Procédure diagnostique en arbre utilisant les tests lisses d'adéquation

Walid AL AKHRAS [1] & Gilles DUCHARME [2]

[1] *Laboratoire de probabilités et statistique cc 051, Université Montpellier, 34095 Montpellier cedex 5, France, walid.al-akhras@univ-montp2.fr*

[2] *Laboratoire de probabilités et statistique cc 051, Université Montpellier, 34095 Montpellier cedex 5, France, gilles.ducharme@univ-montp2.fr*

Résumé. Un test d'adéquation est une procédure d'évaluation de l'hypothèse H_0 : $F = F_0$, où F est la loi, inconnue, d'une variable aléatoire X qui prend ses valeurs dans l'ensemble $\mathcal{S}$, et F_0 est la loi de référence. Cette hypothèse H_0 peut être non rejetée ou rejetée. Dans ce dernier cas, il est alors intéressant de connaître les raisons d'un tel rejet. Pour cela, il faut appliquer des procédures qui s'appellent "Procédures de diagnostic d'adéquation" (PDA). Dans la littérature, il y a deux classes de PDA. La première est locale et basée sur les composantes de la statistique du $\mathcal{X}^2$ de Pearson (1900); elle permet de déterminer des intervalles de $\mathcal{S}$ où le modèle ne colle pas aux données. La deuxième est globale et basée sur les composantes de la statistique du test lisse de Neyman (1937); elle donne des informations sur les écarts entre les moments du modèle posé en H_0 et ceux des données. Il nous a semblé que si on pouvait combiner ces deux méthodes d'une certaine façon, il serait possible d'aller plus loin dans l'extraction des informations diagnostiques. Notre idée consiste à proposer une procédure de diagnostic locale basée sur le test lisse. Il faut donc disposer de tests lisses "locaux", c'est-à-dire restreints à des éléments d'une partition de $\mathcal{S}$. La méthode qu'on utilise est basée sur une structuration en arbre des hypothèses de la famille de tests, cette méthode assure un contrôle fort du taux d'erreur FWER.

Mots-clés. Test lisse d'adéquation, diagnostic de test, procédure hiérarchique.

Abstract. A test of goodness-of-fit is a procedure of evaluation of the hypothesis $H_0 : F = F_0$, where F is the, unknown, distribution, of a random variable X, which takes its values in $\mathcal{S}$, and F_0 is the reference distribution. This hypothesis H_0 can be not rejected or rejected. In the last case, it is interesting to know the reasons for such rejection. For this purpose, one must applies the procedures that are called "Procedure of diagnostic of GoFit" (PDA). In the literature, there are two classes of PDA. The first one is local and based on the components of the statistics $\mathcal{X}^2$ of Pearson (1900); it allows to determine the intervals in $\mathcal{S}$ where the model does not stick to the data. The second

one is global and based on the components of Statistics of smooth GoFit (Neyman 1937); it provides information about deviations between the moments of model under H_0 and those of the data. It seemed to us that if we can combine these two methods, it would be possible to go further in extracting diagnostic information. Our idea is to propose a local diagnostic procedure based on the smooth test. It is therefore necessary to have smooth tests "local", i.e. restricts to elements of a partition of $\mathcal{S}$. The method which we use is based on a tree structure hypothesis of the family of tests, this method assures a strong control of Family wise eror rate FWER.

Keywords. Test of goodness of fit, diagnostic of test, hierarchical procedure.

1 Introduction

Un modèle paramétrique statistique postulé pour une variable aléatoire X qui prend ses valeurs dans l'ensemble $\mathcal{S}$, est une densité $f_0(x, \theta)$ (e.g. $X \sim f_0(x; \theta)$) qui dépend d'un paramètre $\theta \in \Theta$ qui peut être inconnu. Les modèles paramétriques sont des outils importants dans les applications de la statistique, car ils nous aident à comprendre comment le phénomène étudié se comporte.

Cependant, avant d'utiliser ces modèles dans des applications, il faut s'assurer que le modèle *considéré* n'est pas loin du vrai comportement probabiliste de X, c'est-à-dire à sa vraie densité. Dans ce but, les données sont soumises à une procédure de test statistique d'adéquation. Une telle procédure statistique pose la véracité du modèle *considéré* comme une hypothèse nulle H_0 et utilise les données pour essayer de valider cette hypothèse nulle.

