UNE CONDITION NÉCESSAIRE ET SUFFISANTE POUR MINIMISER LE RISQUE D'INFÉRENCE PRÉDICTIVE DANS LES MÉTHODES DE PERTURBATION DE DONNÉES

Cédric WANKO (*), Jean-Marie BERTON (**)

(*)*Membre Laboratoire TRIS-UM (Traitement et Recherche sur l'Information et la Statistique) Montpellier Université.*
(**)*Rattaché Laboratoire TRIS-UM (Traitement et Recherche sur l'Information et la Statistique) Montpellier Université.*

Cedric.wanko@free.fr
Berton.jeanmarie@gmail.com

Mots Clés Distribution bivariée, indépendance, perturbation, inférence prédictive, orthonormalisation de Gram-Schmidt

Résumé

Dans la nécessaire régulation liée au "Contrôle de la Divulgation Statistique" - Statistical Disclosure Control - les méthodes de perturbation des données à l'aide d'un bruit aléatoire permettent de réduire les pertes informationnelles et prennent en compte les données non confidentielles favorisant les risques de divulgation prédictive dans les modèles de codage. Dans ce papier, on définit une condition nécessaire et suffisante montrant que lorsque les variables non confidentielles sont orthonormalisées et perturbées en leur structure alors elles satisfont la "predictive inference risk requirements" qui minimise les risques de divulgation prédictive tout en conservant les "data utility requirements" et "disclosure risk requirements".

Abstract

A Necessary and Sufficient Condition to minimize predictive inference risk for data perturbation methods

In the necessary regulation related to the "Statistical Disclosure Control" perturbation methods using a random noise can reduce information loss and take into account non-confidential data favoring the risks of predictive disclosure in coding models. In this paper, we define a necessary and sufficient condition showing that when the non-confidential variables are orthonormalized and perturbed in their structure then they satisfy the "predictive inference risk requirements" which minimizes the risks of predictive disclosure while preserving the "data utility requirements" and" disclosure risk requirements ".

Keywords : Bivariate distribution, independency, perturbation, predictive inference, Gram-Schmidt orthonormalization

JEL Classification C10, C18, C19, C63, C46, C81
AMS Classification 2010 62-07 . 62B05 . 62G07 . 62H20 . 62J00 . 62P00

1- Introduction

La protection des données confidentielles est un important problème lié à la diffusion de données privées lorsqu'on parle du droit de chacun à déterminer laquelle des informations le concernant est susceptible d'être partagée avec les autres (Fellegi (1972)). Les données utilisées pour la recherche scientifique ou par les décideurs publics pour l'intérêt collectif mais qui sont, néanmoins, trop sensibles pour des raisons éthiques ou légales pour être publiquement divulguées en sont un exemple. Un autre exemple de problèmes liés à la diffusion de données sensibles réside dans la possibilité pour un intrus (hacker ou

snooper)[1] de pouvoir *inférer* certaines informations confidentielles à partir d'une base de données rendue publique par une agence gouvernementale ou commerciale. Une agence doit de fait fournir tous les efforts nécessaires pour maintenir un niveau de *confidentialité* le plus élevé possible concernant les données privées (ou microdonnées) qui lui sont confiées. Cet effort consiste à réduire les risques de divulgation d'attributs, les risques de divulgation d'identité et les risques de divulgation prédictive[2]. Il est important dans les différents modèles de codage de tenir compte des variables non confidentielles face aux problèmes de ré-identification des données confidentielles. On ne peut pas tout masquer il faut de fait prendre en compte l'existence de variables publiques non confidentielles dans le processus de codage et des risques qui leur sont afférents.

Les méthodes de perturbation des données ont gagné une ampleur considérable dans la littérature (Fuller (1993), Muralidhar, Parsa and Sarathy (1999), Willenborg and de Waal (2001), Muralidhar and Sarathy (2003, 2008) pour un bref survey *cf.* Brand (2002)). L'objet des solutions de masquage de données en général est de déterminer un équilibre optimal entre préserver le maximum d'information issue de la base de donnée confidentielle dans la base de donnée masquée et publiquement divulguée tout en garantissant une confidentialité la plus élevée possible par un moyen de contrôle approprié (*cf.* Hundepool and al. (2012) et Matthews and Harel (2011) pour un bref survey). Les agences ont de multiples options dans la divulgation de données "masquées". Elles peuvent ne rien divulguer, divulguer uniquement des données agrégées ou divulguer directement des microdonnées pour chaque individu. Il est évident que dans le cas où elles divulguent des données numériques agrégées ou individuelles, elles doivent conserver les caractéristiques essentielles et suffisantes (covariances, corrélations, moyennes, les caractéristiques des échantillons de population, etc...).

Concernant les méthodes d'anonymisation des répondants par la création de microdonnées synthétiques, les précurseurs en la matière sont Liew and al. (1985). Par la suite différents travaux portant sur la création de données synthétiques ont été développés notamment par des méthodes d'imputation multiple (Rubin (1993), Raghunathan et al. (2003), Drechsler and Reiter (2010), Drechsler (2011), Drechsler (2012), Miranda and Vilhuber (2014), Miranda and Vilhuber (2016)) ou par bootstrap (Fienberg (1994)). Ces méthodes présentent un niveau de sécurité très élevé car les données synthétisées sont totalement indépendantes des données confidentielles[3]. Néanmoins, elles engendrent une perte d'information assez importante[4] et ne permettent pas de contrôler efficacement les risques d'inférences prédictives[5].

Dans le but de lutter contre cette perte d'information, notre attention s'oriente naturellement vers une méthode de *perturbation* des données à l'aide d'un *bruit aléatoire*; une méthode utilisée par Burridge (2003) sous l'appelation de - Information Preserving Statistical Obfuscation (IPSO)[6] et généralisée via un coefficient de similarité par Muralidhar and Sarathy (2008) puis Domingo-Ferrer and González-Nicolás (2010) et enfin Calviño (2017) sous les appelations respectives de Sufficiency Based Noise Addition (SBNA), MicroHybrid method (MH) et Principal Component Analysis method (PCA).

Par ailleurs, on sait que tenir compte des données non confidentielles favorisant les risques de divulgation prédictive dans le modèle de codage contribue à renforcer la sécurité du modèle vis à vis de ces risques de divulgation prédictive. Malgré tout, ces risques sont toujours sensibles aux différentes méthodes de

[1] On parle d'un hacker ou pirate (resp. snooper ou espion) lorsqu'un individu non authorisé (resp. authorisé) à pénétrer dans le système (resp. une partie du système) tente de violer la confidentialité de certaines données.

[2] Lorsque un intrus peut ré-identifier les données confidentielles par l'intermédiaire d'autres variables non confidentielles

[3] Notamment en termes d'inférence d'attributs et d'identité car il est impossible de remonter aux données confidentielles dans ce cas.

[4] Il y a trop de différences entre les données confidentielles et les données synthétisées en termes de similarité.

[5] L'absence de contrôle des risques d'inférence prédictive peut également venir du fait que le terme d'erreur est largement augmenté pour permettre de retrouver les caractéristiques suffisantes de la base de données confidentielles. Malheureusement, cet accroissement du terme d'erreur fournit une information supplémentaire et donc représente un risque d'inférence prédictive.

[6] La méthode IPSO est considérée comme une méthode de synthèse (Calviño (2017)) mais cela est totalement discutable (Muralidhar and Sarathy (2008)). Dans ce papier, on considère la méthode IPSO comme une méthode de perturbation

prédiction relatives au Data Mining (Donoho (2015)) car même si les variables confidentielles sont supprimées du modèle, les variables non confidentielles sont encore présentes dans le processus de codage. Par conséquent, pour minimiser les risques de divulgation prédictive, on propose de minimiser l'influence des variables non confidentielles tout en les conservant au sein du modèle. Les variables non confidentielles doivent donc être peu ou pas significatives dans la spécification du modèle tout en conservant les propriétés théoriques de base du modèle.

Dans ce papier, on définit une condition nécessaire et suffisante permettant de renforcer la sécurité en minimisant les risques de divulgation prédictive dans les méthodes de perturbation. On sait effectivement que le risque de divulgation est minimisé lorsque la divulgation des données ne fournit aucune information supplémentaire aux utilisateurs de données. On montre plus particulièrement que lorsque les variables non confidentielles sont orthonormalisées et perturbées en leur structure alors elles satisfont la "predictive inference risk requirements" qui minimise les risques de divulgation prédictive tout en conservant les "data utility requirements" et "disclosure risk requirements".

On fournit une approche théorique puis calculatoire avant de confronter les résultats à différentes méthodes de perturbation linéaires et non linéaires puis de fournir un exemple empirique.

