Revue littéraire - Econometrics with Privacy Preservation

Thomas

February 2018

1 Définition de l'anonymisation

Il y a trois propriétés à vérifier pour certifier qu'un algorithme est respectueux de la confidentialité des données qu'il manipule :

1. Il doit bénéficier d'un large champs d'application
2. Les erreurs de calculs engendrés sont faibles et mesurables
3. Les temps de calculs correspondant demeurent modérés (au regard de la taille de l'échantillon de données)

Plusieurs méthodes et algorithmes connus pour préserver la confidentialité ne vérifient pourtant pas toutes ces propriétés :

- *Randomized response techniques*, Warner & Greenberg (1965, 1969), une technique simple reposant sur le questionnement indirect mais dont il est difficile de quantifier l'erreur (⇒ (2) non vérifié)
- *Differential Privacy*, Dwork (2006), difficile également de mesurer l'erreur de la méthode. (⇒ (2) non vérifié)
- Ajouter un bruit aléatoire (d'espérance nulle) aux données et travailler sur ses données en ajoutant toutes les données d'un (suffisamment grand) échantillon pour que les bruits se compensent et ainsi retrouver la moyenne originale. Cette méthode ne permet pas de déterminer directement la taille de l'échantillon nécessaire pour que l'erreur soit négligeable, et par ailleurs cela ne fonctionne que pour retrouver la moyenne de l'échantillon (⇒ (1) & (2) non vérifiés)
- Estimer la fonction de densité d'une V.A à partir de V.A iid échantillonnées perturbées avec erreurs (Fan, 1991). Cette méthode est assez simple, néanmoins elle ne permet pas de quantifier les erreurs et sa vitesse de convergence est relativement lente (⇒ (2) & (3) non vérifiés)

L'idée développée par cet article est basée sur un protocole de calcul multipartite simple et rapide permettant de calculer les corrélations et les moyennes en utilisant une distribution uniforme des données. Ce protocole a été présenté pour la première fois dans *Privacy Presering methods for sharing Financial Risk Exposures* (E.A Abbe, A.E Khandani & A.W Lo, 2012).
Cet algorithme garantie la confidentialité des données dans le cadre où les différentes parties prenantes sont **semi-honnêtes**.

2 Rappels et idée générale

2.1 Les types d'adversaires en cryptologie

Il existe deux types d'adversaires en cryptologies :

- **Semi-honnêtes** : Ces adversaires respectent le protocole établi, mais tentent cependant de récupérer des informations privées en essayant d'accéder aux étapes intermédiaires du protocole.
- **Malicieux** : Ces adversaires dévient du protocole afin d'essayer de briser la confidentialité. Ils sont donc à ce titre plus difficile à contrer).

L'algorithme proposé dans cet article est basé sur une approche semi-honnête afin de garantir la confidentialité du processus.

2.2 Présentations des entrées de l'algorithme

On définit m parties ($i = 1, \ldots, m$) ayant chacune d données (pour simplifier ici exclusivement numériques), constituant le vecteur :

$$x_i = (x_{i1}, \ldots, x_{id})^T \quad , m \in \mathbb{N},\ m \geqslant 3,\ d \in \mathbb{N}$$

Il est également supposé dans la suite de l'article que $x_1, \ldots, x_m \in [0, 1[^d$ (si ce n'est pas le cas en réalité, si les données sont bornées une simple normalisation est alors à réaliser).

Soient $x_1, \ldots, x_m$ les données sensibles des m parties et L fonctions $g()$ tel que : $g_k(x) \in [0, 1[, \forall x \in [0, 1[^d$ avec $k = 1, \ldots, L$.

On définit également la notion d'équidistribution invariante et singulière correspondant à une généralisation multidimensionnelle de la distribution uniforme comme le produit cartésien de collections d'hyperplans disjoints.[1]

2.3 Préservation de la confidentialité

On dit qu'un algorithme préserve la confidentialité en présence d'un organisme central si les informations obtenues par ce dernier ont une équidistribution invariante sur l'ensemble de définition et qui dépend seulement de la sortie de l'algorithme. Par ailleurs l'ensemble de définition est symétrique au regard des données de toutes les parties.
L'organisme central ne peut donc pas déduire davantage d'informations auprès des données fournies par les parties qu'elle ne peut en déduire de la sortie de l'algorithme.
Les données des différentes parties sont donc impossibles à distinguer par l'organisme central tant du point de vue de leur distribution que de leurs ensembles de définition.

