

Utilisation de la vraisemblance empirique pour un test d'homogénéité

Flore HARLÉ^{1,2}, Florent CHATELAIN², Cédric GOUY-PAILLER¹, Sophie ACHARD²

¹CEA, LIST, Laboratoire d'Analyse de Données et Intelligence des Systèmes
91191 Gif-sur-Yvette CEDEX, France

²Université Grenoble Alpes et CNRS, GIPSA-Lab
38000 Grenoble, France

flore.harle@cea.fr, florent.chatelain@gipsa-lab.grenoble-inp.fr
cedric.gouy-pailler@cea.fr, sophie.achard@gipsa-lab.grenoble-inp.fr

Résumé – Il existe de nombreux tests d'hypothèses pour la détection d'une différence entre les moyennes des distributions de deux échantillons univariés, mais qui nécessitent souvent un modèle pour les données. Nous proposons ici d'employer la fonction de vraisemblance empirique, qui a l'avantage d'être non paramétrique et s'applique donc à tout type de distributions. Le principe du test d'hypothèse de Vraisemblance Empirique sur une Moyenne Empirique (VEME) est détaillé puis ses performances sont évaluées et comparées à des tests classiques pour plusieurs cas de figure. On met ainsi en valeur l'avantage de l'approche non paramétrique et les propriétés asymptotiques de la statistique proposée.

Abstract – A large variety of statistical hypothesis tests aims at detecting a difference between the means of two univariate samples, but they often require a model on the data. Our idea here is to build a test on the empirical likelihood function, and take advantage of its nonparametrical nature to handle a lot of distributions. The principle of the VEME test is explained and its performances are established for several cases and compared with classical tests. Thus the advantages of this nonparametric method and the asymptotical properties of its statistics are highlighted.

1 Introduction

La comparaison de deux échantillons est un problème très largement rencontré, et pour lequel il existe un grand nombre de méthodes. On le formalise souvent sous la forme d'un test d'hypothèse, dont l'hypothèse nulle H_0 est par exemple que les distributions des deux populations sont les mêmes, contre l'alternative H_1 que ces distributions diffèrent, soit pour la valeur d'un ou plusieurs paramètres, soit qu'elles n'appartiennent pas à la même famille de fonctions. Lorsqu'on dispose de connaissances a priori sur ces distributions, on peut employer des tests bien connus, comme le test t de Student et ses variantes. Ils supposent souvent que les données suivent la loi normale, et peuvent s'avérer mal calibrés pour d'autres distributions. Sans connaissance sur les données, il est donc plus adéquat d'employer des méthodes non paramétriques, par exemple le test de Kolmogorov-Smirnov ou les tests de rang, comme le test de Mann-Whitney-Wilcoxon. La vraisemblance empirique, initialement introduite dans [7], est une méthode d'inférence non paramétrique entièrement basée sur les données, construite à partir d'une fonction de répartition empirique d'une population. Owen lui a consacré un ouvrage très complet [8], elle est souvent utilisée en économétrie [4], et plus récemment dans le domaine du traitement du signal [2, 9].

Diverses méthodes intégrant la vraisemblance empirique ont été proposées, par exemple pour la détection et la localisation de ruptures épidémiques dans des données multivariées [6] ou la régression linéaire [5]. Une série de statistiques pour tester

l'existence de ruptures, la symétrie autour de zéro, l'indépendance, et d'autres propriétés particulières, est également proposée dans [1].

Cette communication présente un test, construit à partir du rapport de vraisemblance empirique, qui permet de détecter un saut de moyenne entre les distributions de deux échantillons de données univariées. Ce nouveau test a l'avantage d'être défini pour des échantillons de tailles petites ou très différentes, contrairement aux tests de vraisemblance empirique déjà proposés dans la littérature. La définition de la vraisemblance empirique et les propriétés des fonctions qui en sont issues sont rappelées dans la partie 2, puis le principe du test est expliqué dans 3. Des simulations dans la partie 4 viennent illustrer les avantages et les limites du test, comparé à des méthodes de référence. Enfin, l'intérêt de ce test est discuté dans la partie 5.

