le cnam

CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS

École Doctorale EDITE
Laboratoire CEDRIC

T H È S E

pour obtenir le grade de

Docteur du Conservatoire National des Arts et Métiers
Spécialité INFORMATIQUE ET STATISTIQUE

Présentée par

Séverine DEMEYER

Approche bayésienne de l'évaluation de l'incertitude de mesure : application aux comparaisons interlaboratoires

sous la direction du Pr. Gilbert **SAPORTA**

Soutenue publiquement le **4 mars 2011** devant le jury composé de

Pr. Vincenzo **ESPOSITO VINZI**	ESSEC	Rapporteur
Dr. Nicolas **FISCHER**	LNE	Co-directeur
Pr. Jean-Louis **FOULLEY**	INRA	Examinateur
Pr. Marc **HIMBERT**	CNAM	Examinateur
Pr. Jean-Michel **MARIN**	Université Montpellier II	Rapporteur
Pr. Eric **PARENT**	AgroParisTech	Examinateur
Dr. Antonio **POSSOLO**	NIST	Examinateur
Pr. Gilbert **SAPORTA**	CNAM	Directeur

Remerciements

Mes premières pensées vont à Gilbert Saporta pour le remercier d'avoir accepté d'encadrer ma thèse puis pour sa patience, sa confiance et sa disponibilité tout au long de ces trois années. J'ai énormément appris à ses côtés. Grâce à lui, j'ai pu vivre une thèse d'exception, avec des événements mémorables, entres autres mon baptème de l'air.

Un grand merci à Nicolas Fischer, qui a vécu cette aventure au plus près, pour tous nos échanges, son soutien permanent, son expérience, ses conseils, ses efforts, qui ont permis d'aboutir à ce dénouement heureux. Pour reprendre une expression de Nicolas, "que de chemin parcouru" également dans notre relation qui s'est transformée en une vraie relation d'amitié.

Je remercie sincèrement Pascal Launey qui m'a permis de progresser sur de nombreux plans avec bienveillance. Je lui en suis très reconnaissante. Au fil du temps, j'ai été de plus en plus friande de nos échanges, de ses conseils. Là aussi, que de chemin parcouru.

J'adresse une pensée chaleureuse à l'équipe Mathématiques et Statistiques du LNE qui m'a accueillie : en plus de Nicolas, je remercie Alexandre Allard, Loïc Coquelin, Géraldine Ebrard, Aurélie Quoix, Catherine Yardin et tout particulièrement Viviane Nal. J'ai trouvé en chacun un soutien et une écoute importants pour moi. Un clin d'oeil à Loïc qui devient le dernier thésard de l'équipe !

Je tiens à remercier Marc Priel qui a permis que ce beau projet autour des statistiques bayésiennes se réalise. Je le remercie également pour son intérêt constant dans mes travaux, jusqu'à la dernière épreuve.

Je remercie tout particulièrement Michèle Désenfant pour nos échanges et son intérêt jusque dans les perspectives de la thèse.

Je tiens à souligner l'influence de Philippe Charlet, notre expert en chimie au LNE, tout au long de la thèse et dans la préparation du sujet de thèse. Il nous a quitté prématurément avant d'avoir vu l'aboutissement des travaux. Sa mémoire est gravée en moi.

Les travaux de thèse ont été enrichis grâce à l'implication de différentes entités du LNE. Ainsi la mise en oeuvre du recueil de l'information auxiliaire sur les laboratoires a été considérablement améliorée grâce à la construction et la mise en ligne sur le site internet du LNE d'un questionnaire destiné aux laboratoires. Pour cela je remercie Karim Benhdech, Julien d'Agostino et Cédric Malherbe du Département Développement du LNE/DRTI. Je remercie également Béatrice Lalere et Véronique Le Diouron du Département Biomédical et Chimie Organique du LNE/DMSI pour la réalisation de mesures de référence dans un but de validation de la méthode statistique proposée.

Je remercie le BAM, Federal Institute for Materials Research and Testing, pour la réalisation de mesures de référence.

Je remercie les partenaires du projet européen : Bertil Magnusson de SP (Technical Research Institute of Sweden), Klaus-Dieter Sommer, Bernd Güttler, Olaf Rienitz et Anna-Lisa Hauswaldt de la PTB (Physikalisch-Technischen Bundesanstalt), pour la richesse d'une première expérience collaborative à l'échelle européenne.

Je tiens à remercier Marie-Philippe Seiller, responsable qualité du BIPEA (Bureau InterProfessionnel d'Etudes Analytiques), pour son intérêt dans les travaux et sa collaboration active qui a permis de développer puis d'illustrer la méthodologie proposée dans la thèse.

Je remercie très sincèrement Marc Himbert de m'avoir fait l'honneur d'accepter d'être le Président de mon jury de thèse. Il m'a dit, oui, à la Chapelle du Musée du CNAM, au détours d'une conversation passionnante un soir de décembre (à l'occasion d'une leçon inaugurale). Ainsi placée sous les meilleures augures, la thèse a vécu un dénouement heureux et j'espère de nombreuses autres conversations.

Je remercie vivement Jean-Michel Marin et Vincenzo Esposito Vinzi d'avoir accepté d'être les rapporteurs de ma thèse. C'est un honneur pour moi. En particulier, je remercie Jean-Michel pour sa lecture critique qui a permis d'améliorer le manuscrit de thèse sur de nombreux points. Je remercie en particulier Vincenzo de m'avoir donné l'occasion de "défendre" les contributions de la thèse et les raisons de la démarche dans le contexte métrologique avant la "défense" officielle.

J'adresse mes vifs remerciements à Antonio Possolo pour m'avoir accueillie au NIST (Maryland, USA), pour son écoute et les vrais échanges que nous avons eus. Je suis honorée qu'il ait accepté de venir des États-Unis pour être membre de mon jury de thèse.

Certaines rencontres marquent plus ou moins, la rencontre avec Jean-Louis Foulley est de celles qui marquent profondément. La première rencontre, à l'initiative de Gilbert, a été suivie d'une semaine passée ensemble dans les calanques de Luminy (sur les modèles à variables latentes ! en 2008), suivie des journées Applibugs. Puis, après ma première présentation à Applibugs où j'évoquais mes soucis d'identifiabilité, Jean-Louis m'a parlé de l'expansion paramétrique. Cette idée géniale de sa part a révolutionné et dépoussiéré l'approche bayésienne que je proposais initialement. Mais bien plus que cela, j'ai appris à connaître Jean-Louis, et j'en suis très heureuse.

La rencontre avec Eric Parent est d'un autre genre. Bien que présent depuis le début dans mon comité de thèse, je crois que c'est la semaine passée sur les sommets à Rochebrune qui m'a permis de mieux le connaître et de l'apprécier totalement. Je ne le remercierai jamais assez de m'avoir invitée à cette semaine studieuse(!). Je remercie aussi Eric pour ses remarques et ses analyses tout au long de la thèse. Je reconnais qu'il m'a fallu à chaque fois un temps d'assimilation pour bien comprendre la portée de ses analyses. Je crois qu'il en reste encore certaines à mettre en oeuvre. Dans une autre vie de la thèse !

Je tiens à remercier chaleureusement l'équipe de la Chaire de Statistique du CNAM pour leur accueil et en particulier Ndeye Niang, Pierre-Louis Gonzalez et Sabine Glodkowski pour le soutien permanent qu'ils m'ont apporté pendant ces trois années.

Je remercie, à l'infini, mes parents, Bertrand et Danielle pour m'avoir montré le chemin

et permis de me réaliser dans ce qui me tient à coeur, avec amour et persévérance.

J'ai également une pensée particulière pour mon grand-père Henri qui m'a transmis la passion de la recherche (et de la cuisine !), et je suis sûre qu'il est très fier du chemin parcouru.

Mes pensées vont enfin à Gilles qui m'accompagne dans la vie depuis 2 ans et demi pour le remercier de sa ténacité et de sa confiance.

Résumé

La modélisation par équations structurelles est très répandue dans des domaines très variés et nous l'appliquons pour la première fois en métrologie dans le traitement de données de comparaisons interlaboratoires. Les modèles à équations structurelles à variables latentes sont des modèles multivariés utilisés pour modéliser des relations de causalité entre des variables observées (les données). Le modèle s'applique dans le cas où les données peuvent être regroupées dans des blocs disjoints où chaque bloc définit un concept modélisé par une variable latente. La structure de corrélation des variables observées est ainsi résumée dans la structure de corrélation des variables latentes. Nous proposons une approche bayésienne des modèles à équations structurelles centrée sur l'analyse de la matrice de corrélation des variables latentes. Nous appliquons une expansion paramétrique à la matrice de corrélation des variables latentes afin de surmonter l'indétermination de l'échelle des variables latentes et d'améliorer la convergence de l'algorithme de Gibbs utilisé. La puissance de l'approche structurelle nous permet de proposer une modélisation riche et flexible des biais de mesure qui vient enrichir le calcul de la valeur de consensus et de son incertitude associée dans un cadre entièrement bayésien. Sous certaines hypothèses l'approche permet de manière innovante de calculer les contributions des variables de biais au biais des laboratoires. Plus généralement nous proposons un cadre bayésien pour l'amélioration de la qualité des mesures.

Nous illustrons et montrons l'intérêt d'une modélisation structurelle des biais de mesure sur des comparaisons interlaboratoires en environnement.

Mots clés : Modèles à équations structurelles, variables latentes, identifiabilité, analyse bayésienne, augmentation des données, expansion paramétrique, algorithme de Gibbs, métrologie, comparaisons interlaboratoires, calcul d'incertitude, connaissances d'experts.

Abstract

Structural equation modelling is a widespread approach in a variety of domains and is first applied here to interlaboratory comparisons in metrology. Structural Equation Models with latent variables (SEM) are multivariate models used to model causality relationships in observed variables (the data). It is assumed that data can be grouped into separate blocks each describing a latent concept modelled by a latent variable. The correlation structure of the observed variables is transferred into the correlation structure of the latent variables. A Bayesian approach of SEM is proposed based on the analysis of the correlation matrix of latent variables using parameter expansion to overcome identifiability issues and improving the convergence of the Gibbs sampler. SEM is used as a powerful and flexible tool to model measurement bias with the aim of improving the reliability of the consensus value and its associated uncertainty in a fully Bayesian framework. The approach also allows to compute the contributions of the observed variables to the bias of the laboratories, under additional hypotheses. More generally a global Bayesian framework is proposed to improve the quality of measurements. The approach is illustrated on the structural equation modelling of measurement bias in interlaboratory comparisons in environment.

Keywords : Structural equation models, latent variables, identifiability, Bayesian analysis, data augmentation, parameter expansion, Gibbs sampler, metrology, interlaboratory comparisons, uncertainty analysis, expert knowledge.

ABSTRACT

Table des matières

Introduction

Cette thèse est le résultat d'une collaboration entre le LNE (Laboratoire National de Métrologie et d'Essais) et le CNAM (Conservatoire National des Arts et Métiers) réalisée dans le cadre d'un contrat CIFRE de février 2008 à février 2011.

La métrologie : la mesure au service des citoyens

La métrologie est la science de la mesure, de la réalisation de la mesure à la capacité de lui associer une incertitude de mesure.

Garantir la qualité d'une mesure est un objectif au quotidien pour les laboratoires qu'ils soient environnementaux, médicaux, ou nationaux de métrologie.

La *mesure* est également notre quotidien, pour nous les *usagers* de la mesure, avec plus ou moins d'impact dans notre vie.

L'amélioration de la qualité de la *mesure* est une réalité reposant sur l'engagement de personnes au service des autres.

Les laboratoires nationaux de métrologie ont un rôle stratégique dans l'amélioration de la qualité de la mesure, dont l'action sera propagée jusqu'à chaque laboratoire du pays. Cette action recouvre plusieurs dimensions

- une action globale (maintien des étalons nationaux,...),
- une action de recherche fondamentale (dématérialisation des étalons,...)
- une action de recherche appliquée (développement d'une méthode de mesure de référence,...)

Le LNE, en tant que laboratoire national de métrologie, assure cette action en France dans le cadre de sa mission de service public, en interaction avec les pouvoirs publics, sur des sujets d'intérêt général.

La contribution de ces travaux de thèse en statistique réalisés au LNE s'inscrit dans le cadre de cette action en réponse à une sollicitation européenne envers les laboratoires nationaux de métrologie européens face à une problématique de santé publique qui dépasse les cadres nationaux.

Apport statistique au projet européen "TRACEBIOACTIVITY"

Le projet "TRACEBIOACTIVITY" rassemble 10 laboratoires nationaux de métrologie européens dans le but d'améliorer les diagnostics médicaux au niveau des laboratoires d'analyses médicales, reposant sur le développement de méthodes de mesure de référence.

Le bénéfice de ces travaux en terme d'amélioration de la qualité de la mesure est transféré des laboratoires nationaux de métrologie aux laboratoires d'analyses médicales en deux temps. Dans un premier temps les laboratoires nationaux se comparent entre eux afin de valider les nouvelles méthodes, puis les laboratoires nationaux participent à des essais d'aptitudes organisés entre laboratoires d'analyses médicales.

Ces deux étapes reposent sur un traitement statistique des résultats de comparaisons entre laboratoires, appelées comparaisons interlaboratoires.

La contribution statistique au projet "TRACEBIOACTIVITY" consiste à proposer une nouvelle modélisation et un nouveau traitement des résultats de comparaisons interlabora-

toires, en tant qu'action support des autres parties du projet, dans un cadre bayésien, qui est l'objet des travaux de thèse présentés dans ce manuscrit.

A terme, la méthode statistique développée a le triple but d'être utilisée en routine par les organisateurs de tests d'aptitude entre laboratoires d'analyses médicales et généralisée à d'autres analytes pour le contrôle qualité, et de permettre d'améliorer la qualité de mesures pour lesquelles on ne peut pas développer de méthode de référence (pour des raisons de coût, de temps,...) à partir de la prise en compte d'expertise.

La métrologie évolue dans un contexte normatif qui nécessite d'introduire un vocabulaire normalisé.

Le vocabulaire de la métrologie

L'ensemble des termes métrologiques est défini dans la norme internationale : *Vocabulaire international de métrologie* (VIM, [JCGM, 2008]).

On en extrait que le *mesurande* est "la grandeur que l'on veut mesurer", le *résultat de mesure* est "un ensemble de valeurs attribuées à un mesurande, complété par toute autre information pertinente disponible". La note 2 de la section 2.9 précise que "le résultat de mesure est généralement exprimé par une valeur mesurée unique et une incertitude de mesure".

L'*incertitude de mesure* est "un paramètre non négatif qui caractérise la dispersion des valeurs attribuées à un mesurande, à partir des informations utilisées." L'*incertitude-type* est "l'incertitude de mesure exprimée sous la forme d'un écart-type".

D'après la note 3 de la section 2.26, "l'incertitude de mesure comprend en général de nombreuses composantes. Certaines peuvent être évaluées par une *évaluation de type A de l'incertitude* à partir de la distribution statistique des valeurs provenant de séries de mesurages et peuvent être caractérisées par des écarts-types. Les autres composantes, qui peuvent être évaluées par une *évaluation de type B de l'incertitude*, peuvent aussi être caractérisées par des écarts-types, évalués à partir de fonctions de densité de probabilité fondées sur l'expérience ou d'autres informations."

Le *modèle de mesure* est la "relation mathématique entre toutes les grandeurs qui interviennent dans un mesurage". D'après la note 1 de la section 2.48, "une forme générale d'un modèle de mesure est l'équation $h(Y, X_1, \ldots, X_n) = 0$ où Y, la grandeur de sortie dans le modèle de mesure, est le *mesurande*, dont la valeur doit être déduite de l'information sur les *grandeurs d'entrée* dans le modèle de mesure $X_1, \ldots, X_n$".

Le calcul d'incertitude : du cadre normatif originel du GUM à l'approche bayésienne

L'incertitude associée à un résultat de mesure peut s'interpréter en tant qu'indicateur de la qualité de la mesure obtenu après un traitement statistique des résultats de mesure.

Le calcul d'incertitude intervient donc de façon directe dans la qualité de la mesure.

