Modélisation statistique pour l'identification et la quantification de manipulations comptables

Marie Chavent[1], Véronique Darmendrail[2], Delphine Feral[1], Hadrien Lorenzo[3], Frédéric Pourtier[2], Jérôme Saracco[1]

[1] *Univ. Bordeaux, CNRS, INRIA, Bordeaux INP, IMB, UMR 5251, F-33400 Talence, France*
marie.chavent@inria.fr, delphine.feral@u-bordeaux.fr, jerome.saracco@inria.fr

[2] *Univ. Bordeaux, IRGO, UR 4190, F-33000 Bordeaux, France*
veronique.darmendrail@u-bordeaux.fr, frederic.pourtier@u-bordeaux.fr

[3] *Univ. Aix-Marseille, I2M, UMR 7373, Marseille, France*
hadrien.lorenzo@univ-amu.fr

Résumé. Cette communication présente une nouvelle méthodologie statistique pour détecter la manipulation comptable autour d'un seuil psychologique (ici, le bénéfice nul d'une entreprise). Un algorithme de type EM est proposé pour estimer les paramètres du modèle de manipulation comptable. Le bon comportement numérique de la méthodologie est illustré sur des données simulées proches des données réelles.

Mots-clés. Statistique appliquée, algorithme EM ; estimation de la densité ; modèle de mélange ; distribution gaussienne tronquée ; manipulation comptable.

Abstract. This communication presents a new statistical methodology to detect earnings management associated with the zero earnings threshold. An EM-type algorithm is proposed to estimate the underlying parameters of the considered earnings management model. The good numerical behavior of the methodology is illustrated on simulated data close to real data.

Keywords. Applied Statistics, EM algorithm; density estimation; Mixture model; Truncated Gaussian distribution; Earnings management.

1 Motivation

La pratique de la manipulation comptable par les entreprises est reconnue depuis longtemps par les chercheurs. La littérature comptable a montré qu'elles sont enclines à gérer leurs bénéfices et leurs pertes comptables de manière à dépasser des seuils spécifiques tels que le bénéfice nul, le bénéfice de l'année précédente ou les prévisions des analystes, voir par exemple Burgstahler & Chuk (2017), ou encore Byzalov & Basu (2019). L'existence de manipulations comptables compromet l'intégrité de l'information financière. Les régulateurs ou les investisseurs s'intéressent donc à la détection de la manipulation comptable, à sa fréquence et à son ampleur. Dans cette communication, une nouvelle méthodologie statistique pour détecter la manipulation comptable associée au seuil de bénéfice nul est présentée. Le modèle statistique de manipulation comptable est présenté à la Section 2. La Section 3 donne une brève description de la procédure d'estimation des paramètres du modèle proposé. Cette dernière est illustrée numériquement à la Section 4 sur des données simulées très similaires aux données réelles.

2 Présentation du modèle statistique de manipulation comptable

Pour déterminer la fréquence et l'ampleur de la manipulation comptable (autour du seuil psychologique correspondant au bénéfice nul) dans un échantillon d'entreprises, la distribution des bénéfices doit être modélisée. Désignons par X_i (resp. Y_i) le bénéfice réel (resp. observé) de l'entreprise i, $i = 1, \ldots, N$. Notons que les revenus réels sont supposés être indépendants et identiquement distribués et suivre un **mélange de deux distributions gaussiennes** de densité