2 Vraisemblance empirique

La vraisemblance empirique repose sur un modèle qui généralise la notion de fonction de répartition (FdR) empirique. La FdR empirique est un estimateur de la FdR qui, pour un échantillon $\mathbf{X} = (X_1, \dots, X_n)$, est défini par

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq x}, \quad (1)$$

où $\mathbb{1}_A$ est la fonction qui vaut 1 lorsque la condition A est vérifiée et 0 sinon. Cet estimateur équivaut à placer une

masse de probabilité $\frac{1}{n}$ en chaque X_i . La vraisemblance empirique est définie comme la probabilité d'obtenir exactement les X_1, \dots, X_n pour une FdR des variables F donnée :

$$L(F) = \prod_{i=1}^n (F(X_i) - F(X_{i-})), \quad (2)$$

où $F(x_{i-}) = \Pr(X < x_i)$ lorsque X est distribuée selon F . Cette vraisemblance est non nulle à condition de placer une masse de probabilité en chaque X_i , ces masses n'étant pas nécessairement équiprobables. En notant, pour $1 \leq i \leq n$,

$$p_i = F(X_i) - F(X_{i-}),$$

la masse de probabilité en X_i , $L(F)$ se réexprime comme $L(F) = \prod_{i=1}^n p_i$, ce qui correspond à modéliser par une loi multinomiale l'échantillon \mathbf{X} . La vraisemblance $L(F)$ étant maximisée lorsque $p_i = \frac{1}{n}$, i.e. pour la FdR empirique F_n , on définit finalement le rapport de vraisemblance empirique pour une FdR donnée F et la FdR empirique F_n :

$$R(F) = \frac{L(F)}{L(F_n)} = \prod_{i=1}^n np_i. \quad (3)$$

L'intérêt de cette approche est de permettre une estimation de l'ensemble p des poids p_i de chaque variable aléatoire en maximisant $R(F)$, sous des contraintes spécifiques au problème considéré, sans nécessiter de connaissances sur la distribution F . Ce rapport est entièrement construit sur les données, et constitue en cela une approche non paramétrique.

Maximiser $R(F)$ revient maintenant à déterminer la fonction de profil de la vraisemblance empirique, qui intègre des contraintes sur les poids p_i et un ou plusieurs paramètres d'intérêt. Pour l'étude de la moyenne μ , le paramètre sur lequel portent les hypothèses de notre test, cette fonction s'écrit :

$$\mathcal{R}(\mu) = \max_p \left\{ \prod_{i=1}^n np_i \mid \sum_{i=1}^n p_i X_i = \mu, p_i > 0, \sum_{i=1}^n p_i = 1 \right\}. \quad (4)$$

Une version non paramétrique du théorème de Wilks, démontrée dans [8], permet alors d'obtenir le résultat clé suivant :

Théorème 2.1. *Soient X_1, \dots, X_n des variables aléatoires indépendantes et de même distribution. Soit μ_0 l'espérance de ces variables, et on suppose que leur variance est finie et non nulle. Alors $-2 \ln(\mathcal{R}(\mu_0))$ converge en distribution quand $n \rightarrow \infty$ vers la loi du χ^2 à un degré de liberté, notée $\chi^2_{(1)}$.*

Ce théorème permet notamment de définir des intervalles de confiance (asymptotiques) et de formuler des tests d'hypothèses pour le paramètre d'intérêt, ici la moyenne. Ce résultat permet aussi de déterminer la loi asymptotique pour notre statistique de test, comme cela est montré dans la partie suivante.

D'un point de vue pratique, le calcul de $\ln(\mathcal{R}(\mu))$ nécessite de résoudre le problème d'optimisation sous contrainte posé par l'expression (4) pour une moyenne μ donnée. Dès lors que μ est un point intérieur à l'enveloppe convexe des données, l'ensemble des poids p_i qui satisfont les contraintes est convexe

et non vide. La fonction objectif $\sum_{i=1}^n \ln(np_i)$ étant strictement concave sur l'ensemble convexe des poids p_i , il existe alors un unique maximum. Le lagrangien du problème dual s'écrit alors pour $\lambda \in \mathbb{R}$ et $\eta \in \mathbb{R}$ comme :

$$H(p, \lambda, \eta) = \sum_{i=1}^n \ln(np_i) - n\lambda \sum_{i=1}^n p_i (X_i - \mu) + \eta \left(\sum_{i=1}^n p_i - 1 \right). \quad (5)$$