Le calcul d'incertitude en métrologie est régi par une norme internationale : le Guide pour l'évaluation de l'incertitude de mesure, ci-après dénommé GUM d'après l'anglais Guide to the expression of uncertainty in measurement, [BIPM *et al.*, 1995]). L'incertitude-type de la grandeur de sortie est un écart-type obtenu par la propagation des incertitudes-types des grandeurs d'entrée dans le modèle de mesure sous l'hypothèse de linéarité du modèle à l'ordre 1 et d'un mesurande unique.

Publié en 1993, le GUM a depuis évolué vers de nouvelles approches publiées sous forme de suppléments au GUM pour prendre en compte des modèles de mesure plus complexes.

Ainsi le supplément 1 permet de lever l'hypothèse de linéarité à l'ordre 1 du modèle de mesure et le supplément 2 traite le cas de modèles à mesurande multiple.

Dans le supplément 1 au GUM [BIPM *et al.*, 2008], la simulation de Monte Carlo est utilisée pour obtenir la densité de probabilité de la grandeur de sortie en propageant les densités de probabilité des grandeurs d'entrée dans le modèle de mesure. L'avantage de cette approche est qu'elle permet de prendre en compte pleinement le modèle mathématique à la différence du GUM qui ne s'appuie que sur une approximation linéaire du modèle.

Ces différents suppléments ont ouvert le chemin à une approche bayésienne du calcul d'incertitude, de plus en plus représentée dans les papiers issus de la recherche récente.

On peut attribuer le point de départ de l'épopée bayésienne à l'article séminal de [Weise et Wöger, 1993] en 1993 et à son application en test de conformité dans [Weise et Wöger, 1994].

[Elster *et al.*, 2007] ont raccordé l'approche bayésienne du calcul d'incertitude au GUM en montrant que le supplément 1 est en fait déjà bayésien ! Ce qui n'étonnera d'ailleurs pas un bayésien.

L'approche bayésienne offre un cadre de modélisation riche et flexible, qui intègre en particulier les avancées déjà permises par les deux suppléments.

L'avancée bayésienne vient de la modélisation simultanée de toutes les grandeurs du modèle et de la prise en compte d'information *a priori* sur les paramètres du modèle. Cette approche fournit, grâce à des simulations dites de Monte Carlo par Chaînes de Markov, la densité de probabilité conjointe des paramètres du modèle *a posteriori* .

Du biais de mesure à l'incertitude

La notion de biais de mesure est définie par rapport à la notion de valeur vraie. En effet, d'après la note 1 de la section 2.17 du VIM [JCGM, 2008] "la valeur de référence pour une erreur systématique est une valeur vraie, une valeur mesurée d'un étalon dont l'incertitude de mesure est négligeable, ou d'une valeur conventionnelle".

D'après le VIM, le *biais de mesure* est "l'estimation d'une erreur systématique", où l'*erreur systématique* est une "composante de l'erreur de mesure qui, dans des mesurages répétés, demeure constante ou varie de façon prévisible".

D'après la note 1 de la section 2.26 du VIM, "parfois, on ne corrige pas des effets systématiques estimés, mais on insère plutôt des composantes associées de l'incertitude".

Les notions d'incertitude et de biais de mesure, piliers de la métrologie, sont les points

clés de la modélisation des résultats de comparaisons interlaboratoires.

Partant du cadre général de contrôle de la qualité, les objectifs précis des comparaisons interlaboratoires se déclinent en pratique selon la nature de la comparaison.

Dans le cadre de cette étude on distingue donc les comparaisons entre laboratoires nationaux de métrologie pour assurer l'équivalence des mesures entre pays, appelées les comparaisons clés et les comparaisons entre laboratoires de routine d'un même pays pour contrôler la continuité de leur performance, appelées tests d'aptitude.

Les comparaisons interlaboratoires : un outil de la qualité

Les comparaisons interlaboratoires sont un outil de validation externe de la qualité des laboratoires et interviennent à ce titre dans le processus d'accréditation des laboratoires.

D'après le VIM ([JCGM, 2008]), une comparaison interlaboratoires est "l'organisation, la réalisation et l'évaluation d'essais ou de mesures sur des objets soumis à l'essai ou à des essais semblables par au moins deux laboratoires différents dans des conditions prédéterminées."

L'objectif est alors d'estimer la valeur vraie du mesurande et son incertitude associée, ainsi que les biais de mesure des laboratoires le cas échéant.

Dans le cas des comparaisons clés, la valeur vraie à estimer est appelée *la valeur de référence*, alors que dans le cas de tests d'aptitude la terminologie est *la valeur assignée* où la *valeur assignée* est soit une *valeur de référence* soit une *valeur consensuelle*. La valeur assignée est définie dans le VIM ([JCGM, 2008]) comme la "valeur attribuée à une grandeur particulière et reconnue, parfois par convention, comme ayant une incertitude appropriée à un usage donné".

Il est donc essentiel que le traitement des résultats des laboratoires fournisse une estimation fiable de la valeur assignée, de son incertitude associée ainsi que des biais de mesure des laboratoires.

Des normes internationales ont été établies pour guider le déroulement d'une comparaison interlaboratoires et proposer des directives pour l'exploitation des résultats.

En amont, la norme ISO/IEC 17025 [ISO-CASCO, 2005] sur les exigences pour la compétence des laboratoires d'essais et d'étalonnage, demande aux laboratoires de participer à des comparaisons interlaboratoires.

La norme ISO/IEC 17043 [ISO-CASCO, 2010] traite des exigences pour les organisateurs de tests d'aptitude et renvoie à la norme NF ISO 13528 [ISO/TC69, 2005] (actuellement en demande de révision) pour la mise en oeuvre pratique de ces exigences. En particulier, le laboratoire organisateur doit être compétent pour

- fabriquer des matériaux, stables et homogènes durant la comparaison, à envoyer aux laboratoires participants (mesurande unique),
- proposer une méthode de calcul de la valeur de référence et de son incertitude associée qui prend en compte les spécificités de la comparaison,
- garantir la confidentialité de leurs résultats.

La norme NF ISO 13528 [ISO/TC69, 2005] donne des indications détaillées sur l'utilisation de méthodes statistiques dans les essais d'aptitude.

Les comparaisons clés

Les comparaisons clés (en anglais *key comparisons*) sont organisées par le CIPM (Comité International des Poids et Mesures) entre laboratoires nationaux de métrologie dans le but de produire des mesures d'équivalences des laboratoires participants, soit de manière unilatérale (par rapport à une valeur de référence) soit de manière bilatérale (les uns par rapport aux autres).

Dans les comparaisons clés, les laboratoires renvoient leur résultat de mesure accompagné de l'incertitude associée. Le calcul de cette incertitude est réalisé selon la démarche du Guide pour l'évaluation de l'incertitude de mesure (GUM).

Un formulaire résumant différentes approches pour calculer la valeur de référence et son incertitude associée a été proposé dans une note technique interne au LNE (voir [Amarouche et Désenfant, 2004]). Ce document de synthèse (en date de 2004) a été motivé par la directive pour les comparaisons clés du CIPM qui laisse une grande liberté dans le choix de la méthode de calcul.

Les tests d'aptitude

D'après la norme NF ISO 13528 [ISO/TC69, 2005], "les essais d'aptitudes par comparaisons interlaboratoires sont utilisés pour déterminer les performances de laboratoires individuels, relatives à des essais ou à des mesures spécifiques et pour surveiller la continuité des performances des laboratoires [...] Dans le langage statistique, les performances des laboratoires peuvent être décrites par trois propriétés : le biais de laboratoire, la stabilité et la répétabilité."

L'ISO/IEC 17 043 [ISO-CASCO, 2010] décrit les cinq façons les plus courantes pour déterminer la valeur assignée selon les spécificités de la comparaison. "Ces modes opératoires nécessitent l'utilisation de :

1. valeurs connues, avec des résultats déterminés par une formulation spécifique de l'entité soumise à l'essai d'aptitude (par exemple fabrication ou dilution) ;
2. valeurs de référence certifiées, déterminées par des méthodes d'essai ou de mesures définitives (pour les essais quantitatifs) ;
3. valeurs de référence, déterminées par analyse, mesurage ou comparaison de l'entité soumise à l'essai d'aptitude avec un matériau de référence ou un étalon traçable à un étalon national ou international ;
4. valeurs consensuelles provenant des participants experts – il convient que les experts (qui peuvent, dans certains cas, être des laboratoires de référence) aient une compétence prouvée dans la détermination du ou des mesurande(s) testé(s), en utilisant des méthodes validées connues pour être hautement exactes et comparables aux méthodes généralement utilisées ;
5. valeurs consensuelles provenant des participants, en utilisant des méthodes statistiques décrites dans la norme NF ISO 13528 [ISO/TC69, 2005] et dans le protocole harmonisé international de l'UICPA, et en tenant compte des effets des observations aberrantes."

Le contexte métrologique et normatif ainsi défini constitue le cadre des travaux de thèse.

Vers la modélisation des biais de mesure lors des tests d'aptitude

Dans ce cadre, la stratégie pour répondre à la problématique du projet européen a été d'orienter les travaux de thèse sur la modélisation des biais de mesure des laboratoires de routine en proposant un cadre de modélisation flexible et généralisable qui repose sur l'intégration d'avis d'experts.

Ce travail de modélisation des biais, si possible en association avec le développement et l'utilisation de méthodes de référence développées par les laboratoires nationaux de métrologie, a le triple objectif suivant :

- amélioration de la fiabilité de la valeur de consensus et de son incertitude associée,
- amélioration du contrôle des biais de mesure,
- identification et quantification des sources de biais de mesure.

Présentation des chapitres

L'exposé des travaux de thèse est divisé en cinq chapitres.

Le premier chapitre présente l'état de l'art de la modélisation des biais de mesure dans le traitement des données de comparaisons interlaboratoires. La revue des méthodes existantes et l'annonce des méthodes alternatives proposées sont réalisées avec la volonté d'unifier l'approche du traitement des données de comparaisons interlaboratoires. La fin du chapitre propose à cet effet un guide d'utilisation des différentes méthodes sous forme d'arbre de décision. Volontairement théorique, ce chapitre introduit les développements des chapitres suivants et prépare au chapitre 4 pour la mise en oeuvre pratique.

Le deuxième chapitre présente le cadre théorique dans lequel les outils statistiques sont développés. Les concepts de statistique bayésienne utilisés sont donc présentés, ainsi qu'un rappel de la théorie des méthodes Monte Carlo par Chaînes de Markov (MCMC).

Le troisième chapitre présente le traitement statistique de l'information auxiliaire par les modèles à équations structurelles (SEM) qui permettent de capturer la structure de cette information (par exemple, un questionnaire). Ces modèles ainsi que les méthodes d'estimation sont traités dans un cadre général, indépendamment de la problématique des comparaisons interlaboratoires (et des résultats de mesure).

Le quatrième chapitre traite de la mise en oeuvre pratique de la méthodologie alternative présentée au chapitre 1. Le traitement des données de comparaisons interlaboratoires est présenté dans un cadre bayésien dans lequel l'information auxiliaire sur les biais de mesure constitue une information *a priori* sur les mesures. Cette information est utilisée dans le calcul de la distribution *a posteriori* de la valeur de consensus, de son incertitude associée et des biais de mesure. On montrera en particulier que la valeur de consensus ainsi estimée est plus robuste et les biais moins sensibles aux résultats des autres laboratoires.

Le cinquième chapitre illustre la mise en oeuvre pratique de la méthodologie développée dans les chapitres précédents, du recueil de l'information auxiliaire sur les biais de mesure à l'interprétation des résultats obtenus. L'approche est appliquée à des données de tests d'aptitude en environnement.

Liste des travaux

Les travaux de thèse ont donné lieu à la rédaction des deux articles suivants

- S. Demeyer, N. Fischer, G. Saporta, Contributions to Bayesian Structural Equation Modelling. In Proceedings of the 19th International Conference on Computational Statistics COMPSTAT'2010, Physica Verlag, Springer, pages 469-476

- S. Demeyer, J.-L. Foulley, N. Fischer, G. Saporta, Bayesian analysis of structural equation models using parameter expansion, *In Statistical Learning and Data Science*, Chapman&Hall/CRC, 2011

Les travaux de thèse ont également donné lieu aux communications orales et au poster suivants

- *Bayesian approach to assign consensus values in PT comparisons*, 255th PTB Seminar : "Emerging Topics in Mathematics for Metrology - From Measurement Uncertainty to Metrology of Complex Systems (MATHMET2010)", Berlin, June 2010

- S. Demeyer, N. Fischer, G. Saporta, *Approche bayésienne des modèles à équations structurelles.* In 42 èmes Journées de Statistique, Marseille, Mai 2010

  S. Demeyer, N. Fischer, G. Saporta, *Un modèle à variables latentes pour le traitement de données hybrides issues de comparaison interlaboratoires.* In 41 èmes Journées de Statistique, Bordeaux, Mai 2009

- S. Demeyer, N. Fischer, G. Saporta, *Un modèle à variables latentes pour le traitement de données hybrides issues de comparaison interlaboratoires*, International Congress of Metrology, Paris, June 2009

Une communication au congrès ASMDA 2011 issue des travaux de cette thèse a également été acceptée.

S. Demeyer, J.-L. Foulley, N. Fischer, G. Saporta, *Bayesian analysis of structural equation models using parameter expansion. Application to expert modelling in uncertainty analysis*, ASMDA 2011, Rome

Chapitre 1

Etat de l'art : modélisation des biais de mesure dans le traitement des données de comparaisons interlaboratoires

1.1 A l'origine de la modélisation des biais de mesure : les valeurs aberrantes

Les observations dites "aberrantes" peuvent être de nature différente : une erreur de manipulation, voire même de recopie du résultat ou une observation exceptionnelle (voir [Rousseeuw, 1991]).

Fréquemment rencontrées (c.f. [Rousseeuw, 1991] et [Müller, 2000]), leur traitement peut avoir différentes implications au niveau de l'analyse et fait des observations aberrantes un problème fondamental dans le traitement des données de comparaisons interlaboratoires.

D'un point de vue statistique, une observation aberrante peut être détectée par un test statistique de consistance (test de Grubbs, test du Khi2). Selon le résultat du test, une observation peut être conservée ou rejetée.

Le rejet des observations aberrantes mène au concept de plus grand ensemble consistant suggéré par Cox et développé dans [Cox, 2007] pour le traitement des résultats de comparaisons clé.

Cependant, rejeter des observations ne fait pas l'unanimité au sein de la communauté. Parmi les détracteurs, [Toman et Possolo, 2009] et [Müller, 2000] recommandent de conserver tous les résultats des laboratoires.

Dans la revue de littérature de ce chapitre et dans les méthodes que nous proposons en fin de chapitre, nous faisons le choix de présenter et développer des méthodes permettant de conserver les résultats de tous les laboratoires participants.

Ce choix se justifie par un effort alternatif de modélisation des biais de mesure, qui est l'essence du travail que nous proposons.

1.2 Modélisation des biais de mesures

On a vu en introduction[1] que la connaissance des biais de mesure est équivalente à la connaissance de la valeur de référence de la comparaison. Or cette dernière étant généralement inconnue, les résultats des laboratoires ne peuvent permettre à eux seuls de connaître les valeurs vraies des biais de mesure des laboratoires.

Les biais de mesure sont donc en pratique estimés comme l'écart du résultat de mesure du laboratoire à la valeur de consensus, calculée à partir des résultats de tous les laboratoires.

Cette pratique est à la base des cartes de contrôle établies par les organisateurs de comparaisons interlaboratoires utilisées pour le suivi des performances des laboratoires.

Le problème se pose alors de la dépendance des biais des laboratoires aux résultats des autres laboratoires.

La modélisation des biais de mesure a pour but de limiter cette dépendance sur la base d'un modèle des résultats de mesure prenant en compte explicitement les biais de mesure.

L'estimation des biais de mesure repose sur les hypothèses suivantes communes à tous les modèles présentés

- une composante du biais de mesure commune à tous les laboratoires n'est pas estimable (c.f. [Chunovkina *et al.*, 2008])
- l'hypothèse (*a priori*) d'un biais moyen nul est largement répandue dans la littérature métrologique (c.f. [Toman et Possolo, 2009])

La modélisation des biais de mesure la plus fréquemment rencontrée dans la littérature est sous la forme d'effet aléatoire, qui consiste à supposer que les écarts entre les résultats des laboratoires et la valeur vraie inconnue sont pris en compte par une augmentation de l'incertitude associée au résultat renvoyé. En d'autres termes, le biais de mesure est confondu avec une composante d'incertitude (ou la somme de plusieurs composantes).

