Un modèle à variables latentes pour le traitement de données hybrides issues de comparaison interlaboratoires

Séverine Demeyer, Nicolas Fischer, Gilbert Saporta

UN MODÈLE À VARIABLES LATENTES POUR LE TRAITEMENT DE DONNÉES HYBRIDES ISSUES DE COMPARAISONS INTERLABORATOIRES

Séverine Demeyer[1,2], Nicolas Fischer[1] & Gilbert Saporta[2]

[1]Severine.Demeyer@lne.fr , Nicolas.Fischer@lne.fr
Laboratoire National de Métrologie et d'Essais,
29 avenue Roger Hennequin, 78197 Trappes Cedex
[2] Gilbert.Saporta@cnam.fr
CEDRIC-CNAM, 292 rue Saint Martin 75141 Paris Cedex 3

Résumé

On propose une nouvelle approche dans l'estimation des biais, c'est à dire des écarts entre résultats de laboratoires et la valeur vraie inconnue d'une grandeur, lors de comparaisons interlaboratoires. Pour cela on combine un modèle de régression hiérarchique des biais et un modèle à équations structurelles pour les caractéristiques des laboratoires obtenues à l'aide d'un questionnaire. L'intégration de ces deux modèles en un seul est possible si des variables latentes expliquent simultanément les résultats de mesures et les réponses au questionnaire. La méthodologie est appliquée à des comparaisons interlaboratoires pour des polluants de l'eau.

Mots-clés : modèle hiérarchique, variables latentes, équations structurelles, variables communes, avis d'experts, incertitude

Abstract

A new approach is developed to estimate biases, which are discrepancies between the results of laboratories and the true unknown value of a quantity, in interlaboratory comparisons. Quantitative results as well as qualitative information from a survey (hybrid input data) are combined within a model integrating both a hierarchical regression model of biases and a structural equation model of the answers. The complete model holds when some latent variables are supposed to explain simultaneously the results and the answers. The methodology is applied to interlaboratory comparisons for water pollutants.

Keywords : hierarchical model, latent variables, structural equations, shared variables, experts knowledge, uncertainty

1 Introduction

1.1 CADRE DES COMPARAISONS INTERLABORATOIRES

Pour garantir la qualité de leurs mesures, les laboratoires d'analyse participent régulièrement à des comparaisons interlaboratoires qui leur permettent de se comparer en mesurant les mêmes grandeurs. Un des critères de performance des laboratoires est leur biais de mesure, qui est en métrologie l'écart du résultat du laboratoire à la valeur vraie de la grandeur, inconnue.

Précisément, les comparaisons interlaboratoires ont pour but d'attribuer une valeur de consensus et une incertitude de mesure à une grandeur à partir des résultats de mesure des laboratoires.
En métrologie, les termes valeur vraie, incertitude de mesure et biais de mesure sont précisément définis dans un guide international, le VIM (2007).
Des normes proposent des méthodes pour estimer les biais de mesures. Cependant, elles ne donnent pas d'indications pour expliquer la raison de ces biais de mesure ni pour prendre en compte des informations qui seraient disponibles sur la raison de ces biais.

1.2 LES MÉTHODES NORMALISÉES

Le calcul de la valeur de consensus et de son incertitude associée sont réalisées dans le cadre de la norme ISO 13528 (2005). On considère les points de la norme où les laboratoires ne sont pas contraints de fournir l'incertitude associée à leur résultat de mesure.
La valeur de consensus peut alors être estimée par la moyenne robuste des résultats fournis par tous les laboratoires participant calculée à l'aide de l'algorithme A : en notant x^* la moyenne robuste recherchée et s^* l'écart-type robuste recherché, on initialise l'algorithme en posant $x^* = \text{médiane des } x_i$ et $s^* = 1,483 \times \text{médiane des } \left|x_i - x^*\right|$.

On met à jour les estimations de x^* et s^* jusqu'à convergence de l'algorithme en calculant :

$$x^* = \frac{1}{p}\sum_{i=1}^{p} x_i^* \qquad s^* = 1,134\sqrt{\frac{\sum_{i=1}^{p}\left(x_i^* - x^*\right)}{p-1}}$$

où $x_i^* = \begin{cases} x^* - \delta & \text{si } x_i < x^* - \delta \\ x^* + \delta & \text{si } x_i > x^* + \delta \\ x_i & \text{sinon} \end{cases}$, $\delta = 1,5s^*$ et p est le nombre de laboratoires participant.