2- Une base théorique pour minimiser le risque d'inférence prédictive : une approche théorique

Idéalement, les données masquées et publiquement divulguées doivent satisfaire deux conditions génériques majeures: *maximum data utility* et *minimum disclosure risk*. Nous rajoutons une troisième condition: *minimum predictive inference risk*. Soit $\widetilde{\mathrm{S}} = (\widetilde{\mathrm{S}}_1, \widetilde{\mathrm{S}}_2)$ le vecteur orthonormalisé et transformé des données non confidentielles $\mathrm{S} = (\mathrm{S}_1, \mathrm{S}_2)$ du modèle[7]. Afin de minimiser les risques de divulgation prédictives lorsque les variables non confidentielles sont utilisées pour la ré-identification des variables confidentielles $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2)$ sachant les variables masquées et publiquement divulguées $\mathrm{Y} = (\mathrm{Y}_1, \mathrm{Y}_2)$, on estime la densité jointe gaussienne résiduelle de sorte qu'elle mette en évidence une *mesure d'indépendance*. Soient $f()$ la densité de probabilité et $F()$ sa primitive.

- La "*data utility or accuracy requirements*" implique que les relations entre Y et $\widetilde{\mathrm{S}}$ sont les mêmes qu'entre X et $\widetilde{\mathrm{S}}$. En effet, si selon Muralidhar and Sarathy (2003) une procédure générale de perturbation implique qu'une observation y_i est générée à partir de la distribution conditionnelle $f(\mathrm{X}/\mathrm{S} = \varsigma_i)$. En appliquant cette condition au vecteur $\widetilde{\mathrm{S}}$ on obtient:

$$y_i : f\left(\mathrm{X}/\widetilde{\mathrm{S}} = \widetilde{\varsigma}_i\right) \tag{1}$$

$$f\left(\mathrm{X},\mathrm{Y}/\widetilde{\mathrm{S}} = \widetilde{\varsigma}_i\right) = f\left(\mathrm{X}/\widetilde{\mathrm{S}} = \widetilde{\varsigma}_i\right) f\left(\mathrm{Y}/\widetilde{\mathrm{S}} = \widetilde{\varsigma}_i\right)$$

Dans ce cas Y est indépendant de X sachant $\widetilde{\mathrm{S}} = \widetilde{\varsigma}_i$. L'observation y_i est une réalisation indépendante de $f\left(\mathrm{X}/\widetilde{\mathrm{S}} = \widetilde{\varsigma}_i\right)$. On en déduit donc que

$$f\left(\mathrm{X}/\widetilde{\mathrm{S}}\right) = f\left(\mathrm{Y}/\widetilde{\mathrm{S}}\right)$$

$$f\left(\mathrm{Y},\widetilde{\mathrm{S}}\right) = f\left(\mathrm{Y}/\widetilde{\mathrm{S}}\right) f(\widetilde{\mathrm{S}}) = f\left(\mathrm{X}/\widetilde{\mathrm{S}}\right) f(\widetilde{\mathrm{S}}) = f\left(\mathrm{X},\widetilde{\mathrm{S}}\right)$$

Par dessus tout, $f(\mathrm{Y}) = \int_{\widetilde{\mathrm{S}}} f\left(\mathrm{Y},\widetilde{\mathrm{S}}\right) d\widetilde{\mathrm{S}} = \int_{\widetilde{\mathrm{S}}} f\left(\mathrm{X},\widetilde{\mathrm{S}}\right) d\widetilde{\mathrm{S}} = f(\mathrm{X})$. On peut de fait résumer les conditions

[7] On évite de créer une variable intermédiaire représentant la variable S orthonormalisée par le processus de Gram-Schmidt, premièrement pour ne pas alourdir les notations et deuxièmement car on considère l'orthonormalisation réalisée dans $\widetilde{\mathrm{S}}$

nécessaires par

$$f(\mathrm{Y}) = f(\mathrm{X}) \ et \ f\left(\mathrm{Y},\tilde{\mathrm{S}}\right) = f\left(\mathrm{X},\tilde{\mathrm{S}}\right) \tag{2}$$

- La "*disclosure risk requirements*" implique que la confidentialité de X est maintenue et les données publiquement accessibles $(\mathrm{Y},\tilde{\mathrm{S}})$ n'augmentent pas le risque de divulgation:

$$f\left(\mathrm{X}/\tilde{\mathrm{S}},\mathrm{Y}\right) = f\left(\mathrm{X}/\tilde{\mathrm{S}}\right) \tag{3}$$

- La "*predictive inference risk requirements*" implique que si selon une orthonormalisation de Gram-Schmidt on a:

$$\tilde{\mathrm{S}}_1 = \mathrm{S}_1 \ et \ \tilde{\mathrm{S}}_2 = \mathrm{S}_2 - \frac{f(\mathrm{S}_1,\mathrm{S}_2)}{f(\mathrm{S}_1,\mathrm{S}_1)}\mathrm{S}_1 \tag{4}$$

et

$$f\left(\tilde{\mathrm{S}}\right) = f(\tilde{\mathrm{S}}_1) f(\tilde{\mathrm{S}}_2) \tag{5}$$

alors d'après (2) on a

$$f(\mathrm{X}-\mathrm{Y},\tilde{\mathrm{S}}) = 0 \tag{6}$$

Or sachant (4) et (6) on peut écrire:

$$f(\mathrm{X}-\mathrm{Y},\tilde{\mathrm{S}}_1) = f(\mathrm{X}-\mathrm{Y},\mathrm{S}_1) = 0$$

$$f(\mathrm{X}-\mathrm{Y},\tilde{\mathrm{S}}_2) = f(\mathrm{X}-\mathrm{Y},\mathrm{S}_2) - \underbrace{f(\mathrm{X}-\mathrm{Y},\mathrm{S}_1)}_{0}\frac{f(\mathrm{S}_1,\mathrm{S}_2)}{f(\mathrm{S}_1,\mathrm{S}_1)}$$

On en conclut donc que $f(\mathrm{X}-\mathrm{Y},\mathrm{S}_2) = 0$ ce qui équivaut à:

$$f(\mathrm{Y}) = f(\mathrm{X}) \ et \ f(\mathrm{Y},\mathrm{S}) = f(\mathrm{X},\mathrm{S}) \tag{7}$$

On retrouve les conditions de Muralidhar and Sarathy (2003) de "*data utility or accuracy requirements*". La réciproque est vraie et se vérifie aisément: à partir de (7) on retrouve (6).

3- Une condition nécessaire et suffisante pour minimiser le risque d'inférence prédictive : une approche computationnelle

Pour une procédure générale calculatoire de minimisation du risque d'inférence prédictive, on considère le vecteur de variables conditionnelles $\mathrm{S} = (\mathrm{S}_1/\mathrm{S}_2 = m_{\varsigma_2}, \mathrm{S}_2/\mathrm{S}_1 = m_{\varsigma_1})$ avec m_{ς_i} la moyenne théorique de la série S pour l'observation i. En outre, par définition, on sait que:

$$\underbrace{Var(\mathrm{S})}_{\sigma^2_{\mathrm{S}}} = \underbrace{E(Var(\mathrm{S}))}_{\sigma^2_{\mathrm{S}}\left(1-(\phi\rho)^2\right)} + \underbrace{Var(E(\mathrm{S}))}_{\sigma^2_{\mathrm{S}}(\phi\rho)^2}$$

On note $\rho \in [-1,+1]$ et $\phi \in \mathsf{R}$ un facteur (ou filtre) qui permet d'introduire une perturbation. Si aucune part de la variance de S_1 (resp. S_2) n'est expliquée par S_2 (resp. S_1), alors on fait l'hypothèse que $\sigma_{\mathsf{S}}^2(\phi\rho)^2 = 0$ et de fait $\sigma_{\mathsf{S}}^2\left(1-(\phi\rho)^2\right) = \sigma_{\mathsf{S}}^2$ pour $\phi = 0$. Si ℓ est une mesure de dépendance lors de la transformation S alors l'estimation de la log vraisemblance de la densité jointe

$$\log \prod_{i=1}^{n} f_{\ell}(\mathsf{S}) = \log \prod_{i=1}^{n} \left\{ c_{\ell}(F_{\mathsf{S}}) f_{\mathsf{S}_1/\mathsf{S}_2=m_{\zeta_2}}(\zeta_1) f_{\mathsf{S}_2/\mathsf{S}_1=m_{\zeta_1}}(\zeta_2) \right\}$$

ou l'estimation de la densité de *copule* gaussienne de transformation associée à $f_{\ell}(\mathsf{S})$ et relative au vecteur S notée $c_{\ell}(F_{\mathsf{S}})$

$$\log \prod_{i=1}^{n} c_{\ell}(F_{\mathsf{S}}) = \log \prod_{i=1}^{n} \left\{ f_{\ell}(\mathsf{S}) f^{-1}_{\mathsf{S}_1/\mathsf{S}_2=m_{\zeta_2}}(\zeta_1) f^{-1}_{\mathsf{S}_2/\mathsf{S}_1=m_{\zeta_1}}(\zeta_2) \right\}$$

nous donne la mesure de dépendance de perturbation:

$$\tilde{\rho} = \phi\rho\left(\left(1-\ell^2\right)\left(1-(\phi\rho)^2\right)+\phi\rho\ell\right)$$

Par ailleurs, si $\phi = 0$ alors $\tilde{\rho} = 0$ nécessairement: dans ce cas la perturbation provenant du filtre est dite *muette*. Si en plus de $\phi = 0$, on a $\ell = 0$ alors la perturbation globale du modèle est dite *muette*: on retrouve le modèle de codage d'origine en l'absence de quelque perturbation que ce soit.