La revue de littérature qui suit s'appuie sur les quatre hypothèses alternatives suivantes sur les biais de mesure

- pas d'hypothèse sur les biais
- pas de biais de mesure
- le biais est un effet fixe
- le biais est un effet aléatoire

qui permet un premier niveau de classification des méthodes utilisées.

Il est crucial de noter que le choix entre les différentes méthodes d'estimation repose sur une analyse préalable de la nature des biais de mesure.

La revue de littérature des méthodes couramment utilisées dans le cas d'un traitement du biais sous la forme d'un effet aléatoire fournira un deuxième niveau de classification des méthodes d'estimation entre méthodes robustes, méthodes bayésiennes et moyennes pondérées.

1. A la section 'Du biais à l'incertitude de mesure'

1.3 Evaluation de la valeur de consensus par moyenne robuste ou moyenne pondérée ?

Le choix d'une évaluation par méthode robuste ou toute autre méthode est conduit par la nature des résultats et une certaine expérience du traitement des résultats de comparaisons interlaboratoires.

Ainsi dans le cas d'une comparaison interlaboratoires où un résultat de mesure unique est renvoyé par chaque laboratoire sans incertitude associée (fréquemment rencontré en pratique et en particulier dans les exemples d'applications de la thèse), la méthode à utiliser est la méthode robuste de la norme NF ISO 13528.

Avec la donnée des réplicats de mesure de chaque laboratoire, l'estimation d'un modèle à effets aléatoires peut être réalisée soit par méthode robuste soit par une autre méthode ; on privilégiera la méthode robuste si on sait par exemple que le processus de mesure est encore globalement mal maîtrisé (à rapprocher de la nature des observations aberrantes).

Avec la donnée supplémentaire de valeurs d'incertitude de chaque laboratoire, l'utilisation de méthodes robustes semble moins justifiée.

1.3.1 Les moyennes pondérées

Les moyennes pondérées sont largement utilisées dans le traitement de données de comparaisons interlaboratoires. Les moyennes pondérées peuvent être obtenues directement ou en sortie d'algorithmes itératifs, selon des hypothèses sur les biais propres à chaque comparaison.

[Rukhin, 2009] propose une revue des principales méthodes de calcul de moyennes pondérées et une discussion sur leur utilisation.

Les méthodes menant à des moyennes pondérées présentées dans la revue de bibliographie sont

- l'estimateur de Graybill-Deal (pas de biais)
- les méthodes bayésiennes (ici dans le cas du biais aléatoire)
- l'algorithme de Mandel Paule (biais aléatoire)

1.3.2 Les moyennes robustes

Les algorithmes robustes reposent sur l'utilisation des statistiques d'ordre. Ces algorithmes s'appliquent donc à des résultats de laboratoires classés par ordre croissant.

On note $X_{(1)}, \ldots, X_{(n)}$ l'échantillon ordonné où

$$X_{(1)} = \min \{X_i\} \tag{1.1}$$
$$X_{(n)} = \max \{X_i\} \tag{1.2}$$
$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)} \tag{1.3}$$

Les estimateurs robustes fournis par ces algorithmes sont des combinaisons linéaires des

statistiques d'ordre de l'échantillon. La théorie des L estimateurs associée à ces méthodes est développée dans [Huber et Ronchetti, 2009].

[Müller, 2000] justifie et suggère l'utilisation des méthodes robustes dans le traitement des comparaisons clés en complément des méthodes traditionnelles.

[AMC, 1989] suggère l'utilisation des méthodes robustes pour les essais coopératifs et les essais collaboratifs.

Enfin la norme NF ISO 13528 [ISO/TC69, 2005] propose une méthode robuste pour le traitement des données de tests d'aptitude.

Les méthodes robustes présentées dans la revue de littérature sont

- la médiane et son incertitude (pas de biais)
- la moyenne winsorizée et ses deux implémentations dans les normes [ISO/TC69, 2005] (pas d'hypothèse sur les biais) et [AMC, 1989] (biais aléatoire)

1.3.3 Vers une approche unifiée du traitement des données de comparaisons interlaboratoires

Dans l'introduction de la thèse on a distingué et présenté les comparaisons clés (entre laboratoires nationaux de métrologie) et les tests d'aptitude (entre laboratoires de routine) parmi l'ensemble des comparaisons interlaboratoires existantes.

Cette distinction s'est opérée sur la base des objectifs où dans le premier cas l'objectif est d'évaluer la comparabilité des mesures entre pays et où dans le deuxième cas l'objectif est d'évaluer les performances de laboratoires au sein d'un pays.

Or dans tous les deux cas les méthodes sont développées selon la nature des résultats et de l'information à disposition dans le but de produire une valeur de référence ou de consensus (respectivement), son incertitude associée et les biais de mesure des laboratoires. Dans le cas des comparaisons clés on produit de plus le degré d'équivalence entre laboratoires, estimé à partir de la différence des résultats des laboratoires.

Typiquement, les méthodes développées pour les comparaisons clés prennent toutes en compte l'incertitude des résultats des laboratoires et on observe que l'évolution des normes et des règlementations pour les tests d'aptitude va dans le sens de demander l'incertitude des résultats des laboratoires de routine.

Il semble donc raisonnable, lorsque l'on dispose des incertitudes des laboratoires, d'appliquer les méthodes développées dans le cadre des comparaisons clés.

Inversement, si on fait le choix de ne pas prendre en compte les incertitudes associées aux résultats dans le cas des comparaisons clés (c.f. [Amarouche et Désenfant, 2004]) la médiane 1.4.2.2 peut être utilisée dans les deux types de comparaisons.

Ces remarques motivent la classification (annoncée à la section 1.2) des différentes méthodes présentées dans la revue de littérature à la section 1.4 selon leurs hypothèses, qui peuvent bien sûr être liées au type de la comparaison.

L'organisation des différents critères relevés dans la revue de littérature est réalisée à la section 1.5 et fournit un outil directement utilisable dans le choix de la méthode sous la forme d'un arbre de décision.

1.4 Revue de littérature

1.4.1 En l'absence d'hypothèse sur les biais : la méthode robuste de la norme NF ISO 13528 [ISO/TC69, 2005]

La norme NF ISO 13528 propose une méthode robuste, implémentation du processus de winsorization rappelé ci-dessous, localement appelée algorithme A, pour calculer une estimation de la valeur de consensus et de son incertitude associée dans le cas de tests d'aptitude. Au contraire des autres méthodes robustes, l'algorithme A est une méthode normalisée. On retrouvera le processus de winsorization dans un algorithme de l'Analytical Methods Committee [AMC, 1989] à la section 1.4.4.1.

Rappel : le processus de winsorization D'un point de vue historique les méthodes robustes sont apparues en 1953 dans un article de [Box, 1953]. Depuis, selon [Kafadar, 2003], la plupart des méthodes dites robustes ont été inspirées par les travaux de John Tukey. Dès 1960, [Tukey, 1960b] justifie l'utilisation de la variance winsorizée à la place de la variance empirique, ainsi nommée après Charles Winsor. Pour en comprendre les raisons, reportons-nous à cette citation de [Tukey, 1960a]

> Finally there is winsorization. Charles Winsor put forward a principle of quite general application, namely : while the numerical value of an apparently wild observation is untrustworthy, the direction of its deviation (e.g. high or low) is worthy of attention. He applied this principle to outliers by taking the largest deviations or largest residuals and decreasing their magnitude, while retaining their sign, until they are equal to the next largest ones, thus making a qualitatively reasonable adjustment. In fact, it is possible to give quite quantitative reasons why this sort of adjustment is not only a convenient approach but also an effective one.
>
> JOHN TUKEY, *Discussion of Anscombe and Daniel papers*, 1960

D'un point de vue mathématique le processus de winsorization consiste à remplacer les valeurs au-dessus d'un certain seuil par la valeur du seuil et les valeurs en dessous d'un autre seuil par la valeur de ce seuil. En notant α_g (resp. α_d) la proportion d'observations aberrantes à gauche (resp. à droite), le processus ramène les $\alpha_g n$ plus petites observations à la valeur $x_{\alpha_g n+1}$ et les $\alpha_d n$ plus grandes observations à la valeur $x_{n-\alpha_g n}$.

Ainsi le processus ne rejette pas les valeurs aberrantes mais les ramène à des valeurs plus crédibles en leur affectant un poids.

Cet algorithme est par exemple utilisé en intégration d'images en astronomie où il permet de moyenner des pixels associés à une même coordonnée en appliquant le processus de winsorization aux pixels aberrants (utilisé dans le logiciel de traitement d'images avancé PixInsight : http ://pixinsight.com).

Implémentation de l'algorithme de winsorization sous la forme de l'algorithme A

L'algorithme A s'applique aux résultats des laboratoires, où le résultat d'un laboratoire est ici la moyenne de l'ensemble des mesures qu'il a effectuées sur le matériau d'essai (voir [ISO/TC69, 2005] paragraphe 5.6.2).

L'algorithme A est détaillé dans l'encadré 1.4.1.

On note p le nombre de laboratoires participants.

Algorithme 1.4.1 *Algorithme A (norme NF ISO 13528)*
Initialisation

$$x^* = \textit{médiane des } x_i \tag{1.4}$$
$$s^* = \textit{médiane des } |x_i - x^*| \tag{1.5}$$

A chaque itération

$$x_i^* = \begin{cases} x^* - 1.5s^* \textit{ si } x_i < x^* - 1.5s^* \\ x^* + 1.5s^* \textit{ si } x_i > x^* + 1.5s^* \\ x_i \textit{ sinon} \end{cases} \tag{1.6}$$

$$x^* = \sum_{i=1}^{p} \frac{1}{p} x_i^* \tag{1.7}$$

$$s^* = 1.134 \sqrt{\sum_{i=1}^{p} \frac{1}{p-1} \left(x_i^* - x^*\right)^2} \tag{1.8}$$

Après convergence la valeur de consensus est estimée par $\hat{\mu} = x^*$, son incertitude associée est estimée par

$$u\left(\hat{\mu}\right) = 1.25 \frac{s^*}{\sqrt{p}} \tag{1.9}$$

et le biais de mesure du laboratoire i est estimé par $\hat{b}_i = x_i - \hat{\mu}$.

Cet algorithme permet en particulier de traiter les résultats en-dessous de la valeur de détection en les ramenant à $\hat{\mu} - 1.5s^*$.

Interprétation L'étape de mise à jour de la valeur de consensus à l'itération $r+1$ à l'expression 1.7 est en fait une moyenne pondérée des valeurs renvoyées à l'itération précédente x_i^r, de poids ϖ_i^{r+1} définis par :

$$\varpi_i^{r+1} = \begin{cases} 1 \text{ si } x_i^{r+1} = x_i^r \\ \dfrac{1.5s^*}{|x_i^r - \hat{\mu}|} \text{ sinon} \end{cases} \tag{1.10}$$

Le processus de winsorization ainsi défini est représenté à la figure 1.1.

Remarque : L'algorithme fait apparaître deux constantes en relation avec une loi normale centrée réduite, pour laquelle on s'intéresse aux réalisations au delà de 1.5 fois l'écart-type (noté 1.5σ) en valeur absolue

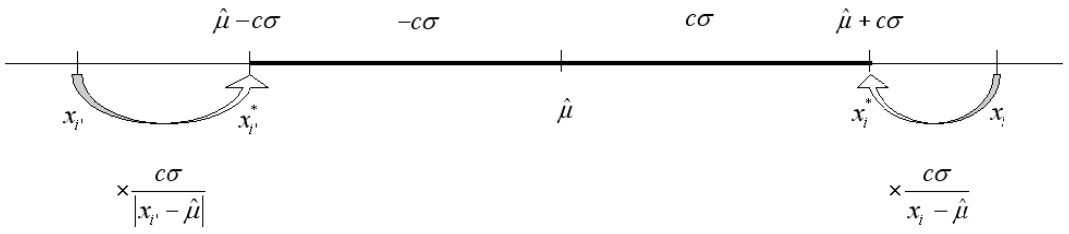


FIGURE 1.1 – Illustration du processus de winsorization

- la valeur 1.5 à l'expression 1.6 est conseillée par [Huber et Ronchetti, 2009], notée c dans la figure 1.1,
- la valeur 1.134 à l'expression 1.8 vise à compenser le fait qu'on remplace toutes les valeurs en dessous et au dessus de 1.5σ à chaque itération par le quantile à 1.5σ, ce qui amène à sous-estimer l'écart-type. La valeur 0.134 est approximativement la fraction de la surface sous la queue d'une distribution gaussienne coupée à $\pm 1.5\sigma$.

1.4.2 En l'absence de biais de mesure

1.4.2.1 Modèle linéaire pour données répétées

Le modèle inéaire pour données répétées s'applique à tous les résultats lorsqu'aucune source de biais de mesure n'est identifiée ou aux données conservées après passage d'un test de consistance des résultats.

Le modèle s'écrit

$$x_{ij} = \mu + \varepsilon_{ij} \tag{1.11}$$

où x_{ij} est la j^{e} mesure du laboratoire i, μ est la valeur vraie de la comparaison et $\varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma_i^2\right)$. La variance des mesures renvoyées par le laboratoire i est donc σ_i^2.

En notant $\bar{x}_{i.}$ la moyenne des mesures du laboratoire i et n_i le nombre de réplicats de mesure du laboratoire i, le meilleur estimateur au sens des moindres carrés est la moyenne pondérée

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w_i \hat{x}_{i.}}{\sum_{i=1}^{n} w_i},\ w_i = \frac{\sigma_i^2}{n_i} \tag{1.12}$$

de variance $u^2\left(\hat{\mu}\right) = \frac{1}{\sum_{i=1}^{n} w_i}$.

Si σ_i^2 est inconnu, l'estimateur de Graybill-Deal (voir [Graybill et Deal, 1959]) est la version plug-in de l'estimateur ci-dessus où $\frac{\sigma_i^2}{n_i}$ est remplacé par son estimateur sans biais s_i^2

$$\hat{\mu}_{GD} = \frac{\sum_{i=1}^{n} s_i^{-2} x_i}{\sum_{i=1}^{n} s_i^{-2}} \tag{1.13}$$

L'estimateur traditionnel de la variance de l'estimateur de Graybill-Deal est connu pour sous-estimer la variance (voir [Zhang, 2006]). [Zhang, 2006] propose une étude de différents estimateurs de la variance de l'estimateur de Graybill-Deal.

1.4.2.2 La médiane et son incertitude associée pour des données non répétées

La médiane $\tilde{m}$ d'un échantillon ordonné est définie par

$$\tilde{m} = \begin{cases} x_{k+1},\ k = \dfrac{n-1}{2} \text{ si } n \text{ est impair} \\ \dfrac{x_k + x_{k+1}}{2},\ k = \dfrac{n}{2} \text{ si } n \text{ est pair} \end{cases} \tag{1.14}$$

Le calcul de l'incertitude associée à la médiane nécessite d'introduire une grandeur basée elle-aussi sur la médiane : la médiane des écarts absolus à la médiane notée MAD (d'après l'anglais, *median absolute deviation*) définie par

$$MAD = \text{med}\left\{|x_i - \tilde{m}|\right\} \tag{1.15}$$

Après calculs, repris dans [Müller, 2000], l'incertitude associée à la médiane s'écrit

$$u\left(\tilde{m}\right) = \frac{1.858}{\sqrt{n-1}} MAD \tag{1.16}$$

1.4.3 Le biais de mesure est un effet fixe : le modèle à effets fixes

Le modèle à effets fixes s'applique dans le cas où on suppose que les biais de mesure ne varient pas d'une expérience à l'autre.

L'hypothèse d'un biais fixe se traduit en moyenne sur les résultats.

Le modèle s'écrit

$$x_{ij} = \mu + \beta_i + \varepsilon_{ij} \tag{1.17}$$

où μ est la valeur vraie de la comparaison, β_i est le biais (fixe) du laboratoire i et $\varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma_i^2\right)$. La variance des mesures renvoyées par le laboratoire i est donc σ_i^2, la même que dans le modèle à moyenne commune de la section 1.4.2.

Le modèle est estimable sous la contrainte de biais moyen nul suivante : $\sum_{i=1}^{n} \beta_i = 0$, par exemple par maximum de vraisemblance (voir [Toman et Possolo, 2009]).