La réponse de cette procédure est binaire: rejet / non rejet du modèle choisi. En cas de non rejet, le test d'adéquation n'a pas trouvé de raison de déclarer que le modèle *considéré* est erroné. Dans ce cas, le modèle *considéré* peut être par la suite utilisé avec un certain degré de confiance, lequel dépend des risques d'erreur (Type 1 et Type 2) qui ont été contrôlés ou calculés.

Cependant, quant le test d'adéquation rejette le modèle *considéré*, on se trouve avec une information souvent difficile à exploiter. Ainsi, au cous des années, les statisticiens ont essayé d'extraire des tests d'adéquation des indices sur les aspects des modèles posés en hypothèses nulles qui ont besoin de corrections. De telles procédures d'extraction sont nommées "procédures de diagnostic d'adéquation". Le but de ces procédures est de fournir des informations qui permettront de touver un nouveau modèle *considéré* qui ne sera pas rejeté et sera donc suffisamment consonant avec les données pour être mis à l'usage avec le degré voulu de confiance.

La première PDA est liée au test de $\mathcal{X}^2$ (Pearson 1900) dont la statistique de test, pour un échantillon $\mathcal{E} = \{X_1, \ldots, X_n\}$ et une partition $I_1, \ldots, I_K$ du support, est donnée par

$$\mathcal{X}^2 = \sum_{k=1}^{K} \frac{(N_k - np_k)^2}{np_k}. \tag{1}$$

Cette PDA (Eye et Bogat, 2004) utilise les valeurs des composantes $C_k = \frac{(N_k - np_k)^2}{np_k}$, pour étudier l'écart entre le nombre d'observation N_k dans chaque intervalle I_k, et ce qui est attendu (soit np_k). L'information statistique ici sera donc une information locale dans chaque intervalle de la partition.

La deuxième PDA est produite par Henze et Klar (1996). Cette PDA est basée sur le test lisse de Neyman (1937), dont la statistique de test d'uniformité $U(0,1)$ est donnée par

$$\mathcal{R}_K = \sum_{k=1}^{K} n \left(\overline{L}_k\right)^2, \tag{2}$$

où $\overline{L}_k = n^{-1} \sum_{i=1}^{n} L_k(X_i)$ et $L_k(\cdot)$ est le polynôme de Legendre orthonormal (sur $(0,1)$) d'ordre k. Il se trouve que les composantes $n\overline{L}_k$ donnent quelques informations diagnostiques. En effet, comme $L_1(x) = 2\sqrt{3}(x - 1/2)$, une grande valeur de $n\left(\overline{L}_1\right)^2$ indique que les données ne sont pas consonantes avec le fait que l'espérance de X sous H_0 doit être égale à $1/2$. De même, $L_2(x) = -6\sqrt{5}\left((x-1/2)^2 - 1/12\right)$ et une grande valeur de $n\left(\overline{L}_2\right)^2$ indique que la variance de X s'écarte de celle de la loi $U(0,1)$ (qui est égale à $1/12$). De même, $L_3(x)$ est lié au coefficient d'asymétrie (ou skewness) et $L_4(x)$ est lié au coefficient d'aplatissement (ou kurtosis). Donc, cette PDA fourni des informations diagnostiques globales sur les écarts entre les moments du modèle posé en H_0 et ceux des données.

Aucune de ces deux PDA, n'est conçue le risque d'amender un modèle qui n'a pas besoin de correction; En plus, aucune n'extrait toute l'information diagnostique. Cette constatation est le point de départ de notre travail car il nous a semblé que si l'on pouvait les combiner d'une certaine façon, et construire une PDA locale liée au test lisse, il serait possible d'aller plus loin dans l'extraction des informations diagnostiques.

2 Eléments constitutifs d'une PDA locale par tests lisses

Pour appliquer cette PDA locale basée sur le test lisse, il faut disposer des tests lisses "locaux", c'est-à-dire restreint à des éléments d'une partition de $\mathcal{S}$. Nous allons considérer le cadre où X est une variable aléatoire *continue* à valeurs dans $\mathcal{S}$ (pas nécessairement l'intervalle $(0,1)$) de densité $f_X(\cdot)$. Sous H_0, la densité posée ne sera plus la loi uniforme et elle pourrait dépendre de paramètres inconnus. On va aussi considérer les cas où l'intervalle d'intérêt $I = (a,b) \subset \mathcal{S}$ est fixe mais pourrait aussi dépendre des paramètres inconnus de la loi.