Condition 1 $\tilde{\rho} = 0$ *est une condition nécessaire et suffisante pour que* $\phi = 0$ *sachant* ρ *donné et* ℓ *fixé.*

Or $\tilde{\rho} = 0$ implique $\phi\rho\left(\left(1-\ell^2\right)\left(1-(\phi\rho)^2\right)+\phi\rho\ell\right) = 0$. Le filtre ϕ est de fait la solution d'un polynôme de degré 3 et peut prendre trois valeurs distinctes. Si $\tilde{\rho}$ est le coefficient de corrélation linéaire de $\tilde{\mathsf{S}}$, on doit vérifier que $r_{\tilde{\mathsf{S}}} = \tilde{\rho} = 0$ sachant ρ donné, ℓ fixé et ϕ déterminé de façon calculatoire, $\forall \phi \in \mathsf{R}$. Il faut donc déterminer les différentes valeurs de $\phi = (\phi_1, \phi_2, \phi_3)$ pour lesquelles $\tilde{\rho} = 0$ sachant $\rho \in [-1;+1]$ donné et ℓ fixé.

1/ On utilise une *orthonormalisation de Gram Schmidt* sur la variable S préalablement centrée et réduite. Cela permet d'obtenir empiriquement $\tilde{\mathsf{S}}$ tel que $r_{\tilde{\mathsf{S}}} = 0$, $\mathsf{E}(\tilde{\mathsf{S}}) = 0$ et $\mathsf{E}(\tilde{\mathsf{S}}^2) = \sigma_{\tilde{\mathsf{S}}}^2$.

2/ Par la suite, on obtient de façon calculatoire ϕ en résolvant le polynôme de degré 3 vérifiant $r_{\tilde{\mathsf{S}}} = \tilde{\rho} = 0$.

$$\phi_1 = \frac{-\ell + (\ell^2 + 4\left(1-\ell^2\right)^2)^{\frac{1}{2}}}{-2\rho\left(1-\ell^2\right)} \;\; et \;\; \phi_2 = 0 \;\; et \;\; \phi_3 = \frac{-\ell - (\ell^2 + 4\left(1-\ell^2\right)^2)^{\frac{1}{2}}}{-2\rho\left(1-\ell^2\right)}$$

Dans ce cas le vecteur S centré réduit et orthonormalisé selon un processus de Gram-Schmidt est une représentation empirique de S. On obtient $\tilde{\mathsf{S}} = \mathsf{S}\tilde{\rho}^{-1}\phi\rho$. On voit bien que pour $\tilde{\rho} = 0$, $\phi \neq 0$, ρ donné et ℓ fixé, le vecteur $\tilde{\mathsf{S}}$ est perturbé de sorte que $\lim_{\phi \neq 0} \tilde{\mathsf{S}} \to \pm\infty$ et $\sigma_{\tilde{\mathsf{S}}}^2 \to +\infty$. Donc si dans un modèle de perturbation de données on assigne un coefficient β à $\tilde{\mathsf{S}}$ pour $\tilde{\rho} = 0$ et $\phi \neq 0$ alors ce coefficient converge vers 0 car il représente le rapport entre la covariance $\sigma_{\mathsf{X}\tilde{\mathsf{S}}}$ et la variance $\sigma_{\tilde{\mathsf{S}}}^2$. De fait $\tilde{\mathsf{S}}$ devient

trés peu voire non significatif dans le modèle.

Condition 2 $\tilde{\rho} = 0$ *et* $\phi \neq 0$ *est une condition nécessaire et suffisante pour minimiser le risque d'inférence prédictive dans les modèles de codage de données par perturbation.*

Soit le vecteur aléatoire X avec x_{1i} (resp. x_{2i}) la valeur prise par l'observation i pour la variable aléatoire X_1 (resp. X_2), $\forall i \in \{1...10\}$. Afin d'agir sur l'influence de $\tilde{\mathsf{S}}$ dans la transformation de X en Y, avec y_{1i} (resp. y_{2i}) la valeur prise par l'observation i pour la variable aléatoire Y_1 (resp. Y_2), $\forall i \in \{1...10\}$, on effectue une transformation de l'information contenue dans S orthonormalisé afin de créer le vecteur aléatoire $\tilde{\mathsf{S}}$ représentant le résultat de la perturbation du vecteur S orthonormalisé. On note ς_{1i} (resp. ς_{2i}) et $\tilde{\varsigma}_{1i}$ (resp. $\tilde{\varsigma}_{2i}$) les valeurs prises par l'observation i respectivement pour les variables aléatoires S_1 (resp. S_2) et $\tilde{\mathsf{S}}_1$ (resp. $\tilde{\mathsf{S}}_2$), $\forall i \in \{1...10\}$ avec $\tilde{\varsigma}_{1i} = \varsigma_{1i}\tilde{\rho}^{-1}\phi\rho$ et $\tilde{\varsigma}_{2i} = \varsigma_{2i}\tilde{\rho}^{-1}\phi\rho$. On retrouve $\sigma^2_{\tilde{\mathsf{S}}} = s^2_{\mathsf{S}}\tilde{\rho}^{-1}\phi\rho$ avec s^2_{S} la variance empirique de S.

Démonstration*.* Le modèle de perturbation de données doit vérifier les conditions (*requirements*) précédemment établies et notamment la condition de "*predictive inference risk requirements"*. Pour cela on considére la transformation croissante sur le vecteur S telle que $\tilde{\mathsf{S}} = \mathsf{S}\tilde{\rho}^{-1}\phi\rho$. Dans le cas *univarié*, le modèle de perturbation de données est :

$$y_i = \gamma + \alpha x_i + \beta\phi\rho\tilde{\rho}^{-1}\varsigma_i + \varepsilon_i, \forall i \in \{1...n\}$$

Donc les conditions recherchées sont telles que

$$\overline{\mathsf{Y}} = \overline{\mathsf{X}}$$
$$Cov(\mathsf{X},\mathsf{S}) = Cov(\mathsf{Y},\mathsf{S})$$
$$V(\mathsf{X}) = V(\mathsf{Y})$$

avec ε les résidus centrés et de variance σ^2_ε. La variable ε est orthogonale à X et S ce qui implique $\mathsf{E}(\varepsilon\mathsf{S}) = 0$ et $\mathsf{E}(\varepsilon\mathsf{X}) = 0$. Le premier résultat est la détermination de la constante (*intercept*) du modèle.

$$\gamma = \overline{\mathsf{Y}} - \alpha\overline{\mathsf{X}} - \beta\phi\rho\tilde{\rho}^{-1}\overline{\mathsf{S}} = (1-\alpha)\overline{\mathsf{X}} - \beta\phi\rho\tilde{\rho}^{-1}\overline{\mathsf{S}} \quad \text{(r1)}$$

En ce qui concerne la seconde condition

$$Cov(\mathsf{Y},\mathsf{S}) = Cov\left(\left(\alpha\mathsf{X} + \beta\phi\rho\tilde{\rho}^{-1}\mathsf{S} + \varepsilon\right), \left(\phi\rho\tilde{\rho}^{-1}\mathsf{S}\right)\right)$$
$$= \alpha\phi\rho\tilde{\rho}^{-1}\mathsf{E}(\mathsf{XS}) + \beta\left(\phi\rho\tilde{\rho}^{-1}\right)^2\mathsf{E}(\mathsf{S}^2) + \underbrace{\phi\rho\tilde{\rho}^{-1}\mathsf{E}(\varepsilon\mathsf{S})}_{0}$$

$$= \alpha\phi\rho\tilde{\rho}^{-1}s_{\mathsf{XS}} + \beta\left(\phi\rho\tilde{\rho}^{-1}\right)^2 s^2_{\mathsf{S}}$$
$$= Cov(\mathsf{X},\mathsf{S})$$

donc

$$\mathsf{E}(\mathsf{XS}) = \alpha\phi\rho\tilde{\rho}^{-1}s_{\mathsf{XS}} + \beta\left(\phi\rho\tilde{\rho}^{-1}\right)^2 s^2_{\mathsf{S}}$$
$$s_{\mathsf{XS}} - \alpha\phi\rho\tilde{\rho}^{-1}s_{\mathsf{XS}} = \beta\left(\phi\rho\tilde{\rho}^{-1}\right)^2 s^2_{\mathsf{S}}$$
$$\beta = \left(1 - \alpha\phi\rho\tilde{\rho}^{-1}\right)s_{\mathsf{XS}}\left(\left(\phi\rho\tilde{\rho}^{-1}\right)^2 s^2_{\mathsf{S}}\right)^{-1}$$