Dans le cas où σ_i^2 est connu la valeur assignée est estimée par la moyenne des résultats $\hat{\mu} = \bar{x}$, son incertitude associée est estimée par l'écart-type des résultats et le biais de mesure est estimé par l'écart à la moyenne $\hat{\beta}_i = \bar{x}_{i.} - \bar{x}$.

1.4.4 Le biais de mesure est un effet aléatoire : le modèle à effets aléatoires

Dans ce modèle le biais de mesure est considéré comme un effet aléatoire (voir section 1.2).

Dans la littérature deux variantes sont proposées, dépendant de l'information à disposition sur les incertitudes renvoyées. Dans le premier cas (Cas 1), on ne dispose pas d'information sur l'incertitude causée par les biais de mesure alors que dans le deuxième cas on dispose d'information détaillée par laboratoire (Cas 2).

1.4.4.1 Cas 1 : Pas d'information détaillée sur l'incertitude

Les biais de mesure b_i sont considérés comme un échantillon issu d'une population de moyenne nulle et de variance commune τ^2 à estimer, s'interprétant comme la variance interlaboratoires. En notant x_{ij} la j^{e} mesure du laboratoire i le modèle à effets aléatoires s'écrit dans ce cas

$$x_{ij} = \mu + b_i + \varepsilon_{ij} \tag{1.18}$$

où μ est la valeur vraie de la comparaison, $b_i \sim \mathcal{N}\left(0, \tau^2\right)$ et $\varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma_i^2\right)$. La variance des mesures renvoyées par le laboratoire i est donc portée à $\tau^2 + \sigma_i^2$.

Le modèle peut s'estimer par maximum de vraisemblance (voir [Toman et Possolo, 2009]).

L'algorithme de Mandel-Paule, [Mandel et Paule, 1970]

Dans le cas particulier où $\sigma_i^2 = \sigma^2$ pour tout i, le modèle peut s'estimer par l'algorithme de Mandel-Paule [Mandel et Paule, 1970] du nom de John Mandel et Robert C. Paule, très utilisé au NIST. Cet algorithme itératif estime les deux composantes de variance τ^2 et σ^2, représentant respectivement la variance interlaboratoire et la variance intralaboratoire, et fournit une valeur de référence sous la forme de la moyenne pondérée suivante

$$\hat{\mu}_{MP} = \frac{\sum_{i=1}^{n} w_i \bar{x}_{i.}}{\sum_{i=1}^{n} w_i},\ w_i = \frac{1}{\lambda + \frac{1}{n_i}} \tag{1.19}$$

$$u^2\left(\hat{\mu}_{MP}\right) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} w_i} \tag{1.20}$$

où n_i est le nombre de réplicats de mesure du laboratoire i, $\lambda = \frac{\hat{\tau}^2}{\hat{\sigma}^2}$ avec $\hat{\tau}^2$ et $\hat{\sigma}^2$ les estimations respectives de τ^2 et σ^2 en sortie d'algorithme.

L'algorithme robuste de l'Analytical Methods Committee, [AMC, 1989]

La méthode robuste de [AMC, 1989] est une alternative robuste pour estimer le modèle à effets aléatoires en l'absence d'information détaillée sur l'incertitude (c.f. section 1.4.4,

cas 1). Cette hypothèse sur les biais de mesure des laboratoires fait de cet algorithme une alternative à celui de Mandel-Paule (section 1.4.4).

On rappelle le modèle à effets aléatoires :

$$x_{ij} = \mu + b_i + \varepsilon_{ij} \tag{1.21}$$

On note p le nombre de laboratoires et r le nombre de réplicats de mesures par laboratoire, supposé identique pour tous les laboratoires.

En notant $\mu_i = \mu + b_i$ on observe que la moyenne des μ_i est μ sous l'hypothèse de biais moyen nul faite au début de ce chapitre.

La procédure itérative repose sur les deux points suivants

- sachant μ_i on peut estimer μ (en tant que moyenne des μ_i) et la variance totale des mesures $\sigma_L^2 + \frac{\sigma_e^2}{r}$
- sachant les seules mesures on peut estimer μ_i et la variance résiduelle σ_e^2

Après ces deux étapes on obtient immédiatement la variance interlaboratoire.

La procédure de [AMC, 1989] alterne donc les deux étapes. Chaque étape est un algorithme itératif implémentant un algorithme semblable à l'algorithme A de la norme NF ISO 13528 [ISO/TC69, 2005], respectivement sur les moyennes $\bar{x}_{i.}$ des laboratoires et les résultats des laboratoires x_{ij}.

La première étape estime les paramètres μ et $\sigma_1^2 = \sigma_L^2 + \frac{\sigma_e^2}{r}$ à partir du modèle 1.18. L'algorithme robuste de la première étape alterne les trois sous-étapes suivantes jusqu'à convergence

$$\mu_i^*\left(\mu^*, \sigma_1^*\right) = \begin{cases} \mu^* - c_1\sigma_1^* \text{ si } \mu_i < \mu^* - c_1\sigma_1^* \\ \mu^* + c_1\sigma_1^* \text{ si } \mu_i > \mu^* + c_1\sigma_1^* \\ \mu_i \text{ sinon} \end{cases} \tag{1.22}$$

$$\mu^* = \sum_{i=1}^{p} \frac{1}{p}\mu_i^* \tag{1.23}$$

$$\sigma_1^* = 1.134\sqrt{\sum_{i=1}^{p} \frac{1}{p-1}\left(\mu_i^* - \mu^*\right)^2} \tag{1.24}$$

avec $c_1 = 1.5\sqrt{1 - \frac{1}{p}}$.

La deuxième étape estime les paramètres μ_i et σ_e^2 à partir du modèle $x_{ij} = \mu_i + \varepsilon_{ij}$. L'algorithme robuste de la deuxième étape alterne les trois sous-étapes suivantes jusqu'à convergence

$$x_{ij}^*\left(\mu_i^*,\sigma_e^*\right)=\begin{cases}\mu_i^*-c_2\sigma_e^* \text{ si } x_{ij}<\mu_i^*-c_2\sigma_e^*\\ \mu_i^*+c_2\sigma_e^* \text{ si } x_{ij}>\mu_i^*+c_2\sigma_e^*\\ x_{ij} \text{ sinon}\end{cases} \tag{1.25}$$

$$\mu_i^*=\frac{1}{r}\sum_{k=1}^{r}x_{ik}^* \tag{1.26}$$

$$\sigma_e^*=1.134\sqrt{\sum_{i=1}^{p}\sum_{k=1}^{r}\frac{1}{p}\frac{1}{r-1}\left(x_{ik}^*-\mu_i^*\right)^2} \tag{1.27}$$

avec $c_2=1.5\sqrt{1-\frac{1}{r}}$.

Après convergence la valeur de consensus est estimée comme dans l'algorithme A (section 1.4.1) par $\hat{\mu}=\mu^*$. Bien que [AMC, 1989] ne précise pas l'incertitude de $\hat{\mu}$ on suppose par analogie avec l'algorithme A section 1.4.1 que l'incertitude est donnée par

$$u\left(\hat{\mu}\right)=1.25\frac{\sigma_1^*}{\sqrt{p}} \tag{1.28}$$

L'initialisation des paramètres de l'algorithme total peut être choisie selon [AMC, 1989]

$$\mu_i=\text{médiane des } x_{ij} \tag{1.29}$$

$$\sigma_e=\text{médiane des } |x_{ij}-\mu_i|/\left(0.6745\sqrt{1-1/r}\right) \tag{1.30}$$

1.4.4.2 Cas 2 : Information détaillée sur l'incertitude

On suppose que chaque laboratoire renvoie séparément l'écart-type σ_i associé à la répétabilité de ses mesures et l'écart-type τ_i associé à l'incertitude restante (évaluée par des méthodes de type A et de type B, voir la section 'Le vocabulaire de la métrologie' en introduction de la thèse). Dans ces conditions la quantité $\sqrt{\tau_i^2+\sigma_i^2}$ correspond à l'incertitude totale du résultat du laboratoire i. Le biais de mesure b_i est donc considéré comme une réalisation d'une variable aléatoire de moyenne nulle et de variance connue τ_i^2.

En notant x_{ij} la j^{e} mesure du laboratoire i le modèle à effets aléatoires s'écrit dans ce cas

$$x_{ij}=\mu+b_i+\varepsilon_{ij} \tag{1.31}$$

où μ est la valeur vraie de la comparaison, $b_i\sim\mathcal{N}\left(0,\tau_i^2\right)$ et $\varepsilon_{ij}\sim\mathcal{N}\left(0,\sigma_i^2\right)$. La variance des mesures renvoyées par le laboratoire i est donc portée à $\tau_i^2+\sigma_i^2$.

[Toman, 2007] et [Chunovkina *et al.*, 2008] ont proposé deux méthodes d'estimation de ce modèle dans un cadre bayésien.

L'étude de [Toman, 2007] repose sur une formulation hiérarchique du modèle nécessitant l'introduction d'hyperparamètres pour modéliser la valeur vraie (voir la remarque en fin de section). Le choix d'une distribution *a priori* gaussienne, $N\left(0, v^2\right)$ avec v^2 qui tend vers l'infini, sur la valeur vraie entraîne la perte de la conjugaison. Dans ces conditions l'estimation de la valeur de consensus et de son incertitude associée est réalisée sous approximation de normalité.

L'étude de [Chunovkina *et al.*, 2008] repose sur la distribution jointe du mesurande et des biais de mesure des laboratoires sous la distribution *a priori*

$$\pi\left(\mu, b_1, \ldots, b_n\right) \propto \pi\left(b_1\right) \ldots \pi\left(b_n\right) \tag{1.32}$$

où $\pi\left(b_i\right)$ est la distribution normale centrée de variance τ_i^2.

Les deux méthodes fournissent les mêmes estimations de la valeur de consensus et de son incertitude associée, à savoir

$$\hat{\mu} = u\left(\hat{\mu}\right)^2 \sum_{i=1}^{N} \frac{\bar{x}_{i.}}{\tau_i^2 + \frac{\sigma_i^2}{n_i}} \tag{1.33}$$

$$u\left(\hat{\mu}\right)^2 = \left(\sum_{i=1}^{N} \frac{1}{\tau_i^2 + \frac{\sigma_i^2}{n_i}} \right)^{-1} \tag{1.34}$$

[Chunovkina *et al.*, 2008] fournit en plus une estimation des biais de mesure et une incertitude associée reposant sur une pondération de la formule traditionnelle du biais, à savoir l'écart des résultats à la moyenne pondérée

$$\hat{b_i} = \frac{\tau_i^2}{\tau_i^2 + \frac{\sigma_i^2}{n_i}} \left(\bar{x}_{i.} - \hat{\mu}\right) \tag{1.35}$$

$$u\left(b_i\right)^2 = \tau_i^2 \frac{\frac{\sigma_i^2}{n_i} + \tau_i^2 \frac{u(\hat{\mu})^2}{\tau_i^2 + \frac{\sigma_i^2}{n_i}}}{\tau_i^2 + \frac{\sigma_i^2}{n_i}} \tag{1.36}$$

Remarque : Le modèle à effets aléatoires peut s'écrire de manière équivalente sous forme d'un modèle hiérarchique, où chacune des variabilités correspond à un niveau de modélisation hiérarchique des mesures. En notant δ_i la moyenne des mesures du laboratoire i, le premier niveau modélise la distribution des résultats x_{ij} de chaque laboratoire autour de δ_i . L'erreur associée à ce niveau correspond à la variabilité des effets aléatoires σ_i^2.

Le deuxième niveau modélise la distribution des moyennes δ_i des laboratoires autour de la valeur vraie μ. L'erreur associée à ce niveau correspond à la variabilité des effets systématiques τ^2 ou τ_i^2 dans les cas 1 et 2 respectivement. La modélisation de ce deuxième niveau est indépendante des résultats x_{ij}.

Le troisième niveau correspond à la spécification bayésienne du modèle où m et v^2 sont les hyperparamètres associés à μ.

La formulation hiérarchique du modèle 1.31 s'écrit

$$x_{ij} \sim \mathcal{N}\left(\delta_i, \sigma_i^2\right) \tag{1.37}$$
$$\delta_i \sim \mathcal{N}\left(\mu, \tau_i^2\right) \tag{1.38}$$
$$\mu \sim \mathcal{N}\left(m, v^2\right) \tag{1.39}$$

où σ_i^2 est la variabilité des effets aléatoires, τ_i^2 est la variabilité des effets systématiques et σ_i^2 et τ_i^2 sont fournies par les laboratoires.

Dans la dernière partie on complète cet arbre en introduisant un critère de décision correspondant à la disponibilité d'information auxiliaire sur les laboratoires pour proposer une méthode alternative lorsque seuls les résultats de mesures des laboratoires sont disponibles pour établir la valeur de consensus

1.5 Classification des méthodes

Comme annoncé à la section 1.3.3, on propose une classification des méthodes rencontrées dans la revue de littérature selon les hypothèses de ces méthodes concernant la nature des biais de mesure et la donnée de l'incertitude de mesure.

La classification est proposée à la figure 1.2.

Toutes les méthodes présentées permettent d'estimer la valeur de consensus et son incertitude associée en tirant profit de toute l'information disponible sur les mesures.

Les différentes méthodes ont également mis en évidence la volonté de modéliser les biais de mesure des laboratoires, notamment au moyen d'informations supplémentaires sur les incertitudes renvoyées, à demander aux laboratoires.

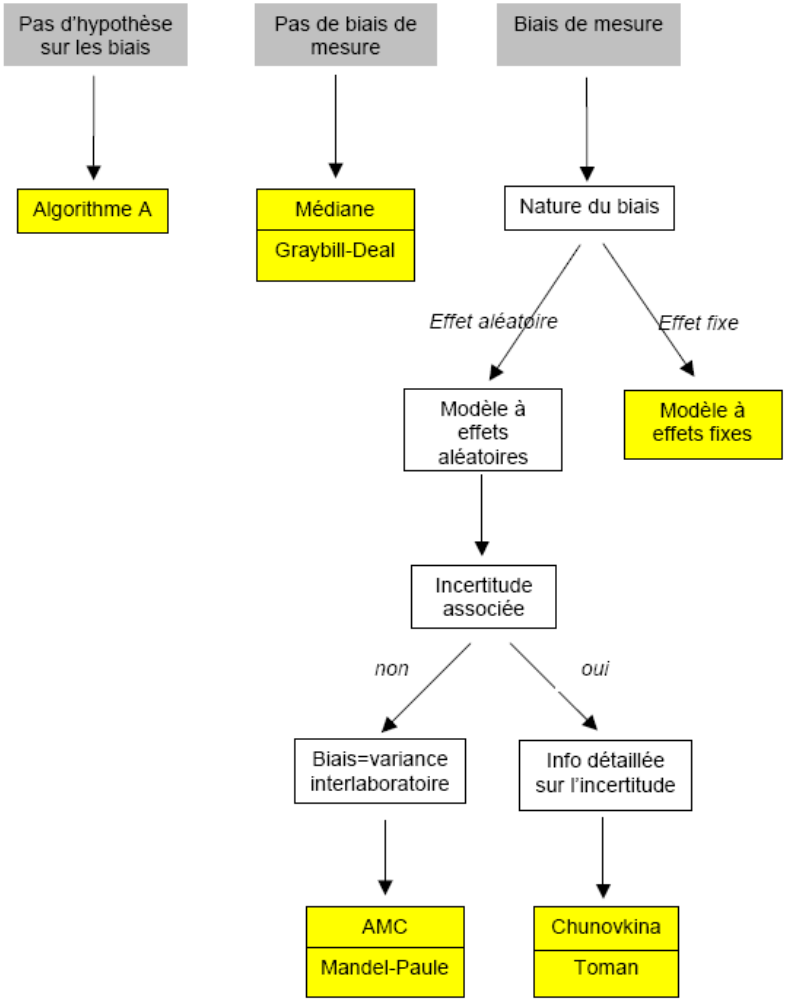


FIGURE 1.2 – Classification des méthodes présentées dans la revue de littérature

1.6 Nouvelle approche de modélisation des biais de mesure

1.6.1 Objectifs d'une approche alternative de modélisation des biais de mesure

Parmi les méthodes présentées, l'existence d'un biais de mesure est prise en compte soit par une augmentation en moyenne des résultats des laboratoires soit par une augmentation de l'incertitude des résultats mais aucune approche rencontrée dans la littérature ne permet d'identifier les sources de biais ni de les quantifier.

