Données manquantes ou censurées : principes de base.

Bernard Delyon *

10 juillet 2013

Ce texte a pour but de donner un aperçu des méthodes existantes et des principes sous-jacents. Une étude assez complète sur le sujet avec de nombreux exemples se trouve dans [2]. Voir aussi [1, 3, 4] pour des exemples d'application.

1 Exemples de données manquantes ; idées générales

Voici quelques exemples de données incomplètes :

- Une étude est basée sur le suivi de l'état de santé d'individus sur plusieurs années ; certains individus peuvent disparaître de la circulation au milieu de l'étude. On a donc un tableau (x_{ij}), i =individu et j =année, incomplet.
- **Données censurées :** une variable z est non-mesurée quand la variable w dépasse une certaine valeur. Exemple : on cherche à estimer la loi de $z =$ « durée de vie d'une ampoule » à partir d'un test de 10 000 heures sur un échantillon. La durée de vie des ampoules encore en état de marche à la fin de l'expérience est inconnue, mais on sait qu'elle est supérieure à la durée de l'expérience ; elles apportent donc de l'information sur la loi de z bien que z ne soit pas réellement observée ; ici $z = w$.

On considérera les situations-type

1. **Tableau incomplet.** Que dire d'un tableau individus/variables $X = (x_{ij})$ incomplet ?
2. **Séries temporelles.** On veut identifier un modèle AR mais la série temporelle observée z_n possède des échantillons manquants.
3. **Données exponentielles censurées.** Les variables z_n sont supposées avoir une distribution exponentielle de paramètre β inconnu. Au lieu d'observer z_n on observe $\check{z}_n = \min(z_n, c)$ où c est une constante fixe. Notons qu'ici le processus de perte dépend de z de manière déterministe.
4. **Régression.** On cherche à faire la régression linéaire de x sur y, i.e. basée sur le modèle régression $y = X\beta+$bruit, mais le tableau individus/variables $X = (x_{ij})$ est incomplet. Même s'il y a une variable manquante pour chaque individu, l'estimation de β reste possible. Il peut s'agir aussi d'un autre type de régression comme la régression logistique.

Il existe trois types de méthodes employées pour étendre les algorithmes d'estimation habituels d'un paramètre β aux données incomplètes :

1. **Élimination :** éliminer les individus incomplets de l'étude, ce qui augmentera la variance d'estimation et risque de biaiser le résultat (cf les données censurées).
2. **Imputation :** remplacer les données manquantes par une valeur, éventuellement estimée à partir des données observées ; cette méthode est à éviter s'il y a beaucoup de données manquantes, car elle va biaiser les estimateurs (par exemple diminuer les variances).

*IRMAR, Université Rennes-I, Campus de Beaulieu, 35042 Rennes cedex, France ; delyon@maths.univ-rennes1.fr

3. Plus rigoureusement, se ramener à un problème paramétrique général :
   (a) **Modélisation** de la loi des données manquantes. Pour cela on crée un modèle augmenté, qui prend en compte les données manquantes. Ce modèle sera généralement paramétré par (β, γ) où β est le paramètre du problème avec données complètes, et γ est connu ou inconnu.
   Dans l'exemple 4, les lignes de X seront souvent considérées i.i.d $\mathcal{N}(\mu, R)$, et $\gamma = (\mu, R)$. Dans l'exemple 3, $\gamma = c$.
   (b) **Estimation** de $\theta = (\beta, \gamma)$. L'estimation β et γ ne pourra pas se faire indépendamment, à moins de faire des approximations. Les formules de vraisemblance étant souvent trop compliquées (*même si γ est connu*), on est souvent contraint de mettre en œuvre des algorithmes itératifs, faisant éventuellement intervenir des méthodes de Monte-Carlo.

Notations. $z = (z_o, z_m)$ désignera dans la suite l'ensemble des données observées et manquantes.

2 Méthode générale

On postule l'existence d'un paramètre θ tel que les données complètes z suivent la loi P_θ. Une méthode naturelle serait d'estimer à la fois le paramètre et les données manquantes au maximum de vraisemblance :

$$\max_{\theta, z_m} P_\theta(z_o, z_m). \tag{1}$$

On va voir qu'il vaut bien mieux maximiser en θ la probabilité des données observées :

$$p_\theta(z_o) = \int p_\theta(z_o, z_m) dz_m. \tag{2}$$

Exemple : données exponentielles censurées. On reprend l'exemple 3. On suppose c connu et donc $\theta = \beta$. Comme on a directement la probabilité de la variable $\check{z}_i$:

$$p(\check{z}) = \theta e^{-\theta \check{z}} \quad \text{si} \quad \check{z} \neq c$$
$$P(\check{z} = c) = \int_{z>c} \theta e^{-\theta z} dz = e^{-\theta c} \quad \text{sinon}$$

qui est un mélange d'une loi continue et d'une discrète, on obtient la maximisation la log-vraisemblance (les variables censurées sont les m_0 dernières) :

$$\hat{\theta} = \arg\max_\theta \sum_{i=1}^{n-m_0} (\log(\theta) - \theta z_i) - m_0 \theta c = (n - m_0) \left(\sum_{i=1}^{n} \check{z}_i \right)^{-1}.$$