La norme ISO 13528 autorise d'autres méthodes de calcul à la place de l'Algorithme A, dans la mesure où elles sont fondées sur une **base statistique éprouvée** et où le rapport décrit la méthode utilisée.

L'incertitude associée est estimée par $u_x = 1,25 \times \frac{s^*}{\sqrt{p}}$ dans le cas où la valeur de consensus est calculée avec l'algorithme A.
Le biais de mesure d'un laboratoire est estimé par l'écart du résultat rendu par le laboratoire à la valeur de consensus.

1.3 PROPOSER UNE NOUVELLE APPROCHE

Notre approche consiste à identifier de l'information auxiliaire sur les mesures et sur les laboratoires et à la combiner aux résultats de mesure fournis par les laboratoires. Plus précisément, l'objectif du travail présenté ici est de combiner les informations caractéristiques de la participation d'un laboratoire à une comparaison interlaboratoires de natures quantitative et qualitative, au sein d'un modèle à variables latentes destiné à expliquer les biais de mesure des laboratoires lors des comparaisons. La méthodologie est généralisable aux autres comparaisons interlaboratoires.
La validation du modèle sera réalisée en comparant la sortie du modèle à la valeur métrologique et son incertitude associée, fournies par une procédure ou méthode primaire (cf. VIM) lors d'une comparaison.

2 Choix de l'application

L'application choisie est celle de la mesure de micropolluants organiques dans l'eau douce par des laboratoires spécialisés dans l'analyse des eaux lors de leur participation aux comparaisons interlaboratoires organisées par le Bureau InterProfessionnel d'Etudes Analytiques (BIPEA).

Chaque année le BIPEA organise 3 comparaisons interlaboratoires impliquant une trentaine de ces laboratoires, pour lesquelles les échantillons sont fabriqués à 3 niveaux de concentration (niveau bas _ niveau moyen _ niveau haut), à chaque comparaison correspondant un unique niveau. Les comparaisons qui ont même niveau ont des valeurs vraies différentes car les micropolluants ne sont pas dosés précisément par le fabricant des échantillons.

Notre choix a été motivé par la particularité de ces comparaisons pour lesquelles chaque laboratoire renvoie **un unique résultat sans incertitude associée**. La parcimonie des données favorise en effet une approche plus générale dans la modélisation.

3 Recueil de l'information auxiliaire

3.1 RÉALISATION D'UN QUESTIONNAIRE

L'information auxiliaire dont nous avons besoin n'est pas recueillie systématiquement avec les résultats des laboratoires. Nous avons donc opté pour la réalisation d'un questionnaire à destination des laboratoires, portant sur la participation de ces laboratoires aux 9 dernières comparaisons. Le recueil des réponses est en cours.

3.2 EXPLOITATION DU QUESTIONNAIRE

L'exploitation du questionnaire nécessite l'interprétation des questions en terme de contribution au biais. En raison du nombre important de questions par rapport au nombre de laboratoires nous avons demandé à un collège d'experts de discuter le regroupement des questions en thèmes contribuant au biais. Les dimensions suivantes ont été identifiées : CONTEXTE DU LABORATOIRE, PRÉPARATION DES ÉCHANTILLONS, MÉTHODE ANALYTIQUE, ETALONNAGE, EFFET DE MATRICE, VALIDATION DES RÉSULTATS. Avec les notations et le vocabulaire du paragraphe 4.2 , le premier bloc est défini par :

Variables latentes	Variables manifestes	
QUESTIONS GÉNÉRALES (Z_1)	y_{11}	Le laboratoire est-il accrédité ISO 17025 ?
	.	Est-il accrédité pour le programme 100-1 (physico-chimie) ?
	.	Le laboratoire utilise-t-il les méthodes normalisées suivantes ?
	.	Comment estimez-vous l'incertitude de mesure ?
	y_{15}	Rendez-vous l'incertitude à vos clients ?

4 Modélisation

4.1 HYPOTHÈSES

Hypothèse1. : Le biais de mesure d'un laboratoire dépend du niveau de concentration de l'échantillon.

Hypothèse2. : Chaque bloc de questions est unidimensionnel, résumé par une unique variable latente.
Hypothèse3. : Les variables latentes expliquent simultanément les réponses des laboratoires au questionnaire et les biais de mesure des laboratoires.