Nous proposons un effort supplémentaire de modélisation des biais de mesure avec le triple objectif suivant

- identifier et quantifier les sources de biais à l'aide d'expertise
- proposer une estimation des biais encore moins dépendante des résultats des autres

laboratoires
- améliorer la fiabilité de la valeur de consensus et de son incertitude associée

Pour cela nous supposons l'existence d'**information auxiliaire explicative des biais de mesure** accessible au statisticien sous forme de **variables explicatives des biais de mesure** grâce à des experts des processus de mesure.

1.6.2 Mise en oeuvre de la nouvelle approche de modélisation des biais de mesure

La mise en oeuvre de l'approche que nous proposons se décompose en cinq étapes
- **Etape 1 :** recueil d'information auxiliaire pertinente sur les sources de biais
- **Etape 2 :** transformation de l'information auxiliaire en variables explicatives du biais de mesure
- **Etape 3 :** formulation du modèle des biais de mesure
- **Etape 4 :** analyse du modèle des biais de mesure
- **Etape 5 :** raccordement du modèle des biais de mesure aux résultats de mesure

L'étape 1 consiste en le recueil d'une information auxiliaire sur les biais de mesure. Cette information peut être déjà à disposition de l'organisateur de la comparaison interlaboratoires ou demandée en routine en accompagnement du résultat de mesure ou encore obtenue après renseignement d'un questionnaire validé par des experts.

L'étape 2 consiste à modéliser chaque source de biais identifiée à l'étape 1 par une variable dont la réalisation sur les laboratoires est soit continue, soit binaire, soit catégorielle ordonnée telle que **l'ordre des modalités ou l'intensité des réalisations aient une interprétation naturelle en terme de qualité de la pratique du laboratoire**.

Les étapes 1 et 2 sont illustrées à partir de notre propre expérience au chapitre 5.

L'étape 3 est véritablement le coeur de la nouvelle approche que nous proposons et consiste à structurer les variables de biais construites à l'étape 2 au sein d'un modèle des biais. La modélisation des biais de mesure est réalisée grâce à une modélisation par équations structurelles (SEM) des variables de biais, reposant sur des avis d'expert, où chaque variable de biais observée est associée à une unique variable "résumée" latente représentant une composante du biais et où la structure de corrélation des variables observées est transférée dans la structure de corrélation des variables latentes. Le modèle structurel ainsi construit représente la structure des biais.

En pratique il est courant de considérer les trois composantes suivantes du biais de mesure : la préparation de la mesure, la mesure et le contrôle qualité dont les relations de causalité sont représentées à la figure 1.3.

L'étape 4 d'estimation bayésienne (justifiée à la section 1.6.3) du modèle des biais apporte l'information supplémentaire pour étudier les contributions des différentes sources de biais au biais des laboratoires, et pour améliorer la fiabilité de la valeur de consensus, de son incertitude associée et des estimations des biais de mesure.

L'étape 5 repose sur une nouvelle modélisation et un nouveau traitement des résultats de comparaisons interlaboratoires. Le cadre bayésien de l'analyse des biais de l'étape 4 permet la modélisation et l'estimation des indicateurs de la comparaison interlaboratoires

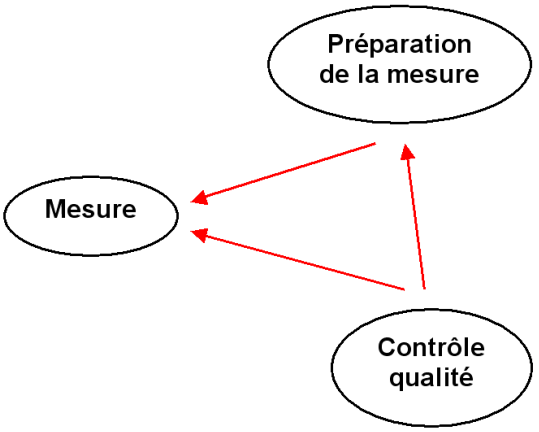


FIGURE 1.3 – Relations de causalité entre trois composantes (latentes) du biais de mesure.

conditionnellement au modèle des biais.

1.6.3 Pourquoi une approche bayésienne ?

La démarche s'interprète directement en termes d'information *a priori* et d'information *a posteriori* relativement à la donnée des résultats de comparaisons interlaboratoires définissant de fait un cadre bayésien au traitement des données de comparaisons interlaboratoires.

Ainsi l'information auxiliaire sur les biais a le rôle d'une information *a priori* sur les résultats de mesure, obtenue indépendamment des résultats de mesure.

Cette information est combinée aux résultats de mesure dans le but de mettre à jour la connaissance de la valeur de consensus, de son incertitude associée et des biais de mesure et d'étudier les contributions des variables de biais au biais des laboratoires *a posteriori* .

Le choix d'un cadre bayésien d'estimation du modèle des biais permet une démarche pleinement bayésienne du traitement des données de comparaisons interlaboratoires bénéficiant de l'étude conjointe (i.e. simultanée) de toutes les grandeurs en jeu.

Le chapitre 2 présente les outils théoriques de la statistique bayésienne utilisés dans ces travaux.

Le chapitre 3 développe la méthode d'estimation du modèle structurel des biais dans un cadre bayésien.

Le chapitre 4 décline l'utilisation du modèle des biais dans le traitement des données de comparaisons interlaboratoires selon la typologie de la comparaison établie à la section

Le chapitre 5 illustre les cinq étapes nécessaires à la mise en oeuvre de la nouvelle approche sur deux applications.

Chapitre 2

Les méthodes bayésiennes

L'objectif de ce chapitre est de présenter les outils théoriques utilisés dans le développement de l'approche alternative introduite à la fin du premier chapitre et développée au chapitre 3. Précisément, ces outils interviennent dans l'estimation des modèles permettant d'interpréter l'information auxiliaire sur les laboratoires.

Pour des raisons qui apparaîtront au chapitre suivant nous avons choisi de mettre en oeuvre les méthodes de Monte Carlo par Chaînes de Markov (MCMC), qui reposent sur une spécification bayésienne des modèles à estimer.

Les méthodes MCMC sont en fait les descendantes de la méthode de Monte-Carlo qui a en son temps marqué un renouveau dans les méthodes d'échantillonnage statistique. Pour l'anecdote, on notera même que c'est l'un des protagonistes à l'origine de la méthode de Monte Carlo que l'on retrouvera à l'origine des méthodes MCMC : le fameux Nicholas Metropolis.

On ne peut pas séparer la présentation de ces méthodes du contexte historique qui a mené à leur développement : la seconde guerre mondiale et les problématiques liées à la mise au point de la bombe atomique. C'est l'objet de la première section de ce chapitre.

La deuxième section rappelle les fondements de la statistique bayésienne en développant plus particulièrement la spécification bayésienne des modèles linéaires, utilisée au chapitre 3.

La troisième section présente les notions et les résultats principaux relatifs aux méthodes MCMC permettant de comprendre et justifier l'utilisation de telles méthodes de simulations. On développera en particulier les algorithmes de Gibbs et de Metropolis-Hastings.

La dernière partie présente une méthode permettant d'améliorer les propriétés de convergence des algorithmes de Gibbs, l'expansion paramétrique, et donne les idées principales pour comprendre la validité de la méthode.

De nos jours les statisticiens font appel aux méthodes de simulations MCMC dans des applications variées mais il faut garder à l'esprit que ces méthodes ont été développées à partir d'applications en physique. Nous sommes même persuadés que pour bien comprendre ces méthodes il est nécessaire de se reporter au problème physique initial, à savoir l'étude de l'équilibre des systèmes qui sera le fil rouge de ce chapitre.

2.1 Historique : les méthodes de Monte-Carlo

2.1.1 Cadre

Les méthodes de Monte Carlo ont été développées à Los Alamos (Nouveau Mexique) pendant la seconde guerre mondiale pour répondre au besoin de simulations dans le cadre du projet Manhattan de développement de la première arme nucléaire aux Etats-Unis par les Alliés.

Cependant les méthodes Monte Carlo ont commencé à être étudiées en profondeur suite à l'arrivée du premier ordinateur électronique en 1945, l'ENIAC, pour répondre au nouveau contexte international de guerre froide.

2.1.2 Années 40 : le premier ordinateur électronique et les premières simulations de Metropolis

Le premier ordinateur électronique l'ENIAC (Electronic Numerical Integrator and Computer) a été développé à l'Université de Pennsylvanie à Philadelphie pour l'armée américaine à Aberdeen Proving Grounds (Maryland) à partir de l'année 1943 avec l'objectif de calculer les trajectoires en balistique dix fois plus vite que les machines existantes.

C'est John von Neumann, alors consultant à Los Alamos et à Aberdeen, qui a compris les perspectives que cette machine pouvait apporter aux chercheurs travaillant sur la bombe atomique à Los Alamos, en particulier Edward Teller et son groupe.

Peu avant, en 1942, Nicholas Metropolis travaillait avec Edward Teller à l'université de Chicago au développement d'un réacteur, quand J. Robert Oppenheimer leur a offert de continuer leurs travaux à Los Alamos. C'est ainsi que Metropolis a rejoint le projet Manhattan dans la division théorique. Ses premiers travaux ont été de développer les équations d'états pour des matériaux à hautes température, pression et densité.

John von Neumann a alors demandé à Nicholas Metropolis et à Stanislaw Frankel de préparer des modèles de réactions thermonucléaires à calculer par l'ENIAC, comme tests de l'ordinateur.

La guerre s'est terminée avant la fin des tests et le contexte incertain de l'après-guerre a motivé la poursuite des recherches avec l'ENIAC. Entre temps l'ENIAC avait pleinement suscité fascination et excitation autour de lui.

La présentation des résultats au printemps 1946 a non seulement suscité optimisme concernant la faisabilité de l'arme nucléaire mais aussi marqué le début d'une nouvelle ère en simulation incarnée en la personne de Stanislaw Ulam.

2.1.3 1946 : Stanislaw Ulam et la naissance des méthodes Monte carlo

Jusqu'aux années, 1940 les méthodes de simulations et d'échantillonnage statistique utilisaient les simulations pour tester un problème déterministe déjà bien compris. Stanislaw Ulam a apporté un regard nouveau [Metropolis, 1987]

> Stan's extensive mathematical background made him aware that statistical sampling techniques had fallen into desuetude because of the length and tediousness of

> the calculations. But with this miraculous development of the ENIAC–along with the applications Stan must have been pondering–it occurred to him that statistical techniques should be resuscitated, and he discussed this idea with von Neumann.
>
> NICHOLAS METROPOLIS, *The beginning of the Monte Carlo method*, Los Alamos Science, 1987

L'idée de Stanislaw Ulam est de traiter un problème déterministe en le transformant en un problème probabiliste dont la résolution repose sur l'introduction de nombres pseudo aléatoires.

L'origine de cette idée remonte à une période de convalescence de Stanislaw Ulam en 1946 pendant laquelle il s'occupait en jouant au solitaire [Eckhardt, 1987]

> The first thoughts and attempts I made to practice [the Monte Carlo method] were suggested by a question which occurred to me in 1946 as I was convalescing from an illness and playing solitaires. The question was what are the chances that a Canfield solitaire laid out with 52 cards will come out successfully ? After spending a lot of time trying to estimate them by pure combinatorial calculations, I wondered whether a more practical method than "abstract thinking" might not be to lay it out say one hundred times and simply observe and count the number of successful plays. This was already possible to envisage with the beginning of the new era of fast computers, and I immediately thought of problems of neutron diffusion and other questions of mathematical physics, and more generally how to change processes described by certain differential equations into an equivalent form interpretable as a succession of random operations. Later... [in 1946, I] described the idea to John von Neumann and we began to plan actual calculations.
>
> ROGER ECKHARDT, *Stan Ulam, John von Neumann, and the Monte Carlo Method*, Los Alamos Science, 1987

Les travaux, secrets, nécessitaient un nom de code : Monte Carlo [Metropolis, 1987]

> It was at that time that I suggested an obvious name for the statistical method-a suggestion not unrelated to the fact that Stan had an uncle who would borrow money from relatives because he "just had to go to Monte Carlo."The name seems to have endured.
>
> NICHOLAS METROPOLIS, *The beginning of the Monte Carlo method*, Los Alamos Science, 1987

2.1.4 Influences des méthodes développées à Los Alamos

Les travaux conjoints de Metropolis et Ulam ont donné lieu au premier papier sur les simulations de Monte Carlo en 1949 [Metropolis et Ulam, 1949].

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 247 SEPTEMBER 1949 *Volume 44*

THE MONTE CARLO METHOD

NICHOLAS METROPOLIS AND S. ULAM
Los Alamos Laboratory

We shall present here the motivation and a general description of a method dealing with a class of problems in mathematical physics. The method is, essentially, a statistical approach to the study of differential equations, or more generally, of integro-differential equations that occur in various branches of the natural sciences.

Peu après en 1953 Metropolis propose l'algorithme de Metropolis conjointement avec les Teller et les Rosenbluth [Metropolis *et al.*, 1953].

THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.

En conclusion, les recherches menées à Los Alamos sont aujourd'hui considérées comme les premières simulations informatiques et l'algorithme de Metropolis a récemment été considéré comme l'un des dix algorithmes ayant eu la plus grande influence sur le développement et la pratique des sciences et de l'ingénierie [Beichl et Sullivan, 2000].

2.1.5 Fil rouge : étude statistique d'un système à l'équilibre

Comme annoncé en introduction de ce chapitre, c'est l'application qui a motivé l'algorithme de Metropolis [Metropolis *et al.*, 1953] qui sera le fil rouge de ce chapitre.

Le fil rouge est donc l'étude de la distribution des états d'énergie des particules d'un système en fonction de la température du système.

En particulier on s'intéresse à la proportion de particules dans les différents états d'énergie **lorsque le système est à l'équilibre**.

On sera en particulier attentif au vocabulaire théorique qui est une transposition directe du phénomène physique. Ainsi le terme probabilité de mouvement propre aux chaînes de Markov est dans cet exemple exactement la probabilité que les particules d'un système bougent d'un état d'énergie à un autre.

Pour rappel utile, la constante de Boltzmann k_B relie les unités de température aux unités d'énergie.

La probabilité p_i d'occuper un état d'énergie E_i est proportionnelle au facteur de Boltzmann

$$p_i = \frac{1}{Z(T)} \exp -\frac{E_i}{k_B T} \tag{2.1}$$

où $Z(T) = \sum_i \exp -\frac{E_i}{k_B T}$ est la constante de normalisation

Le facteur de Boltzmann $\exp -\frac{E_i}{k_B T}$ est donc un poids affecté aux différents états d'énergie qui détermine la probabilité des particules d'être à l'état d'énergie i.

La distribution p_i est la distribution de Boltzmann, qui a la plus grande entropie pour un état d'énergie moyen E_i.

2.2 Cadre bayésien et formule de Bayes

2.2.1 Définitions de statistique classique

Un **paramètre** est une constante qui permet de contrôler un phénomène modélisé.

Fil rouge 1 *La température est un paramètre et on cherche à estimer la température à l'équilibre.*

Une **variable aléatoire** est une fonction définie sur l'ensemble Ω des résultats possibles d'une expérience aléatoire, telle qu'il soit possible de déterminer la probabilité pour qu'elle prenne une valeur donnée ou qu'elle prenne une valeur dans un intervalle donné. Les variables aléatoires peuvent être représentées par leur loi de probabilité ou leur densité de probabilité.

Formellement, une variable aléatoire réelle est une application mesurable de (Ω, C, P) dans $\mathbb{R}$ muni de sa tribu borélienne [Saporta, 2006].

Fil rouge 2 *Le niveau d'énergie est une variable aléatoire. Chaque particule peut atteindre n'importe quel niveau d'énergie en faisant éventuellement varier la température. On s'intéresse ici aux valeurs possibles de cette variable à la température d'équilibre.*

La **densité de probabilité** d'une variable aléatoire absolument continue représente la vraisemblance relative d'observer une réalisation de cette variable à un point donnée de l'espace d'observation.

Fil rouge 3 *La répartition des particules dans les différents états d'énergie, encore appelée distribution des particules, en référence à la façon dont les particules sont distribuées dans*

les niveaux d'énergie, donne une approximation de la densité de probabilité des différents états d'énergie.