En annulant les dérivées partielles par rapport aux p_i , il vient :

$$p_i = \frac{1}{n} \frac{1}{1 + \lambda(X_i - \mu)}, \quad (6)$$

pour $1 \leq i \leq n$, avec $\eta = -n$, λ étant obtenu numériquement comme la solution de $\frac{1}{n} \sum_{i=1}^n \frac{X_i - \mu}{1 + \lambda(X_i - \mu)} = 0$, voir [8]. Notons enfin que le profil $\mathcal{R}(\mu)$ n'est plus défini lorsque μ n'est pas à l'intérieur de l'enveloppe convexe des données, les contraintes n'étant plus satisfaites. Par convention, on pose alors $\mathcal{R}(\mu) = 0$.

3 Test d'hypothèse VEME

La formulation du problème est la suivante : on compare deux échantillons $\mathbf{X}_1 = (X_1, \dots, X_{n_1})$ et $\mathbf{X}_2 = (X_{n_1+1}, \dots, X_{n_1+n_2})$, de tailles respectives n_1 et n_2 . On note $n = n_1 + n_2$ la taille totale des deux échantillons et $\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i$ et $\bar{X}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^n X_i$ leurs moyennes empiriques respectives. On suppose que les variables X_i , pour $1 \leq i \leq n$ sont indépendantes, et que, pour chacun des échantillons, les variables sont identiquement distribuées selon une loi inconnue notée respectivement F_1 et F_2 . Soient μ_1 et μ_2 les moyennes des distributions F_1 et F_2 . Détecter une différence de moyenne entre les deux échantillons conduit au test d'hypothèse suivant

$$\begin{aligned} H_0: & \mu_1 = \mu_2, \\ H_1: & \mu_1 \neq \mu_2. \end{aligned} \quad (7)$$

Le test VEME, pour Vraisemblance Empirique sur une Moyenne Empirique, repose sur la vraisemblance empirique de l'échantillon global $\mathbf{X} = (X_1, \dots, X_{n_1+n_2})$ de taille $n = n_1 + n_2$ évaluée en la moyenne empirique \bar{X}_1 ou \bar{X}_2 de l'un des deux échantillons. Sans perte de généralité, on peut supposer que la moyenne empirique est celle du premier échantillon \bar{X}_1 , quitte à intervertir les échantillons \mathbf{X}_1 et \mathbf{X}_2 . La statistique de test est alors le profil de la vraisemblance empirique

$$\mathcal{R}(\bar{X}_1) = \max_p \left\{ \prod_{i=1}^n np_i \mid \sum_{i=1}^n p_i X_i = \bar{X}_1, p_i > 0, \sum_{i=1}^n p_i = 1 \right\}.$$

Cette statistique présente l'avantage d'être toujours définie quelles que soient les tailles $n_1 \geq 1$ et $n_2 \geq 1$ des échantillons testés, puisque par construction \bar{X}_1 appartient toujours à l'enveloppe convexe de l'échantillon global \mathbf{X} . C'est une différence majeure par rapport aux tests de vraisemblance empirique proposés dans la littérature [3, 6]. Ces derniers reposent classiquement sur le profil du produit des vraisemblances empiriques des deux échantillons \mathbf{X}_1 et \mathbf{X}_2 sous la contrainte d'égalité des moyennes. Ils s'avèrent mal calibrés, voire indéfinis dès lors qu'au moins un des deux échantillons est trop petit.

Rejeter l'hypothèse nulle revient maintenant à déterminer si $\mathcal{R}(\bar{X}_1) < z$ pour un seuil z donné. Afin de contrôler le risque de première espèce à un niveau de significativité α donné, il est nécessaire de connaître la loi de $\mathcal{R}(\bar{X}_1)$ sous H_0 . Sous l'hypothèse d'homoscédasticité, ce contrôle est assuré par le résultat asymptotique suivant :

