tous les éléments se trouvant sur Ω^c. Ce sont les diagonales les plus aux extrémités de la matrice A^* qui seront déterminées en premier, et ainsi de suite en terminant par la diagonale principale. Comme chacun de ces éléments sont déterminés de manière unique, nous avons trouvé une matrice unique A^* de rang r_1.

En conclusion, nous avons démontré qu'il existait une manière unique de compléter la matrice A_1^L et donc que

$$A_1^L = A_2^L (= A^*) \quad \text{et} \quad B_1^L = B_2^L,$$

presque partout sur $\mathcal{T}_L$. □

Théorème 2. *Soient $\mathcal{A} : L^2[0,1] \longrightarrow L^2[0,1]$ un opérateur de covariance de rang $r < \infty$ et de noyau a tel que les fonctions propres associées sont réelles analytiques et $\mathcal{B} : L^2[0,1] \longrightarrow L^2[0,1]$ un opérateur de covariance de largeur de bande δ et de noyau b.*

Pour $(t_1, \ldots, t_L) \in \mathcal{T}_L$, soit

$$A^L(i,j) := a(t_i, t_j) \qquad et \qquad B^L(i,j) := b(t_i, t_j),$$

avec $i, j = 1, \ldots, L$ et

$$R^L = A^L + B^L.$$

Supposons que $\delta < \frac{1}{4}$ et $L \geq 4r + 4$ et définissons la matrice $P^L \in \mathbb{R}^{L \times L}$ telle que pour $i, j = 1, \ldots, L$,

$$P^L(i,j) = \mathbb{1}(|i-j| > [L/4]).$$

Alors, presque partout sur $\mathcal{T}_L$,

1. *A^L est l'unique solution au problème d'optimisation*

$$\min_{\theta \in \mathbb{R}^{L \times L}} \text{rang}(\theta) \text{ tel que } \left|\left|P^L \circ (R^L - \theta)\right|\right|_F^2 = 0.$$

2. *De manière équivalente, $\forall \tau > 0$ suffisamment petit,*

$$A^L = \arg\min_{\theta \in \mathbb{R}^{L \times L}} \left\{ \left|\left|P^L \circ (R^L - \theta)\right|\right|_F^2 + \tau \text{ rang}(\theta) \right\}.$$

Démonstration. Nous allons tout d'abord appliquer le théorème 1. Pour rappel, ce théorème nous dit qu'il existe, sous certaines hypothèses, une seule et unique manière d'écrire la somme

$$R^L = A^L + B^L.$$

Vérifions les différentes hypothèses permettant la mise en oeuvre du théorème 1. Les deux opérateurs $\mathcal{A}$ et $\mathcal{B}$ sont bien de rang fini r par hypothèse, $\mathcal{B}$ a bien une largeur de bande $\delta < \frac{1}{4} < \frac{1}{2}$ et les fonctions propres de $\mathcal{A}$ sont bien réelles analytiques par hypothèse. Il nous reste maintenant à prouver que $L \geq \dfrac{2r+2}{1-2\delta}$. Comme

$$\delta < \frac{1}{4} \Longrightarrow 1 - 2\delta > \frac{1}{2},$$

nous avons que

$$\frac{2r+2}{1-2\delta} < \frac{2r+2}{\frac{1}{2}} = 4r + 4.$$

Comme nous supposons par hypothèse que $L \geq 4r+4$, nous avons grâce à l'inégalité calculée précédemment que $L \geq \dfrac{2r+2}{1-2\delta}$. Nous pouvons donc bel et bien appliquer le théorème 1. Ainsi, la matrice R^L peut être décomposée de manière unique en

$$R^L = A^L + B^L,$$

avec A^L de rang au plus r.

Ecrivons

$$\begin{aligned} P^L \circ R^L &= P^L \circ \left(A^L + B^L\right) \\ &= P^L \circ A^L + P^L \circ B^L. \end{aligned} \tag{3.11}$$

Comme la matrice B^K est une matrice avec une largeur de bande $\delta < \frac{1}{4}$, nous savons en appliquant le Lemme 1 que

$$B^L(i,j) = 0 \ \text{ si } \ |i-j| > [\delta L].$$

Nous savons également par définition que

$$P^L(i,j) = 0 \ \text{ si } \ |i-j| \leq [L/4].$$

Ainsi,

$$(P^L \circ B^L)(i,j) = 0 \ \text{ si } \ |i-j| > [\delta L] \ \text{ ou } \ |i-j| \leq [L/4].$$

Nous concluons enfin que

$$P^L \circ B^L = 0$$

par le fait que $\delta < \dfrac{1}{4}$ implique que $[\delta L] < [L/4]$.

Revenons à l'égalité (3.11) en écrivant, grâce à ce que nous venons de trouver et par définition d'une norme,

$$\begin{aligned} P^L \circ R^L = P^L \circ A^L &\iff P^L \circ R^L - P^L \circ A^L = 0 \\ &\iff P^L \circ (R^L - A^L) = 0 \\ &\iff \left|\left|P^L \circ (R^L - A^L)\right|\right|_F^2 = 0. \end{aligned}$$

La fonction objectif

$$\min_{\theta \in \mathbb{R}^{L \times L}} \text{rang}(\theta) \;\; \text{tel que} \;\; \left|\left|P^L \circ (R^L - \theta)\right|\right|_F^2 = 0$$

a donc pour valeur minimale r avec $\theta = A^L$.

Nous allons maintenant montrer que $\forall \tau > 0$,

$$A^L = \arg \min_{\theta \in \mathbb{R}^{L \times L}} \left\{ \left|\left|P^L \circ (R^L - \theta)\right|\right|_F^2 + \tau \text{ rang}(\theta) \right\},$$

c'est-à-dire que A^L est la matrice au rang le plus bas qui annule la fonction objectif.

Comme A^L est l'unique solution du problème

$$\min_{\theta \in \mathbb{R}^{L \times L}} \text{rang}(\theta) \;\; \text{tel que} \;\; \left|\left|P^L \circ (R^L - \theta)\right|\right|_F^2 = 0,$$

cela implique que $\forall \tau > 0$ et $\forall \theta \in \mathbb{R}^{L \times L}$ de rang au moins r, nous avons

$$\left|\left|P^L \circ (R^L - A^L)\right|\right|_F^2 + \tau \text{ rang}(A^L) < \left|\left|P^L \circ (R^L - \theta)\right|\right|_F^2 + \tau \text{ rang}(\theta). \tag{3.12}$$

Ainsi, il n'existe pas de matrice au rang supérieur à r fournissant un meilleur résultat de la fonction objectif que A^L.

Passons maintenant au cas des matrices de rang au plus $r-1$ avec $r > 1$. Soient

$$\mu = \min_{\substack{\theta \in \mathbb{R}^{L \times L}, \\ \text{rang}(\theta) \leq r-1}} \left\{ \left|\left|P^L \circ (R^L - \theta)\right|\right|_F^2 \right\} \quad \text{et} \quad \tau_* = \frac{\mu}{r-1},$$

comme

$$\left|\left|P^L \circ (R^L - A^L)\right|\right|_F^2 + \tau \text{ rang}(A^L) = 0 + \tau \; r = \tau \; r,$$

en prouvant que $\forall \tau < \tau_*$ et $\forall \theta \in \mathbb{R}^{L \times L}$ de rang au plus $r-1$,

$$\tau \; r < \mu + \tau \leq \left|\left|P^L \circ (R^L - \theta)\right|\right|_F^2 + \tau \text{ rang}(\theta), \tag{3.13}$$

nous aurons montré que

$$\left|\left|P^L \circ (R^L - A^L)\right|\right|_F^2 + \tau \text{ rang}(A^L) < \left|\left|P^L \circ (R^L - \theta)\right|\right|_F^2 + \tau \text{ rang}(\theta) \tag{3.14}$$

pour n'importe quelle matrice θ de dimension $L \times L$ de rang strictement inférieur à r, ce qui prouve bien que la matrice A^L donne un meilleur résultat pour la fonction objectif que n'importe quelle matrice de rang au plus $r-1$.

Soient $\tau < \tau_*$ et $\theta \in \mathbb{R}^{L\times L}$ telle que $\text{rang}(\theta) \leq r-1$, démontrons la première inégalité de (3.13), c'est-à-dire

$$\tau\ r < \mu + \tau.$$

En effet, par définition de τ_*,

$$\begin{aligned}\tau\ r - \tau < \mu &\Longleftrightarrow \tau\ (r-1) < \tau_*\ (r-1)\\ &\Longleftrightarrow \tau < \tau_*.\end{aligned}$$

La seconde inégalité, c'est-à-dire

$$\mu + \tau \leq \left|\left|P^L \circ (R^L - \theta)\right|\right|_F^2 + \tau\ \text{rang}(\theta),$$

est démontrée grâce aux points suivants :

- $\left|\left|P^L \circ (R^L - \theta)\right|\right|_F^2 \geq \min\limits_{\substack{\widetilde{\theta}\in\mathbb{R}^{L\times L},\\ \text{rang}(\widetilde{\theta})\leq r-1}} \left\{\left|\left|P^L \circ \left(R^L - \widetilde{\theta}\right)\right|\right|_F^2\right\} = \mu$;
- $\tau\ \text{rang}(\theta) \geq \tau$ car $\text{rang}(\theta) \geq 1$.