On a distinguer cinq cas, selon la connaissance que l'on a des paramètres de la densité posée en H_0 et des bornes de l'intervalle I. Les deux premiers cas sont les tests lisses globales, et les trois qui suivent sont des nouveaux résultats, qui aident à effectuer des tests lisses locaux.

1-Cas $H_0 : X \sim f_0(.,\boldsymbol{\lambda})$, $\boldsymbol{\lambda}$ connu. Dans ce cas, si l'on applique la transformation $Y = F_0(X)$, la variable aléatoire Y est uniformément distribuée sur l'intervalle $(0,1)$ et le problème de tester $H_0 : f_X(\cdot) = f_0(\cdot)$ est équivalent au problème "canonique" de tester $H_0^{can} : Y \sim U(0,1)$. La statistique de test sera celle de Neyman (1937):

$$\mathcal{R}_K = n\left(\overline{L}_1\right)^2 + n\left(\overline{L}_2\right)^2 + \cdots + n\left(\overline{L}_K\right)^2 \xrightarrow{L} \mathcal{X}^2_K. \tag{3}$$

2-Cas $H_0 : X \sim f_0(\cdot;\boldsymbol{\lambda})$, $\boldsymbol{\lambda}$ inconnu et estimé par le EVM $\widehat{\boldsymbol{\lambda}}$. On suppose ici que la densité $f_0(\cdot;\boldsymbol{\lambda})$ satisfait aux conditions de régularité assurant que $\widehat{\boldsymbol{\lambda}} \xrightarrow{PS} \boldsymbol{\lambda}$ et $\sqrt{n}\left(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\right) \xrightarrow{L} \mathcal{N}_m(\mathbf{0}, \mathcal{I}_{\boldsymbol{\lambda}})$, où $\mathcal{I}_{\boldsymbol{\lambda}}$ est l'information de Fisher pour $\boldsymbol{\lambda}$. Si $\bar{\hat{L}}_k = n^{-1}\sum_{i=1}^n L_k\left(F_0\left(X_i;\widehat{\boldsymbol{\lambda}}\right)\right)$ et $\bar{\hat{L}}^T = \left(\bar{\hat{L}}_1, \cdots, \bar{\hat{L}}_K\right)$, on a $\sqrt{n}\bar{\hat{L}} \xrightarrow{L} \mathcal{N}\left(0, I_K - J_{\widehat{\boldsymbol{\lambda}}}\mathcal{I}^{-1}_{\widehat{\boldsymbol{\lambda}}}J^T_{\widehat{\boldsymbol{\lambda}}}\right)$. La statistique de test dans ce cas est calculée par (Kopecky et Pierce 1979):

$$\hat{\mathcal{R}}_K = n\bar{\hat{L}}^T\left(I_K - J_{\widehat{\boldsymbol{\lambda}}}\mathcal{I}^{-1}_{\widehat{\boldsymbol{\lambda}}}J^T_{\widehat{\boldsymbol{\lambda}}}\right)^{-1}\bar{\hat{L}} \xrightarrow{L} \mathcal{X}^2_K,$$

où $J_{\widehat{\boldsymbol{\lambda}}} = Cov\left[L^T\left(F_0(X;\lambda)\right), \frac{\partial}{\partial\lambda^T}\log f_0(X;\lambda)\right]$.

3-Cas $H_0 : X \sim f_0(\cdot;\boldsymbol{\lambda})$, $\boldsymbol{\lambda}$ connu; données tronquées à un intervalle $I = (a,b)$. Ici, on veut tester si les données, qui appartiennent à l'intervalle (a,b), suivent la troncation de $f_0(\cdot;\boldsymbol{\lambda})$ sur cet intervalle. Pour assurer qu'il y aura des données dans chaque intervalle, on suppose que $0 < p_0^{(a,b)} < 1$ où $p_0^{(a,b)} = P_0[X \in (a,b)]$ (Probabilité sous H_0). La difficulté est que le nombre des données $N^{(a,b)}$ dans l'intervalle (a,b) est aléatoire ($N^{(a,b)} \sim B\left(n, p_0^{(a,b)}\right)$), mais le fait qu'il est binomiale nous a aidé à montrer que $\sqrt{N^{(a,b)}}\bar{L}^*_k = \frac{1}{\sqrt{N^{(a,b)}}}\sum_{i=1}^{N^{(a,b)}} L_k\left(F_0^{(a,b)}(X_i)\right) \xrightarrow{L} \mathcal{X}^2$. La statistique de test locale sera donc

$$\mathcal{R}_K^{(a,b)} = N^{(a,b)}\left(\bar{L}^*_1\right)^2 + N^{(a,b)}\left(\bar{L}^*_2\right)^2 + \ldots + N^{(a,b)}\left(\bar{L}^*_K\right)^2 \xrightarrow{L} \mathcal{X}^2_K.$$