on voit bien qu'en développant β on a

$$\beta = \left(s_{XS} \left(\left(\phi\rho\tilde{\rho}^{-1} \right)^2 s_S^2 \right)^{-1} - s_{XS}\alpha \left(\left(\phi\rho\tilde{\rho}^{-1} \right) s_S^2 \right)^{-1} \right) \quad \text{(r2)}$$

le coefficient β est non significatif lorsque

$$\lim_{\phi\rho\tilde{\rho}^{-1} \to \pm\infty} s_{XS} \left(\left(\phi\rho\tilde{\rho}^{-1} \right)^2 s_S^2 \right)^{-1} = 0$$

$$\lim_{\phi\rho\tilde{\rho}^{-1} \to \pm\infty} s_{XS}\alpha \left(\left(\phi\rho\tilde{\rho}^{-1} \right) s_S^2 \right)^{-1} = 0$$

et significatif[8] comme dans le modèle SBNA dès que

$$\lim_{\phi\rho\tilde{\rho}^{-1} \to 1} s_{XS} \left(\left(\phi\rho\tilde{\rho}^{-1} \right)^2 s_S^2 \right)^{-1} = s_{XS} \left(s_S^2 \right)^{-1}$$

$$\lim_{\phi\rho\tilde{\rho}^{-1} \to 1} s_{XS}\alpha \left(\left(\phi\rho\tilde{\rho}^{-1} \right) s_S^2 \right)^{-1} = s_{XS}\alpha \left(s_S^2 \right)^{-1}$$

On en conclut que

$$\lim_{\phi\rho\tilde{\rho}^{-1} \to \pm\infty} \beta = \left(1 - \alpha\phi\rho\tilde{\rho}^{-1} \right) s_{XS} \left(\left(\phi\rho\tilde{\rho}^{-1} \right)^2 s_S^2 \right)^{-1} = 0$$

$$\lim_{\phi\rho\tilde{\rho}^{-1} \to 1} \beta = (1 - \alpha) s_{XS} \left(s_S^2 \right)^{-1}$$

Enfin la troisième condition est telle que:

$$V(Y) = E\left(\left(\alpha X + \beta\phi\rho\tilde{\rho}^{-1} S + \varepsilon \right), \left(\alpha X + \beta\phi\rho\tilde{\rho}^{-1} S + \varepsilon \right) \right)$$

$$= \alpha^2 E\left(X^2 \right) + 2\alpha\beta\phi\rho\tilde{\rho}^{-1} E(XS) + \underbrace{\alpha E(X\varepsilon)}_{0} + \beta^2 \left(\phi\rho\tilde{\rho}^{-1} \right)^2 E\left(S^2 \right)$$

$$+ \underbrace{\beta \left(\phi\rho\tilde{\rho}^{-1} \right) E(S\varepsilon)}_{0} + \underbrace{\alpha E(X\varepsilon)}_{0} + \underbrace{\beta \left(\phi\rho\tilde{\rho}^{-1} \right) E(S\varepsilon)}_{0} + E\left(\varepsilon^2 \right)$$

$$= V(X)$$

donc

$$s_X^2 = \alpha^2 s_X^2 + 2\alpha\beta\phi\rho\tilde{\rho}^{-1} s_{XS} + \beta^2 \left(\phi\rho\tilde{\rho}^{-1} \right)^2 s_S^2 + s_\varepsilon^2$$

$$s_\varepsilon^2 = \left(1 - \alpha^2 \right) s_X^2 - 2\alpha\beta\phi\rho\tilde{\rho}^{-1} s_{XS} - \beta^2 \left(\phi\rho\tilde{\rho}^{-1} \right)^2 s_S^2$$

en remplaçant β par sa valeur en **r2**

$$s_\varepsilon^2 = \left(1 - \alpha^2 \right) s_X^2 - 2\alpha \underbrace{\left(1 - \alpha\phi\rho\tilde{\rho}^{-1} \right) s_{XS} \left(\left(\phi\rho\tilde{\rho}^{-1} \right)^2 s_S^2 \right)^{-1}}_{\beta} \phi\rho\tilde{\rho}^{-1} s_{XS}$$

$$- \underbrace{\left(1 - \alpha\phi\rho\tilde{\rho}^{-1} \right)^2 s_{XS}^2 \left(\left(\phi\rho\tilde{\rho}^{-1} \right)^2 s_S^2 \right)^{-2}}_{\beta^2} \left(\phi\rho\tilde{\rho}^{-1} \right)^2 s_S^2$$

$$s_\varepsilon^2 = \left(1 - \alpha^2 \right) s_X^2 - \left(1 - \alpha\phi\rho\tilde{\rho}^{-1} \right) s_{XS}^2 \left(\left(\phi\rho\tilde{\rho}^{-1} \right)^2 s_S^2 \right)^{-1} \left(1 + \alpha\phi\rho\tilde{\rho}^{-1} \right)$$

$$s_\varepsilon^2 = \left(1 - \alpha^2 \right) s_X^2 - \left(1 - \alpha^2\phi\rho\tilde{\rho}^{-2} \right) s_{XS}^2 \left(\left(\phi\rho\tilde{\rho}^{-1} \right)^2 s_S^2 \right)^{-1}$$

[8] Comme dans Muralidhar et Sarathy (2008).

et en développant on obtient

$$s_\varepsilon^2 = \left(1-\alpha^2\right)s_{\mathrm{X}}^2 - \left(\left(\phi\rho\tilde{\rho}^{-1}\right)^{-2} - \alpha^2\right)s_{\mathrm{XS}}^2\left(s_{\mathrm{S}}^2\right)^{-1} \tag{r3}$$

Là encore on constate que

$$\lim_{\phi\rho\tilde{\rho}^{-1}\to\pm\infty} s_\varepsilon^2 = \left(1-\alpha^2\right)s_{\mathrm{X}}^2 + \alpha^2 s_{\mathrm{XS}}^2\left(s_{\mathrm{S}}^2\right)^{-1}$$

$$\lim_{\phi\rho\tilde{\rho}^{-1}\to 1} s_\varepsilon^2 = \left(1-\alpha^2\right)\left(s_{\mathrm{X}}^2 - s_{\mathrm{XS}}^2\left(s_{\mathrm{S}}^2\right)^{-1}\right)$$

La variance de l'erreur en **r3** est la même que dans le modèle SBNA lorsque $\phi\rho\tilde{\rho}^{-1}\to 1$ mais comme le vecteur S est préalablement centré réduit et orthonormalisé alors $s_{\mathrm{XS}}^2\left(s_{\mathrm{S}}^2\right)^{-1}\to 0$ car $s_{\mathrm{XS}}^2\to 0$. Par contre, dés que $\phi\rho\tilde{\rho}^{-1}\to\pm\infty$ on a également $s_{\mathrm{XS}}^2\left(s_{\mathrm{S}}^2\right)^{-1}\to 0$ car $s_{\mathrm{S}}^2\to+\infty$. Donc on peut en conclure que

$$s_\varepsilon^2 = \lim_{\phi\rho\tilde{\rho}^{-1}\to\pm\infty} s_\varepsilon^2 = \lim_{\phi\rho\tilde{\rho}^{-1}\to 1} s_\varepsilon^2 \tag{r4}$$

La variance de l'erreur reste inchangée.

Condition 3 *La condition 2 est robuste au cas multivarié*

Démonstration. Dans le cas *multivarié*, les représentations multivariées des variance-covariances et moyennes sont notées Σ et μ. on a

$$y_i = \gamma + x_i\alpha^T + \phi\rho\tilde{\rho}^{-1}\varsigma_i\beta^T + \varepsilon_i, \forall i\in\{1\ldots n\}$$

et les conditions nécessaires deviennent

$$\mu_{\mathrm{X}} = \mu_{\mathrm{Y}}$$

$$\Sigma_{\mathrm{X}\phi\rho\tilde{\rho}^{-1}\mathrm{S}} = \Sigma_{\mathrm{Y}\phi\rho\tilde{\rho}^{-1}\mathrm{S}}$$

$$\Sigma_{\mathrm{XX}} = \Sigma_{\mathrm{YY}}$$

ce qui nous donne les résultats suivants

$$\beta^T = \Sigma^{-1}_{\left(\phi\rho\tilde{\rho}^{-1}\right)^2\mathrm{S}^2}\Sigma_{\phi\rho\tilde{\rho}^{-1}\mathrm{SX}}\left(I-\alpha^T\right) \tag{r5}$$

$$\gamma = (I-\alpha)\mu_{\mathrm{X}} - \mu_{\beta\phi\rho\tilde{\rho}^{-1}\mathrm{S}} \tag{r6}$$

$$\Sigma_{\varepsilon\varepsilon} = \left(\Sigma_{\mathrm{XX}} - \Sigma_{\mathrm{X}\phi\rho\tilde{\rho}^{-1}\mathrm{S}}\Sigma^{-1}_{\left(\phi\rho\tilde{\rho}^{-1}\right)^2\mathrm{S}^2}\Sigma_{\phi\rho\tilde{\rho}^{-1}\mathrm{SX}}\right)$$

$$-\alpha\left(\Sigma_{\mathrm{XX}} - \Sigma_{\mathrm{X}\phi\rho\tilde{\rho}^{-1}\mathrm{S}}\Sigma^{-1}_{\left(\phi\rho\tilde{\rho}^{-1}\right)^2\mathrm{S}^2}\Sigma_{\phi\rho\tilde{\rho}^{-1}\mathrm{SX}}\right)\alpha^T$$

avec I la matrice identité.