En conclusion, nous avons montré que

$$A^L = \arg\min_{\theta\in\mathbb{R}^{L\times L}} \left\{\left|\left|P^L \circ (R^L - \theta)\right|\right|_F^2 + \tau\ \text{rang}(\theta)\right\},\quad \forall\tau \in]0, \tau_*[.$$

Remarquons que, par définition, τ_* va dépendre de r. Cela ne veut pas dire pour autant que la fonction objectif va dépendre d'une inconnue. En effet, comme

$$L \geq 4r+4 \Longleftrightarrow r \leq \frac{L}{4} - 1 < [L/4],$$

il est possible de déterminer r en calculant le rang de la sous-matrice formée des $[L/4]$ premières lignes et $[L/4]$ dernières colonnes de la matrice R^L à l'image de la preuve du théorème 1, ce qui nous permettra immédiatement de déterminer τ_*. □

Théorème 3. *Supposons que $E(||X||^4) < \infty$, $E(||U||^4) < \infty$, $E(\varepsilon^2) < \infty$ et $\delta < \frac{1}{4}$. Soit $L_* \geq 4(r+1)$ entier fixé et supposons que $\tau_n \longrightarrow 0$, $n\tau_n \longrightarrow \infty$ et $L^{-2} = \mathcal{O}(n^{-1})$.*

Alors, lorsque $n \longrightarrow \infty$,

1. $P(\hat{r}_{L_*} = r) \longrightarrow 1$,
2. $\left|\left|\left|\hat{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} = \mathcal{O}_P\left(n^{-1/2}\right)$ *et* $\left|\left|\left|\hat{\mathcal{K}}_X - \mathcal{K}_X\right|\right|\right|_{HS} = \mathcal{O}_P\left(n^{-1/2}\right)$,
3. $\left|\left|\hat{\beta}_X - \beta_X\right|\right| = \mathcal{O}_P\left(n^{-1/2}\right)$.

Démonstration.
Preuve du point 1 : Nous allons tout d'abord appliquer la Proposition 4 énoncée précédemment. Nous avons bien des moments finis d'ordre 4 pour la variable aléatoire X par hypothèse, nous avons également que $\delta < \frac{1}{4}$, $\tau_n \longrightarrow 0$ et une taille de grille $L_* \geq 4(r+1)$. Nous obtenons alors par la Proposition 4 que

$$L_*^{-2}|\hat{r}_{L_*} - r| = \mathcal{O}_P\left(\frac{1}{n\tau_n}\right).$$

Ce résultat peut être traduit grâce à la définition de convergence en probabilité par

$$P\left(L_*^{-2}|\hat{r}_{L_*} - r| > \frac{\gamma}{n\tau_n}\right) \xrightarrow{n\to\infty} 0, \ \forall\gamma > 0$$

ou encore

$$P\left(|\hat{r}_{L_*} - r| \leq \frac{\gamma}{n\tau_n}L_*^2\right) \xrightarrow{n\to\infty} 1, \ \forall\gamma > 0.$$

Soit $\gamma > 0$, en posant $\varepsilon := \frac{\gamma}{n\tau_n}L_*^2 > 0$, nous obtenons

$$P(|\hat{r}_{L_*} - r| \leq \varepsilon) \xrightarrow{n\to\infty} 1,$$

qui correspond à la convergence en probabilité de $\hat{r}_{L_*}$ vers r. Puisque cette convergence est vérifiée quel que soit ε, elle est également valable pour $\varepsilon \in]0,1[$. Comme $\hat{r}_{L_*}$ et r sont des entiers, nous pouvons réécrire la convergence en probabilité comme

$$P(\hat{r}_{L_*} = r) \xrightarrow{n\to\infty} 1.$$

Preuve du point 2 : Soit $M > 0$,

$$\begin{aligned}
& P\left(n^{1/2}\left|\left|\left|\widehat{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} > M\right) \\
= \ & P\left(n^{1/2}\left|\left|\left|\widehat{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} > M, \hat{r}_{L_*} = r\right) + P\left(n^{1/2}\left|\left|\left|\widehat{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} > M, \hat{r}_{L_*} \neq r\right) \\
\leq \ & P\left(n^{1/2}\left|\left|\left|\widetilde{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} > M, \hat{r}_{L_*} = r\right) + P\left(\hat{r}_{L_*} \neq r\right) \\
\leq \ & P\left(n^{1/2}\left|\left|\left|\widetilde{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} > M\right) + P\left(\hat{r}_{L_*} \neq r\right), \qquad (3.15)
\end{aligned}$$

où $\widetilde{\mathcal{K}}_X$ est l'estimateur obtenu dans l'algorithme (voir section 3.4) lors de l'estimation du rang r par $\hat{r}_{L_*}$. Il s'agit donc de l'opérateur associé à la matrice

$$\widetilde{K}_X = \arg\min_{\theta:\mathrm{rang}(\theta)=r} ||P_L \circ (\widehat{K}_W - \theta)||.$$

L'objectif dès à présent est de montrer que les deux termes de (3.15) convergent vers 0 lorsque $n \longrightarrow \infty$. Ainsi, par l'inégalité, nous aurons montré que

$$P\left(\left|\left|\left|\widehat{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} > n^{-1/2}M\right) \xrightarrow{n\to\infty} 0, \tag{3.16}$$

ce qui est équivalent à montrer que

$$\left|\left|\left|\widehat{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} = \mathcal{O}_P\left(n^{-1/2}\right).$$

Par le point 1 du théorème (qui a été démontré ci-dessus), le second terme de (3.15) tend vers 0 lorsque $n \longrightarrow \infty$ puisque

$$P(\hat{r}_{L_*} \neq r) = 1 - P(\hat{r}_{L_*} = r) \xrightarrow{n\to\infty} 1 - 1 = 0.$$

Ainsi, pour terminer la démonstration du point 2 de ce théorème, il nous suffit maintenant de montrer que le premier terme de (3.15) tend également vers 0 lorsque $n \longrightarrow \infty$, c'est-à-dire

$$P\left(n^{1/2}\left|\left|\left|\widetilde{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} > M\right) \xrightarrow{n\to\infty} 0. \tag{3.17}$$

Notons $\widetilde{\lambda}_j$ et $\widetilde{\eta}_j$ les valeurs propres et fonctions propres (orthonormés) de $\widetilde{\mathcal{K}}_X$, $j = 1, \ldots, r$. Notons également λ_j et η_j les valeurs propres et fonctions propres (orthonormés) de $\mathcal{K}_X$, $j = 1, \ldots, r$. Nous supposons que les valeurs propres de $\widetilde{\mathcal{K}}_X$ et $\mathcal{K}_X$ sont ordonnées, c'est-à-dire

$$\widetilde{\lambda}_1 \leq \ldots \leq \widetilde{\lambda}_r \quad \text{et} \quad \lambda_1 \leq \ldots \leq \lambda_r.$$

En écrivant $\widetilde{\mathcal{K}}_X^{-1}$ et $\mathcal{K}_X^{-1}$ sous leur représentation avec en valeurs et vecteurs propres, nous avons

$$\widetilde{\mathcal{K}}_X^{-1} = \sum_{j=1}^{r} \widetilde{\lambda}_j^{-1}\left(\widetilde{\eta}_j \otimes \widetilde{\eta}_j\right) \quad \text{et} \quad \mathcal{K}_X^{-1} = \sum_{j=1}^{r} \lambda_j^{-1}\left(\eta_j \otimes \eta_j\right).$$

Nous pouvons ainsi calculer

$$
\begin{aligned}
& \left|\left|\left|\widetilde{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} \\
= & \left|\left|\left|\sum_{j=1}^{r} \left(\widetilde{\lambda}_j^{-1}\left(\widetilde{\eta}_j \otimes \widetilde{\eta}_j\right) - \lambda_j^{-1}\left(\eta_j \otimes \eta_j\right)\right)\right|\right|\right|_{HS} \\
\leq & \sum_{j=1}^{r} \left|\left|\left|\widetilde{\lambda}_j^{-1}\left(\widetilde{\eta}_j \otimes \widetilde{\eta}_j\right) - \lambda_j^{-1}\left(\eta_j \otimes \eta_j\right)\right|\right|\right|_{HS} \\
= & \sum_{j=1}^{r} \left|\left|\left|\widetilde{\lambda}_j^{-1}\left(\widetilde{\eta}_j \otimes \widetilde{\eta}_j\right) - \lambda_j^{-1}(\widetilde{\eta}_j \otimes \widetilde{\eta}_j) - \lambda_j^{-1}\left(\eta_j \otimes \eta_j\right) + \lambda_j^{-1}(\widetilde{\eta}_j \otimes \widetilde{\eta}_j)\right|\right|\right|_{HS} \\
= & \sum_{j=1}^{r} \left|\left|\left|\left(\widetilde{\lambda}_j^{-1} - \lambda_j^{-1}\right)\left(\widetilde{\eta}_j \otimes \widetilde{\eta}_j\right) + \lambda_j^{-1}(\widetilde{\eta}_j \otimes \widetilde{\eta}_j - \eta_j \otimes \eta_j)\right|\right|\right|_{HS} \\
\leq & \sum_{j=1}^{r} \left|\widetilde{\lambda}_j^{-1} - \lambda_j^{-1}\right| |||\widetilde{\eta}_j \otimes \widetilde{\eta}_j|||_{HS} + \sum_{j=1}^{r} \lambda_j^{-1} |||\widetilde{\eta}_j \otimes \widetilde{\eta}_j - \eta_j \otimes \eta_j|||_{HS} \\
\leq & \sum_{j=1}^{r} \widetilde{\lambda}_j^{-1}\lambda_j^{-1}\left|\widetilde{\lambda}_j - \lambda_j\right| + 2\sum_{j=1}^{r} \lambda_j^{-1}\, ||\widetilde{\eta}_j - \eta_j|| \qquad (3.18) \\
\leq & \widetilde{\lambda}_r^{-1}\widetilde{\lambda}_r r \max_{1\leq j\leq r}\left|\widetilde{\lambda}_j - \lambda_j\right| + 2\lambda_r^{-1} r \max_{1\leq j\leq r} ||\widetilde{\eta}_j - \eta_j|| \\
\leq & \widetilde{\lambda}_r^{-1}\widetilde{\lambda}_r\, r \left|\left|\left|\widetilde{\mathcal{K}}_X - \mathcal{K}_X\right|\right|\right|_{HS} + 4\sqrt{2}\lambda_r^{-1} r\; a_r^{-1} \left|\left|\left|\widetilde{\mathcal{K}}_X - \mathcal{K}_X\right|\right|\right|_{HS} \qquad (3.19) \\
= & \left(\widetilde{\lambda}_r^{-1}\widetilde{\lambda}_r\, r + 4\sqrt{2}\lambda_r^{-1} r\; a_r^{-1}\right) \left|\left|\left|\widetilde{\mathcal{K}}_X - \mathcal{K}_X\right|\right|\right|_{HS}.
\end{aligned}
$$