Remarque :
Les hypothèses 1 et 3 ont été validées par les métrologues. Seule l'hypothèse d'unidimensionnalité des blocs n'est pas *a priori* validée et nécessitera d'être testée dès l'obtention des données. Cependant cette hypothèse est cruciale. En effet, de cette hypothèse dépend la modélisation des réponses au questionnaire proposée au paragraphe 4.5.

4.2 NOTATIONS

Données quantitatives : X_{ijh} modélise le résultat de mesure fourni par le laboratoire j lors de la comparaison i du niveau h.
Données qualitatives : y_{jkl} est la variable manifeste qui modélise la réponse du laboratoire j à la question l du groupe k.
Variables latentes : Z_k est la variable latente associée au bloc *k*
Paramètres à estimer : μ_{ih} modélise la valeur vraie de la comparaison i du niveau h et γ_{jh} modélise le biais du laboratoire j dans le niveau h, variables continues.

4.3 MODÉLISATION DES RÉSULTATS DE MESURE DES LABORATOIRES

D'après l'hypothèse 1 et la définition du biais de mesure, on modélise le résultat de mesure en fonction de la valeur vraie de la comparaison et du biais du laboratoire par le modèle mixte suivant :

$$X_{ijh} = \mu_{ih} + \gamma_{jh} + \varepsilon_{ijh}$$

où les erreurs de mesures ε_{ijh} sont supposées indépendantes.

4.4 MODÉLISATION HIÉRARCHIQUE DES BIAIS DE MESURE

D'après l'hypothèse 1, on modélise les biais de mesure par un modèle de régression à coefficients variables où le coefficient de régression de chaque variable latente est supposé dépendre du niveau de concentration :

$$\gamma_{jh} = \sum_{k=1}^{K} \beta_{kh} Z_{kj} + \eta_{jh}$$

Le paramètre β_{kh} quantifie l'influence de la variable latente *k* sur le biais à un niveau *h*.

4.5 MODÉLISATION DES RÉPONSES AU QUESTIONNAIRE PAR UN MODÈLE À ÉQUATIONS STRUCTURELLES

4.5.1 Modèle de mesure

Puisque les réponses y_{jkl} au questionnaire sont de nature qualitative et que le nombre de modalités diffère pour chaque question, on modélise les probabilités des modalités de chaque question par un modèle logistique multinomial en suivant Skrondal (2004).
En notant $\pi_{jklm} = \mathrm{P}\left(y_{jkl} = m\right)$ la probabilité que le laboratoire j choisisse la modalité m à la question kl et M_{kl} le nombre de modalités de la question kl, on a donc

$$\pi_{jklm} = \frac{\exp\left(\alpha_{klm} Z_{kj}\right)}{\sum_{m'=1}^{M_{kl}} \exp\left(\alpha_{klm'} Z_{kj}\right)} \text{ avec } \sum_{m} \pi_{jklm} = 1.$$

En posant $\alpha_{klM_{kl}} = 0$ on peut interpréter le coefficient de régression α_{klm} comme l'effet de la variable latente Z_k sur la probabilité pour un laboratoire de choisir la modalité m.

4.5.2 Modèle structurel

Les corrélations entre variables latentes sont modélisées par :

$$Z_k = \sum_{k'=1}^{K} \delta_{kk'} \omega_{kk'} Z_{k'} + \tau_k$$

où $\delta_{kk'} = \begin{cases} 1 \text{ si } k \neq k' \text{ et si il existe un lien entre } Z_k \text{ et } Z_{k'} \\ 0 \text{ sinon} \end{cases}$

Les poids $\omega_{kk'}$ sont les coefficients de régression de $Z_{k'}$ dans la régression multiple de Z_k sur l'ensemble des $Z_{k'}$ qui lui sont liées.