∎

Condition 4 La condition 2 et la condition 3 sont robustes aux prévisions

On voit bien que si $y_i = \gamma + \alpha x_i + \beta\phi\rho\tilde{\rho}^{-1}\varsigma_i + \varepsilon_i, \forall i\in\{1\ldots n\}$

alors après simplifications

$$x_i = (y_i - \gamma)\alpha^{-1}, \forall i\in\{1\ldots n\} \tag{r7}$$

avec y_i la valeur masquée et publiquement divulguée et x_i la valeur décodée. Si on note $prév$ le résultat de la prévision on a :

$$prév(x_i) = (prév(y_i) - \gamma)\alpha^{-1}, \forall i \in \{1...n\} \quad \text{(r8)}$$

Exemple 1 *Si on prend une valeur initiale* $y_2 = 2{,}13808569$ *dont on cherche à faire une prévision* y_1^*. *Alors* $prév(y_2) = y_1^* = 12{,}58336345$. *En utilisant r7, pour* $\alpha = 0.7$ *on obtient le décodage de la valeur initiale* $x_2^{décodé} = 2{,}138117106$ *et le décodage de la prévision* y_1^* *en* $x_1^{décodé} = 12{,}58336797$. *Maintenant si on fait une prévision* $prév\left(x_2^{décodé}\right) = x_1^* = 12{,}58336797$ *on vérifie que* $x_1^* = x_1^{décodé}$.

4- Une base de simulation pour minimiser le risque d'inférence prédictive dans les méthodes de perturbation : approche empirique

Dans ce qui suit, on propose une application de la méthode à différents modèles de codage. L'objectif est que les variables non confidentielles aient une très faible influence dans la transformation des variables confidentielles en variables masquées et publiquement divulguées Y.

Perturbation Approaches Based on Linear Models - Le modèle de Franconi et Stander (2002) utilise une méthode de perturbation avec variables S. Ce modèle représente un cas particulier de Muralidhar et al. (1999). Les auteurs recherchent différentes possibilités permettant d'obtenir les conditions selon lesquelles les caractéristiques $\{\mathrm{Y}, \mathrm{S}\}$ sont les mêmes que $\{\mathrm{X}, \mathrm{S}\}$. Le modèle GADP (Muralidhar et al. (2001)) (General Additive Data Perturbation) généralise en terme d'inférence prédictive les modèles de perturbation précédents. En ce qui nous concerne, en appliquant la transformation S, les paramètres sont:

$$\gamma = \mu_{\mathrm{X}} - \Sigma_{\mathrm{X}\tilde{\mathrm{S}}}\Sigma_{\tilde{\mathrm{S}}\tilde{\mathrm{S}}}^{-1}\mu_{\tilde{\mathrm{S}}}$$
$$\beta = \Sigma_{\mathrm{X}\tilde{\mathrm{S}}}\Sigma_{\tilde{\mathrm{S}}\tilde{\mathrm{S}}}^{-1}$$
$$\Sigma_\varepsilon = \left(\Sigma_{\mathrm{XX}} - \Sigma_{\mathrm{X}\tilde{\mathrm{S}}}\Sigma_{\tilde{\mathrm{S}}\tilde{\mathrm{S}}}^{-1}\Sigma_{\tilde{\mathrm{S}}\mathrm{X}}\right)$$

La contante du modèle est représentée par γ, β est le paramètre relatif à S transformée en $\tilde{\mathrm{S}}$ et Σ_ε la matrice des variance-covariances de l'erreur. Le problème de ces méthodes est qu'elles fonctionnent très bien sur de large bases de données mais ne sont plus aussi efficaces sur des bases plus petites. Le modèle IPSO (Burridge (2003)) vient palier à ce problème en générant les variables perturbées en fonction de la distribution conditionnelle de Y sachant S. La même spécificité des paramètres est conservée pour γ, β et Σ_ε. Une amélioration supplémentaire de Muralidhar and Sarathy (2005a) assure une minimisation des risques de divulgation. Cette amélioration est référencée sous l'appelation de EGADP (Exact General Additive Data Perturbation). Par la suite, pour solutionner le problème de perte informationnelle entre X et Y Muralidhar et Sarathy (2008) proposent la méthode SBNA. Cette méthode généralise la méthode IPSO par l'intermédiaire d'un coefficient de similarité α. En tenant compte de ce coefficient de similarité, les paramètres précédents sont:

$$\gamma = (I - \alpha)\mu_{\mathrm{X}} - \beta\mu_{\tilde{\mathrm{S}}}$$
$$\beta^T = \Sigma_{\tilde{\mathrm{S}}\tilde{\mathrm{S}}}^{-1}\Sigma_{\tilde{\mathrm{S}}\mathrm{X}}\left(I - \alpha^T\right)$$
$$\Sigma_\varepsilon = \left(\Sigma_{\mathrm{XX}} - \Sigma_{\mathrm{X}\tilde{\mathrm{S}}}\Sigma_{\tilde{\mathrm{S}}\tilde{\mathrm{S}}}^{-1}\Sigma_{\tilde{\mathrm{S}}\mathrm{X}}\right) - \alpha\left(\Sigma_{\mathrm{XX}} - \Sigma_{\mathrm{X}\tilde{\mathrm{S}}}\Sigma_{\tilde{\mathrm{S}}\tilde{\mathrm{S}}}^{-1}\Sigma_{\tilde{\mathrm{S}}\mathrm{X}}\right)\alpha^T$$

Perturbation Approaches Based on Nonlinear Models - Sarathy et al. (2002) ont récemment proposé une approche alternative basée sur l'approximation de la distribution jointe de l'ensemble des variables par une copule gaussienne multivariée. La méthode utilise une modélisation GADP sur les variables issues de la distribution d'une copule multivariée afin de créer une variable masquée. Pour créer les variables masquées et publiquement divulguées, les auteurs ré-utilisent la fonction quantile afin de conserver les liaisons

monotones et les densités marginales. Muralidhar and Sarathy (2005b) ont proposé une autre méthode de mélange de données (*data shuffling*) combinant cette fois perturbation et échange de données (*data swapping*). La méthode utilise à nouveau une modélisation GADP.

Results from the Minimization of Predictive Inference Risk in Perturbation Methods - Dans chacune des méthodes de perturbations linéaires ou non linéaires précédentes, on analyse les résultats selon deux groupes:

Concernant les modèles GADP, EGADP et IPSO d'une part, si $\ell = 0$ et ρ donné alors soit $\phi = 0$ et dans ce cas γ, β et Σ_ε restent inchangés car $\tilde{\rho}(\phi\rho)^{-1} = 1$, soit $\phi \neq 0$ et $\tilde{\rho}(\phi\rho)^{-1} \approx 0$ et dans ce cas

$$\Sigma_{\tilde{S}\tilde{S}} \to \pm\infty$$
$$\gamma \to \mu_X$$
$$\beta \to 0$$
$$\Sigma_\varepsilon \to \Sigma_{XX}$$

Si $\ell \neq 0$ et ρ donné alors soit $\phi = 0$ et dans ce cas γ, β et Σ_ε varient car $\tilde{\rho}(\phi\rho)^{-1} = \left(1 - \ell^2\right)$, soit $\phi \neq 0$ et $\tilde{\rho}(\phi\rho)^{-1} \approx 0$ et dans ce cas

$$\Sigma_{\tilde{S}\tilde{S}} \to \pm\infty$$
$$\gamma \to \mu_X$$
$$\beta \to 0$$
$$\Sigma_\varepsilon \to \Sigma_{XX}$$

Concernant les modèles SBNA, MicroHybrid method et PCA method d'autre part, $\forall \ell \in [-1;1]$ et pour ρ donné si $\phi = 0$ alors on retrouve les mêmes résultats que précédemment à $(I - \alpha)$ prés. Par contre si $\phi \neq 0$ alors on obtient

$$\gamma \to (I - \alpha)\mu_X$$
$$\beta^T \to 0$$
$$\Sigma_\varepsilon \to \Sigma_{XX} - \alpha\Sigma_{XX}\alpha^T$$

Dès que $\phi \neq 0$ on a $\beta \to 0$ et $\tilde{S}$ n'est quasiment pas influent dans la transformation de X en Y. De plus, le terme d'erreur ε reste inchangé en ce sens que $\Sigma_\varepsilon \to \Sigma_{XX}$ et la constante ne dépend plus que de μ_X. Tout cela à $(I - \alpha)$ prés si l'on prend en compte le coefficient de similarité α, $\forall \ell \in [-1;1]$ et pour ρ donné.