Le passage à la ligne (3.18) par rapport à la ligne précédente dans le calcul ci-dessus est justifié pour le premier terme par

$$
\begin{aligned}
|||\widetilde{\eta}_j \otimes \widetilde{\eta}_j|||_{HS}^2 &\overset{\text{déf}}{=} \sum_{i=1}^{\infty} ||(\widetilde{\eta}_j \otimes \widetilde{\eta}_j)(\widetilde{\eta}_i)||^2 \\
&= \sum_{i=1}^{\infty} ||\widetilde{\eta}_j\, \langle \widetilde{\eta}_j, \widetilde{\eta}_i\rangle||^2 \\
&= ||\widetilde{\eta}_j||^2 = 1.
\end{aligned}
$$

Pour prouver le passage à la ligne (3.18) pour le deuxième terme par rapport à la ligne qui précède, montrons que

$$
|||\widetilde{\eta}_j \otimes \widetilde{\eta}_j - \eta_j \otimes \eta_j|||_{HS} \leq 2\, ||\widetilde{\eta}_j - \eta_j||.
$$

Calculons tout d'abord

$$
\begin{aligned}
&|||\widetilde{\eta}_j \otimes \widetilde{\eta}_j - \eta_j \otimes \eta_j|||_{HS}^2 \\
&\overset{\text{déf}}{=} \sum_{i=1}^{\infty} ||(\widetilde{\eta}_j \otimes \widetilde{\eta}_j - \eta_j \otimes \eta_j)(\widetilde{\eta}_i)||^2 \\
&= \sum_{i=1}^{\infty} ||\widetilde{\eta}_j \ \langle \widetilde{\eta}_j, \widetilde{\eta}_i \rangle - \eta_j \ \langle \eta_j, \widetilde{\eta}_i \rangle||^2 \\
&= \sum_{i=1}^{\infty} \Big(\langle \widetilde{\eta}_j \ \delta_{ij} - \eta_j \ \langle \eta_j, \widetilde{\eta}_i \rangle, \widetilde{\eta}_j \ \delta_{ij} - \eta_j \ \langle \eta_j, \widetilde{\eta}_i \rangle \rangle \Big) \\
&= \sum_{i=1}^{\infty} \Big(\langle \widetilde{\eta}_j \ \delta_{ij}, \widetilde{\eta}_j \ \delta_{ij} \rangle - \langle \widetilde{\eta}_j \ \delta_{ij}, \eta_j \ \langle \eta_j, \widetilde{\eta}_i \rangle \rangle - \langle \eta_j \ \langle \eta_j, \widetilde{\eta}_i \rangle, \widetilde{\eta}_j \ \delta_{ij} \rangle + \langle \eta_j \ \langle \eta_j, \widetilde{\eta}_i \rangle, \eta_j \ \langle \eta_j, \widetilde{\eta}_i \rangle \rangle \Big) \\
&= \langle \widetilde{\eta}_j, \widetilde{\eta}_j \rangle - \langle \widetilde{\eta}_j, \eta_j \rangle \langle \eta_j, \widetilde{\eta}_j \rangle - \langle \eta_j, \widetilde{\eta}_j \rangle \langle \eta_j, \widetilde{\eta}_j \rangle + \langle \eta_j, \eta_j \rangle \sum_{i=1}^{\infty} \langle \eta_j, \widetilde{\eta}_i \rangle \langle \eta_j, \widetilde{\eta}_i \rangle \\
&= ||\widetilde{\eta}_j||^2 - 2\langle \eta_j, \widetilde{\eta}_j \rangle^2 + \sum_{i=1}^{\infty} \langle \eta_j, \widetilde{\eta}_i \rangle^2 \\
&= ||\widetilde{\eta}_j||^2 - 2\langle \eta_j, \widetilde{\eta}_j \rangle^2 + ||\eta_j||^2 \qquad \text{(par l'identité de Parseval)} \\
&= 2 - 2\langle \eta_j, \widetilde{\eta}_j \rangle^2.
\end{aligned}
$$

Enfin, pour prouver l'inégalité, montrons que

$$
\begin{aligned}
& 2 - 2\langle \eta_j, \widetilde{\eta}_j \rangle^2 \overset{?}{\leq} 4\,||\widetilde{\eta}_j - \eta_j||^2 \\
\iff \quad & 2 - 2\langle \eta_j, \widetilde{\eta}_j \rangle^2 \overset{?}{\leq} 4\langle \widetilde{\eta}_j - \eta_j, \widetilde{\eta}_j - \eta_j \rangle \\
\iff \quad & 2 - 2\langle \eta_j, \widetilde{\eta}_j \rangle^2 \overset{?}{\leq} 4\,(\langle \widetilde{\eta}_j, \widetilde{\eta}_j \rangle - 2\langle \widetilde{\eta}_j, \eta_j \rangle + \langle \eta_j, \eta_j \rangle) \\
\iff \quad & 2 - 2\langle \eta_j, \widetilde{\eta}_j \rangle^2 \overset{?}{\leq} 4\,||\widetilde{\eta}_j||^2 - 8\langle \widetilde{\eta}_j, \eta_j \rangle + 4\,||\eta_j||^2 \\
\iff \quad & 2 - 2\langle \eta_j, \widetilde{\eta}_j \rangle^2 \overset{?}{\leq} 4 - 8\langle \widetilde{\eta}_j, \eta_j \rangle + 4 \\
\iff \quad & -6 + 8\langle \widetilde{\eta}_j, \eta_j \rangle - 2\langle \eta_j, \widetilde{\eta}_j \rangle^2 \overset{?}{\leq} 0. \qquad (3.20)
\end{aligned}
$$

Or, en posant $x = \langle \widetilde{\eta}_j, \eta_j \rangle$ dans la TABLE 3.1 et comme $-1 \leq \langle \widetilde{\eta}_j, \eta_j \rangle \leq 1$, l'inégalité (3.20) est bien vérifiée, ce qui prouve le passage à la ligne (3.18).

Enfin, le passage à la ligne (3.19) par rapport à la ligne qui précède est justifié par l'utilisation de la proposition 5, où $\widetilde{\mathcal{K}}_X$ et $\mathcal{K}_X$ sont par définition symétriques (ce

x		1		3	
$-6+8x-2x^2$	$-$	0	$+$	0	$-$

TABLE 3.1 – Tableau de signe de la fonction $f(x) = -6 + 8x - 2x^2$.

sont des opérateurs de covariance), définis positifs et Hilbert-Schmidt par hypothèse. Ainsi, dans le premier terme, nous utilisons le point 1 de la proposition 5, c'est-à-dire que

$$\forall j \geq 1, \quad |\widetilde{\lambda}_i - \lambda_j| \leq \left|\left|\widetilde{\mathcal{K}}_X - \mathcal{K}_X\right|\right|_{\mathcal{L}},$$

ce qui implique que

$$\max_{1\leq j\leq r} |\widetilde{\lambda}_j - \lambda_j| \leq \left|\left|\widetilde{\mathcal{K}}_X - \mathcal{K}_X\right|\right|_{\mathcal{L}} \leq \left|\left|\left|\widetilde{\mathcal{K}}_X - \mathcal{K}_X\right|\right|\right|_{\mathcal{HS}}.$$

La dernière inégalité est justifiée par le fait que $||.||_{\mathcal{L}} \leq |||.|||_{HS}$ (voir [10]). Dans le second terme de la ligne (3.19), c'est le point 2 de la proposition 5 que nous utilisons pour justifier le passage d'une ligne à l'autre, c'est-à-dire que

$$||\widetilde{\eta}_j - \eta_j|| \leq 2\sqrt{2}\ a_r^{-1} \left|\left|\widetilde{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|_{\mathcal{L}} \leq 2\sqrt{2}\ a_r^{-1} \left|\left|\left|\widetilde{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS},$$

où $a_r := \min_{k\neq j} |\widetilde{\lambda}_j - \lambda_k|$.

En conclusion, nous avons bel et bien prouvé que

$$\left|\left|\left|\widetilde{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} \leq \left(\widetilde{\lambda}_r^{-1}\widetilde{\lambda}_r\ r + 4\sqrt{2}\lambda_r^{-1} r\ a_r^{-1}\right) \left|\left|\left|\widetilde{\mathcal{K}}_X - \mathcal{K}_X\right|\right|\right|_{HS}. \tag{3.21}$$

Ainsi, il nous suffit désormais de démontrer que

$$\left|\left|\left|\widetilde{\mathcal{K}}_X - \mathcal{K}_X\right|\right|\right|_{HS} = \mathcal{O}_P\left(n^{-1/2}\right),$$

ce qui impliquera par l'inégalité (3.21) que

$$\left|\left|\left|\widetilde{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} = \mathcal{O}_P\left(n^{-1/2}\right).$$

Cela prouvera alors (3.17) qui prouve à son tour (3.16).