Suite : comparaison avec (1). On vérifie facilement que la formule (1) conduit à

$$\hat{\theta}^{-1} = \frac{1}{n} \left(m_0 c + \sum_{i=1}^{n-m_0} z_i \right)$$

C'est la moyenne empirique où l'on a imputé la valeur c aux données manquantes. C'est une valeur plus faible que la précédente car on a imputé la valeur la plus vraisemblable, c'est-à-dire la plus petite, c, au lieu de prendre en compte la possibilité que ces données ont d'avoir toute valeur supérieure. Noter que si presque toutes les données sont tronquées, cet estimateur donne c, ce qui est peu vraisemblable, car si on avait $E[z] = c$ seule environ la moitié des données aurait dépassé c.

L'algorithme EM. La maximisation de l'intégrale (2) est assez difficile à mettre en œuvre. Considérons la fonction

$$Q(\theta, \theta_0) = \int \log p_\theta(z_o, z_m) p_{\theta_0}(z_m|z_o) dz_m = E_{\theta_0}[\log p_\theta(z)|z_o] \tag{3}$$

C'est la log-vraisemblance moyenne sachant les variables observées sous l'hypothèse que les données manquantes ont été tirées sous θ_0. Il se trouve que le maximum $\hat{\theta}$ de (2) satisfait $\hat{\theta} = \arg\max_\theta Q(\theta, \hat{\theta})$ ce qui, sans être évident, est assez logique. D'où l'algorithme itératif :

$$\theta_n = \arg\max_\theta Q(\theta, \theta_{n-1}).$$

On montre que cet algorithme converge de manière assez générale vers $\hat{\theta}$. L'étape E (expectation) de l'algorithme est le calcul de $Q(., \theta_n)$ et l'étape M est la maximisation. *Expérimentalement cet algorithme a une vitesse de convergence très lente, surtout si l'initialisation n'est pas soignée.* Si l'étape E est difficle à réaliser, on peut la remplacer par une **simulation** :

$$\hat{\theta}_n \quad = \quad \arg\max_\theta \sum_{k=1}^{K} \log(p_\theta(z_o, z_{mk})) \tag{4}$$

où les z_{mk} sont générés sous la loi conditionnelle $p_{\hat{\theta}_{k-1}}(z_m|z_o)$. Noter le caractère intuitif de l'algorithme et que pour avoir convergence, K doit tendre vers l'infini avec n. L'algorithme **SEM** fait $K = 1$ mais moyenne les $\hat{\theta}_n$ obtenus. Une autre variante est l'**algorithme stochastique** :

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \gamma_{n+1} \frac{d \log p_{\hat{\theta}_n}(z_o, z_{m,n})}{d\theta} \tag{5}$$

où la variable $z_{m,n}$ est simulée selon la loi $p_{\hat{\theta}_n}(z_m|z_o)$, et γ_n est un gain, typiquement $\gamma_n = 1/n$. **L'imputation** multiple est une autre manière, très approximative, et non-itérative, de résoudre ce problème à partir d'une loi pilote $p_0(z_m|z_o)$ pour les données manquantes (cette loi appartiendra souvent à la famille p_θ). On simule sous p_0 plusieurs données $z_{m1}, \ldots z_{mK}$, puis l'on pose

$$\hat{\theta}_{imp1} \quad = \quad \frac{1}{K} \sum \hat{\theta}_k, \quad \hat{\theta}_k = \arg\max_\theta(\log(p_\theta(z_o, z_{mk}))$$

ce qui fait une itération de SEM. Une variante est $\hat{\theta}_{imp2} = \hat{\theta}_1$ donné par une itération de (4).

Une généralisation. Soit un algorithme d'estimation qui sur les données complètes s'écrit, $H(\hat{\theta}, z) = 0$, son estimation sur données incomplètes sera alors $\int H(\hat{\theta}, z) p_{\hat{\theta}}(z_m|z_o) dz_m = 0$, et les méthodes précédentes se généralisent sans problème.

