Utilisation de tests de structure en régression sur variable fonctionnelle.

Laurent Delsol, Frédéric Ferraty, Philippe Vieu

Utilisation de tests de structure en régression sur variable fonctionnelle.

Laurent DELSOL[1] & Frédéric FERRATY[2] & Philippe VIEU [2]

[1] *Université d'Orléans, MAPMO, Fédération Denis Poisson. Route de Chartres, B.P. 6759 - 45067 Orléans cedex 2 FRANCE*

[2] *Université Paul Sabatier, Institut de Mathématiques de Toulouse. 118 route de Narbonne, 31062 TOULOUSE Cedex 9 FRANCE*

Abstract: This work focuses on recent advances on the way general structural testing procedures can be constructed in regression on functional variable. Our test statistic is constructed from a specific estimator adapted to the specific model to be checked and uses recent advances concerning kernel smoothing methods for functional data. A general theoretical result states the asymptotic normality of our statistic under the null hypothesis and diverges under the local alternatives. This result opens interesting prospects about tests for no-effect, for linearity, or for reduction dimension of the covariate. Bootstrap methods are then proposed to compute the threshold value of our test. Finally, we present some applications to spectrometric datasets and discuss interesting prospects for the future.

Mots-clés: test de structure, régression, variable fonctionnelle, rééchantillonnage, non-effet, linéaire, multivarié, spectrométrie.

1 Introduction

Certains phénomènes évoluent au cours du temps ou des conditions du milieu dans lequel l'expérience est réalisée. Il n'est donc pas rare d'être amené à prendre en compte des observations discrétisées de leur évolution (pouvant être modélisée par une courbe) afin d'étudier de manière plus pertinente un problème concret. Les récents progrès technologiques permettent fréquemment de disposer de données discrétisées sur des grilles assez fines qui reflètent de manière appropriée la nature fonctionnelle de ces phénomènes. De nombreuses méthodes ont été proposées afin de sélectionner parmi l'ensemble de ces observations discrétisées un petit nombre de points permettant de répondre aussi bien que possible au problème posé. Cependant, il est souvent intéressant d'attacher également de l'importance à la dynamique de ce type de phénomènes ainsi qu'à leur structure particulière. Une manière adaptée d'y parvenir consiste à modéliser les données dont nous disposons comme la discrétisation d'une variable fonctionnelle (c'est à dire de dimension infinie). C'est une manière de généraliser l'approche multivariée qui permet d'obtenir une représentation plus synthétique des données prenant en compte la régularité et la nature intrinsèque du phénomène dont proviennent nos observations.

La branche de la statistique consacrée à l'étude de données fonctionnelles est actuellement en plein essor en raison des perspectives pratiques et théoriques qu'elle propose. De nombreux modèles et méthodes de statistique multivariée ont été généralisées afin de s'adapter à ce nouveau type de modélisation. On pourra notamment consulter les ouvrages de références de Ramsay et Silverman (1997, 2002, 2005), Bosq (2000), Ferraty et Vieu (2006), ainsi que Ferraty et Romain (2010) . Nous nous intéressons plus particulièrement dans ce travail à l'étude de problèmes de régression sur variable fonctionnelle:

$$Y = r(\mathcal{X}) + \epsilon,$$

où Y est une varable aléatoire réelle, $\mathcal{X}$ une variable aléatoire à valeurs dans un espace semi-métrique $(\mathcal{E}, d)$ et $\mathbb{E}[\epsilon|\mathcal{X}] = 0$.
De nombreux auteurs ont déjà considéré l'estimation de l'opérateur de régression r au travers de variantes de ce modèle correspondant à différentes hypothèses sur la structure de l'opérateur r. On peut notamment évoquer le modèle linéaire fonctionnel introduit par Ramsay et Dalzell (1991):

$$Y = \alpha_0 + < \alpha, \mathcal{X} >_{\mathbb{L}^2([0;1])} + \epsilon,\ (\alpha_0, \alpha) \in \mathbb{R} \times \mathbb{L}^2([0;1]).$$