Par [8] (cité dans [7]), en utilisant la norme de Frobenius des matrices de covariance $\widetilde{K}_X$ et K_X, il nous suffit pour arriver à ce résultat de montrer que

$$L^{-2}\left|\left|\widetilde{K}_X - K_X\right|\right|_F^2 = \mathcal{O}_P\left(n^{-1}\right). \tag{3.22}$$

Définissons les fonctionnelles suivantes

- $\mathbb{S}_{n,L}(\theta) := L^{-2}||P_L \circ (\widehat{K}_W - \theta)||_F^2$,

- $S_{n,L}(\theta) := L^{-2}\,||P_L \circ (K_W - \theta)||_F^2\,,$

et la distance

$$d_n(\theta_1, \theta_2) := L^{-1}||\theta_1 - \theta_2||_F,$$

où $\theta_1, \theta_2 \in \mathbb{R}^{L\times L}$ de rang r.

Vérifions les hypothèses du théorème 2 et appliquons le. Par définition, l'opérateur $\mathcal{K}_X$ est bien de rang fini r avec des fonctions propres analytiques et $K_W = K_X + K_U$, avec K_U, la matrice de covariance issue du noyau de l'opérateur de covariance des erreurs de mesure U, une matrice bande avec une largeur de bande $\delta < \frac{1}{4}$. Par hypothèse, nous avons bien que la taille de la grille vaut $L(\geq L_*) \geq 4(r+1)$. Ainsi, par le théorème 2, la matrice K_X est l'unique matrice de rang r qui minimise la quantité $S_{n,L}(.)$ définie précédemment.

Réalisons le développement de Taylor au deuxième ordre autour de K_X

$$S_{n,L}(\theta) = S_{n,L}(K_X) + \langle S'_{n,L}(K_X), \theta - K_X\rangle_F + \frac{1}{2}\langle S''_{n,L}(\theta_*)(\theta - K_X), \theta - K_X\rangle_F,$$

où $\theta_* = \alpha\ \theta + (1-\alpha)K_X$ pour $\alpha \in [0,1]$.

En calculant la dérivée première de $S_{n,L}$ et $\mathbb{S}_{n,L}$ ainsi que la dérivée seconde de $S_{n,L}$, nous obtenons

- $\mathbb{S}'_{n,L}(\breve{\theta}) = -2L^{-2}P_L \circ \left(\widehat{K}_W - \breve{\theta}\right),$
- $S'_{n,L}(\breve{\theta}) = -2L^{-2}P_L \circ \left(K_W - \breve{\theta}\right),$
- $S''_{n,L}(\widetilde{\theta})\breve{\theta} = -2L^{-2}P_L \circ \breve{\theta},$

où $\widetilde{\theta}, \breve{\theta} \in \mathbb{R}^{L\times L}$ de rang r.

En utilisant le fait que

$$\mathbb{S}_{n,L}(\theta) = L^{-2}\,\mathrm{Tr}((P_L \circ (\widehat{K}_W - \theta))^t (P_L \circ (\widehat{K}_W - \theta)))$$

par définition de la norme de Frobenius (voir définition 4) et que, par [16], pour toute matrice A et X,

$$\frac{\partial\,\mathrm{Tr}(AX)}{\partial X} = \frac{\partial\,\mathrm{Tr}(XA)}{\partial X} = A^t,$$
$$\frac{\partial\,\mathrm{Tr}(AX^t)}{\partial X} = \frac{\partial\,\mathrm{Tr}(X^tA)}{\partial X} = A,$$
$$\frac{\partial\,\mathrm{Tr}(X^tX)}{\partial X} = 2X,$$

nous pouvons justifier que

$$\begin{aligned}
\mathbb{S}'_{n,L}(\breve{\theta}) &= L^{-2}\frac{\partial}{\partial\breve{\theta}}\left(\left|\left|P_L\circ\left(\widehat{K}_W-\breve{\theta}\right)\right|\right|_F^2\right)\\
&= L^{-2}\frac{\partial}{\partial\breve{\theta}}\operatorname{Tr}\left(\left(P_L\circ(\widehat{K}_W-\breve{\theta})\right)^t\left(P_L\circ(\widehat{K}_W-\breve{\theta})\right)\right)\\
&= L^{-2}\frac{\partial}{\partial\breve{\theta}}\operatorname{Tr}\Big((P_L\circ\widehat{K}_W)^t(P_L\circ\widehat{K}_W)-(P_L\circ\widehat{K}_W)^t(P_L\circ\breve{\theta})\\
&\qquad\qquad -(P_L\circ\breve{\theta})^t(P_L\circ\widehat{K}_W)+(P_L\circ\breve{\theta})^t(P_L\circ\breve{\theta})\Big)\\
&= L^{-2}\Big(0-2(P_L\circ\widehat{K}_W)+2(P_L\circ\breve{\theta})\Big)\\
&= -2L^{-2}P_L\circ(\widehat{K}_W-\breve{\theta}).
\end{aligned}$$

Les calculs des deux autres dérivées $S'_{n,L}(\breve{\theta})$ et $S''_{n,L}(\widetilde{\theta})$ se font de manière similaire et ne sont pas détaillés pour des raisons de redondance.

Comme $S(K_X)=0$, cela implique que $P_L\circ(K_W-\breve{\theta})=0$ et donc que $S'_{n,L}(K_X)=0$. Par conséquent,

$$\begin{aligned}
|\Delta(\theta)| &:= |S_{n,L}(\theta)-S_{n,L}(K_X)|\\
&= \left|\langle S'_{n,L}(K_X),\theta-K_X\rangle_F+\frac{1}{2}\langle S''_{n,L}(\theta_*)(\theta-K_X),\theta-K_X\rangle_F\right|\\
&= \left|0+\frac{1}{2}\langle -2L^{-2}P_L\circ(\theta-K_X),\theta-K_X\rangle_F\right|\\
&\leq L^{-2}\,||P_L\circ(\theta-K_X)||_F\,||\theta-K_X||_F\,, \text{ par l'inégalité de Cauchy-Schwarz}\\
&\leq L^{-2}\,||\theta-K_X||_F^2\\
&= \left(d_n(\theta,K_X)\right)^2.
\end{aligned}$$

Cette inégalité implique que

$$\sup_{\substack{\theta:\operatorname{rang}(\theta)=r\\ d_n(\theta,K_X)<\xi}}|\Delta(\theta)|\leq\xi^2.$$

Réalisons maintenant le développement de Taylor au premier ordre autour de K_X

$$\mathbb{S}_{n,L}(\theta)=\mathbb{S}_{n,L}(K_X)+\langle\mathbb{S}'_{n,L}(\theta_{**}),\theta-K_X\rangle_F$$

et de la même manière

$$S_{n,L}(\theta)=S_{n,L}(K_X)+\langle S'_{n,L}(\theta_{**}),\theta-K_X\rangle_F,$$

où $\theta_{**}=\beta\theta+(1-\beta)K_X$, pour $\beta\in[0,1]$.

Ainsi, nous pouvons calculer

$$\begin{aligned}
|D(\theta)| &:= |\mathbb{S}_{n,L}(\theta) - S_{n,L}(\theta) - \mathbb{S}_{n,L}(K_X) + S_{n,L}(K_X)| \\
&= \left|\langle \mathbb{S}'_{n,L}(\theta_{**}), \theta - K_X\rangle_F - \langle S'_{n,L}(\theta_{**}), \theta - K_X\rangle_F\right| \\
&= \left|\langle -2L^{-2}P_L \circ (\widehat{K}_W - \theta_{**}), \theta - K_X\rangle_F - \langle -2L^{-2}P_L \circ (K_W - \theta_{**}), \theta - K_X\rangle_F\right| \\
&= 2L^{-2}\left|\langle P_L \circ (\widehat{K}_W - \theta_{**}), \theta - K_X\rangle_F - \langle P_L \circ (K_W - \theta_{**}), \theta - K_X\rangle_F\right| \\
&= 2L^{-2}\left|\langle P_L \circ (\widehat{K}_W - \theta_{**}) - P_L \circ (K_W - \theta_{**}), \theta - K_X\rangle_F\right| \\
&= 2L^{-2}\left|\langle P_L \circ (\widehat{K}_W - K_W), \theta - K_X\rangle_F\right| \\
&\leq 2L^{-2}\left|\left|\widehat{K}_W - K_W\right|\right|_F ||\theta - K_X||_F, \text{ par l'inégalité de Cauchy-Schwarz} \\
&= 2L^{-1}\left|\left|\widehat{K}_W - K_W\right|\right|_F d_n(\theta, K_X).
\end{aligned}$$

Cette inégalité implique que

$$\sup_{\substack{\theta:\text{rang}(\theta)=r \\ d_n(\theta,K_X)<\xi}} |D(\theta)| \leq 2L^{-1}\xi \left|\left|\widehat{K}_W - K_W\right|\right|_F.$$

Nous savons via [10] que

$$\mathrm{E}\left(L^{-2}\left|\left|\widehat{K}_W - K_W\right|\right|_F^2\right) \leq Cn^{-1},$$

où $C = \sup_{s,t\in[0,1]} \mathrm{Var}(W(s)W(t))$. Ainsi, par la croissance de l'espérance et par

$$\begin{aligned}
E\left(\left|\left|\widehat{K}_W - K_W\right|\right|_F\right) &= \sqrt{\mathrm{E}\left(\left|\left|\widehat{K}_W - K_W\right|\right|_F^2\right) - \mathrm{Var}\left(\left|\left|\widehat{K}_W - K_W\right|\right|_F\right)} \\
&\leq \sqrt{\mathrm{E}\left(\left|\left|\widehat{K}_W - K_W\right|\right|_F^2\right)} \\
&\leq \sqrt{\frac{Cn^{-1}}{L^{-2}}} = L\sqrt{Cn^{-1}},
\end{aligned}$$

nous obtenons que

$$\begin{aligned}
E\left(\sup_{\substack{\theta:\text{rang}(\theta)=r \\ d_n(\theta,K_X)<\xi}} |D(\theta)|\right) &\leq 2L^{-1}\xi\ \mathrm{E}\left(\left|\left|\widehat{K}_W - K_W\right|\right|_F\right) \\
&\leq 2\xi\sqrt{\frac{C}{n}}.
\end{aligned}$$

Grâce à ce résultat, nous satisfaisons aux hypothèses d'un théorème que nous considérons comme acquis, détaillé dans [2]. La conclusion de ce théorème affirme que le minimiseur $\widetilde{K}_X$ de $\mathbb{S}_{n,L}$ satisfait

$$n\ d_n^2\left(\widetilde{K}_X, K_X\right) = \mathcal{O}_P(1),$$

ce qui est équivalent à dire que

$$L^{-1}\left|\left|\widetilde{K}_X - K_X\right|\right|_F = \mathcal{O}_P\left(n^{-1/2}\right),$$

lorsque $n \longrightarrow \infty$.