Proposition 3.1. *On suppose que sous H_0 les variances des lois F_1 et F_2 existent et sont identiques. On note $\gamma = \lim_{n \rightarrow +\infty} \frac{n_2}{n_1}$, et l'on suppose que $0 < \gamma < +\infty$. Alors, sous H_0 , $-2 \frac{n_1}{n_2} \ln(\mathcal{R}(\bar{X}_1))$ converge en distribution quand $n \rightarrow \infty$ vers la loi du χ^2 à un degré de liberté, notée $\chi_{(1)}^2$.*

Preuve. La preuve reprend les éléments de celle du théorème 2.1 donnée dans [8, pp. 219-222] : une majoration du multiplicateur de Lagrange λ introduit dans (5) permet d'obtenir d'après l'expression des probabilités p_i donnée en (6) le développement suivant lorsque $n \rightarrow +\infty$:

$$-2 \ln(\mathcal{R}(\bar{X}_1)) = n \frac{(\bar{X} - \bar{X}_1)^2}{S} + o(1), \quad (8)$$

où $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ est la moyenne empirique de l'échantillon global \mathbf{X} et $S = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_1)^2$. Il suffit donc de montrer

$$\frac{n_1}{n_2} \frac{n(\bar{X} - \bar{X}_1)^2}{S} \xrightarrow[n \rightarrow +\infty]{Loi} \chi_{(1)}^2. \quad (9)$$

On note $\sigma^2 = \text{var}(X_i)$ la variance des variables sous l'hypothèse nulle. La variable $N = \sqrt{n}(\bar{X} - \bar{X}_1) = \frac{n_2}{\sqrt{n}}(\bar{X}_2 - \bar{X}_1)$ est centrée et de variance $\text{var}(N) = \frac{n_2}{n_1} \sigma^2$. Les moyennes empiriques \bar{X}_1 et \bar{X}_2 sont indépendantes et convergent en distribution vers des lois normales lorsque n_1 et n_2 tendent vers l'infini d'après le théorème central limite. Par conséquent $\sqrt{\frac{n_1}{n_2}} \frac{N}{\sigma}$ converge en distribution vers la loi normale standard, et donc $\frac{n_1}{n_2} \frac{n(\bar{X} - \bar{X}_1)^2}{\sigma^2}$ converge en distribution vers la loi $\chi_{(1)}^2$. De plus S est une moyenne empirique de n variables telle que

$$\begin{aligned} E[S] &= \frac{1}{n} \left(E \left[\sum_{i=1}^{n_1} (X_i - \bar{X}_1)^2 \right] + E \left[\sum_{i=n_1+1}^n (X_i - \bar{X}_1)^2 \right] \right), \\ &= \frac{1}{n} \left((n_1 - 1)\sigma^2 + \left(n_2 + \frac{n_2}{n_1} \right) \sigma^2 \right) = \sigma^2 + O\left(\frac{1}{n}\right). \end{aligned}$$

Sous des conditions assez faibles de régularité, S converge en probabilité vers σ^2 lorsque n tend vers l'infini - c'est par exemple une conséquence de l'inégalité de Tchebychev dès lors que $E[X_i^4] < +\infty$, pour $1 \leq i \leq n$, car dans ce cas $\text{var}(S) = O\left(\frac{1}{n}\right)$. Le théorème de Slutsky assure alors la convergence écrite en (9), ce qui prouve le résultat énoncé. \square

Il faut noter qu'un résultat similaire peut également être obtenu sous l'hypothèse d'hétéroscédasticité, en injectant les variances empiriques de chaque segment dans le facteur de calibration. L'étude d'une telle calibration dépasse cependant le cadre de cette communication.

Finalement la statistique de test obtenue n'est pas symétrique et dépend du choix du premier échantillon dont la moyenne \bar{X}_1

sert à évaluer le profil $\mathcal{R}(\bar{X}_1)$. Déterminer l'échantillon qui garantit la plus grande puissance quelle que soit l'alternative est un problème théorique difficile. En pratique, plusieurs stratégies sont possibles : choix de l'échantillon dont la moyenne empirique donne la statistique minimale ou maximale, échantillon dont la variance empirique est minimale ou maximale... Dans le cas homoscédastique, un choix simple qui est adopté dans la suite consiste à opter pour l'échantillon le plus grand, appelé \mathbf{X}_1 : $n_1 \geq n_2$. Ce choix s'avère en effet puissant et robuste pour un large ensemble d'alternatives testées et est indépendant des variables des échantillons, la calibration de la proposition 3.1 donne donc directement la loi de $\mathcal{R}(\bar{X}_1)$ sous H_0 . Notons que dans le cas d'échantillons appariés ($n_1 = n_2$), ce choix peut s'effectuer de manière arbitraire ou aléatoire.