Ce modèle a été abondamment étudié au cours des dernières années comme en témoignent notamment les travaux de Cardot *et al.* (1999,2000,2007), Ramsay et Silverman (1997, 2005), Preda et Saporta (2005), Hall et Cai (2006), Crambes *et al.* (2009) ainsi que Ferraty et Romain (2010, Chapitre 2).
Divers autres modèles basés sur une certaine structure de r ont été considérés comme on peut notamment le voir dans les travaux de Sood *et al.* (2009) concernant un modèle additif multivarié basé sur les premiers coefficients d'une A.C.P. fonctionnelle, Ait Saidi *et al.* (2008) à propos du modèle à indice simple fonctionnel, ou Aneiros-Perez et Vieu (2009) pour le modèle partiellement linéaire fonctionnel. Cela illustre la grande diversité des modélisations que l'on peut proposer, d'autant plus qu'il est vraisemblable que de nouveaux exemples de modèles "structurels" soient considérés dans les années à venir (modèles additifs fonctionnels, partiellement fonctionnel, ...).
D'autre part, Ferraty et Vieu (2000) ont considéré un modèle non-pararamétrique fonctionnel dans lequel aucune hypothèse n'est faite sur la structure de r, mais simplement sur sa régularité (de type Hölder). De nombreuses références sont données à ce propos dans les travaux de Ferraty *et al.* (2002), Masry (2005), Ferraty et Vieu (2006), Delsol (2007,2009) ainsi que Ferraty et Romain (2010, Chapitres 1, 4, et 5).

2 Tests de structure

2.1 Généralités

Comme nous venons de le voir, la littérature concernant les méthodes d'estimation en régression sur variable fonctionnelle est assez conséquente. L'objectif de cet exposé est

sensiblement différent puisque l'on ne désire pas estimer l'opérateur r mais construire des outils statistiques permettant de tester si il a une certaine structure (e.g. constant, linéaire, multivarié, ...). La littérature consacrée à ce type de problèmes se limite, autant que nous le sachions, aux travaux de Cardot *et al.* (2003,2004) dans le cas particulier du modèle linéaire, Gadiaga et Ignaccolo (2005) qui proposent des tests de non effet basés sur des méthodes de projections ainsi que Chiou et Müller (2007) qui introduisent une approche heuristique pour construire un test d'adéquation. Il semble donc qu'il n'existe pas de méthode générale permettant de tester la validité des différents modèles évoqués au paragraphe précédant. Notons dans ce qui suit $\mathcal{R}$ une famille d'opérateurs de carré intégrables et w une fonction de poids. Au travers de cet exposé nous souhaitons présenter une approche générale permettant de tester l'hypothèse nulle

$$\mathcal{H}_0 : \{\exists r_0 \in \mathcal{R}, P(r(\mathcal{X}) = r_0(\mathcal{X})) = 1\}$$

contre des alternatives locales de la forme

$$\mathcal{H}_{1,n} : \{ \inf_{r_0 \in \mathcal{R}} \|r - r_0\|_{\mathbb{L}^2(wdP_{\mathcal{X}})} \geq \eta_n\}.$$

Notre statistique de test est construite, de manière similaire à l'approche utilisée par Härdle et Mammen (1993), à partir d'un estimateur $\hat{r}$ spécifique au modèle que l'on veut tester (donc à la famille $\mathcal{R}$) et de méthodes d'estimation à noyau (noté K):

$$T_n = \int (\sum_{i=1}^{n} (Y_i - \hat{r}(\mathcal{X}_i)) K\left(\frac{d(\mathcal{X}_i, x)}{h_n}\right))^2 w(x) dP_{\mathcal{X}}(x).$$

Pour des raisons techniques, on fait l'hypothèse que l'estimateur $\hat{r}$ est construit sur un échantillon D_1 indépendant de $D = (\mathcal{X}, Y_i)_{1 \leq i \leq n}$ Un résultat donné par Delsol *et al.* (2010) montre la normalité asymptotique de T_n sous l'hypothèse nulle et sa divergence sous l'alternative sous des hypothèses générales. Ce résultat permet d'envisager l'utilisation de ce type de statistique de test dans un gand nombre de situations pour lesquelles les hypothèses peuvent être vérifiées comme par exemple:

- test d'un modèle a priori: $\mathcal{R} = \{r_0\}$, $\hat{r} = r_0$.
- test de non effet: $\mathcal{R} = \{r : \exists C \in \mathbb{R},\, r \equiv C\}$, $\hat{r} = \overline{Y}_n$.
- test de modèle multivarié: $\mathcal{R} = \{r : r = g \circ V,\, V : \mathcal{E} \to \mathbb{R}^p \text{ connu}, g : \mathbb{R}^p \to \mathbb{R}\}$, $\hat{r}$ estimateur multivarié à noyau construit à partir de $(Y_i, V(\mathcal{X}_i))_{1 \leq i \leq n}$.
- test de linéarité: $\mathcal{R} = \{r : r = \alpha_0 + < \alpha, . >, (\alpha_0, \alpha) \in \mathbb{R} \times \mathbb{L}^2[0;1]\}$, $\hat{r}$ estimateur fonctionnel spline (voir Crambes *et al.* 2009).
- test de modèle à indice simple fonctionnel: $\mathcal{R} = \{r : r = g(< \alpha, . >), \alpha \in \mathcal{E},\, g : \mathbb{R} \to \mathbb{R}\}$, $\hat{r}$ estimateur proposé par Ait Saidi *et al.* (2008).

D'autres situations peuvent également être considérées dès lors que l'on est en mesure de fournir un estimateur $\hat{r}$ ayant de bonnes propriétés.

2.2 Utilisation concrète

La mise en oeuvre de la procédure de test décrite plus haut nécessite de calculer la valeur seuil du test. On pourrait penser l'estimer à partir de la loi asymptotique. Cependant, les termes dominants du bias et de la variance sont difficiles à estimer, c'est pourquoi on préfère utiliser des méthodes de rééchantillonnage. L'idée est de générer, à partir de l'échantillon original, B échantillons pour lesquels l'hypothèse nulle est approximativement vérifiée. Ensuite, on calcule sur chacun de ces échantillons la valeurs de la statistique de test et on prend comme valeur seuil la quantile empirique d'ordre $1-\alpha$ des valeurs obtenues.

Nous proposons la procédure de rééchantillonnage suivante dans laquelle les étapes 1-4 sont réalisées séparément sur les échantillons $D : (\mathcal{X}_i, Y_i)_{1\leq i\leq n}$ et $D_1 : (\mathcal{X}_i, Y_i)_{n+1\leq i\leq N}$. Dans les lignes suivantes $\hat{r}_K$ représente l'estimateur à noyau fonctionnel de l'opérateur de régression calculé à partir de l'échantillon considéré (D or D1).

Procédure de rééchantillonnage:

Pré-traitement:

1. $\hat{\epsilon}_i = Y_i - \hat{r}_K(X_i)$
2. $\tilde{\epsilon}_i = \hat{\epsilon}_i - \bar{\hat{\epsilon}}$

Répéter B fois les étapes 3-5:

3. Générer les résidus (3 méthodes différentes NB, SNB ou WB)
   - NB• $(\epsilon_i^b)_{1\leq i\leq n}$ tirés avec remise parmi $(\tilde{\epsilon}_i)_{1\leq i\leq n}$
   - SNB• $(\epsilon_i^b)_{1\leq i\leq n}$ générés à partir d'une version lissée $\tilde{F}_n$ de la fonction de répartition empirique de $(\tilde{\epsilon}_i)_{1\leq i\leq n}$ $(\epsilon_i^b = \tilde{F}_n^{-1}(U_i),\ U_i \sim \mathcal{U}(0,1))$
   - WB• $(\epsilon_i^b) = \tilde{\epsilon}_i V_i$ où $V_i \sim P_W$ vérifie les conditions suivantes: $E[V_i] = 0$, $E[V_i^2] = 1$ et $E[V_i^3] = 1$.
4. Générer des réponses "correspondant" à $\mathcal{H}_0$
$$Y_i^b = \hat{r}(X_i) + \epsilon_i^b$$
5. Calculer la statistique de test T_n^b à partir de l'échantillon généré $(\mathcal{X}_i, Y_i^b)_{1\leq i\leq N}$

Calculer la valeur empirique du seuil

6. Pour un test de niveau α, prendre comme valeur le quantile empirique d'ordre $1-\alpha$ de la famille $(T_n^b)_{1\leq b\leq B}$.