Comme par hypothèse de ce théorème, nous considérons que $L^{-2} = \mathcal{O}\left(n^{-1}\right)$, nous avons démontré le résultat (3.22).

Preuve du point 3 : Notons par définition la version matricielle de la covariance empirique issue de l'opérateur de covariance entre les variables fonctionnelles Y et W comme

$$\widehat{C}_{Y,W} := \widehat{\mathrm{Cov}}(Y,W) = \frac{1}{n}\sum_{i=1}^{n} Y_i W_i - \bar{Y}\bar{W},$$

où $\bar{Y}$ et $\bar{W}$ est la moyenne empirique de la variable Y et W respectivement. Par la Proposition 6, nous savons lorsque $n \longrightarrow \infty$ que

$$\left|\left|\widehat{C}_{YW} - \mathrm{Cov}(Y,W)\right|\right| = \mathcal{O}_P\left(n^{-1/2}\right).$$

Enfin comme $\left|\left|\left|\widehat{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS} = \mathcal{O}_P\left(n^{-1/2}\right)$ par le point 2 de ce théorème, et comme $\widehat{\beta}_X = \widehat{\mathcal{K}}_X^{-1}\widehat{C}_{Y,W}$ et $\beta_X = \mathcal{K}_X^{-1}\mathrm{Cov}(Y,W)$,

$$\begin{aligned}
\left|\left|\widehat{\beta}_X - \beta_X\right|\right| &= \left|\left|\widehat{\mathcal{K}}_X^{-1}\widehat{C}_{Y,W} - \mathcal{K}_X^{-1}\mathrm{Cov}(Y,W)\right|\right| \\
&= \left|\left|\widehat{\mathcal{K}}_X^{-1}\widehat{C}_{Y,W} - \mathcal{K}_X^{-1}\widehat{C}_{Y,W} + \mathcal{K}_X^{-1}\widehat{C}_{Y,W} - \mathcal{K}_X^{-1}\mathrm{Cov}(Y,W)\right|\right| \\
&= \left|\left|\left(\widehat{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right)\widehat{C}_{Y,W} + \mathcal{K}_X^{-1}\left(\widehat{C}_{Y,W} - \mathrm{Cov}(Y,W)\right)\right|\right| \\
&\leq \left|\left|\left|\widehat{\mathcal{K}}_X^{-1} - \mathcal{K}_X^{-1}\right|\right|\right|_{HS}\left|\left|\widehat{C}_{Y,W}\right|\right| + \left|\left|\left|\mathcal{K}_X^{-1}\right|\right|\right|_{HS}\left|\left|\widehat{C}_{Y,W} - \mathrm{Cov}(Y,W)\right|\right| \\
&= \mathcal{O}_P\left(n^{-1/2}\right),
\end{aligned}$$

lorsque $n \longrightarrow \infty$.

□

Remarquons que dans le cas d'une réponse fonctionnelle, le théorème 3 reste valable avec une légère adaptation. Ainsi, sous les mêmes hypothèses avec $\mathrm{E}(||\varepsilon||^2) < \infty$, nous avons que

$$|||\widehat{\mathcal{B}}_X - \mathcal{B}_X|||_{HS} = \mathcal{O}_P(n^{-1/2}),$$

lorsque $n \longrightarrow \infty$.

3.4 Description de l'algorithme

L'algorithme se décompose en deux étapes (voir [7]). La première se concentre sur l'estimation du rang de $\mathcal{K}_X$ tanids que la seconde calcule l'estimateur de $\mathcal{K}_X$.

1. Pour l'itération b, $b = 1, \ldots, B$, nous sélectionnons aléatoirement une grille $s_1, \ldots, s_{L_*}$ de taille entière $L_* = L/m \geq 4(r+1)$, avec $m > 1$, au sein de la grille complète $t_1, \ldots, t_L$. Pour la sélection de la sous-grille, le i^e point est choisi aléatoirement parmi $t_{m(i-1)+1}, \ldots, t_{mi}$, pour $i = 1, \ldots, L_*$.

2. Nous calculons la matrice de covariance empirique de $(W(s_1), \ldots, W(s_{L_*}))$ sur la sous-grille de taille L_* que nous notons $\widehat{K}_{W_*}$.

3. Pour $j = 1, \ldots, M$, où $1 \leq M \leq L_*/4 - 1$ est un entier, nous calculons la valeur de la fonction

$$f_{L_*}(j) := \min_{\substack{\theta \in \mathbb{R}^{L_* \times L_*} \\ \text{rang}(\theta) = j}} ||P_{\delta_*}^{L_*} \circ (\widehat{K}_{W_*} - \theta)||_F^2,$$

   où $P_{\delta_*}^{L_*} = \mathbb{1}(|i-j| > [\delta_* L_*])$ avec $0 \leq \delta_* \leq 1/4$. La minimisation de $f_{L_*}(j)$ peut être calculée par une méthode de quasi-Newton en commençant au pas initial par une projection de rang j de la matrice de $\widehat{K}_W$ par une décomposition en valeurs singulières (voir [7]).

4. Calculons $\widetilde{r}_{b,L_*} = \min\{j : f_{L_*}(j) \leq c\}$, où $c > 0$ est un paramètre permettant de réaliser un compromis entre la minimisation du rang et la fonction $f_{L_*}(j)$ (voir [7]).

5. L'estimateur du rang, noté $\widehat{r}_{L_*}$, est calculé par le mode des valeurs de $\widetilde{r}_{b,L_*}$.

6. Nous calculons la matrice de covariance empirique de $(W(t_1), \ldots, W(t_L))$ sur l'ensemble de la grille de taille L qui correspond à $\widehat{K}_W$ et calculons l'estimateur de la matrice de covariance de $(X(t_1), \ldots, X(t_L))$, notée $\widehat{K}_X$, comme

$$\arg \min_{\substack{\theta \in \mathbb{R}^{L \times L} \\ \text{rang}(\theta) = \widehat{r}_{L_*}}} ||P^L \circ (\widehat{K}_W - \theta)||_F^2,$$

   où $P^L = \mathbb{1}(|i-j| > [\delta L])$. La minimisation est également réalisée grâce à une méthode de quasi-Newton.

7. Nous calculons les valeurs propres $\widehat{\lambda}_j$ et les vecteurs propres $\widehat{\eta}_j$ de l'estimateur $\widehat{K}_X$. Comme les valeurs propres de la matrice inverse sont $\widehat{\lambda}_j^{-1}$ et les vecteurs propres de la matrice inverse sont également $\widehat{\eta}_j$,

$$\widehat{\mathcal{K}}_X^{-1} = \sum_{j=1}^{\widehat{r}_{L_*}} \widehat{\lambda}_j^{-1} (\widehat{\eta}_j \otimes \widehat{\eta}_j).$$

Ainsi, grâce à cet algorithme, nous parvenons à trouver un estimateur $\widehat{\mathcal{K}}_X^{-1}$ nécessaire à l'estimation du paramètre β_X qui, rappelons le, est défini par

$$\widehat{\mathcal{K}}_X^{-1}\widehat{\mathcal{C}}_{W,Y}.$$

Notons que pour le cas d'une réponse fonctionnelle,

$$\widehat{\mathcal{C}}_{W,Y} = n^{-1}\sum_{i=1}^{n} Y_i \otimes W_i - \bar{Y} \otimes \bar{W}.$$

Chapitre 4

Illustration via des simulations numériques

Ce chapitre a pour objectif d'illustrer la mise en œuvre de l'algorithme présenté en section 3.4 dont le but est de déterminer un estimateur convergent du paramètre fonctionnel $\beta_X(t)$ dans un modèle de régression contenant des erreurs de mesure fonctionnelles. Les résultats sont présentés pour deux modèles différents dans le cadre d'une régression à variable explicative fonctionnelle et à réponse scalaire. Ces exemples de modèles s'inspirent de ceux présentés dans [7].

4.1 Modèle 1

Le premier modèle sur lequel nous allons nous pencher est le suivant. Les observations X sans erreur de mesure sont générées, pour $i = 1, \ldots, n$, où n est la taille de l'échantillon, par

$$X_i = \sum_{j=1}^{r} \lambda_j^{1/2} \kappa_j \eta_j,$$

où le rang $r = 3$, $\lambda = [1.5, 0.9, 0.3]$, $\eta_1 \equiv 1$, $\eta_2 = \sqrt{2}\sin(2\pi t)$, $\eta_3 = \sqrt{2}\cos(2\pi t)$, avec $t \in [0, 1]$, et $\kappa_j \sim \mathcal{N}(0, 1)$.