Un **modèle statistique paramétrique** consiste en l'observation d'une variable aléatoire X, distribuée selon la densité de probabilité $p(X|\theta)$, où seul le paramètre θ est inconnu et appartient à un espace vectoriel de dimension finie.

La *vraisemblance* peut être vue comme le résultat de la modélisation des réalisations d'une variable aléatoire dans le cadre d'un modèle statistique paramétrique. Formellement, soit un n-échantillon d'une variable aléatoire X. On notera $p(x_1, \ldots, x_n)$ soit la densité jointe de $X_1, \ldots, X_n$ si X est absolument continue, soit la probabilité jointe $P(X_1 = x_1, \ldots, X_n = x_n)$ si X est discrète. $p(x_1, \ldots, x_n)$ considéré comme fonction de $\boldsymbol{\theta}$ seul est appelé vraisemblance de $\boldsymbol{\theta}$ ([Saporta, 2006]).

2.2.2 La nature des paramètres en statistique bayésienne

Au contraire de la statistique classique, en statistique bayésienne les paramètres sont considérés comme des variables aléatoires auxquelles on affecte une densité de probabilité.

La définition d'un modèle statistique paramétrique s'en trouve donc modifiée : il faut affecter des densités de probabilité aux paramètres au moment de la définition du modèle pour pouvoir prendre en compte leur nature aléatoire. Cette distribution est la distribution *a priori* des paramètres car elle est définie avant d'observer les résultats.

Un modèle statistique bayésien se compose donc d'un modèle statistique paramétrique $f(\boldsymbol{y}|\boldsymbol{\theta})$ et d'une distribution *a priori* sur le paramètre $\boldsymbol{\theta}$ [Robert, 2001].

Plus généralement les paramètres peuvent ne pas être les seules grandeurs inconnues du modèle, il faut également compter avec les données manquantes, les variables latentes, etc, auxquelles on affecte une distribution *a priori* . Le cadre bayésien se prête ainsi volontiers à l'imputation des valeurs manquantes ou non observées.

2.2.3 Principe général de l'inférence des paramètres en statistique bayésienne

L'inférence des paramètres consiste à déterminer la densité de probabilité conjointe des grandeurs inconnues (paramètres, etc) à partir de toute l'information disponible sur les paramètres apportée par les données.

Cette remarque est connue sous le nom de *principe de vraisemblance* qui stipule que l'information apportée par une observation $\boldsymbol{x}$ sur le paramètre $\boldsymbol{\theta}$ est entièrement contenue dans la fonction de vraisemblance $l(\boldsymbol{\theta}|\boldsymbol{x})$ ([Robert, 2001]).

L'inférence repose donc sur les notions de distribution conjointe et de distributions conditionnelles qui sont l'essence des méthodes bayésiennes et que l'on définit ci-après.

2.2.4 Distributions jointes et distributions conditionnelles

Soit $X_1, \ldots, X_n$ des variables aléatoires réelles. On définit la *fonction de répartition* multidimensionnelle $F : \mathbb{R}^n \to [0, 1]$ par

$$F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) \tag{2.2}$$

On définit la *densité jointe* de $X_1, \ldots, X_n$, $f_{X_1,\ldots,X_n}$ comme la densité multidimensionnelle associée à la fonction de répartition multidimentionnelle

$$F_{X_1,\ldots,X_n}(X_1, \ldots, X_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f_{X_1,\ldots,X_n}(u_1, \ldots, u_n)\, \mathrm{d}u_1 \ldots \mathrm{d}u_n \tag{2.3}$$

En pratique on factorise la loi jointe en un produit de grandeurs indépendantes. Ainsi en présence d'observations indépendantes identiquement distribuées (iid) la vraisemblance se factorise en le produit des densités individuelles.

La *densité marginale* d'une variable s'obtient en intégrant la densité conjointe sur les autres variables

$$f_{X_i}(x_i) = \int \ldots \int f_{X_1,\ldots,X_n}(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)\, \mathrm{d}x_1 \ldots \mathrm{d}x_{i-1} \mathrm{d}x_{i+1} \ldots \mathrm{d}x_n \tag{2.4}$$

La *densité conditionnelle* $f_{X_n|X_1,\ldots,X_{n-1}}$ d'une variable X_n conditionnellement à toutes les autres variables est définie par

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_n|X_1,\ldots,X_{n-1}}(x_n|x_1, \ldots, x_{n-1})\, f_{X_1,\ldots,X_{n-1}}(x_1, \ldots, x_{n-1}) \tag{2.5}$$

En poursuivant la décomposition de $f_{X_1,\ldots,X_{n-1}}(x_1, \ldots, x_{n-1})$ on obtient la factorisation suivante de la densité conjointe

$$\begin{aligned} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) &= f_{X_n|X_1,\ldots,X_{n-1}}(x_n|x_1, \ldots, x_{n-1}) && (2.6)\\ & f_{X_{n-1}|X_1,\ldots,X_{n-2}}(x_{n-1}|x_1, \ldots, x_{n-2}) && (2.7)\\ & \ldots && (2.8)\\ & f_{X_2|X_1}(x_2|x_1) && (2.9)\\ & f_{X_1}(x_1) && (2.10) \end{aligned}$$

En réécrivant l'expression 2.5 et en renversant les densités conditionnelles on obtient

$$\begin{aligned} f_{X_n|X_1,\ldots,X_{n-1}} &= \frac{f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)}{f_{X_1,\ldots,X_{n-1}}(x_1, \ldots, x_{n-1})} && (2.11)\\ &= \frac{f_{X_1,\ldots,X_{n-1}|X_n}(x_1, \ldots, x_{n-1}|x_n)\, f_{X_n}(x_n)}{f_{X_1,\ldots,X_{n-1}}(x_1, \ldots, x_{n-1})} && (2.12) \end{aligned}$$

qui n'est autre que la formule de Bayes présentée à la section 2.2.6, exprimée en termes de distributions *a priori* et *a posteriori* .

Dans la suite, en notant $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ le vecteur de paramètres à estimer, l'inférence bayésienne consiste à déterminer la densité conjointe de tous les paramètres conditionnellement aux données $\boldsymbol{y}$, à savoir la densité $\pi(\theta_1, \ldots, \theta_K|\boldsymbol{y})$.

2.2.5 Notations

Pour plus de clarté dans ce chapitre on prend les notations suivantes pour les différentes distributions rencontrées

- $\pi(\boldsymbol{\theta})$ pour la distribution *a priori* de $\boldsymbol{\theta}$
- $f(\boldsymbol{y}|\boldsymbol{\theta})$ pour la vraisemblance des données $\boldsymbol{y}$
- $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ pour la distribution *a posteriori* de $\boldsymbol{\theta}$
- $m(\boldsymbol{y})$ pour la distribution marginale de $\boldsymbol{y}$

Dans le chapitre suivant les densités de probabilité seront désignées par la notation générique entre crochets [ ].

2.2.6 La formule de Bayes

La formule de Bayes est un mécanisme d'apprentissage qui combine l'information *a priori* sur les paramètres et l'information apportée par les données en une information *a posteriori* sur les paramètres.

Précisément la formule de Bayes s'applique à la densité conjointe *a priori* des paramètres et renvoie la densité conjointe des paramètres *a posteriori* c'est-à-dire conditionnellement aux observations.

La formule de Bayes appliquée à des densités de probabilité s'écrit

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{m(\boldsymbol{y})} \tag{2.13}$$

où $m(\boldsymbol{y})$ est la distribution marginale des données qui a ici le rôle d'une constante de normalisation pour que la densité *a posteriori* s'intègre bien à 1.

Puisque $\boldsymbol{y}$ et $m(\boldsymbol{y})$ sont connus *a posteriori* , la distribution *a posteriori* de $\boldsymbol{\theta}$ évaluée en $\boldsymbol{y}$ est proportionnelle à $f(\boldsymbol{y}|\boldsymbol{\theta})\,p(\boldsymbol{\theta})$. La formule 2.13 peut donc être réécrite sous la forme suivante, où $\propto$ est un symbole de proportionnalité

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}) \tag{2.14}$$

Cette expression montre que l'inférence bayésienne vérifie le principe de vraisemblance : *a posteriori* , l'information provenant des données provient exclusivement de la vraisemblance $f(\boldsymbol{y}|\boldsymbol{\theta})$.

La formule de Bayes et le calcul des différentes lois marginales qui en découle font appel à de nombreux calculs d'intégrales (c.f. 2.2.4) multidimensionnelles pour lesquelles la forme analytique peut être très complexe. Les méthodes de simulations de Monte Carlo permettent alors d'approcher ces intégrales.

2.2.7 Les distributions conjuguées

Le choix de modéliser la distribution *a priori* par une distribution conjuguée avec la vraisemblance permet une interprétation claire de la formule de Bayes, en tant qu'un

processus de mise à jour des paramètres de la distribution *a priori* qui donne la distribution *a posteriori* , formalisé dans la définition suivante ([Robert, 2001]).

Définition 2.2.1 *Une famille $\mathcal{F}$ de distributions de probabilité sur Θ est conjuguée pour une fonction de vraisemblance $f(\boldsymbol{y}|\boldsymbol{\theta})$ si, pour tout $\pi \in \mathcal{F}$, la distribution a posteriori $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ appartient aussi à $\mathcal{F}$.*

Une distribution *a priori* conjuguée est également un choix objectif pour la distribution *a priori* où seuls les hyperparamètres restent à déterminer.

Cependant ce choix est basé sur des considérations de commodité de calcul qui peuvent être discutables selon les applications.

2.2.8 Grandeurs d'intérêt *a posteriori*

La distribution *a posteriori* $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ comprend toute l'information disponible sur les paramètres.

Lorsque θ est unidimensionnel la représentation graphique de $\pi(\theta|\boldsymbol{y})$ permet de visualiser l'incertitude associée à l'estimation du paramètre.

La distribution *a posteriori* peut en particulier fournir des estimations ponctuelles des paramètres, sous la forme d'une moyenne *a posteriori* et d'une variance *a posteriori* obtenues à partir de calculs d'espérance et de variance (resp.) de lois conditionnelles.

L'espérance *a posteriori* de θ est la densité de probabilité *a posteriori* de θ moyennée sur toutes les valeurs possibles de θ sachant que $\boldsymbol{y}$ a été observé et s'écrit

$$E(\theta|Y=\boldsymbol{y}) = \int \theta\pi(\theta|\boldsymbol{y})\,\mathrm{d}\theta \tag{2.15}$$

Remarque : On peut montrer ([Robert, 2001]) que la moyenne *a posteriori* est l'estimateur qui minimise une fonction de coût quadratique.

La variance *a posteriori* de θ s'écrit

$$V(\theta|\boldsymbol{y}) = E^{\theta|\boldsymbol{y}}\left(\theta^2\right) - \left[E^{\theta|\boldsymbol{y}}(\theta)\right]^2 \tag{2.16}$$

$$= \int \theta^2\pi(\theta|\boldsymbol{y})\,\mathrm{d}\theta - \left[\int \theta\pi(\theta|\boldsymbol{y})\,\mathrm{d}\theta\right]^2 \tag{2.17}$$

Ces quantités peuvent être obtenues de manière analytique dans le cas de distributions conjuguées et de modèles simples. En pratique ces quantités sont calculées à partir de simulations qui permettent de s'affranchir de la conjugaison et de calculer dans des modèles complexes.

La distribution *a posteriori* permet aussi de construire la distribution de probabilité de nouvelles observations $\tilde{\boldsymbol{y}}$ issues du même modèle conditionnellement aux données déjà observées $\boldsymbol{y}$, appelée *distribution predictive a posteriori* et définie par

$$p\left(\tilde{\boldsymbol{y}}|\boldsymbol{y}\right) = \int p\left(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}|\boldsymbol{y}\right) \mathrm{d}\boldsymbol{\theta} \tag{2.18}$$

2.2.9 La distribution *a posteriori* : un compromis entre les données et l'information *a priori*

Gelman [Gelman *et al.*, 2004] fait remarquer que l'on pourrait s'attendre à ce que la variance *a posteriori* des paramètres soit plus petite que la variance *a priori* parce que l'on prend en compte de l'information sur les paramètres à travers les données.

Cette remarque est vraie en moyenne car la variance *a priori* de $\boldsymbol{\theta}$ se décompose selon la formule

$$V\left(\boldsymbol{\theta}\right) = E^{\boldsymbol{y}}\left(V\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right) + V^{\boldsymbol{y}}\left(E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right) \tag{2.19}$$

qui implique que la variance *a posteriori* $V\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$ est en moyenne plus petite que la variance *a priori* $V\left(\boldsymbol{\theta}\right)$. La différence $V^{\boldsymbol{y}}\left(E\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right)$ dépend de la variabilité de la moyenne *a posteriori* sur la distribution des données possibles. Gelman en déduit que plus cette variabilité est grande plus la variabilité *a posteriori* de $\boldsymbol{\theta}$ est petite.

Plus généralement, en ne raisonnant plus sur les espérances, la moyenne *a posteriori* et la variance *a posteriori* peuvent être vus comme des compromis entre l'information dans les données et l'information *a priori* . Les poids associés à ces deux sources d'informations dépendent respectivement de la variabilité des données observées et de la variabilité de l'information *a priori* . Plus la taille de l'échantillon augmente plus le poids relatif de l'*a priori* diminue.

Le processus de mise à jour bayésien est illustré dans l'exemple 2.2.1.

Exemple 2.2.1 *Calcul de la moyenne a posteriori d'un échantillon gaussien de variance connue Soit* $\boldsymbol{y} = (y_1, \ldots, y_n)$ *où* $y_i \sim N\left(\mu, \sigma^2\right)$ *avec* σ^2 *connu. Le choix d'une distribution a priori conjuguée à la vraisemblance nous amène à choisir distribué selon une loi normale, par exemple* $\mu \sim N\left(\mu_0, \sigma_0^2\right)$ *où* μ_0 *et* σ_0^2 *sont des hyperparamètres connus, mais conservés sous leur forme littérale pour le propos.*

La distribution a posteriori est donc normalement distribuée $\mu|\boldsymbol{y} \sim N\left(\mu_n, v_n\right)$. *Un rapide calcul nous donne*

$$\mu_n = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}} = p_0\mu_0 + p_n\bar{y} \tag{2.20}$$

$$\frac{1}{v_n} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2/n} \tag{2.21}$$

où les quantités $p_0 = \frac{\frac{1}{\sigma_0^2}}{\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2/n}}$ *et* $p_n = \frac{\frac{1}{\sigma^2/n}}{\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2/n}}$ *représentent respectivement le poids de l'a priori et le poids des observations dans la construction de la moyenne a posteriori .*

La quantité $\frac{1}{v_n}$ *est la précision a posteriori , et s'exprime comme la somme de la précision a priori* $\frac{1}{\sigma_0^2}$ *et de la précision de l'échantillon observé* $\frac{1}{\sigma^2/n}$.

On observe que quand le nombre d'observations n tend vers l'infini ou lorsque la variance a priori tend vers l'infini à n fixé, le poids p_0 *de l'a priori tend vers* 0 *et* p_n *tend vers 1 de telle sorte que la distribution a posteriori peut être approchée par* $\mu|\boldsymbol{y} \sim N\left(\bar{y}, \sigma^2/n\right)$, *qui est un résultat de statistique fréquentielle.*

De la même façon, à n fixé, lorsque l'a priori se resserre à savoir σ_0^2 *tend vers* 0, *alors* p_0 *tend vers* 1 *et* p_n *tend vers 0, supprimant de fait l'influence des données dans le calcul de la distribution a posteriori .*

En conclusion, le mécanisme bayésien permet de s'adapter aux différents cas rencontrés en pratique, en jouant sur la variabilité de l'information *a priori* . En particulier, un petit nombre d'observations peut être compensé par une croyance plus forte dans l'*a priori* . Réciproquement, l'information *a priori* s'efface en présence d'un grand nombre de données suffisamment précises.

2.3 Spécification bayésienne du modèle de régression linéaire

Cette section a pour but de présenter le cadre bayésien des modèles de régression linéaire standard.

En particulier on montre comment obtenir la spécification de la distribution *a priori* des paramètres conjuguée avec la vraisemblance dans le cas du modèle linéaire standard, qui se généralise immédiatement au modèle linéaire généralisé.

Les distributions *a posteriori* sont ensuite calculées sous ces distributions *a priori* .

Enfin on aborde la formulation hiérarchique des modèles de régression qui correspond en statistique classique au modèle mixte.