$$\boldsymbol{\pi\varphi_A(x) + (1-\pi)\varphi_B(x)}, \tag{1}$$

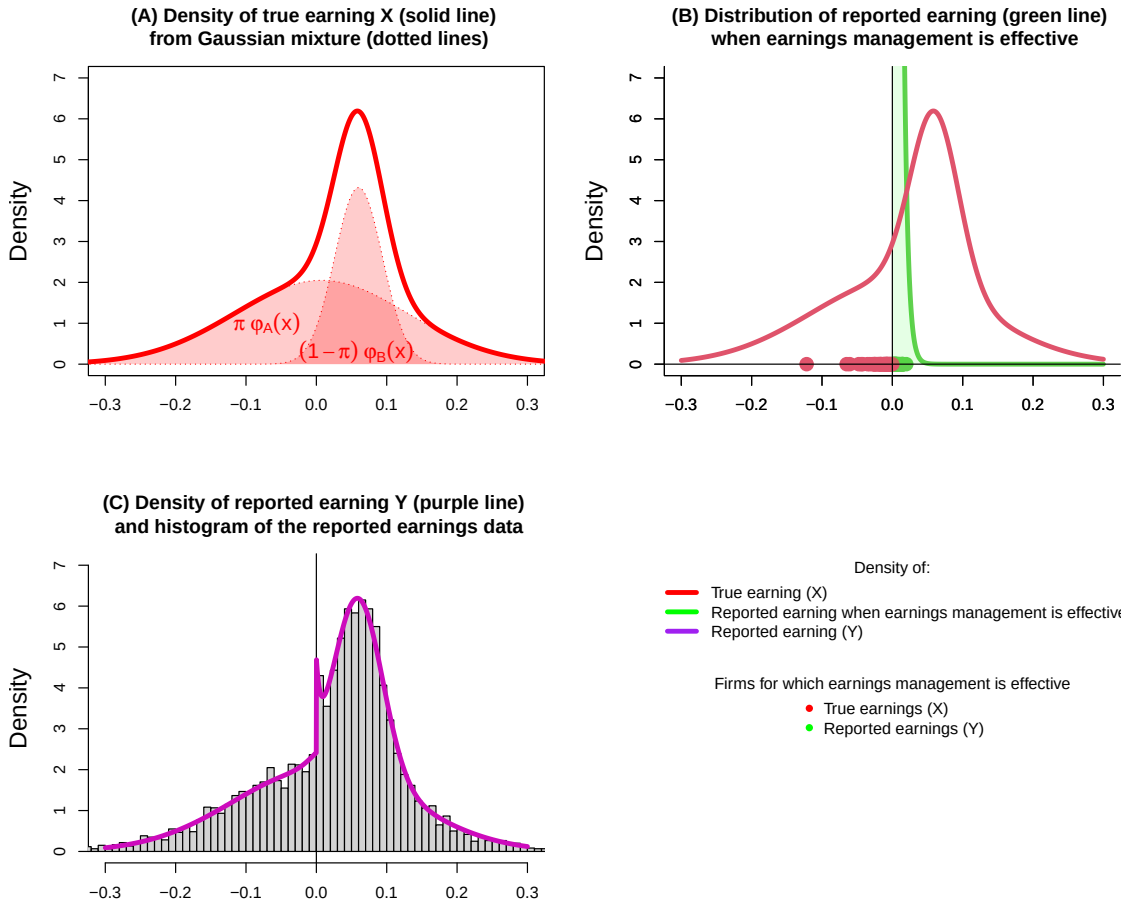
où $\pi \in]0,1[$ et φ_A (resp. φ_B) est la densité de la loi normale d'espérance μ_A (resp. μ_B) et de variance σ_A^2 (resp. σ_B^2). Le choix d'un mélange gaussien permet de modéliser une distribution avec des queues de poids différents, comme cela est souvent observé dans les données comptables que l'on va considérer.

Lorsqu'une entreprise a une valeur de bénéfice réel inférieure au seuil zéro, elle peut s'engager ou pas dans une manipulation comptable et le bénéfice déclaré sera alors supérieur au seuil zéro. Lorsque cette manipulation comptable est effective pour une entreprise i, le revenu déclaré Y_i est supposé suivre la **loi exponentielle $\mathcal{E}(\lambda)$ de paramètre λ**, sinon $Y_i = X_i$ le revenu réel.