La procédure du test VEME ainsi défini se résume par :

- on calcule la moyenne empirique des réalisations de \mathbf{X}_1 : $\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$,
- on résout numériquement le problème dual (5), (6), afin de calculer la réalisation de la statistique de test $t = -2 \frac{n_1}{n_2} \ln(\mathcal{R}(\bar{x}_1))$,
- on rejette H_0 au niveau α lorsque $t \geq z_{1-\alpha}$ où $z_{1-\alpha}$ est le quantile $1 - \alpha$ de la loi $\chi_{(1)}^2$.

4 Résultats expérimentaux

Afin de d'établir les performances du test VEME et de les comparer avec d'autres approches plus classiques, on réalise une série de simulations où les échantillons sont de taille ou de variance très différentes, ou bien ne suivent pas une loi normale. Les résultats sont présentés sous la forme de courbes de sensibilité et spécificité (COR) pour la puissance du test, donnant la probabilité de détection (PD) en fonction de la probabilité de fausse alarme (PFA), et de diagrammes quantile-quantile pour la calibration, où l'on vérifie l'adéquation de la statistique de test $T = -2 \frac{n_1}{n_2} \ln(\mathcal{R}(\bar{X}_1))$ avec la loi $\chi_{(1)}^2$.

Dans un premier temps, on trace le diagramme quantile-quantile de notre statistique de test, pour des observations suivant la loi normale standard, sous H_0 . Les figures 1(a) et 1(b) sont obtenues pour $n_1 = 95$ et $n_2 = 5$, puis $n_1 = 5$ et $n_2 = 5$ respectivement. On observe bien que la distribution asymptotique de T est la loi $\chi_{(1)}^2$, même pour des tailles d'échantillons très différentes, ou très petites.

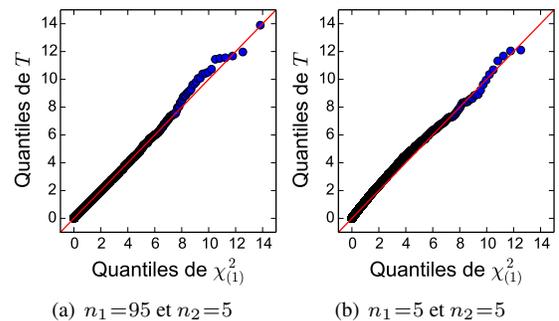


FIGURE 1 – Calibration du test : diagrammes quantile-quantile.

On compare ensuite les performances du test VEME avec le test t de Student, pour lequel on suppose que les échantillons ont la même variance, le test de Mann-Whitney-Wilcoxon, noté MWW, qui est un test non paramétrique basé sur les rangs des observations, et un test qui intègre un rapport de vraisemblance empirique. Ce dernier, que l'on notera VE seg, est présenté dans [10], et repose sur le profil proposé dans [3] des vraisemblances empiriques des deux échantillons. Notons que ce test n'est pas applicable à de trop petits échantillons, par exemple lorsque $n_2 \leq 10$: la moyenne proposée pour évaluer les vraisemblances empiriques n'appartient plus nécessairement à l'enveloppe des échantillons, contrairement au test VEME proposé où cette condition est toujours vérifiée. Dans nos simulations, sous l'hypothèse gaussienne pour des échantillons de même taille et de même variance, tous ces tests ont des puissances similaires. On s'intéresse donc à des cas non gaussiens : les échantillons sont générés selon les lois exponentielles de paramètre $a=1.0$, avec $n_1=n_2=50$, et les observations de \mathbf{X}_2 sont décalées de la valeur 0.3. Les courbes COR résultantes sont tracées dans la figure 2(a). Le test MWW donne les meilleurs résultats, tandis que les courbes des tests VEME, VE seg et de Student sont confondues. On trace également les résultats obtenus pour le test VEME lorsque la moyenne \bar{X}_1 est celle de l'échantillon qui a la plus grande variance empirique, et non plus la plus grande taille. Avec ce critère, on observe une légère amélioration de la puissance du test.