On considèrere notamment trois exemples de lois P_W données par Mammen (1993). Les différentes méthodes utilisées pour générer les résidus ont des propriétés différentes. Au vu des simulations il semble intéressant d'utiliser des méthodes de type "bootstrap sauvage" (WB) qui produisent des tests plus puissants et sont par nature plus robustes à l'hétéroscédasticité des résidus.

Enfin, l'intégrale par rapport à $P_{\mathcal{X}}$ qui apparait dans la définition de T_n est approchée par une moyenne empirique sur un troisième échantillon indépendant de D_1 et D_2.

2.3 Application en spectrométrie

Les courbes spectrométriques constituent un exemple intéressant de données de nature fonctionnelle. Elles correspondent à la mesure de l'absorption d'une lumière émise en direction d'un produit en fonction de sa longeur d'onde. Les courbes spectrométriques peuvent notamment être utilisées pour connaître le contenu d'un produit sans avoir besoin de réaliser une analyse chimique (voir par exemple Borggaard et Thodberg, 1992). Il est courant, en chimie quantitative, de faire une transformation des courbes originales (correspondant en quelque sorte à des dérivations). L'approche que nous venons de présenter peut être appliquée dans ce contexte pour apporter des éléments de réponse à des questions portant sur

- la validité d'un modèle proposé par des spécialistes.
- l'existence d'un lien entre une des dérivées de la courbe spectrométrique et la quantité que l'on cherche à prédire.
- la nature du lien reliant les dérivées de la courbe spectrométrique et le contenu chimique du produit
- la validité d'un modèle ne prenant en compte que certaines portions ou points de la courbe spectrométrique (ou de ses dérivées) dont on suppose qu'ils résument l'information apportée par la courbe spectrométrique.

Nous illustrerons brièvement la manière dont ces questions peuvent être addréssées en étudiant des jeux de données concrets.

3 Discussion

L'approche générale que nous venons de présenter constitue une première méthode pour construire des tests de structure de nature assez variée en régression sur variable fonctionnelle (se rapporter à Delsol (2008) pour une discussion plus complète). L'utilisation de ces tests sur des données spectrométriques nous fournit des informations pertinentes sur la structure du lien entre la courbe spectrométrique et le contenu chimique du produit. De tels outils peuvent également s'avérer intéressants lorsque l'on cherche à extraire les informations pertinentes de la courbe explicative, ce qui permet souvent d'améliorer la qualité d'estimations. Il semble toutefois intéressant d'essayer d'améliorer notre approche et de proposer d'autres statistiques de test. Il serait notamment important de proposer une alternative qui ne nécessite pas de découper notre échantillon original en trois sous-échantillons, ce qui peut se révéler génant en pratique. Toutefois, il est à noter que notres approche offre un large spectre d'applications possibles. Elle pourrait être utilisée de manière intéressante dans un algorithme permettant de sélectionner les portions ou les points informatifs de la variable explicative fonctionnelle. Elle pourrait aussi se révéler intéressante dans le cadre du choix de la semi-métrique car elle peut permettre de tester la régularité de r par rapport à une semi-métrique d_1 contre la régularité de r par rapport à

une semi-métrique d_2 vérifiant $d_1 \leq d_2$. Nous discuterons les éventuelles améliorations qui peuvent être effectuées pour améliorer notre approche et conclurons sur les perspectives à venir.

Bibliographie

[1] Ait-Saïdi, A., Ferraty, F., Kassa, R. et Vieu, P. (2008) Cross-validated estimations in the single functional index model. soumis

[2] Aneiros-Perez, G. and Vieu, P. (2008) Time series prediction: a semi-functional partial linear model. *Journal of Multivariate Analysis*, Accepté.

[3] Borggaard, C. et Thodberg, H.H. Optimal minimal neural interpretation of spectra, *Analytical chemistry*, 64, (5), 545–551.

[4] Bosq, D. (2000) *Linear Processes in Function Spaces : Theory and Applications*, Lecture Notes in Statistics, 149, Springer-Verlag, New York.