Afin de modéliser la transition vers la manipulation comptable (des bénéfices) pour une entreprise i, la variable aléatoire T_i est introduite : lorsque X_i est inférieur à la valeur seuil de zéro, T_i sachant $X_i = x$ suit une distribution de Bernoulli de paramètre $\tau(x)$. Cette fonction $\tau(.)$ définie pour $x < 0$ est croissante, de manière à ce que la probabilité de manipulation soit d'autant plus forte que le bénéfice de l'entreprise est proche du seuil psychologique zéro.

L'idée sous-jacente est de se concentrer uniquement sur le sous-échantillon d'entreprises

Figure 1: Les densités sous-jacentes à la modélisation de la manipulation comptable.



associées aux $Y_i \geq 0$, c'est à dire celles ayant déclaré un bénéfice positif. Le travail ci-après est donc implicitement développé conditionnellement à $Y \geq 0$. Les bénéfices déclarés correspondants de ces entreprises peuvent alors être considérés comme un mélange entre les bénéfices "réels" et les bénéfices "manipulés", qui peuvent être modélisés comme suit :

$$\forall \boldsymbol{y \geq 0}, \quad \boldsymbol{f(y) = q f_1(y) + (1-q) f_2(y)}, \tag{2}$$

où $q \in]0,1[$, f_1 est la densité de la loi $\mathcal{E}(\lambda)$, et f_2 est la densité du mélange:

$$\forall y \geq 0, \quad f_2(y) = \tilde{\pi}\varphi_A^{(+)}(y) + (1-\tilde{\pi})\varphi_B^{(+)}(y), \tag{3}$$

avec $\varphi_A^{(+)}$ (resp. $\varphi_B^{(+)}$) la densité de la loi normale tronquée (à zéro) d'espérance μ_A (resp. μ_B) et de variance σ_A^2 (resp. σ_B^2). Le vecteur des paramètres sous-jacents du modèle simplifié[1] est donc le suivant :

$$\theta = \left(q, \lambda, \tilde{\pi}, \mu_A, \sigma_A^2, \mu_B, \sigma_B^2\right). \tag{4}$$

3 Procédure d'estimation des paramètres

Étant donné l'échantillon observé des revenus (positifs) déclarés $\mathcal{D}_n = \{y_1, \ldots, y_n\}$, l'objectif est d'estimer θ par maximum de vraisemblance

$$\hat{\theta} = \left(\hat{q}, \hat{\lambda}, \hat{\tilde{\pi}}, \hat{\mu}_A, \hat{\sigma}_A^2, \hat{\mu}_B, \hat{\sigma}_B^2\right) := \arg\max_{\theta} \ \ell(\theta; \mathcal{D}_n) \tag{5}$$

où $\ell(\theta, \mathcal{D}_n) = \prod_{i=1}^n f(y_i)$.

A cette fin, un algorithme de type EM (Expectation–Maximization) (voir Dempster et al., 1977) a été mis au point. Cet algorithme repose sur la vraisemblance complétée par deux variables latentes dichotomiques : $\tilde{T}_i$ (resp. $\tilde{Z}_i$) qui suit une distribution de Bernoulli de paramètre q (resp. $\tilde{\pi}$). La variable latente $\tilde{T}_i$ indique si l'entreprise i a manipulé ses bénéfices ou non, tandis que la variable $\tilde{Z}_i$ est utilisée pour gérer le mélange des deux distributions gaussiennes tronquées. La densité "complétée", élément de base pour mettre en œuvre l'algorithme EM, est donnée par :

$$(q f_1(y))^t (1-q)^{1-t} \left(\tilde{\pi}\varphi_A^{(+)}(y)\right)^{(1-t)z} \left((1-\tilde{\pi})\varphi_B^{(+)}(y)\right)^{(1-t)(1-z)}. \tag{6}$$

Plus de détails techniques sont disponibles dans Chavent et al. (2023).

[1] dans le sens où l'on ne se focalise que sur la sous-échantillon d'entreprises ayant déclaré un bénéfice positif

4 Illustration numérique

Une rapide description des données simulées est fournie à la Section 4.1. Les résultats obtenus sont commentés brièvement à la Section 4.2.