4.6 INTÉGRATION AU SEIN D'UN MÊME MODÈLE

L'hypothèse 3 se traduit par l'existence de variables aléatoires communes à l'explication des biais et des réponses au questionnaire, ces variables étant les variables latentes. Cette hypothèse nous permet de regrouper les équations des paragraphes précédents selon le modèle hiérarchique :

$$X_{ijh} = \mu_{ih} + \gamma_{jh} + \varepsilon_{ijh} \qquad \text{(niveau 1)}$$

$$\gamma_{jh} = \sum_{k=1}^{K} \beta_{kh} Z_{kj} + \eta_{jh} \qquad \text{(niveau 2)}$$

$$\pi_{jklm} = \frac{\exp\left(\alpha_{klm} Z_{kj}\right)}{\sum_{m'=1}^{M_{kl}} \exp\left(\alpha_{klm'} Z_{kj}\right)} \qquad \text{(niveau 3)}$$

$$\mathrm{Z}_k = \sum_{k'=1}^{K} \omega_{kk'} Z_{k'} + \tau_k \qquad \text{(niveau 4)}$$

4.7 PARAMÈTRES D'INTÉRÊT POUR LA PROBLÉMATIQUE ÉNONCÉE

A partir des notations précédentes, les paramètres directement importants pour les laboratoires sont μ_{ih} (auxquels sont liées les estimations des valeurs de consensus et de leurs incertitudes associées), γ_{jh} et β_{kh}. On s'intéresse aussi aux paramètres associés aux modalités et aux questions qui contribuent le plus aux biais de mesure, qui devront apparaître en sortie du modèle.

5 Résolution du modèle par une approche bayésienne

L'hypothèse de corrélation entre variables latentes modélisée au niveau 4 détermine la structure de corrélation des coefficients de régression au niveau 2. L'hypothèse 3 permet donc de n'utiliser qu'une fois les informations apportées par l'hypothèse de corrélation.

Pour cette raison une approche dans un cadre bayésien nous semble particulièrement appropriée pour notre cas d'étude. En effet la distribution *a posteriori* des paramètres intègre en elle seule toute l'information disponible dans le sens où elle est calculée à partir de toutes les données d'entrée quelle que soit leur nature et en prenant en compte les liens entre tous les paramètres et les variables latentes. L'idée est alors de combiner les deux approches bayésiennes suivantes :

1- Régression bayésienne avec variables latentes
On s'inspire des travaux de Chen *et al* (2007) qui proposent un modèle de simulation par échantillonnage de Gibbs dans le cadre de modèles de régression où les variables expliquées et les variables explicatives dépendent des mêmes variables latentes sous hypothèses de normalité.

2- Approche bayésienne des modèles à équations structurelles
Lee (2007) a proposé la résolution de modèles à équations structurelles à réponses dichotomiques dans un cadre bayésien.

Le modèle sera estimé par simulation au moyen de méthodes Monte Carlo par Chaînes de Markov implémentées sous Winbugs et R. On s'aidera des codes fournis par Gelman et Hill (2007) pour le calcul des distributions conditionnelles dans la partie hiérarchique du modèle et de Albert (2007) pour des exemples dans un cadre plus général.

6 Conclusion et perspectives

Le modèle présenté considère implicitement que les réponses des laboratoires au questionnaire ne dépendent pas de la comparaison. Ce n'est cependant pas toujours le cas et une prochaine version du modèle prendra en compte ce point.
Plus généralement, cette approche peut être non seulement étendue à d'autres comparaisons interlaboratoires mais aussi à d'autres domaines de l'industrie où les données d'entrée du modèle sont hybrides (quantitatives et qualitatives) (cf. Chen (2007)).

Bibliographie

Albert J., 2007, *Bayesian Computation with R*, Springer, Berlin
Chen H., Bakshi B.R. et Goel P.K., 2007, Bayesian latent variable regression via Gibbs sampling : methodology and practical aspects, *Journal of Chemometrics* 21: 578-591
Chen H., 2007, *Sampling-based bayesian latent variable regression methods with applications in process engineering*, Ph.D. thesis, The Ohio State University
Gelman A. et Hill J., *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University Press, New-York
Lee S.-Y., 2007, *Structural Equation Modeling : A Bayesian Approach*, Wiley, New-York
Skrondal A. et Rabe-Hesketh S., 2004, *Generalized latent variable modeling : Multilevel, longitudinal and structural equation models*, Chapman & Hall/CRC, London
ISO 13528, *Méthodes statistiques utilisées dans les essais d'aptitude par comparaisons interlaboratoires*, International Organization for Standardisation, Geneva
ISO/IEC GUIDE 99, 2007: *Vocabulaire International de Métrologie – Concepts fondamentaux et généraux et termes associés (VIM)*, International Organization for Standardisation, Geneva