Empirical Examples: The SBNA Multivariate case - Le tableau 1 (resp. 2) en Appendice 1 récapitule les matrices des variance-covariances de trois simulations pour $\ell = -2{,}00E-10$ (resp. $\ell = 0.6$) et[9] $\rho = 0{,}87038828$ lorsque $\alpha = \begin{pmatrix} 0{,}7 & 0 \\ 0 & 0{,}7 \end{pmatrix}$ et après transformation de S préalablement orthonormalisé par un processus de Gram-Schmidt en $\tilde{S}$. Avant tout, on récapitule les résultats de $\Sigma_{\tilde{S}\tilde{S}}$, β^T et Σ_ε pour les 3 simulations:

[9] Avec $E \pm k = 10^{\pm k}$

★ d'une part $r_{\tilde{S}} = -1{,}30713E-16$ et $\tilde{\rho} = 3{,}35E-10$ avec $\tilde{\rho}(\phi_1\rho)^{-1} = 3{,}35E-10$ tel que:

Simulation A $\rightarrow \Sigma_{\tilde{S}\tilde{S}} = \begin{pmatrix} 8{,}88558E+17 & -153{,}6 \\ -153{,}6 & 1{,}55402E+18 \end{pmatrix}$ *et* $\beta^T = \begin{pmatrix} -6{,}10435E-11 & -1{,}53232E-10 \\ 4{,}61911E-12 & 1{,}34641E-10 \end{pmatrix}$

★ d'autre part $r_{\tilde{S}} = -1{,}46914E-16$ et $\tilde{\rho} = 2E-10$ avec $\tilde{\rho}(\phi_2\rho)^{-1} = 1$ tel que:

Simulation B $\rightarrow \Sigma_{\tilde{S}\tilde{S}} = \begin{pmatrix} 0{,}1 & -1{,}94289E-17 \\ -1{,}94289E-17 & 0{,}174892348 \end{pmatrix}$ *et* $\beta^T = \begin{pmatrix} -0{,}181962612 & -0{,}456764668 \\ 0{,}013768967 & 0{,}401348488 \end{pmatrix}$

★ puis $r_{\tilde{S}} = -1{,}43976E-16$ et $\tilde{\rho} = -7{,}35E-10$ avec $\tilde{\rho}(\phi_3\rho)^{-1} = 7{,}35E-10$ tel que:

Simulation C $\rightarrow \Sigma_{\tilde{S}\tilde{S}} = \begin{pmatrix} 1{,}8487E+17 & -35{,}2 \\ -35{,}2 & 3{,}23324E+17 \end{pmatrix}$ *et* $\beta^T = \begin{pmatrix} -1{,}33829E-10 & -3{,}35938E-10 \\ 1{,}01267E-11 & 2{,}95181E-10 \end{pmatrix}$

★★ par ailleurs $r_{\tilde{S}} = -9{,}66269E-17$ et $\tilde{\rho} = 1{,}11E-09$ avec $\tilde{\rho}(\phi_1\rho)^{-1} = 7{,}06516E-10$ tel que:

Simulation D $\rightarrow \Sigma_{\tilde{S}\tilde{S}} = \begin{pmatrix} 2{,}00335E+17 & -25{,}6 \\ -25{,}6 & 3{,}5037E+17 \end{pmatrix}$ *et* $\beta^T = \begin{pmatrix} -1{,}28559E-10 & -3{,}22711E-10 \\ 9{,}72799E-12 & 2{,}83559E-10 \end{pmatrix}$

★★ ensuite $r_{\tilde{S}} = -1{,}20352E-16$ et $\tilde{\rho} = 1{,}00E-09$ avec $\tilde{\rho}(\phi_2\rho)^{-1} = 0{,}640000001$ tel que:

Simulation E $\rightarrow \Sigma_{\tilde{S}\tilde{S}} = \begin{pmatrix} 0{,}244140624 & -3{,}88578E-17 \\ -3{,}88578E-17 & 0{,}426983271 \end{pmatrix}$ *et* $\beta^T = \begin{pmatrix} -0{,}116456072 & -0{,}292329388 \\ 0{,}008812139 & 0{,}256863033 \end{pmatrix}$

★★ et enfin $r_{\tilde{S}} = -1{,}75574E-16$ et $\tilde{\rho} = 9{,}15E-10$ avec $\tilde{\rho}(\phi_3\rho)^{-1} = -1{,}43984E-09$ tel que:

Simulation F $\rightarrow \Sigma_{\tilde{S}\tilde{S}} = \begin{pmatrix} 4{,}8236E+16 & -11{,}2 \\ -11{,}2 & 8{,}43611E+16 \end{pmatrix}$ *et* $\beta^T = \begin{pmatrix} 2{,}61997E-10 & 6{,}57668E-10 \\ -1{,}98251E-11 & -5{,}77878E-10 \end{pmatrix}$

On note que; au même titre que dans les simulations B et E; dans les simulations A, C, D et F, la covariance des erreurs est

$$\Sigma_\varepsilon = \begin{pmatrix} 0{,}088580701 & 0{,}299591506 \\ 0{,}299591506 & 2{,}172061748 \end{pmatrix}$$

Lorsque l'on sélectionne le modèle de régression permettant éventuellement de tester si notre filtre fournit les résultats escomptés, il est nécessaire de bien spécifier le modèle de régression. Une spécification incomplète est équivalente à une spécification erronée. En d'autres termes, si une perturbation a lieu il faut la prendre en considération dans le modèle. Pour une perturbation *d'ordre 1* on a:

$$y_i = \gamma + \alpha x_i + \beta_2\phi_2\rho\tilde{\rho}^{-1}\varsigma_i + \beta_1\phi_1\rho\tilde{\rho}^{-1}\varsigma_i + \varepsilon_i, \forall i \in \{1...n\} \quad \text{(r9)}$$

pour une perturbation *muette d'ordre 2*:

$$y_i = \gamma + \alpha x_i + \beta_2\phi_2\rho\tilde{\rho}^{-1}\varsigma_i + \varepsilon_i, \forall i \in \{1...n\} \quad \text{(r10)}$$

et pour une perturbation *d'ordre 3:*

$$y_i = \gamma + \alpha x_i + \beta_2\phi_2\rho\tilde{\rho}^{-1}\varsigma_i + \beta_3\phi_3\rho\tilde{\rho}^{-1}\varsigma_i + \varepsilon_i, \forall i \in \{1...n\} \quad \text{(r11)}$$

Si on génère une perturbation ϕ_1 ou ϕ_3, une utilisation de r10 au lieu de r9 ou r11 donnera des β probablement non nuls mais le modèle sera mal spécifié. Si par contre, on utilise le modèle r9 ou r11 selon le filtre utilisé alors le modèle sera bien spécifié mais les β seront non significatifs de sorte que $\beta \to 0$.

Exemple 2 *En appendice 2 les résultats montrent des coefficients β non nuls lorsque la perturbation est muette (simulations B et E) compte tenu d'une régression linéaire du type r10. Par contre les autres β sont clairement nuls. Les variables non confidentielles deviennent absolument non significatives dans la détermination de y_i lorsque les régressions linéaires sont de type r9 ou r11 et par conséquent dans la déduction de x_i.*

En appendice 3 on peut vérifier que les covariances et corrélations sont maintenues même aprés une perturbation des données.

5- CONCLUSION

On définit une condition nécessaire et suffisante montrant que les variables non confidentielles orthonormalisées et perturbées en leur structure satisfont la "*predictive inference risk requirements*" qui minimise les risques de divulgation prédictive tout en conservant les "*data utility requirements*" et "*disclosure risk requirements*".

L'indépendance des variables favorisant les risques d'inférence prédictive est générée par une *orthonormalisation de Gram Scmidt*. Par conséquent, quelque soit la linéarité ou la non linéarité de la liaison entre les variables favorisant les risques d'inférence prédictive, le coefficient de corrélation linéaire de perturbation $\tilde{\rho}$ est égale à 0 au même titre que ceux de spearman $\tilde{\rho}_{spearman}$ et Kendall $\tilde{\tau}$ dans le cas gaussien.

En ce qui concerne les risques de divulgation prédictive par inférence, le modèle répond bien aux méthodes de prédiction relatives au Data Minig. Considérant l'arbitrage nécessaire entre biais et approximation d'une part face à variance et estimation d'autre part, la construction même du modèle permet de jouer sur les deux aspects. On ajoute un moyen de contrôle supplémentaire face aux tentative d'inférence prédictive par l'intermédiaire du filtre ϕ qui représente la perturbation. On en déduit qu'une méthode permettant de complexifier le modèle et de fait permettant de diminuer l'approximation (tel que le modèle inféré *fit* parfaitement avec les données sensibles) engendre une augmentation trés importante de la variance ce qui pénalise l'estimation et toute possibilité de prévision. A contrario, un modèle moins complexe (disposant de moins de variables explicatives) diminue la variance et améliore l'estimation mais tout ça au détriment d'une trop grande approximation et par conséquent d'un biais trop important (qui ne permet pas de *fitter* correctement les données sensibles) (*cf.* Exemple 2).