4.1 Paramètres des simulations

Deux scénarios ont été envisagés. Deux échantillons de taille $n = 2000$ ont été générés à partir du modèle (2) pour les deux paramètres θ décrits dans la Table 1.

Table 1: Paramètres utilisés dans les simulations.

	q	λ	$\tilde{\pi}$	μ_A	σ_A	μ_B	σ_B
Scénario 1	0.1	12	0.3	1	1	4	0.5
Scénario 2	0.05	185.3	0.632	0.007	0.123	0.06	0.034

Commentaires rapides sur les deux scénarios.

- Dans le premier scénario, il est relativement facile d'identifier et d'estimer toutes les composantes de θ.
- Le second scénario est plus complexe, mais correspond à des situations plus réalistes dans le contexte de la manipulation comptable. Il est fortement inspiré du premier jeu de données réelles qui a été utilisé par Chen et al. (2010).

4.2 Résultats

Les résultats numériques obtenus sont respectivement présentés dans les figures 2 et 3. Afin d'avoir une visibilité sur la variabilité des estimateurs, $B = 100$ échantillons Bootstrap ont été générés, et les paramètres associés ont ensuite été estimés. La variabilité de la densité estimée est fournie via l'intervalle de confiance Bootstrap à 90% (en rouge), ainsi que celles de l'estimation des composantes de θ via les boxplots correspondants.

On observe clairement que la procédure d'estimation permet de retrouver correctement les composantes du paramètre θ et donc la vraie densité f des Y_i, aussi bien avec les données issues du scénario 1 (voir Figure 2) qu'avec les données issues du scénario 2 (voir Figure 3).

Figure 2: Résultats de la simulation dans le cadre du scénario 1 avec la vraie densité f (en bleu) et la densité estimée (en marron), ainsi que les valeurs réelles et estimées de θ, avec la variabilité bootstrap (en orange pâle).

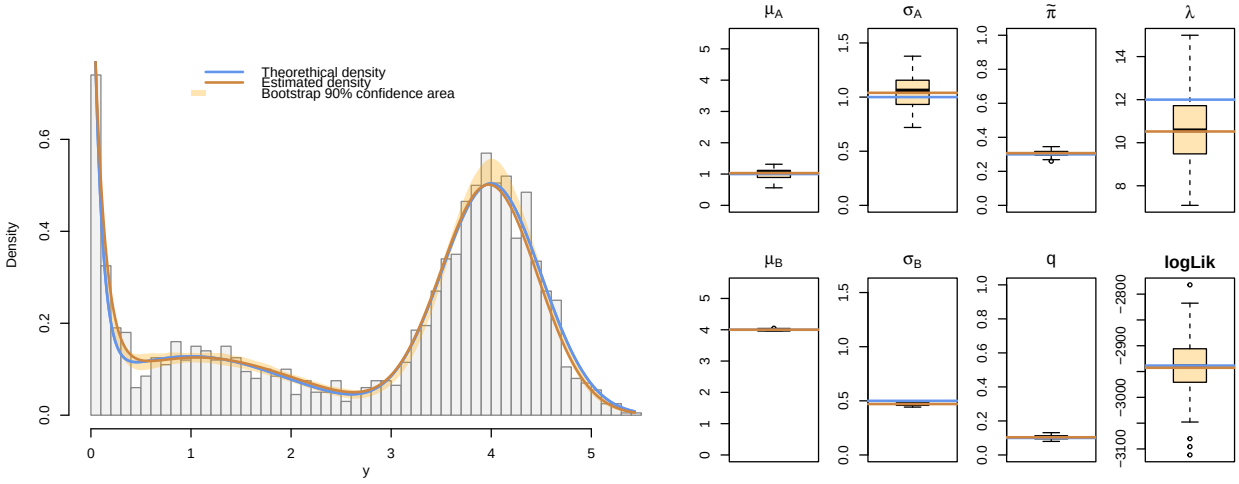
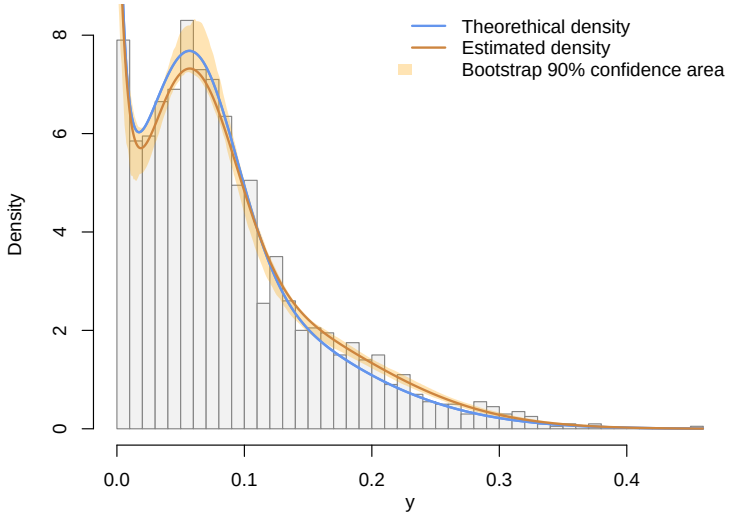