Les erreurs de mesure fonctionnelles U sont définies, pour $i = 1, \ldots, n$, par

$$U_i = \sum_{l=1}^{D} \gamma_l^{-1/2} \kappa_l \phi_l,$$

où $\gamma_1 = 0.09$, γ_l prend des valeurs équidistantes entre 0.04 et 0.01 ($l = 2, \ldots, D$), ϕ_j est une fonction triangulaire unitaire définie sur $[(j-1)\delta, j\delta]$ et $\delta = 0.05$ avec $D = [1/\delta]$. Ainsi, les erreurs U sont corrélées sur un intervalle de temps de 0.05.

La variable observée W est donc définie, pour $i = 1, \ldots, n$, par

$$W_i = X_i + U_i.$$

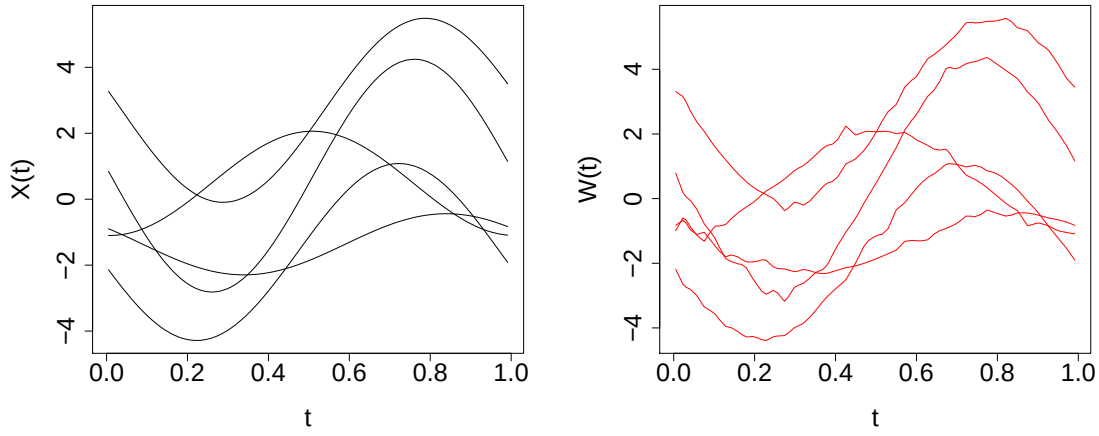

FIGURE 4.1 – Graphique des 5 premiers individus de l'échantillon pour la variable sans erreur de mesure X (à gauche) et des 5 individus correspondant pour la variable avec des erreurs de mesure W (à droite) dans le modèle 1.

La FIGURE 4.1 montre un exemple des 5 premiers individus de l'échantillon. Sur le graphe de gauche, nous retrouvons la variable fonctionnelle sans erreur de mesure X définie sur l'intervalle $[0, 1]$. Sur le graphe de droite, nous avons la variable fonctionnelle W, c'est-à-dire la variable correspondante X perturbée par des erreurs de mesure. Nous avons choisi une taille de grille $L = 100$ dans cet exemple.

Enfin, le paramètre β_X est défini par

$$\beta_X = \eta_1 + \eta_2 - \eta_3.$$

Le premier objectif de l'algorithme est pour rappel de déterminer un estimateur du rang r. Nous avons choisi dans notre exemple $n = 100$ et $L_* = 25$ qui est la taille de la grille sur laquelle va se réaliser l'estimation de r. Les autres constantes utilisées dans l'algorithme décrit en section 3.4 sont $B = 100$, $M = 10$ et $c = 0.01L_*^2$ (voir [7]). La FIGURE 4.2 nous montre l'évolution du pourcentage d'estimateurs qui estiment correctement la valeur de r en fonction de la taille de la sous-grille L_*. Nous observons qu'au plus la taille de l'échantillon n est importante, au plus vite nous atteignons la proportion de 100% avec l'augmentation de L_*. Cela fait sens avec le théorème 3 dans lequel il est dit que, pour $L_* \geq 4(r + 1)$, la probabilité que $\widehat{r}_{L_*}$ soit égal à r converge vers 1 lorsque n tend vers l'infini. Dans notre cas, L_* doit être égal à au mois $4(r + 1) = 16$. Nous pouvons donc prendre une valeur proche de 16 mais pas trop, auquel cas nous devrions faire tendre la taille de l'échantillon vers de très grandes valeurs. Dans notre cas, où $n = 100$, 100% des estimateurs calculés ont bien estimé le rang r à partir de $L_* > 20$. Ainsi, le choix de $L_* = 25$ est pertinent.

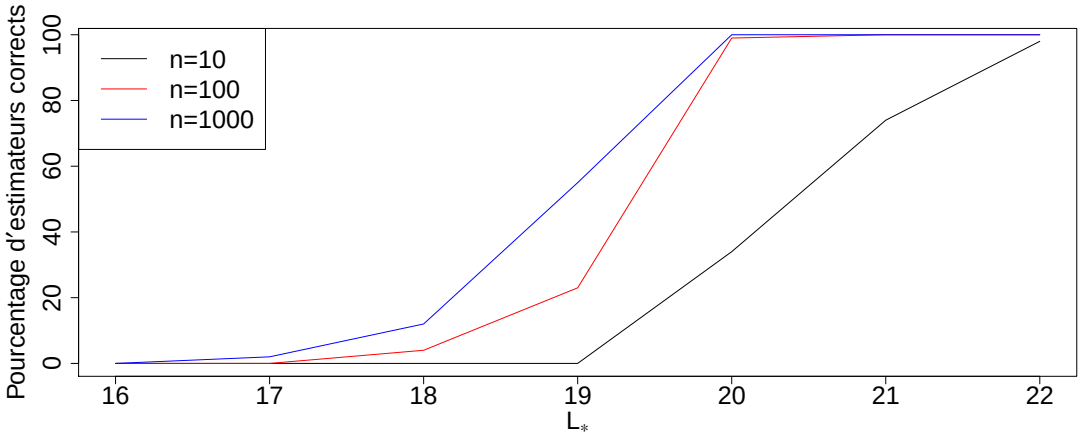

FIGURE 4.2 – Graphique du pourcentage d'estimateurs $\widehat{r}_{b,L_*}$ calculés par l'algorithme égaux à la vraie valeur du rang r en fonction de la taille de la sous-grille L_*.

Dans la FIGURE 4.3, nous retrouvons en rouge la fonction du vrai paramètre β_X dans le modèle sans erreur de mesure. Dans la FIGURE 4.4, nous avons le graphe de l'estimateur $\hat{\beta}_W$, c'est-à-dire l'estimateur de la régression naïve avec erreurs de mesure. Nous observons sur la première illustration que l'algorithme a réussi à trouver un bon estimateur du vrai β_X tandis que le graphe de la seconde illustration nous permet de voir que la régression naïve nous donne des résultats qui ne correspondent pas à l'allure du graphe de β_X.

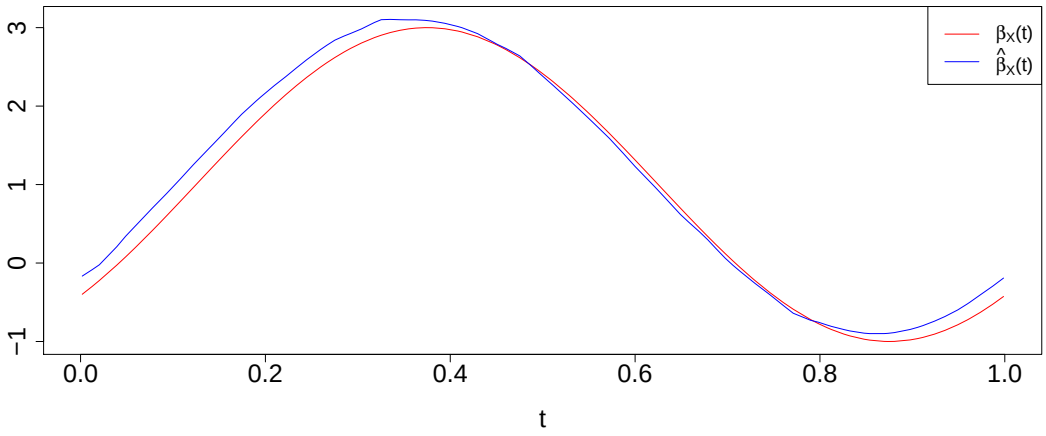

FIGURE 4.3 – Graphe pour le modèle 1 du vrai paramètre fonctionnel $\beta_X(t)$ en rouge et de son estimateur $\hat{\beta}_X(t)$ obtenu par l'algorithme corrigeant les effets de la présence d'erreurs de mesure en bleu.

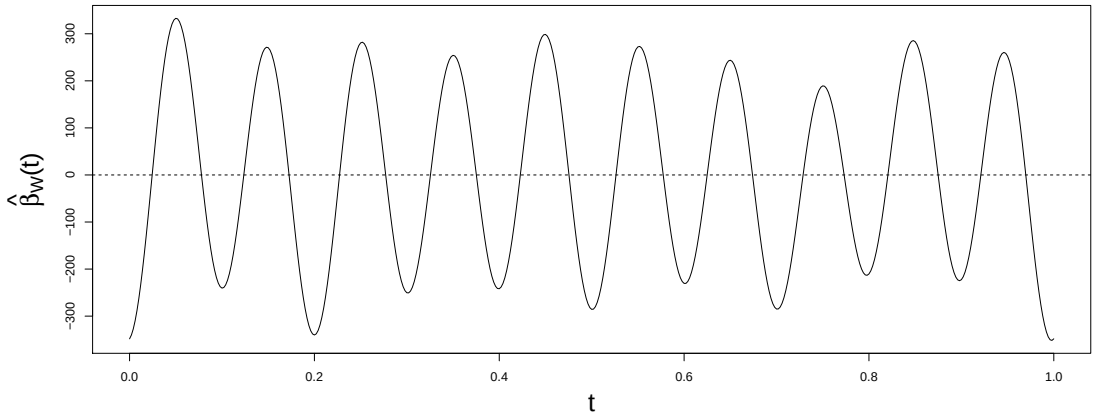

FIGURE 4.4 – Graphe pour le modèle 1 de l'estimateur naïf $\hat{\beta}_W(t)$ obtenu grâce à la fonction *fRegress* en *R*. Le nombre de termes utilisés dans la base de la série de Fourier est de 25.