3 L'algorithme de cryptage et de recouvrement

3.1 Le cryptage

1. Chaque partie calcule les L fonctions g: $g_1(x_i), \ldots, g_L(x_i)$ avec $i \in [\![1, m]\!]$
2. $\forall i, \hat{i} = 1, \ldots, m$ où $i \neq \hat{i}$, la partie i fournit à $\hat{i}$ L nombres aléatoires uniformément distribués sur $[0, m[$: $R_{i\hat{i}}(1), \ldots, R_{i\hat{i}}(L)$ iid.
3. Pour $i = 1, \ldots, m$ et $k = 1, \ldots, L$ la partie i ne révèle que $S_i(k)$ à l'organisme central tel que :

$$S_i(k) := \left\{ g_k(x_i) + \sum_{\hat{i} \neq i} (R_{\hat{i}i}(k) - R_{i\hat{i}}(k)) \right\} \bmod m$$

$S_i(k)$ pour $i = 1, \ldots, m$ et $= 1, \ldots, L$ est appelé **données cryptées** et représente toute l'information que l'organisme central peut obtenir.

[1] Pour plus d'information sur cette distribution, voir l'article

3.2 Le recouvrement

L'organisme central calcule pour $k = 1, \ldots, L$:

$$Q(k) := \left\{ \sum_{i=1}^{m} S_i(k) \right\} \bmod m$$

$Q(1), \ldots, Q(k)$ est appelé **information recouverte**.

Dans l'étape de recouvrement, l'organisme central peut recouvrir les informations suivantes à partir des données cryptées :

$$Q(1) = \sum_{i=1}^{m} g_1(x_i), \ldots, Q(L) = \sum_{i=1}^{m} g_L(x_i)$$

Puisque l'information recouverte dépend des fonctions g il est donc nécessaire d'apporter un soin particulier à leurs formulations ainsi qu'à leurs calculs afin de pouvoir procéder à l'inférence statistique souhaitée (régression linéaire, logistique, ...)

4 Pour résumer

4.1 Précision

Le recouvrement des informations s'effectue sans erreurs.

4.2 Efficacité

La complexité de l'algorithme est en $O(m^2 L)$, on peut également découper les m parties en J blocs de tailles M tel que la complexité par bloc soit de $O(m^2 L/J)$.

4.3 Flexibilité

Cet algorithme peut s'appliquer à plusieurs méthodes de statistiques inférentielles, notamment la régression linéaire, l'estimation par maximum de vraisemblance, la régression logistique, déterminer les distributions empiriques, calculs de quantiles empiriques,...

4.4 Confidentialité

La confidentialité des données est préservée dans un cadre semi-honnête, en présence d'un organisme central et même en cas de collusions majoritaire parmi les parties lors d'une cyber-attaque ayant réussi a récupérer les sorties de l'algorithme auprès de l'organisme central.

5 Un exemple concret : la régression linéaire

On considère un exemple classique de régression linéaire simple multivariée avec n variables tel que :

$$y_i = \beta_1 z_{i1} + \cdots + \beta_n z_{in} + \epsilon_i = Z_i^T \beta + \epsilon_i$$

Soit $y = Z\beta + \epsilon$, où :

$$Z_i = (z_{i1}, \ldots, z_{in})^T, \beta = (\beta_1, \ldots, \beta_n)^T, \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \text{ et } y = (y_1, \ldots, y_n)^T$$

Ainsi $Z = \begin{bmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & \ddots & \vdots \\ z_{m1} & \cdots & z_{mn} \end{bmatrix}$, une matrice $m \times n$.

On suppose également que $z_{ij} \in [0,1[, \forall i = 1,\dots,m$ et $\forall j = 1,\dots,n$, sinon normaliser simplement les données si les z_{ij} et les y_i sont bornés.
En définissant les données tel que :

$$x_i = (x_{i1},\dots,x_{in},x_{in+1})^T := (z_{i1},\dots,z_{in},y_i)^T \quad (d = n+1)$$