Bibliographie

[1] BRAND R. (2002). Microdata Protection through Noise Addition. In: Domingo-Ferrer J. (eds) *Inference Control in Statistical Databases*. Lecture Notes in Computer Science, 2316. Springer, Berlin, Heidelberg.

[2] BURRIDGE, J. (2003). Information preserving statistical obfuscation. *Statistics and Computing*, 13, 321-327.

[3] CALVIÑO, A. (2017). A Simple Method for Limiting Disclosure in Continuous Microdata Based on Principal Component Analysis. *Journal of Official Statistics*, 33, 1, 15–41

[4] DOMINGO-FERRER, J. and U. GONZÁLEZ-NICOLÁS, (2010). Hybrid Microdata Using Microaggregation. *Information Sciences*, 180, 2834–2844.

[5] DONOHO D. (2015). *50 years of Data Science*. Retrieved from http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf.

[6] DRECHSLER, J., REITER J.P., (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105, 1347-1357.

[7] DRECHSLER, J. (2011). *Synthetic datasets for statistical disclosure control: theory and implementation*. Springer Science Business Media.

[8] DRECHSLER, J. (2012). New Data Dissemination Approaches in Old Europe-Synthetic Datasets for a

German Establishment Survey. *Journal of Applied Statistics*, 39, 243–265.

[9] FELLEGI, I. P. (1972). On the question of statistical confidentiality. *J. Amer. Statist. Assoc.*, 67, 7-18.

[10] FIENBERG, S. (1994). *A Radical Proposal for the Provision of Micro-Data Samples and the Preservation of Confidentiality*. Technical Report 611, Department of Statistics, Carnegie Mellon University.

[11] FRANCONI, L., STANDER J. (2002) A model based method for disclosure limitation of business microdata. *The Statistician*, 51, 1–11.

[12] FULLER, W. A. (1993) Masking procedures for microdata disclosure limitation. *Journal of Official Statistics*, 9, 383-406.

[13] HUNDEPOOL, A., J. DOMINGO-FERRER, L. FRANCONI, S. GIESSING, R. LENZ, J. NAYLOR, E. SCHULTE NORDHOLT, G. SERI, P. De WOLF, (2010). *Handbook on Statistical Disclosure Control*. ESSNet SDC.

[14] LEE, S., GENTON M. G., ARELLANO-VALLE, R. B., (2010). Perturbation of numerical confidential data via skew-t distributions. *Management Science*, 56, 318-333.

[15] LIEW, C. K., CHOI, U. J, LIEW C. J., (1985). A data distortion by probability distribution. *ACM Trans. Database Systems*, 10, 395-411.

[16] MATTHEWS G. J., HAREL O., (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, 5, 1-29.

[17] MIRANDA, J., VILHUBER, L. (2014). Using Partially Synthetic Data to Replace Suppression in the Business Dynamics Statistics: Early Results. In: Domingo-Ferrer J. (eds) *Privacy in Statistical Databases*. PSD 2014. Lecture Notes in Computer Science, vol 8744. Springer, Cham 232-242.

[18] MIRANDA, J., VILHUBER, L.,(2016). *Using Partially Synthetic Microdata to Protect Sensitive Cells in Business Statistics*, Working Papers 16-10 Center for Economic Studies, U.S. Census Bureau.

[19] MURALIDHAR, K., PARSA, R., SARATHY , R., (1999). A general additive data perturbation method for database security. *Management Science*, 45, 1399-1415.

[20] MURALIDHAR, K., PARSA, R., SARATHY , R., (2001). An improved security requirement for data perturbation with implications for e-commerce. *Decision Sciences*, 32, 683-698.

[21] MURALIDHAR, K., SARATHY , R., (2003). A theoretical basis for perturbation methods. *Statistics and Computing*, 13, 329-335.

[22] MURALIDHAR, K., SARATHY , R. (2005a), An enhanced data perturbation approach for small data sets. *Decision Sciences*, 36, 513-529.

[23] MURALIDHAR, K., SARATHY , R. (2005b), *Data shuffling: A new masking approach for numerical data.* Working Paper - University Of Kentucky, Lexington KY.

[24] MURALIDHAR, K., SARATHY , R. (2008), Generating Sufficiency Based Non-synthetic Perturbed Data. *Transactions on Data Privacy*, 1, 17-33.

[25] PALLEY, M. A., SIMONOFF, J. S. (1987), The use of regression methodology for the compromise of confidential information in statistical databases. *ACM Trans. Database Systems*, 12, 593-608.

[26] RAGHUNATHAN, T.E., REITER , J., RUBIN, D. (2003), Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics*, 19, 1–16.

[27] RUBIN, D. (1993), Statistical Disclosure Limitation. *Journal of Official Statistics*, 9, 461–468.

[28] SARATHY , R., MURALIDHAR, K. (2002), The Security of Numerical Confidential Data in Databases. *Information Systems Research*, 13, 389-403.

[29] SARATHY , R., MURALIDHAR, K., PARSA R. (2002), Perturbing non-normal confidential variables: The copula approach. *Management Science*, 48, 1613-1627.

[30] SARATHY , R., MURALIDHAR , K. (2006), Secure and Useful Data Sharing. *Decision Support Systems*, 42, 204- 220.

[31] WILLENBORG, L., de WAAL, T. (2001), *Elements of Statistical Disclosure Control*. New York: Springer

APPENDICE 1

Table 1. Covariances between Original, Masked and Perturbed non confidential data.

				SIMULATION A		SIMULATION B		SIMULATION C	
				$\phi_1 = 1,148912529$		$\phi_2 = 2,29782E-10$		$\phi_3 = -1,148912529$	
$\mathcal{X}_1$	$\mathcal{X}_2$	$\mathcal{Y}_1$	$\mathcal{Y}_2$	$\tilde{\mathcal{S}}_1$	$\tilde{\mathcal{S}}_2$	$\tilde{\mathcal{S}}_1$	$\tilde{\mathcal{S}}_2$	$\tilde{\mathcal{S}}_1$	$\tilde{\mathcal{S}}_2$
13,1336116	3,405629582	12,99177362	0,798138311	−1476822990	−1611185857	−0,495433694	−0,540508759	−673626014,9	−734913199,2
12,21902051	0,381739203	12,38209476	2,185487635	−1148640103	153446272,1	−0,385337318	0,051477025	−523931344,9	69991733,26
12,45489329	3,918321291	12,8913032	5,311166729	−820457216,5	652146656,4	−0,27524941	0,218777355	−374236674,9	297464866,3
12,65588765	2,607636941	12,29699151	0,828069752	−492274329,9	241677785	−0,16514565	0,810763138	−224542005	1102369799
13,1550851	1,343879314	13,48421775	4,103615276	−164091443,3	383615680,2	−0,055048188	0,128692562	−74847334,99	174979333,1
12,91117111	4,329667476	12,65936243	4,422694477	164091443,3	−1649547425	0,055048188	−0,553378015	74847334,99	−752411132,5
12,40024436	−0,743247498	12,15790086	−1,139169272	492274329,9	−1150847041	0,16514565	−0,386077685	224542005	−524937999,4
12,66972201	3,28571431	12,16437319	2,288800287	820457216,5	613785088,3	0,275240941	0,205908099	374236674,9	279966933
12,72553848	4,881391268	12,89529343	3,857115875	1148640103	111248547 3	0,385337318	0,373208429	523931344,9	507440066,1
11,50835506	−2,030608038	11,91021843	−1,27579522	147682290	−920677632,5	0,495433694	−0,308862148	673626014,9	−419950399,5
$\Sigma_{\mathcal{X}\mathcal{X}}$				Σ_{yy}		$\Sigma_{\mathcal{X}\tilde{\mathcal{S}}} = \Sigma_{y\tilde{\mathcal{S}}}$			
0,210845383	0,690522006	0,210845383	0,690522006	−180802242,3	23927306,2	−0,060654204	0,008026956	−82469662,81	10914006,65
0,690522006	4,803779739	0,690522006	4,803779739	−453851894,9	697451625,4	−1,94289E−17	0,174892348	−207016308,3	318129906,1

Table 2. Covariances between Original, Masked and Perturbed non confidential data.