Pour calculer la proximité de l'estimateur obtenu par l'algorithme, nous allons calculer la norme 2 de la différence entre $\hat{\beta}_X$ et β_X. Pour illustrer la vitesse de convergence de cet estimateur, nous allons calculer cette norme de l'erreur en fonction de la valeur de la taille de l'échantillon n. Nous allons le faire pour $n = 10, 50, 100, 200$. A chaque valeur de n, nous allons générer le modèle 20 fois et calculer la norme de l'erreur pour réaliser une moyenne. Les résultats obtenus sont présentés dans la FIGURE 4.5 où la courbe noire représente les valeurs calculées tandis que la courbe rouge est tracée pour mettre en évidence l'allure du graphe qui suit un tracé en $n^{-1/2}$. Cette figure nous rappelle ainsi le point 3 du théorème 3 qui nous dit que

$$||\hat{\beta}_X - \beta_X|| = \mathcal{O}_P(n^{-1/2}).$$

4.2 Modèle 2

Le second modèle sur lequel nous allons nous pencher est le suivant. Les observations X sans erreur de mesure sont générées, pour $i = 1, \ldots, n$, où n est la taille de l'échantillon, par

$$X_i = \sum_{j=1}^{r} \lambda_j^{1/2} \kappa_j \eta_j,$$

où le rang $r = 5$, $\lambda = [1.5, 1.2, 0.9, 0.6, 0.3]$, les η_j sont les r premiers polynômes de Legendre $f_1 \equiv 1$, $f_2 = 2t - 1$, $f_3 = 6t^2 - 6t + 1$, $f_4 = 20t^3 - 30t^2 + 12t + 1$, $f_5 = 70t^4 - 140t^3 + 90t^2 - 20t + 1$ normalisés, avec $t \in [0, 1]$, et $\kappa_j \sim \mathcal{N}(0, 1)$.
Les erreurs de mesure fonctionnelles U sont définies, pour $i = 1, \ldots, n$, par

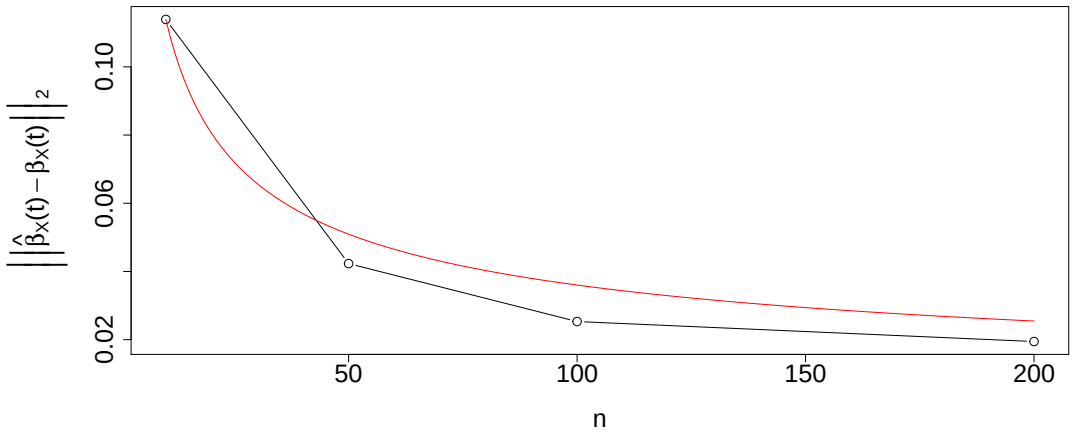

FIGURE 4.5 – Graphe pour le modèle 1 de l'erreur moyenne $||.||_2$ entre le vrai paramètre fonctionnel $\beta_X(t)$ et son estimateur $\hat{\beta}_X(t)$ obtenu par l'algorithme en fonction de la taille de l'échantillon n en noir. La vitesse de convergence est $\mathcal{O}(n^{-1/2})$ (représentée en rouge). Le modèle a été généré 20 fois pour chaque valeur de n.

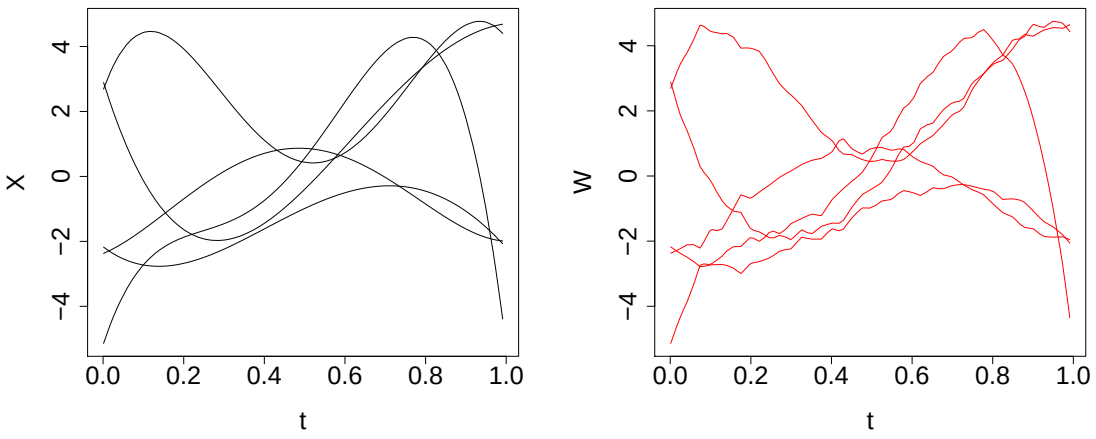

FIGURE 4.6 – Graphique des 5 premiers individus de l'échantillon pour la variable sans erreur de mesure X (à gauche) et des 5 individus correspondant pour la variable avec des erreurs de mesure W (à droite) dans le modèle 2.

$$U_i = \sum_{l=1}^{D} \gamma_l^{-1/2} \kappa_l \phi_l,$$

où $\gamma_1 = 0.09$, γ_l prend des valeurs équidistantes entre 0.04 et 0.01 ($l = 2, \ldots, D$), ϕ_j est une fonction triangulaire unitaire définie sur $[(j-1)\delta, j\delta]$, avec $\delta = 0.05$ et $D = [1/\delta]$.

La variable observée W est donc définie, pour $i = 1, \ldots, n$, par

$$W_i = X_i + U_i.$$

La FIGURE 4.6 montre un exemple des 5 premiers individus de l'échantillon. Sur le graphe de gauche, nous retrouvons la variable fonctionnelle sans erreur de mesure X définie sur l'intervalle $[0, 1]$. Sur le graphe de droite, nous avons la variable fonctionnelle W, c'est-à-dire la variable correspondante X perturbée par des erreurs de mesure. Nous avons choisi $L = 100$ également dans cet exemple ainsi que les mêmes valeurs pour les constantes B, M et c que le modèle 1 (voir section 3.4).

Enfin, le paramètre β_X est défini par

$$\beta_X = 0.7\eta_1 + 3\eta_2 - \eta_4 + 0.5\eta_5.$$

Dans la FIGURE 4.7, nous retrouvons en rouge la fonction du vrai paramètre β_X dans le modèle sans erreur de mesure. Dans la FIGURE 4.8, nous avons le graphe de l'estimateur $\hat{\beta}_W$, c'est-à-dire l'estimateur de la régression naïve avec erreurs de mesure. Nous observons sur la première illustration que l'estimateur de β_X correspond

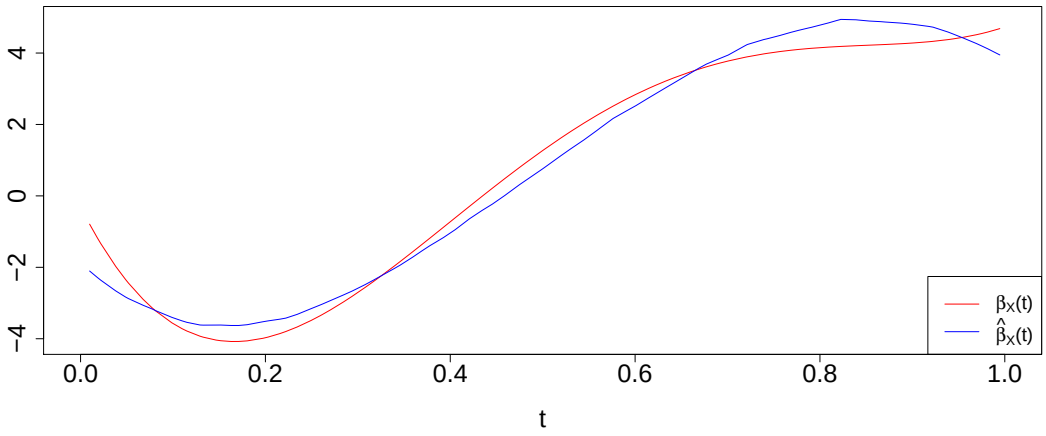

FIGURE 4.7 – Graphe pour le modèle 2 du vrai paramètre fonctionnel $\beta_X(t)$ en rouge et de son estimateur $\hat{\beta}_X(t)$ obtenu par l'algorithme corrigeant les effets de la présence d'erreurs de mesure en bleu.