Figure 3: Résultats de la simulation dans le cadre du scénario 2 avec la vraie f densité (en bleu) et la densité estimée (en marron) et la variabilité bootstrap (en orange pâle).



5 Observations finales

- Dans le cadre de ce résumé, le bon comportement numérique de la procédure d'estimation n'a été évalué que sur deux jeux de données (associés à deux scénarios différents) à la Section 4. Des simulations numériques de plus grande ampleur seront présentées lors de la communication orale.

  Une application sur données réelles servira également d'illustration.

- À partir du paramètre estimé $\hat{\theta}$ du modèle simplifié, il est possible d'obtenir une estimation des paramètres du modèle statistique initial[2] et de retrouver la fréquence et l'ampleur de la manipulation comptable sur l'ensemble de la population des entreprises.

- En se basant sur la variabilité Bootstrap, il sera également possible de comparer des modèles basés sur deux populations, comme des pays différents, des sous-périodes différentes, ou différentes catégories d'entreprises.

  Ce point sera illustré sur des données réelles avec une comparaison entre deux pays et une comparaison avec deux méthodes comptables.

- De plus, il existe d'autres domaines d'application pour ces méthodes de détection et de quantification de manipulation de données autour d'un seuil psychologique. Dans différents domaines (autres que celui de la comptabilité), un décideur a parfois l'opportunité et l'incitation de passer d'un niveau juste inférieur ou à un niveau juste supérieur à un point de référence, grâce à de la manipulation de données.

- Un package R est actuellement en cours de développement.

Bibliographie

Burgstahler, D. and Chuk, E. (2017). What have we learned about earnings management? Integrating discontinuity evidence. *Contemp. Account. Res.*, 34(2), 726749.

Byzalov, D. and Basu, S. (2019). Modeling the determinants of meet-or-just-beat behavior in distribution discontinuity tests. *Journal of Accounting and Economics*, 68(23), 101266.

Chavent, M., Darmendrail, V., Feral, D., Lorenzo, H. , Pourtier, F. and Saracco, J. (2023) A new statistical methodology to detect earnings management. *Proceedings of the 37th International Workshopon Statistical Modelling (IWSM 2023)*, July 17-21, 2023-Dortmund, Germany, 394 – 399.

Chen, S.K., Lin, B-X., Wang, Y. and Wu, L. (2010). The frequency and magnitude of earnings managements: Time-series and multi-threshold comparisons. *International Review of Economics and Finance*, **19**, 671 – 685.

Dempster, A.P., Laird, N.M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc., Ser. B*, **39**, 1 – 38.

[2] c'est-à-dire du modèle qui ne se limite pas uniquement aux bénéfices positifs.