L'estimateur des moindres carrés ordinaires des coefficients de la régression linéaire, noté $\hat{\beta}$ est donné par :

$$\hat{\beta} = (Z^T Z)^{-1} Z^T y$$

et l'estimateur non biaisé de la variance de ϵ_i vaut :

$$\hat{S^2} = \frac{(y - Z\hat{\beta})^T (y - Z\hat{\beta})}{m-n}$$

que l'on peut également écrire :

$$\begin{aligned}
\hat{S^2} =& \frac{1}{m-n}[(y^T - \hat{\beta}^T Z^T)(y - Z\hat{\beta})] && (1)\\
=& \frac{1}{m-n}[y^T y - y^T Z\hat{\beta} - \hat{\beta} Z^T y + \hat{\beta^T} Z^T Z\hat{\beta}] && (2)\\
=& \frac{1}{m-n}[y^T y - y^T Z(Z^T Z)^{-1} Z^T y - y^T Z(Z^T Z)^{-1} Z^T y + y^T Z \underbrace{(Z^T Z)^{-1} Z^T Z}_{=I} (Z^T Z)^{-1} Z^T y] && (3)\\
=& \frac{1}{m-n}[y^T y - y^T Z(Z^T Z)^{-1} Z^T y] = \boxed{\frac{y^T y - (Z^T y)^T (Z^T Z)^{-1} Z^T y}{m-n}} && (4)
\end{aligned}$$

Dans notre cas de figure les données étant privées et sensibles, pour appliquer une régression linéaire sur celles-ci tout en conservant leurs confidentialités on peut utiliser notre algorithme en déterminant simplement les L fonctions g nécessaires afin de calculer les deux estimateurs précédents.

En prenant $L = d(d+1)/2 = (n+1)(n+2)/2$ tel que $g_{kl}(x) = x(k)x(l)$ pour $k,l = 1,\dots,n+1$ et $k \leqslant l$ avec $x = (x(1),\dots,x(n+1))^T$. On peut définir $\hat{\beta}_e$ et ${\hat{S^2}}_e$ tels que :

$$\begin{aligned}
\hat{\beta}_e :=& (Z^T Z)_e^{-1} (Z^T y)_e && (5)\\
{\hat{S^2}}_e :=& \frac{(y^T t)_e - (Z^T y)_e^T (Z^T Z)_e^{-1} (Z^T y)_e}{m-n} && (6)
\end{aligned}$$

On a alors :

$$\begin{cases}
(y^T y)_e := & Q(n+1,n+1) = \left\{ \sum_{i=1}^{m} S_i(n+1,n+1) \right\} \bmod m \\
(Z^T Z)_e := & (Z^T Z)_e(k,l) = Q(k,l) = \left\{ \sum_{i=1}^{m} S_i(k,l) \right\} \bmod m \\
(Z^T y)_e(k) := & Q(k,n+1) = \left\{ \sum_{i=1}^{m} S_i(k,n+1) \right\} \bmod m
\end{cases} \quad (7)$$

On remarque en effet qu'en appliquant la procédure définie précédemment on a :

$$(Z^T Z)(k,l) = \sum_{i=1}^{m} x_{ik} x_{il} \quad \text{et} \quad (Z^T y)(k) = \sum_{i=1}^{m} x_{ik} y_i$$

$$\begin{cases} Q(n+1,n+1) = & \sum_{i=1}^{m} g_{n+1,n+1}(x_i) = \sum_{i=1}^{m} x_{i,n+1}^2 = y^T y \\ Q(k,l) = & \sum_{i=1}^{m} g_{k,l}(x_i) = \sum_{i=1}^{m} x_{i,k} x_{i,l} = (Z^T Z)(k,l) \\ Q(k,n+1) = & \sum_{i=1}^{m} g_{k,n+1}(x_i) = \sum_{i=1}^{m} x_{i,k} y_i = (Z^T y)(k) \end{cases} \tag{8}$$

avec $k, l = 1, \ldots, n$ où $k < l$ et $(Z^T Z)(l,k) = (Z^T Z)(k,l)$.

6 Transposition de la méthode à un cadre Assureur-Prestataire externe classique

On se place dans le cadre où une entreprise a recours à un prestataire externe afin d'externaliser certains calculs ou méthodes statistiques (*Cloud computing*).
On peut alors imaginer un aménagement de la méthode définie au sein de cet article en considérant chaque individu de la base de données comme étant une unique donnée d'une des m parties de la méthode. Où m serait donc également le nombre de lignes dans le jeu de données.

References

[1] N. Cai & S. Kou, *Econometrics with Privacy Preservation.* , 2017.