3 Exemples

3.1 Le tableau incomplet.

On reprend l'exemple 1. La manière la plus simple est de postuler une distribution gaussienne pour les variables du tableau. Il s'agit donc de retrouver les paramètres $\theta = (\mu, R)$ de cette distribution. L'algorithme EM donne les deux étapes suivantes pour chaque itération [1] :

- Exprimer pour chaque individu la loi (gaussienne) de x_m sachant x_o :
$$E[x_m|x_o] = \mu_m + R_{m,o} R_{o,o}^{-1}(x_o - \mu_o), \quad Cov(x_m|x_o) = R_{m,m} - R_{m,o} R_{o,o}^{-1} R_{o,m}$$

1. On note $R_{m,o}$ la sous-matrice de R obtenue en sélectionnant les paires d'indices dont le premier est manquant dans x et le second présent ; de même pour $R_{m,m}$ et $R_{o,o}$. Cette sous-matrice *dépend de l'individu.*

- Réestimer β par les moyennes et variances empiriques conditionnelles aux données observées :

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_i, \quad R = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i-\mu)^T(\hat{x}_i-\mu) + Cov(x_i|x_o)$$

où $\hat{x}_i$ vaut x_o pour les variables observées et $E[x_m|x_o]$ pour les variables manquantes, et $Cov(x_i|x_o)$ est la matrice dont les seuls coefficients non-nuls sont ceux dont les deux indices correspondent à des données manquantes, formée des coefficients de $Cov(x_{m,i}|x_o)$.

3.2 La régression

Revenons à l'exemple 4. S'il s'agit d'une régression linéaire et que l'on accepte l'hypothèse de distribution gaussienne, la méthode plus directe est d'appliquer l'algorithme du § 3.1 au tableau (y, X) pour en déduire directement $E[y|X] = \mu_y + (X-\mu_x)R_{xx}^{-1}R_{xy}$, ce qui donne $\hat{\beta} = R_{xx}^{-1}R_{xy}$ et pour le facteur de la constante $\hat{\beta}_0 = \mu_y - \mu_x R_{xx}^{-1}R_{xy}$. Ceci a l'avantage de fonctionner même si y est vectoriel incomplet.

Si la régression n'est pas linéaire, on va considérer le cas où y est complet et X suit une loi gaussienne. Ici $\theta = (\gamma, \beta)$ et $z = (z_o, z_m) = ((y, X_o), X_m)$, où γ contient les paramètres de la loi gaussienne pour la matrice X, β est le vecteur de régression et X_m les données manquantes de X. L'approche rigoureuse est de tout estimer d'un bloc, car y apportant de l'information sur X, on ne devrait pas estimer γ à partir de X_o seulement. Notons que

$$p_\theta(z_o, z_m) = p_\beta(y|X_m, X_o)p_\gamma(X_m|X_o)p_\gamma(X_o).$$

On voit que l'algorithme EM est difficile à mettre en œuvre, tandis que la réalisation de l'algorithme (4) ou (5) ne pose pas trop de problème car on sait simuler les données manquantes sous $p_\gamma(X_m|X_o)$ et l'utilisation d'une méthode de rejet permet d'obtenir alors une réalisation de z_m sous la loi $p_\theta(z_o, z_m)$ (comme ici tout est gaussien, on peut même simuler directement z_m).

Une autre approche est de considérer que $\gamma = (\mu, R)$ a été correctement estimé simplement grâce à X_o et à l'algorithme du § 3.1, c.-à-d. que l'apport d'information de y sur γ est très faible ; il est désormais considéré connu et donc maintenant $\theta = \beta$. On peut alors appliquer la méthode d'imputation avec $p_0 = p(X_m|X_o)$ (ce qui revient à considérer que $p(X_m|X_o, y) \simeq p(X_m|X_o)$). On simule alors K exemplaires du tableau gaussien sous la loi $p(X_m|X_o)$ et les deux estimées sont

$$\hat{\beta}_{imp1} = \frac{1}{K}\sum_{k=1}^{K}(X_k^TX_k)^{-1}X_k y, \quad \hat{\beta}_{imp2} = \left(\frac{1}{K}\sum_{k=1}^{K}X_k^TX_k\right)^{-1}\left(\frac{1}{K}\sum_{k=1}^{K}X_k^T\right)y.$$

Références

[1] C.C. Clogg, D.B. Rubin, N. Schenker, B. Schultz, L. Weidman, "Multiple Imputation of Industry and Occupation codes in Census Public-use Samples Using Bayesian Logistic Regression", *JASA*, 86, 413, 68-78, 1991.

[2] J.A. Little, D.B. Rubin, *Statistical analysis with missing data*, Wiley, 1987.

[3] E. Raghunathan, D.S. Siscovick, "A multiple-imputation analysis of a case-control study..." *Appl. Statist.*, 45 , 3, 335-352, 1996.

[4] D.B. Rubin, "Multiple imputation after 18+ years", *JASA* 91, No 434, 473-489, 1996.

[5] P. Zhang "Multiple Imputation : Theory and Method", *Internat. Statist. Rev*, 71, No 3 (2003), 581-592.