Enfin, on génère des échantillons selon des lois inverses-gammas de paramètres de forme 4.0 et d'échelle 0.5, avec un décalage de 0.3 pour les observations de \mathbf{X}_2 , et $n_1=n_2=20$. Les courbes COR obtenues sont données dans la figure 2(b). Le test VE seg est maintenant moins efficace que les tests VEME et de Student. Le critère du maximum de la variance permet de nouveau une amélioration de la détection, mais sans atteindre les performances du test MWW.

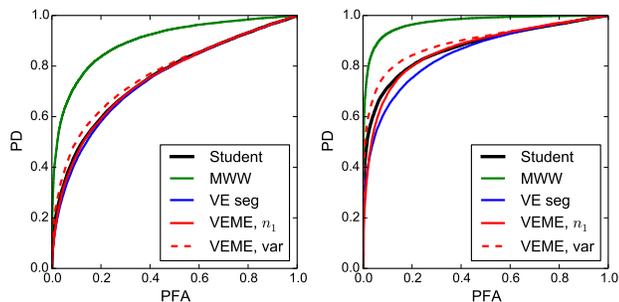


FIGURE 2 – Courbes COR pour les tests de Student (noir), MWW (vert), VE seg (bleu) et VEME (rouge), avec comme critère le segment le plus grand (n_1 , trait plein) et la variance la plus grande (var, tirets).

5 Discussion et conclusion

On a présenté le test d'hypothèse VEME pour la détection d'une différence entre les moyennes de deux populations, basé

sur la fonction de profil de la vraisemblance empirique. La calibration asymptotique sous H_0 obtenue pour la statistique de test est vérifiée expérimentalement même lorsque n est petit. Dans le cas gaussien, la puissance du test est comparable à celle des tests classiques de Student et le test non paramétrique MWW, ainsi que celle du test présenté dans [10] défini uniquement pour des échantillons suffisamment grands. En revanche dans le cas non gaussien, le test MWW demeure plus performant, bien que présentant des seuils discrets dans les probabilités de détection. Pour améliorer les résultats du VEME dans le cas non gaussien, il faut rechercher un autre critère pour le choix de la moyenne empirique, actuellement arrêté à la moyenne de l'échantillon le plus grand. Expérimentalement le critère de la variance la plus grande peut conduire à de meilleurs résultats, mais la calibration de la proposition 3.1 doit alors être adaptée. Enfin, ce test peut être étendu au traitement de données multivariées, ce qui s'avère plus compliqué avec le test MWW.

Références

- [1] J.H.J. Einmahl and I.W. McKeague. Empirical likelihood based hypothesis testing. *Bernoulli*, 9(2) :267–290, 2003.
- [2] H. Harari-Kermadec and F. Pascal. On the use of empirical likelihood for non-gaussian clutter covariance matrix estimation. In *Proc. of the IEEE-RADAR-08*, Roma, Italy, may 2008.
- [3] B.-Y. Jing. Two-sample empirical likelihood method. *Statistics & probability letters*, 24(4) :315–319, 1995.
- [4] Y. Kitamura. Empirical Likelihood Methods in Econometrics : Theory and Practice. Cowles Foundation Discussion Papers 1569, Cowles Foundation for Research in Economics, Yale University, June 2006.
- [5] Z. Liu and L. Qian. Changepoint estimation in a segmented linear regression via empirical likelihood. *Communications in Statistics - Simulation and Computation*, 39(1) :85–100, 2009.
- [6] W. Ning, J. Pailden, and A. Gupta. Empirical likelihood ratio test for the epidemic change model. *Journal of Data Science*, 10 :107–127, 2012.
- [7] A.B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2) :237–249, 1988.
- [8] A.B. Owen. *Empirical Likelihood*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2001.
- [9] F. Pascal, H. Harari-Kermadec, and P. Larzabal. The empirical likelihood method applied to covariance matrix estimation. *Signal Processing*, 90(2) :566–578, 2010.
- [10] Changliang Z., Yukun L., Peng Q., and Zhaojun W. Empirical likelihood ratio test for the change-point problem. *Statistics & Probability Letters*, 77(4) :374 – 382, 2007.