4-Cas $H_0 : X \sim f(x;\boldsymbol{\lambda})$, $\boldsymbol{\lambda}$ estimé par le EVM $\widehat{\boldsymbol{\lambda}}$; données tronquées à $I = (a,b)$. On note ici que l'estimateur $\hat{\boldsymbol{\lambda}}$ est calculé à partir de toutes les données, non seulement de celles qui appartiennent à (a,b), donc l'information de Fisher $\mathcal{I}_{\widehat{\boldsymbol{\lambda}}}$ ne changera pas. Un travail similaire que le cas 2, en prenant compte du fait que $N^{(a,b)} \sim B\left(n, p_0^{(a,b)}\right)$, on a $\sqrt{N^{(a,b)}}\bar{\hat{L}} \xrightarrow{L} \mathcal{N}\left(0, I_K - p_0^{(a,b)}J^{(a,b)}_{\boldsymbol{\lambda}}\mathcal{I}^{-1}_{\boldsymbol{\lambda}}\left(J^{(a,b)}_{\boldsymbol{\lambda}}\right)^T\right)$, où $J^{(a,b)}_{\boldsymbol{\lambda}} = Cov\left[L^T\left(F_0^{(a,b)}(X;\lambda)\right), \frac{\partial}{\partial\lambda^T}\log f_0^{(a,b)}(X;\lambda)\right]$ est calculée à partir des données tronquées. En fin la statistique de test sera

$$\hat{\mathcal{R}}_K^{(a,b)} = N^{(a,b)}(\bar{\hat{L}})^T\left(I_K - \hat{p}_0^{(a,b)}J^{(a,b)}_{\widehat{\boldsymbol{\lambda}}}\mathcal{I}^{-1}_{\widehat{\boldsymbol{\lambda}}}(J^{(a,b)}_{\widehat{\boldsymbol{\lambda}}})^T\right)^{-1}\bar{\hat{L}} \xrightarrow{L} \mathcal{X}^2_K.$$

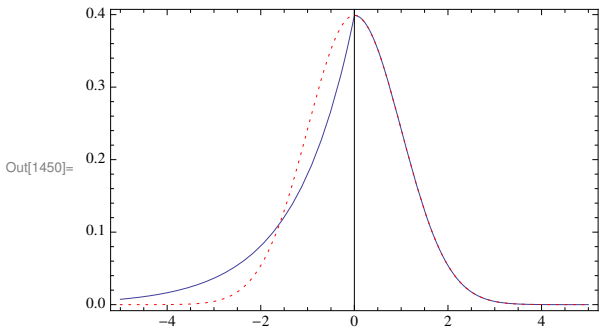

Figure 1: Exemple: La densité de Laplace$(0, 1.25331)$ à gauche et $\mathcal{N}(0,1)$ à droite.

5-Cas $X \sim f(x;\boldsymbol{\lambda})$, $\boldsymbol{\lambda}$ estimé par $\widehat{\boldsymbol{\lambda}}$; données tronquées à un intervalle $\boldsymbol{I} = (a(\boldsymbol{\lambda}), b(\boldsymbol{\lambda}))$.
Dans ce cas, on a trois difficultés ensemble, la taille d'échantillon aléatoire, les paramètres estimés et les bornes de l'intervalle qui dépendent des estimateurs. Le résultat que l'on a eu après un long travail, était surprenant; Le calcul de la statistique de test ne change pas et c'est exactement pareil que dans le cas 4.

Dans tous ces cas, on rejette l'hypothèse nulle $H_0^{(a,b)}$ si la statistique de test dépasse $x^2_{K,1-P_0[(a,b)]\alpha}$, le quantile d'ordre $1 - P_0[(a,b)]\alpha$ de la loi $\mathcal{X}^2_K$.

Le choix de ces niveaux de test $P_0[(a,b)]\alpha$, nous permet de résoudre le problème de test multiple dans cette PDA. On a montré qu'avec ce choix, le taux d'erreur FWER (Family wise error rate) est fortement contrôlé.