2.3.1 Le modèle de régression linéaire univarié

2.3.1.1 Rappel

La régression est utilisée pour comprendre la relation entre les manifestations d'un phénomène et des variables explicatives de ce phénomène. On se place dans le cadre où le phénomène observé pour chaque individu peut être modélisé par une unique variable aléatoire réelle. On appelle alors y_i la variable aléatoire modélisant le phénomène observé pour l'individu i et on suppose que les valeurs de chaque individu sur cette variable sont indépendantes et ont été générées par le même processus (identiquement distribuées). On appelle x_{ik} la valeur observée pour l'individu i sur la variable explicative k où chaque variable explicative est ici considérée continue.

Le modèle de régression linéaire suppose une relation linéaire entre les variables explicatives et la variable à expliquer, de la forme

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \varepsilon_i \tag{2.22}$$
$$= \boldsymbol{x_i^t}\boldsymbol{\beta} + \varepsilon_i,\ i = 1 \ldots n \tag{2.23}$$

avec $\varepsilon_i \sim N\left(0, \sigma^2\right)$ où σ^2 représente la variabilité des données non expliquée par le modèle, appelée de fait variance résiduelle.

Les coefficients de régression $\beta_1, \ldots, \beta_K$ caractérisent la relation entre les différentes variables en jeu. La précision des estimations des coefficients de régression dépend de la variance résiduelle.

Le modèle de régression linéaire est souvent représenté sous la forme matricielle suivante

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.24}$$

où $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$, $\boldsymbol{X}$ est la matrice de terme général x_{ik}, $i = 1, \ldots, n$, $k = 1, \ldots, K$ et $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$.

La présence d'une constante dans le modèle se traduit par une colonne de 1 dans la première colonne de x, c'est-à-dire $x_{i1} = 1$ pour tout i.

2.3.2 Spécification bayésienne et estimation des paramètres du modèle de régression linéaire univarié

2.3.2.1 Construction de la distribution jointe *a priori* conjuguée

La spécification bayésienne du modèle de régression linéaire consiste à affecter des distributions *a priori* aux paramètres $\beta_1, \ldots, \beta_K, \sigma^2$, c'est-à-dire à construire la distribution $\pi\left(\boldsymbol{\beta}, \sigma^2\right)$.

Une approche possible est de construire une distribution *a priori* conjuguée avec la vraisemblance. Pour cela on cherche une distribution jointe $\pi\left(\boldsymbol{\beta}, \sigma^2\right)$ qui ait la même forme que la vraisemblance.

La vraisemblance s'exprime sous la forme

$$f\left(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right) \propto \frac{1}{\left(\sigma^2\right)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^t \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right) \tag{2.25}$$

On remarque que la vraisemblance est une forme quadratique en $\boldsymbol{\beta}$.

En introduisant $\hat{\boldsymbol{\beta}}$, l'estimateur des moindres carrés de $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X^t X}\right)^{-1} \boldsymbol{X^t y}$, la forme quadratique dans 2.25 se décompose selon

$$\left(y - \boldsymbol{X}\boldsymbol{\beta}\right)^t \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^t \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^t \boldsymbol{X^t X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \tag{2.26}$$
$$= \nu s^2 + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^t \boldsymbol{X^t X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \tag{2.27}$$

où $\nu = n - K$ et $\nu s^2 = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^t \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)$ est la somme des carrés résiduels.

En remplaçant la forme quadratique de l'expression 2.25 par sa décomposition 2.27, la vraisemblance se factorise sous la forme suivante

$$f\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2\right) \propto \frac{1}{\left(\sigma^2\right)^{(n-K)/2}} \exp\left(-\frac{vs^2}{2\sigma^2}\right) \frac{1}{\left(\sigma^2\right)^{K/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)^t \boldsymbol{X^t X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right) \tag{2.28}$$

Cette factorisation nous amène à la factorisation suivante de la distribution jointe *a priori*

$$\pi\left(\boldsymbol{\beta},\sigma^2\right) = \pi\left(\sigma^2\right)\pi\left(\boldsymbol{\beta}|\sigma^2\right) \tag{2.29}$$

où on reconnaît de plus que $\pi\left(\sigma^2\right)$ est distribué selon une distribution inverse-χ^2 et $\pi\left(\boldsymbol{\beta}|\sigma^2\right)$ est distribué selon une loi normale multivariée.

La spécification de ces distributions nécessite la donnée des hyperparamètres v_0, s_0^2, $\boldsymbol{\beta_0}$ et H_0 conduisant au choix suivant des distributions *a priori*

$$\pi\left(\sigma^2\right) \sim \text{Inv-}\chi^2\left(\nu_0, s_0^2\right) \tag{2.30}$$

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right) \sim N\left(\boldsymbol{\beta_0}, \sigma^2\boldsymbol{H_0^{-1}}\right) \tag{2.31}$$

Le choix des hyperparamètres peut s'avérer complexe. Dans un cadre non informatif on pourra prendre $\nu_0 = 2$, s_0^2 égal à une valeur centrale, de même que $\boldsymbol{\beta_0}$. Quant à $\boldsymbol{H_0}$, le choix de $\boldsymbol{H_0} = \boldsymbol{Id_K}$ correspond à une régression Ridge et le choix $\boldsymbol{H_0}^{-1} = g\left(\boldsymbol{X^t X}\right)^{-1}$, où g est une constante, renvoie au g-prior de Zellner.

Pour plus de détails on renvoie le lecteur à l'ouvrage [Marin et Robert, 2007].

2.3.2.2 Calcul de la distribution jointe *a posteriori*

D'après la formule de Bayes la distribution *a posteriori* de $\left(\boldsymbol{\beta},\sigma^2\right)$ est proportionnelle au produit de la vraisemblance et de la distribution *a priori*

$$\begin{aligned}\pi\left(\boldsymbol{\beta},\sigma^2|\boldsymbol{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X^t\beta}\right)^t\left(\boldsymbol{y}-\boldsymbol{X^t\beta}\right)\right) \\ \frac{1}{\left(\sigma^2\right)^{\nu_0/2}} \exp\left(-\frac{\nu_0 s_0^2}{2\sigma^2}\right) \\ \frac{1}{\left(\sigma^2\right)^{K/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\boldsymbol{\beta_0}\right)^t \boldsymbol{H_0}\left(\boldsymbol{\beta}-\boldsymbol{\beta_0}\right)\right)\end{aligned} \tag{2.32}$$

Lorsque la variance σ^2 est connue la distribution *a posteriori* de $\boldsymbol{\beta}$ est un compromis entre l'information *a priori* et l'information apportée par les données qui généralise les expressions obtenues dans l'exemple 2.2.1 à savoir

$$\pi\left(\boldsymbol{\beta}|\boldsymbol{y},\sigma^2\right) \sim N\left(\left(\boldsymbol{X^tX}+\boldsymbol{H_0}\right)^{-1}\left(\boldsymbol{X^tX}\hat{\boldsymbol{\beta}}+\boldsymbol{H_0}\boldsymbol{\beta}_0\right),\sigma^2\left(\boldsymbol{X^tX}+\boldsymbol{H_0}\right)^{-1}\right) \tag{2.33}$$

Dans la suite on note $\tilde{\boldsymbol{\beta}} = \left(\boldsymbol{X^tX}+\boldsymbol{H_0}\right)^{-1}\left(\boldsymbol{X^tX}\hat{\boldsymbol{\beta}}+\boldsymbol{H_0}\boldsymbol{\beta}_0\right)$.

Dans cette section σ^2 est supposé inconnu ce qui généralement le cas en pratique.

L'expression 2.33 donne la distribution conditionnelle de $\boldsymbol{\beta}$ sachant σ^2 qui est donc une partie de la distribution jointe *a posteriori* de $\left(\boldsymbol{\beta},\sigma^2\right)$. D'après la factorisation de la distribution *a priori* 2.29 il reste à calculer la distribution *a posteriori* de σ^2.

La distribution *a posteriori* s'obtient en reconstituant la forme quadratique en $\boldsymbol{\beta}$ dans 2.32. Cette étape nécessite la décomposition de Cholesky de $\boldsymbol{H_0} = \boldsymbol{U^tU}$

$$\begin{aligned}\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)^t\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)+\left(\boldsymbol{\beta}-\boldsymbol{\beta_0}\right)^t\boldsymbol{H_0}\left(\boldsymbol{\beta}-\boldsymbol{\beta_0}\right) &= \\ \left(\boldsymbol{y}-\boldsymbol{X\beta}\right)^t\left(\boldsymbol{y}-\boldsymbol{X\beta}\right)+\left(\boldsymbol{U\beta}-\boldsymbol{U\beta_0}\right)^t\left(\boldsymbol{U\beta}-\boldsymbol{U\beta_0}\right) &= \\ \left(\boldsymbol{\xi}-\boldsymbol{W\beta}\right)^t\left(\boldsymbol{\xi}-\boldsymbol{W\beta}\right) &\end{aligned} \tag{2.34}$$

où $\boldsymbol{\xi} = \left(\boldsymbol{y},\boldsymbol{U\beta_0}\right)$ et $W = \left(\boldsymbol{X},\boldsymbol{U}\right)$.

En faisant apparaître $\tilde{\boldsymbol{\beta}}$ dans la dernière ligne on a

$$\left(\boldsymbol{\xi}-\boldsymbol{W\beta}\right)^t\left(\boldsymbol{\xi}-\boldsymbol{W\beta}\right) = \left(\boldsymbol{\xi}-\boldsymbol{W}\tilde{\boldsymbol{\beta}}\right)^t\left(\boldsymbol{\xi}-\boldsymbol{W}\tilde{\boldsymbol{\beta}}\right)+\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)^t\boldsymbol{W^tW}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right) \tag{2.35}$$

$$= n\tilde{s}^2+\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)^t\boldsymbol{W^tW}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right) \tag{2.36}$$

où $n\tilde{s}^2 = \left(\boldsymbol{\xi}-\boldsymbol{W}\tilde{\boldsymbol{\beta}}\right)^t\left(\boldsymbol{\xi}-\boldsymbol{W}\tilde{\boldsymbol{\beta}}\right)$ est la somme des carrés résiduels de $\boldsymbol{\xi} = \left(\boldsymbol{y},\boldsymbol{U\beta}_0\right)$.

En reportant dans 2.32 on identifie les facteurs de la distribution *a posteriori* ; on retrouve bien la distribution conditionnelle *a posteriori* de $\boldsymbol{\beta}$ et on obtient de plus la distribution *a posteriori* de σ^2 sous la forme

$$\pi\left(\sigma^2|\boldsymbol{y}\right) = \frac{1}{\left(\sigma^2\right)^{\frac{n+\nu_0}{2}}}\exp\left(-\frac{1}{2\sigma^2}n\tilde{s}^2+\nu_0s_0^2\right) \tag{2.37}$$

où on reconnaît

$$\pi\left(\sigma^2|\boldsymbol{y}\right) \sim \text{Inv}-\chi^2\left(n+\nu_0,\frac{\nu_0s_0^2+n\tilde{s}^2}{n+\nu_0}\right) \tag{2.38}$$

2.3.2.3 Calcul de la distribution marginale *a posteriori* des coefficients de régression

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) = \int \pi\left(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2\right) \pi\left(\sigma^2|\boldsymbol{y}\right) \mathrm{d}\sigma^2 \tag{2.39}$$

$$= \int \frac{1}{(\sigma^2)^{\frac{n+\nu_0+K}{2}}} \exp\left(-\frac{1}{2\sigma^2} n\tilde{s}^2 + \nu_0 s_0^2 + \left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right)^t \boldsymbol{W}^t \boldsymbol{W} \left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right)\right) \mathrm{d}\sigma^2 \tag{2.40}$$

On reconnaît dans l'intégrale la forme non normalisée d'une distribution inverse-gamma.

La distribution marginale *a posteriori* est donc proportionnelle à

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) \propto \left[1 + \frac{\left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right)^t \boldsymbol{W}^t \boldsymbol{W} \left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right)}{(n+\nu_0)\, \frac{n\tilde{s}^2 + \nu_0 s_0^2}{n+\nu_0}}\right]^{-\frac{n+\nu_0+K}{2}} \tag{2.41}$$

qui est une distribution de Student multivariée de paramètres $n+\nu_0$, $\tilde{\boldsymbol{\beta}}$, $\frac{1}{\lambda^2}\boldsymbol{W}^t\boldsymbol{W}$ où $\lambda^2 = \frac{n\tilde{s}^2+\nu_0 s_0^2}{n+\nu_0}$ d'où

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) \sim t_{n+\nu_0}\left(\tilde{\boldsymbol{\beta}}, \frac{1}{\lambda^2}\left(\boldsymbol{X}^t\boldsymbol{X} + \boldsymbol{H_0}\right)\right) \tag{2.42}$$

2.3.3 Le modèle de régression linéaire multivarié

2.3.3.1 Rappel

Chaque unité statistique est désormais un vecteur de p observations correspondant à des réalisations de p variables à expliquer (par le modèle de régression)

$$\boldsymbol{y}_i = \left(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{ip}\right) \tag{2.43}$$

On suppose que chaque variable $\boldsymbol{y_j}$, $j = 1 \ldots p$ est reliée au même ensemble de prédicteurs $\boldsymbol{X} = \boldsymbol{X_1}, \ldots, \boldsymbol{X_k}$. Sous cette hypothèse le modèle de régression linéaire multivariée s'écrit

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{2.44}$$

où $\boldsymbol{y}$ est une matrice $n \times p$ d'observations p-variées, $\boldsymbol{X}$ est une matrice $n \times k$ de prédicteurs, $\boldsymbol{\theta}$ est la matrice $k \times p$ des coefficients de régression et $\boldsymbol{\varepsilon}$ est la matrice $n \times p$ des erreurs distribuée $\varepsilon \sim N(0, \boldsymbol{\Sigma})$.

Le fait de considérer les p régressions au sein du même modèle signifie que l'on s'intéresse de plus aux relations de dépendance entre les variables observées.

En effet la matrice de covariance marginale des variables observées est $cov(\boldsymbol{y_i}) = \boldsymbol{\theta}^t cov(\boldsymbol{X})\,\boldsymbol{\theta} + \boldsymbol{\Sigma}$ où les paramètres $\boldsymbol{\theta}$ et $\boldsymbol{\Sigma}$ sont estimés conditionnellement à la distribution jointe des variables observées.

Pour une représentation plus simple on introduit l'opérateur de vectorisation vec qui transforme une matrice en le vecteur issu de la concaténation verticale de ses colonnes. Ainsi $\text{vec}(\boldsymbol{y}) = \left(\boldsymbol{y_1^t}, \ldots, \boldsymbol{y_p^t}\right)^t$.

L'application de l'opérateur de vectorisation au modèle 2.44 donne

$$\text{vec}(\boldsymbol{y}) = \left(\boldsymbol{\theta}^t \otimes I_n\right) \text{vec}(\boldsymbol{X}) + \text{vec}(\boldsymbol{\varepsilon}) \tag{2.45}$$

où $\boldsymbol{\theta}^t \otimes I_n$ est le produit tensoriel entre $\boldsymbol{\theta}$ et la matrice identité en dimension n, I_n, c'est-à-dire la matrice $\boldsymbol{\theta}^t$ reportée n fois sur la diagonale de I_n, donnant une matrice de dimension $np \times nk$, $\text{vec}(\boldsymbol{y})$ et $\text{vec}(\boldsymbol{\varepsilon})$ sont des vecteurs de taille np et $\text{vec}(X)$ est un vecteur de taille nk.

Le modèle de régression linéaire multivarié est traité dans un cadre plus général dans [**?**] qui traitent en particulier le cas où $\boldsymbol{X}$ n'est pas commun à toutes les variables à expliquer.