[5] Cardot, H., Ferraty, F., Mas, A. and Sarda, P. (2003) Testing Hypothesys in the Functional Linear Model, *Scandinavian Journal of Statistics*, 30, 241–255.

[6] Cardot, H., Ferraty, F. and Sarda, P. (1999) Functional Linear Model *Statist. and Prob. Letters*, 45, 11–22.

[7] Cardot, H., Ferraty, F. et Sarda, P. (2000) Etude asymptotique d'un estimateur spline hybride pour le modèle linéaire fonctionnel. (French) [Asymptotic study of a hybrid spline estimator for the functional linear model] *C. R. Acad. Sci. Paris*, 330, (6), 501–504.

[8] Cardot, H., Goia, A. et Sarda, P. (2004) Testing for no effect in functional linear regression models, some computational approaches. *Comm. Statist. SimulationComput.* 33, (1), 179–199.

[9] Cardot, H., Crambes, C., Kneip, A. and Sarda, P (2007) Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics and Data Analysis, special issue on functional data analysis*, 51, (10), 4832–4848.

[10] Chiou, J.M. and Müller H.-G. (2007) Diagnostics for functional regression via residual processes. *Computational Statistics & Data Analysis*, 51, (10), 4849–4863.

[11] Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Annals of Statistics*, 37, 35–72.

[12] Delsol, L. (2007) Régression non-paramétrique fonctionnelle : Expressions asymptotiques des moments. *Annales de l'I.S.U.P.*, LI, (3), 43–67.

[13] Delsol, L. (2008) Régression sur variable fonctionnelle: Estimation, Tests de structure et Applications. *Thèse de doctorat de l'Université de Toulouse*.

[14] Delsol, L. (2009) Advances on asymptotic normality in nonparametric functional Time Series Analysis. *Statistics*, 43, (1), 13–33.

[15] Delsol, L., Ferraty, F., and Vieu, P. (2010) Structural test in regression on functional variables. soumis

[16] Ferraty F., Goia A. and Vieu P. (2002b) Functional nonparametric model for time series : a fractal approach for dimension reduction. *Test*, 11, (2), 317–344.

[17] Ferraty, F. et Romain, Y. (2010) *Oxford Handbook on Statistics and FDA* To appear.

[18] Ferraty, F. and Vieu, P. (2000) Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés, *Compte Rendus de l'Académie des Sciences*, Paris, 330, 403–406.

[19] Ferraty, F. and Vieu, P. (2006) *Nonparametric Functional Data Analysis: Theory and Practice* , Springer-Verlag, New York.

[20] Gadiaga, D. and Ignaccolo, R.(2005) Test of no-effect hypothesis by nonparametric regression. *Afr. Stat.*, 1, (1), 67–76.

[21] Hall, P. and Cai, T.T. (2006) Prediction in functional linear regression. (English summary) *Ann. Statist.*, 34, (5), 2159–2179.

[22] Härdle, W. and Mammen, E. (1993) Comparing Nonparametric Versus Parametric Regression Fits *Annals of Statistics*, 21, (4), 1926–1947.

[23] Masry, E. (2005) Nonparametric regression estimation for dependent functional data : asymptotic normality, *Stochastic Process. Appl.*, 115, (1), 155-177.

[24] Mammen, E. (1993) Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.*, 21, (1), 255–285.

[25] Preda, C. et Saporta, G. (2005) PLS regression on a stochastic process. *Comput. Statist. Data Anal.*, 48, (1), 149–158.

[26] Ramsay, J. and Dalzell, C. (1991) Some tools for functional data analysis, *J.R. Statist. Soc. B.*, 53, 539–572.

[27] Ramsay, J. and Silverman, B. (1997) *Functional Data Analysis*, Springer-Verlag, New York.

[28] Ramsay, J. and Silverman, B. (2002) *Applied functional data analysis : Methods and case studies*, Spinger-Verlag, New York.

[29] Ramsay, J. and Silverman, B. (2005) *Functional Data Analysis (Second Edition)*, Spinger-Verlag, New York.

[30] Sood, A., James, G. and Tellis, G. (2009) Functional Regression: A New Model for Predicting Market Penetration of New Products *Marketing Science*, 28, 36–51.