3 Application

Pour expliquer le fonctionnement de notre PDA, on a pris l'exemple où la variable aléatoire X suit une loi Laplace$(0, 1.25331)$ à gauche et $\mathcal{N}(0,1)$ à droite (Figure 1). On a généré 10 000 échantillons de taille $n = 400$ selon ce modèle et essayer de diagnostiquer pour déterminer la cause d'un rejet éventuel de l'hypothèse $H_0 : X \sim \mathcal{N}(0,1)$ ou $\mathcal{N}(\overline{X}, S)$. On a utilisé l'arbre de partition

$$\begin{array}{ccccccc} & & & \mathbb{R} & & & \\ & & (-\infty,\mu) & & & (\mu,+\infty) & \\ (-\infty,\mu-\sigma) & & & (\mu-\sigma,\mu) & (\mu,\mu+\sigma) & & (\mu+\sigma,+\infty) \end{array}$$

La stratégie de notre PDA consiste à appliquer la règle "tester jusqu'à l'acceptation". C'est-à-dire, elle commence à tester l'adéquation globale dans le noeud racine, si l'on accepte, la procédure s'arrête et donc le modèle n'a pas besoin de correction. Mais, si l'on rejette, la PDA va descendre pour l'adéquation locale dans les intervalles $(-\infty, \mu)$ et $(\mu, +\infty)$. Encore, la PDA s'arrête en cas d'acception, et continue en cas de rejet pour tester dans les niveaux suivant de l'arbre.

On a appliqué notre exemple d'abord, dans le cas des paramètres connues. La PDA a rejeté la normalité globale dans toutes les répétions de la procédure, puis elle est descendu pour tester la normalité tronquée sur les intervalles $(-\infty, \mu)$ et $(\mu, +\infty)$. La PDA accepte 97.64% la normalité tronqué sur $(\mu, +\infty)$ donc elle arrête, et contient justement à tester les enfants de $(-\infty, \mu)$ après l'avoir rejeté 100% des fois. À la fin, notre PDA a déterminé que l'intervalle $(-\infty, \mu - \sigma)$ qui cause le rejet de la normalité globale.

Après, on a appliqué les cas des paramètres inconnues (arbres fixes et estimés). On a remarqué que les résultats sont comparables. La PDA a rejeté l'hypothèse nulle globale ($\mathbb{R}$) et locale ($(-\infty, \overline{X})$ et $(\overline{X}, +\infty)$). Le test a rejeté la normalité tronquée sur $(\overline{X}, +\infty)$, car il teste si, les données prises de la loi $\mathcal{N}(0, 1)$, suivent la loi $\mathcal{N}(\overline{X}, S)$ avec $\overline{X} \simeq -0.228$ et $S^2 \simeq 2.019$. La PDA nous a dit que la cause de rejet de normalité globale, est dans les intervalles, $(\overline{X} + S, +\infty)$ (rejeté 62.49%) et $\left(-\infty, \overline{X} - S\right)$ (rejeté 99.14%).

4 Perspectives

Dans l'exemple de Section précédent, on a déterminé l'endroit du problème qui a causé le rejet de l'hypothèse nulle globale, mais on n'a pas dit quelle est ce problème. On va essayer d'utiliser les propriétés diagnostiques des polynômes de Legendre évoquées dans la Section 1, pour déterminer à la fois, l'endroit et le genre du problème qui a causé le rejet.

Notons que l'on a utilisé des arbres qui dépendent uniquement des estimateurs EVMs des paramètres, on voudrais connaître si la PDA fonctionne toujours dans le cas où l'arbre dépend d'autres estimateurs.

En fin, il faudra chercher à appliquer notre PDA sur d'autres lois comme la loi exponentielle; avec des éventuelles censures.

Bibliographie

[1] Henze, N. and Klar, B. (1996). Properly rescaled components of smooth tests of fit are diagnostic. *Australian Journal of Statistics* 38(1) :61–74.
[2] Kopecky, K. J. and Pierce, D. A. (1979). Efficiency of smooth goodness-of-fit tests. Journal of the *American Statistical Association* 74(366) :393–397.
[3] Neyman, J. (1937). "smooth" test for goodness of fit. *Skand. Aktuarietidskr* 20 :150–199.
[4] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series* 50(302) :157–175.
[5] Von eye, A. and Bogat, G. A. (2004). Testing the assumption of multivariate normality. *Psychology Science* 46(2) :243–258.