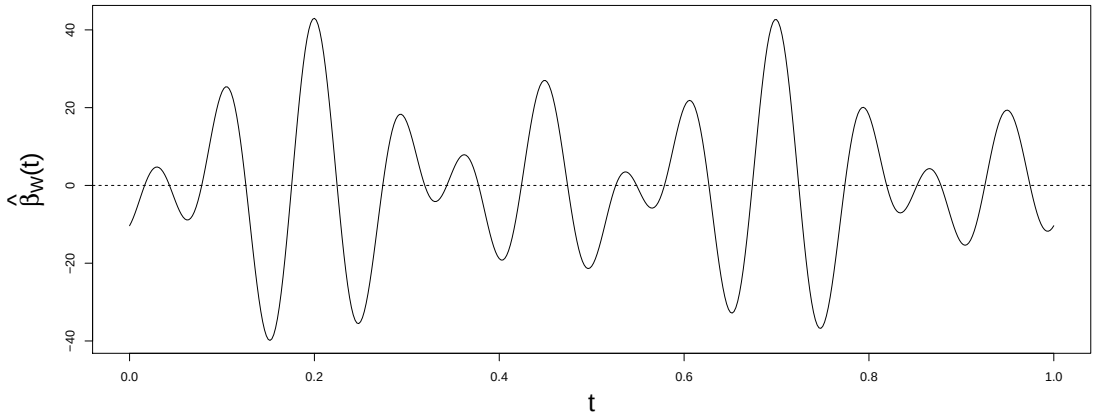

FIGURE 4.8 – Graphe pour le modèle 2 de l'estimateur naïf $\hat{\beta}_W(t)$ obtenu grâce à la fonction *fRegress* en *R*. Le nombre de termes utilisés dans la base de la série de Fourier est de 25.

au tracé de la vraie fonction tandis que le graphe de seconde illustration nous permet de voir que la régression naïve nous donne de mauvais résultats avec une fonction fortement oscillante.

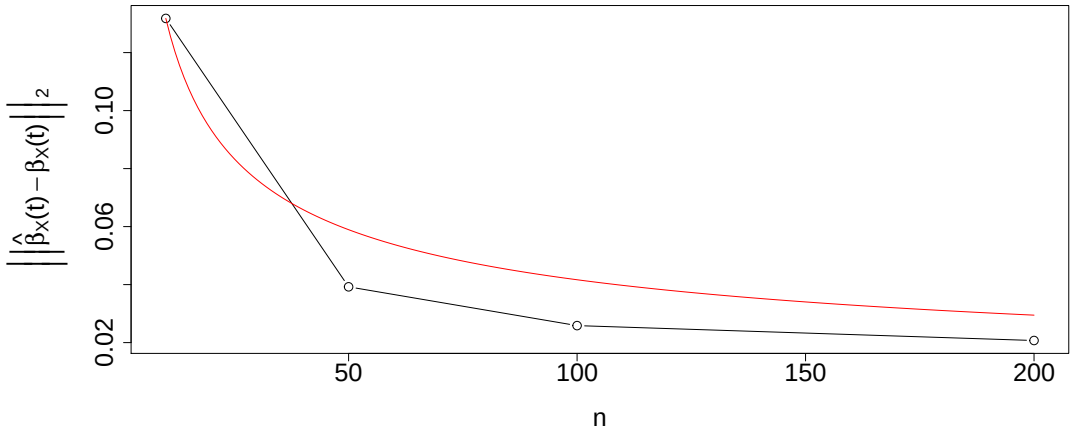

FIGURE 4.9 – Graphe pour le modèle 2 de l'erreur moyenne $||.||_2$ entre le vrai paramètre fonctionnel $\beta_X(t)$ et son estimateur $\hat{\beta}_X(t)$ obtenu par l'algorithme en fonction de la taille de l'échantillon n en noir. La vitesse de convergence est $\mathcal{O}(n^{-1/2})$ (représentée en rouge). Le modèle a été généré 20 fois pour chaque valeur de n.

Comme pour le modèle 1 décrit précédemment, nous allons calculer la norme 2 de la différence entre $\hat{\beta}_X$ et β_X pour évaluer la qualité de l'estimateur. Pour illustrer la vitesse de convergence de cet estimateur, nous allons calculer cette norme de l'erreur en fonction de la valeur de la taille de l'échantillon n. Nous allons également le faire pour $n = 10, 50, 100, 200$. A chaque valeur de n, nous allons générer le modèle 20 fois et calculer la norme de l'erreur pour réaliser une moyenne. Les résultats obtenus sont présentés dans la FIGURE 4.9 où la courbe noire représente les valeurs calculées tandis que la courbe rouge est tracée pour mettre en évidence l'allure du graphe qui suit un tracé en $n^{-1/2}$ et qui nous rappelle ainsi le point 3 du théorème 3.

Conclusion

Dans la première partie de ce mémoire, nous avons introduit la notion d'erreurs de mesure dans des modèles de régression. Tout d'abord, nous nous y sommes intéressés dans le cadre de la régression linéaire multivariée. Nous avons pu constater que certaines études de santé par exemple pouvaient être confrontées à la présence de ces erreurs au sein de certains régresseurs du modèle. Les conséquences qui en découlent peuvent être considérables pour les conclusions d'études statistiques, notamment une mauvaise identification des variables significatives causée par un biais dans l'estimation des coefficients de régression.

Pour corriger ces effets néfastes, nous avons, dans la suite de cette première partie, explicité deux méthodes permettant de retrouver un estimateur qui converge vers la vraie valeur du paramètre β. La première, le SIMEX, est une méthode d'extrapolation des valeurs de l'estimateur obtenues en générant des observations ayant des erreurs de mesure de plus en plus grande. La seconde quant à elle se base sur l'utilisation d'une autre variable, appelée variable instrumentale, possédant des hypothèses importantes dont celle d'être liée au régresseur. En réalisant une régression des moindres carrés à deux étapes, cela nous permet de trouver un estimateur du paramètre β du vrai modèle de régression.

La seconde partie de ce mémoire s'est ensuite concentrée sur la généralisation de la théorie des erreurs de mesure au cas fonctionnel de la régression. Nous avons vu qu'un algorithme nous permettait de retrouver là aussi un estimateur convergent du vrai paramètre qui est devenu, comme pour les régresseurs, fonctionnel. La théorie sous-jacente a été décrite et les théorèmes qui fondent cette méthode ont été démontrés en détails. La suite de cette seconde partie s'est traduite par l'illustration de la mise en pratique de cet algorithme sur des données simulées. Deux modèles différents ont été présentés ainsi que les résultats obtenus sur l'estimateur du paramètre fonctionnel et la vitesse de convergence de celui-ci.

Enfin, je me permets de proposer quelques perspectives à ce mémoire. Dans la régression fonctionnelle, nous avons considéré que les erreurs de mesure étaient des fonctions qui contaminaient la variable explicative fonctionnelle. Il pourrait être pertinent de développer le cas où les erreurs de mesure seraient *iid* comme évoqué dans [7], c'est-à-dire dans le cas limite où $\delta = 0$, et où le SIMEX, méthode utilisée dans ce mémoire dans le cadre de la régression multivariée, est d'ailleurs citée comme

méthode de correction. Ce mémoire se concentre également exclusivement sur une régression linéaire. Il pourrait être adéquat de se pencher sur un autre type de régression, quadratique par exemple. Ce mémoire est donc une première approche au sujet de la présence d'erreurs de mesure dans la régression.

Bibliographie

[1] Pokropek A. *Introduction to instrumental variables and their application to large-scale assessment data. Springer*, 2016.

[2] Wan derVaart A. et Wellner J. *Weak Convergence and Empirical Process.* Springer, 1996.

[3] Raymond J C et al. *Measurement Error in Nonlinear Models : A Modern Perspective, Second Edition.* Taylor & Francis Group, 2006.

[4] Wentzel K. et Asher S. *The Academic Lives of Neglected, Rejected, Popular, and Controversial Children. Child Development*, 1995.

[5] Wentzel K. et Caldwell K. *Friendships, Peer Acceptance, and Group Membership : Relations to Academic Achievement in Middle School. Child Development*, 1997.

[6] Lederer W. et Küchenhoff H. *A short Introduction to the SIMEX and MCSIMEX.* 2006.

[7] Chakraborty A. et Panaretos V. *Regression with genuinely functional errors-in-covariates. arXiv*, 2017".

[8] Descary M-H. et Panaretos V. *Functional Data Analysis by Matrix Completion. arXiv*, 2016.

[9] Stock J. H. et Yogo M. *Testing for weak instruments in linear IV regression. Cambridge University Press*, 2005.

[10] Van Bever G. *SMATM120 : Cours de statistiques avancées - Chapter II : Functional Data Analysis.* Université de Namur, 2019-2020.

[11] Stoehr J. *Méthodes de Monte Carlo. Université Paris Dauphine*, 2020-2021.

[12] Green S. B. Block G. Brinton L. A. Ziegler R. G. Hoover R. et Taylor P. R. Jones D. Y., Schatzkin A. *Dietary fat and breast cancer in the National Health and Nutrition Survey I : Epidemiologic follow-up study. Journal of the National Cancer Institute*, page 465–471, 1987.

[13] Wanjoya A. Njuguna Karomo J., Musili Mwalili S. *Power of Simulation Extrapolation in Correction of Covariates Measured with Errors. International Journal of Data Science and Analysis*, 2019.

[14] Wright J. H. et Yogo M. Stock J. H. *A survey of weak instruments and weak identification in generalized method of moments. Journal of Business and Economic Statistics*, page 518–529, 2002.

[15] Wikipedia. *Analytic function.* `https://en.wikipedia.org/wiki/Analytic_function`. Consulté le 13 août 2022.

[16] Wikipedia. *Matrix calculus.* `https://en.wikipedia.org/wiki/Matrix_calculus`. Consulté le 9 août 2022.