				SIMULATION D		SIMULATION E		SIMULATION F	
				$\phi_1 = 1,807425894$		$\phi_2 = 1,79518E-09$		$\phi_3 = -0,7303204$	
$\mathcal{X}_1$	$\mathcal{X}_2$	$\mathcal{Y}_1$	$\mathcal{Y}_2$	$\tilde{\mathcal{S}}_1$	$\tilde{\mathcal{S}}_2$	$\tilde{\mathcal{S}}_1$	$\tilde{\mathcal{S}}_2$	$\tilde{\mathcal{S}}_1$	$\tilde{\mathcal{S}}_2$
13,1336116	3,405629582	12,99177362	0,798138311	−701235298,5	−765034403,6	−0,774115146	−0,844544934	344089442,6	375395052,3
12,21902051	0,381739203	12,38209476	2,185487635	−545405232,1	72860419,39	−0,602089558	0,080432851	267625122	−35751909,74
12,45489329	3,918321291	12,8913032	5,311166729	−389575165,8	309656782,4	−0,43006397	0,341839616	191160801,4	−151945616,4
12,65588765	2,607636941	12,29699151	0,828069752	−233745099,5	1147551605	−0,258038382	1,266817402	114696480,9	−563092578,4
13,1550851	1,343879314	13,48421775	4,103615276	−77915033,16	182151048,5	−0,086012794	0,201082127	38232160,29	−89379774,35
12,91117111	4,329667476	12,65936243	4,422694477	77915033,16	−783249508,5	0,086012794	−0,864653147	−38232160,29	384333029,7
12,40024436	−0,743247498	12,15790086	−1,139169272	233745099,5	−546453145,5	0,258038382	−0,603246382	−114696480,9	268139323
12,66972201	3,28571431	12,16437319	2,288800287	389575165,8	291441677,6	0,43006397	0,321731404	−191160801,4	−143007639
12,72553848	4,881391268	12,89529343	3,857115875	545405232,1	528238040,6	0,602089558	0,583138169	−267625122	−259201345,6
11,50835506	−2,030608038	11,91021843	−1,27579522	701235298,5	−437162516,4	0,774115146	−0,482597105	−344089442,6	214511458,4
$\Sigma_{\mathcal{X}\mathcal{X}}$				Σ_{yy}		$\Sigma_{\mathcal{X}\tilde{\mathcal{S}}} = \Sigma_{y\tilde{\mathcal{S}}}$			
0,210845383	0,690522006	0,210845383	0,690522006	−85849770,25	11361328,9	−0,094772194	0,012542119	42125659,76	−5574895,239
0,690522006	4,803779739	0,690522006	4,803779739	−215501093,3	331168800,9	−0,237898264	0,365587393	105744321,8	−162501357,8

APPENDICE 2

★

	Dependent variable :					
	Y1	Y2	Y1	Y2	Y1	Y2
	(1)	(2)	(3)	(4)	(5)	(6)
X1	0.700* (0.292)		0.700* (0.292)		0.700* (0.292)	
X2		0.700* (0.292)		0.700* (0.292)		0.700* (0.292)
S1A	-0.000(0.000)	-0.000(0.000)				
S2A	0.000(0.000)	0.000(0.000)				
S1B			-0.182(0.423)	-0.457(1.954)		
S2B			0.014(0.291)	0.401(1.491)		
S1C					-0.000(0.000)	-0.000(0.000)
S2C					0.000(0.000)	0.000(0.000)
Constant	3.775(3.671)	0.641(0.866)	3.775(3.671)	0.641(0.866)	3.775(3.671)	0.641(0.866)
Observations	10	10	10	10	10	10
R^2	0.580	0.548	0.580	0.548	0.580	0.548
Adjusted R^2	0.370	0.322	0.370	0.322	0.370	0.322
Residual Std. Error (df = 6)	0.384	1.903	0.384	1.903	0.384	1.903
F Statistic (df = 3 ; 6)	2.761	2.423	2.761	2.423	2.761	2.423

Note : *p<0.1 ; **p<0.05 ; ***p<0.01

★★

	Dependent variable :					
	Y1	Y2	Y1	Y2	Y1	Y2
	(1)	(2)	(3)	(4)	(5)	(6)
X1	0.700* (0.292)		0.700* (0.292)		0.700* (0.292)	
X2		0.700* (0.292)		0.700* (0.292)		0.700* (0.292)
S1D	-0.000(0.000)	-0.000(0.000)				
S2D	0.000(0.000)	0.000(0.000)				
S1E			-0.116(0.271)	-0.292(1.250)		
S2E			0.009(0.186)	0.257(0.954)		
S1F					0.000(0.000)	0.000(0.000)
S2F					-0.000(0.000)	-0.000(0.000)
Constant	3.775(3.671)	0.641(0.866)	3.775(3.671)	0.641(0.866)	3.775(3.671)	0.641(0.866)
Observations	10	10	10	10	10	10
R^2	0.580	0.548	0.580	0.548	0.580	0.548
Adjusted R^2	0.370	0.322	0.370	0.322	0.370	0.322
Residual Std. Error (df = 6)	0.384	1.903	0.384	1.903	0.384	1.903
F Statistic (df = 3 ; 6)	2.761	2.423	2.761	2.423	2.761	2.423

Note : *p<0.1 ; **p<0.05 ; ***p<0.01

APPENDICE 3

★

Cov	*Y1*	*Y2*	*X 1*	*X2*	*S1A*	*S2A*	*S1B*	*S2B*	*S1C*	*S2C*
Y1	0,21084538									
Y2	0,69052201	4,80377974								
X 1	0,15873909	0,51429171	0,21084538							
X2	0,51429171	3,52609636	0,69052201	4,80377974						
S1A	-180802242	-453851895	-180802242	-453851895	8,8856E+17					
S2A	23927306,2	69745162 5	23927306,2	69745162 5	-153,6	1,554E+18				
S1B	-0,0606542	-0,15225489	-0,0606542	-0,15225489	298086910	-3,5763E-08	0,1			
S2B	0,00802696	0,23397593	0,00802696	0,23397593	7,7486E-08	521331197	3,0531E-17	0,17489235		
S1C	-8246966 2,8	-207016308	-8246966 2,8	-207016308	4,053E+17	-76,8	135966936	2,9802E-08	1,8487E+17	
S2C	10914006,6	318129906	10914006,6	318129906	-76,8	7,0884E+17	-2,3842E-08	23795768	-35,2	3,2332E+17

★★

Cov	*Y1*	*Y2*	*X 1*	*X2*	*S1D*	*S2D*	*S1E*	*S2E*	*S1F*	*S2F*
Y1	0,21084538									
Y2	0,69052201	4,80377974								
X 1	0,15873909	0,51429171	0,21084538							
X2	0,51429171	3,52609636	0,69052201	4,80377974						
S1D	-8584977 0,3	-215501093	-8584977 0,3	-215501093	2,0033E+17					
S2D	11361328,9	331168801	11361328,9	331168801	-25,6	3,5037E+17				
S1E	-0,09477219	-0,23789826	-0,09477219	-0,23789826	221155760	-3,5763E-08	0,24414062			
S2E	0,01254212	0,36558739	0,01254212	0,36558739	-3,5763E-08	386784502	-3,8858E-17	0,42698327		
S1F	42125659,8	105744322	42125659,8	105744322	-9,8302E+16	19,2	-108519013	2,3842E-08	4,8236E+16	
S2F	-557489 5,24	-162501358	-557489 5,24	-162501358	19,2	-1,7192E+17	2,0862E-08	-189791450	-11,2	8,4361E+16

*

Cor	Y1	Y2	X 1	X2	S1A	S2A	S1B	S2B	S1C	S2C
Y1	1									
Y2	0,68612609	1								
X 1	0,75286964	0,51101769	1							
X2	0,51101769	0,7340254	0,68612609	1						
S1A	-0,41771383	-0,21967453	-0,41771383	-0,21967453	1					
S2A	0,04180069	0,25526674	0,04180069	0,25526674	-1,3071E-16	1				
S1B	-0,41771383	-0,21967453	-0,41771383	-0,21967453	1	-9,072E-17	1			
S2B	0,04180069	0,25526674	0,04180069	0,25526674	1,9656E-16	1	2,3086E-16	1		
S1C	-0,41771383	-0,21967453	-0,41771383	-0,21967453	1	-1,4328E-16	1	1,6574E-16	1	
S2C	0,04180069	0,25526674	0,04180069	0,25526674	-1,4328E-16	1	-1,3259E-16	1	-1,4398E-16	1

*
*

Cor	Y1	Y2	X 1	X2	S1D	S2D	S1E	S2E	S1F	S2F
Y1	1									
Y2	0,68612609	1								
X 1	0,75286964	0,51101769	1							
X2	0,51101769	0,7340254	0,68612609	1						
S1D	-0,41771383	-0,21967453	-0,41771383	-0,21967453	1					
S2D	0,04180069	0,25526674	0,04180069	0,25526674	-9,6627E-17	1				
S1E	-0,41771383	-0,21967453	-0,41771383	-0,21967453	1	-1,2228E-16	1			
S2E	0,04180069	0,25526674	0,04180069	0,25526674	-1,2228E-16	1	-1,2035E-16	1		
S1F	0,41771383	0,21967453	0,41771383	0,21967453	-1	1,4769E-16	-1	1,6613E-16	1	
S2F	-0,04180069	-0,25526674	-0,04180069	-0,25526674	1,4769E-16	-1	1,4536E-16	-1	-1,7557E-16	1