2.3.3.2 Spécification bayésienne du modèle de régression linéaire multivarié

La vraisemblance de la distribution normale multivariée s'écrit

$$f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} \exp\left(-\frac{1}{2}\text{tr}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)^t \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)\right) \tag{2.46}$$

En notant $\hat{\beta} = \left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\boldsymbol{X}^t\boldsymbol{y}$ l'estimateur des moindres carrés et faisant apparaître la forme quadratique en $\boldsymbol{\beta}$ on a

$$\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)^t \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right) = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^t \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^t \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) \tag{2.47}$$

En reportant dans l'expression de la vraisemblance

$$f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto \frac{1}{|\boldsymbol{\Sigma}|^{\frac{n-k}{2}}} \exp\left(-\frac{1}{2}\text{tr}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^t \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\boldsymbol{\Sigma}^{-1}\right) \frac{1}{|\boldsymbol{\Sigma}|^{\frac{k}{2}}} \exp\left(-\frac{1}{2}\text{tr}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^t \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right) \tag{2.48}$$

Or $\text{tr}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^t \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right) = \text{vec}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^t \boldsymbol{\Sigma}^{-1} \otimes \left(\boldsymbol{X}^t\boldsymbol{X}\right)\text{vec}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)$

d'où la forme suivante de la vraisemblance

$$f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto \frac{1}{|\boldsymbol{\Sigma}|^{\frac{n-k}{2}}} \exp\left(-\frac{1}{2}\text{tr}\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^t \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\boldsymbol{\Sigma}^{-1}\right) \frac{1}{|\boldsymbol{\Sigma}|^{\frac{k}{2}}} \exp\left(-\frac{1}{2}\text{vec}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^t \boldsymbol{\Sigma}^{-1} \otimes \left(\boldsymbol{X}^t\boldsymbol{X}\right)\text{vec}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right) \tag{2.49}$$

L'observation de l'expression précédente fournit une factorisation similaire à celle obtenue à la section 2.3.2

$$\pi\left(\mathrm{vec}\left(\boldsymbol{\beta}\right),\boldsymbol{\Sigma}\right)=\pi\left(\boldsymbol{\Sigma}\right)\pi\left(\mathrm{vec}\left(\boldsymbol{\beta}\right)|\boldsymbol{\Sigma}\right) \tag{2.50}$$

où on reconnaît que $\pi\left(\boldsymbol{\Sigma}\right)$ est distribué selon une distribution Inverse Wishart et $\pi\left(\mathrm{vec}\left(\boldsymbol{\beta}\right)|\boldsymbol{\Sigma}\right)$ est distribué selon une loi normale multivariée.

On en déduit la spécification suivante des distributions *a priori*

$$\pi\left(\boldsymbol{\Sigma}\right)\sim\mathrm{Inv}-\mathrm{Wishart}_{\nu_0}\left(\left(\nu_0\boldsymbol{S_0}\right)^{-1}\right) \tag{2.51}$$

$$\pi\left(\mathrm{vec}\left(\boldsymbol{\beta}\right)|\boldsymbol{\Sigma}\right)\sim N_{kp}\left(\boldsymbol{\beta_0},\Sigma_0\otimes\boldsymbol{H_0}\right) \tag{2.52}$$

où la distribution Inverse Wishart est définie à l'annexe A.2.6 avec la convention que $\boldsymbol{S_0}$ est une matrice de covariance de dimension $k\times k$, $\boldsymbol{H_0}$ est une matrice de covariance $k\times k$, $\boldsymbol{\Sigma_0}$ est une matrice de covariance $p\times p$ et $\boldsymbol{\beta_0}$ est un vecteur de taille kp.

2.3.4 Modélisation hiérarchique

La modélisation hiérarchique consiste en une modélisation de la distribution *a priori* qui vérifie la définition suivante

Définition 2.3.1 ***Modèle hiérarchique*** *[Robert, 2001] Un modèle bayésien hiérarchique est un modèle statistique bayésien $f\left(\boldsymbol{x}|\boldsymbol{\theta}\right),\pi\left(\boldsymbol{\theta}\right)$ où la distribution a priori $\pi\left(\boldsymbol{\theta}\right)$ se décompose en distributions conditionnelles*

$$\pi_1\left(\boldsymbol{\theta}|\boldsymbol{\theta}_1\right),\pi_2\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2\right),\ldots,\pi_n\left(\boldsymbol{\theta}_{n-1}|\boldsymbol{\theta}_n\right) \tag{2.53}$$

et en une distribution marginale $\pi_{n+1}\left(\boldsymbol{\theta}_n\right)$ telle que

$$\pi\left(\boldsymbol{\theta}\right)=\int\pi_1\left(\boldsymbol{\theta}|\boldsymbol{\theta}_1\right),\pi_2\left(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2\right),\ldots,\pi_n\left(\boldsymbol{\theta}_{n-1}|\boldsymbol{\theta}_n\right)\pi_{n+1}\left(\boldsymbol{\theta}_n\right)\mathrm{d}\boldsymbol{\theta}_1\ldots\mathrm{d}\boldsymbol{\theta}_n \tag{2.54}$$

Les paramètres $\boldsymbol{\theta}_i$ sont les hyperparamètres du niveau $i, i=1\ldots n$.

La définition d'un modèle hiérarchique est liée à la notion d'échangéabilité, qui reflète l'indépendance conditionnelle des paramètres d'un niveau k par rapport aux paramètres du niveau $k+1$, obtenue par exemple dans le cas où $k=1$ lorsque $\pi_1\left(\boldsymbol{\theta}|\boldsymbol{\theta}_1\right)=\prod_{i=1}^{K}\left[\theta_i|\boldsymbol{\theta}_1\right]$. La spécification du modèle linéaire sous la forme

$$x_i\sim N\left(\theta_i,\sigma^2\right) \tag{2.55}$$

$$\theta_i\sim N\left(\mu,\tau^2\right) \tag{2.56}$$

vérifie donc l'échangéabilité en posant $\boldsymbol{\theta}_1=\mu$.

Plus généralement l'échangéabilité signifie que la distribution jointe d'un vecteur ne dépend pas de l'ordre des éléments de ce vecteur : ainsi en cas d'indépendance des éléments

du vecteur par rapport à un hyperparamètre, les distributions marginales individuelles conditionnellement à l'hyperparamètre sont égales. Ceci est formalisé dans la définition suivante

Définition 2.3.2 ***Echangéabilité*** *[Robert, 2001] Une suite $x_1, \ldots, x_n$ de variables aléatoires est échangeable si la distribution jointe $p(x_1, \ldots, x_n)$ est invariante sous n'importe quelle permutation des indices des variables aléatoires, c'est-à-dire*

$$p(x_1, \ldots, x_n) = p\left(x_{(1)}, \ldots, x_{(n)}\right) \tag{2.57}$$

En pratique il est confortable de construire une modélisation hiérarchique de la distribution *a priori* qui vérifie l'échangéabilité à chaque niveau de la modélisation ou qui vérifie l'échangéabilité dans des sous-populations.

Pour cela on peut s'appuyer sur le théorème de représentation de de Finetti, rappelé dans [Parent et Bernier, 2007] : *s*i l'on dispose d'une famille de lois de probabilité indicée par un paramètre ϕ (pour tout ϕ, f_ϕ est une densité de probabilité) et, pour ce paramètre, d'une densité de probabilité $g(\phi)$ alors une représentation de la loi de l'occurrence conjointe de toute collection de taille K de $\boldsymbol{\theta}$ sous la forme

$$p(\boldsymbol{\theta}) = \int_{\phi=0}^{1} \left(\prod_{i=1}^{K} f_\phi(\theta_i) \right) g(\phi) \, \mathrm{d}\phi \tag{2.58}$$

constitue un modèle d'occurrence conjointe des θ_i qui vérifie la propriété d'échangéabilité.

Ce théorème est immédiatement généralisable à une expression conditionnelle $p(\boldsymbol{\theta}|\boldsymbol{\theta}_1)$ qui vérifie l'échangéabilité par rapport à l'hyperparamètre $\boldsymbol{\theta}_1$

$$p(\boldsymbol{\theta}|\boldsymbol{\theta}_1) = \int_{\phi=0}^{1} \left(\prod_{i=1}^{K} f_\phi(\theta_i|\boldsymbol{\theta}_1) \right) g(\phi) \, \mathrm{d}\phi \tag{2.59}$$

et ainsi de suite jusqu'au dernier niveau.

La formulation hiérarchique apparaît par exemple lors de l'inclusion d'effets aléatoires dans un modèle linéaire [Robert, 2001]. La formulation suivante

$$\boldsymbol{y}|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \Sigma_1) \tag{2.60}$$
$$\boldsymbol{\theta}|\boldsymbol{\beta} \sim N(X\boldsymbol{\beta}, \Sigma_2) \tag{2.61}$$

correspond ainsi à une décomposition de la moyenne $\boldsymbol{\theta}$ en la somme d'effets fixes $X\boldsymbol{\beta}$ et d'effets aléatoires $Z\boldsymbol{\eta}$ où η suit une loi normale centrée de matrice de covariance Σ_2, caractéristique d'une formulation en terme de modèle mixte.

Pour illustrer cette double approche (hiérarchique et échangeable) le lecteur peut se reporter à l'article [Lindley et Smith, 1972] qui traite le cas des modèles linéaires pour lesquels les paramètres du modèle linéaire ont eux-mêmes une structure linéaire sur les hyperparamètres. Le cas d'échangéabilité sur des sous-populations est également étudié.

Pour une revue détaillée et illustrée de ces modèles et leur mise en oeuvre dans un cadre plus général on renvoie le lecteur à l'ouvrage [Gelman et Hill, 2006].

Jusqu'à présent dans ce chapitre nous avons pu calculer les distributions *a posteriori* à la main à partir de l'application directe de la formule de Bayes au problème posé.

Cependant les problèmes rencontrés en pratique sont souvent plus complexes et il est très difficile voire même impossible d'obtenir des distributions *a posteriori* explicites (voir la remarque à la fin de la section 2.2.6).

C'est la raison pour laquelle on fait appel pour de tels problèmes à des méthodes de simulation de la distribution jointe *a posteriori* . Les méthodes de simulations que nous avons choisi d'utiliser sont les méthodes Monte Carlo par Chaînes de Markov (MCMC).

La partie suivante a pour objectif de montrer les mécanismes des algorithmes MCMC et de décrire les deux principaux algorithmes MCMC utilisés en pratique.

2.4 Les méthodes de Monte Carlo par Chaînes de Markov (MCMC)

2.4.1 Point historique

Depuis quelques dizaines années les statisticiens font appel presque automatiquement aux méthodes de simulations MCMC (Markov Chain Monte Carlo) pour simuler des distributions multivariées complexes.

Comme on l'a vu en introduction de ce chapitre ces méthodes ont initialement été développées par des physiciens pour répondre à des problématiques bien précises.

Ainsi, l'algorithme de Metropolis remonte aux années cinquante avec l'article de [Metropolis *et al.*, 1953] appliqué aux distributions de Boltzmann et il faudra attendre les années soixante-dix pour une première généralisation de la méthode par [Hastings, 1970] (toujours à partir d'un problème physique sur l'énergie d'un système) et son nom actuel d'algorithme de Metropolis-Hastings . Pour comprendre l'engouement de Hastings, reportons nous à cette citation relevée dans [Rosenthal, 2004]

> When I returned to the University of Toronto, after my time at Bell Labs, I focused on Monte Carlo methods and at first on methods of sampling from probability distributions with no particular area of application in mind. [University of Toronto Chemistry professor] John Valleau and his associates consulted me concerning their work. They were using Metropolis's method to estimate the mean energy of a system of particles in a defined potential field. With 6 coordinates per particle, a system of just 100 particles involved a dimension of 600. When I learned how easy it was to generate samples from high dimensional distributions using Markov chains, I realised how important this was for Statistics, and I devoted all my time to this method and its variants which resulted in the 1970 paper.

L'algorithme de Gibbs a été développé dans les années quatre-vingt dans l'article séminal de [Geman et Geman, 1984] dans le cas particulier des distributions de Gibbs.

L'appropriation des méthodes MCMC par les statisticiens et la large diffusion de ces méthodes a attendu la montée en puissance des microprocesseurs.

De nos jours une grande variété de méthodes MCMC est à la disposition du statisticien, ne se réduisant pas aux algorithmes évoqués ci-dessus, même si ceux-ci sont encore largement utilisés dans leur forme originale. Ainsi ces algorithmes peuvent se combiner au besoin dans des algorithmes hybrides ou bien être adaptatifs.

En résumé, le monde des MCMC bouillonne d'algorithmes et d'astuces suivant la libre imagination du statisticien forcé de répondre aux contraintes des applications, pourvu que ces algorithmes satisfassent les critères de convergence vers la bonne distribution.

Une de ces astuces est l'expansion paramétrique qui permet d'améliorer la convergence des algorithmes de Gibbs en réduisant les autocorrélations dans les chaînes, qui sera l'objet de la dernière partie de ce chapitre.

2.4.2 Les méthodes MCMC

Les méthodes dites MCMC reposent sur la construction de chaînes de Markov ayant la propriété de converger vers la distribution que l'on cherche à simuler.

Cette section a pour but d'introduire et présenter les résultats essentiels justifiant l'utilisation de ces méthodes. L'exposé de la théorie est en particulier réalisée dans [Robert et Casella, 2004] et on trouvera une présentation plus accessible dans [Parent et Bernier, 2007] sur laquelle s'appuie la rédaction de cette section.

2.4.2.1 La propriété markovienne

Les chaînes de Markov sont des processus itératifs aléatoires qui génèrent des suites de réalisations dépendantes, qui vérifient la propriété suivante, dite propriété markovienne. Pour un vecteur de paramètres $\boldsymbol{\theta} \in \mathbb{R}^d$ et pour $A \in \mathcal{B}$, où $\mathcal{B}$ est un borélien de $\mathbb{R}^d$ on a

$$\left[\boldsymbol{\theta}^{(t+1)} \in A | \boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t-1)}, \ldots, \boldsymbol{\theta}^{(0)}\right] = \left[\boldsymbol{\theta}^{(t+1)} \in A | \boldsymbol{\theta}^{(t)}\right] := K\left(\boldsymbol{\theta}, A\right) \tag{2.62}$$

La propriété markovienne signifie que la probabilité conditionnelle de l'état futur $\boldsymbol{\theta}^{(t+1)}$ sachant tous les états passés $\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t-1)}, \ldots, \boldsymbol{\theta}^{(0)}$ ne dépend que de l'état présent $\boldsymbol{\theta}^{(t)}$.

Cette propriété est en particulier très appréciable lors de la programmation des algorithmes car seules les dernières valeurs des itérations ont besoin d'être stockées. L'aspect aléatoire de ces algorithmes traduit par le noyau de transition est concrètement mis en oeuvre grâce aux générateurs de nombres aléatoires des logiciels.

2.4.2.2 La transition entre états de la chaîne

L'expression $K\left(\boldsymbol{\theta}, A\right)$ qui apparaît à droite de l'expression 2.62 est le noyau de transition qui représente la probabilité de passer de l'état $\boldsymbol{\theta}$ de la chaîne à un état de l'ensemble A. Le noyau K est alors représenté sous sa forme intégrale et on peut le relier à sa forme différentielle k par la relation suivante

$$K(\boldsymbol{\theta}, A) = \int_A k(\boldsymbol{\theta}, \boldsymbol{t})\, \mathrm{d}t \tag{2.63}$$

L'expression différentielle intervient par exemple dans la définition de la densité de probabilité π_r d'un état $\tilde{\boldsymbol{\theta}}$ à l'itération r en fonction de la densité de probabilité π_{r-1} de l'état x à l'itération $r-1$

$$\pi_r\left(\tilde{\boldsymbol{\theta}}\right) = \int \pi_{r-1}(\boldsymbol{\theta})\, k\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}\right) \mathrm{d}\boldsymbol{\theta} \tag{2.64}$$

Plus généralement, l'expression différentielle sert à établir les équations de Chapman-Kolmogorov.

2.4.2.3 La distribution invariante de la chaîne

Pour garantir la convergence vers la distribution cible, la distribution cible doit être la distribution invariante (dite également stationnaire) de la chaîne de Markov ainsi construite.

La distribution invariante, notée ici π^*, satisfait la situation d'équilibre probabiliste suivante exprimée sur la forme intégrale du noyau de transition

$$\pi^*(A) = \int K(\boldsymbol{\theta}, A)\, \pi^*(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{2.65}$$

La même égalité exprimée sur la forme différentielle du noyau de transition donne

$$\pi^*\left(\tilde{\boldsymbol{\theta}}\right) = \int k\left(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}\right) \pi^*(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{2.66}$$

qui signifie que si une distribution est invariante pour une chaîne alors dès qu'un état $\boldsymbol{\theta}$ est distribué selon cette distribution l'état suivant est également distribué selon cette distribution.

Il reste alors à rappeler sous quelles conditions la distribution cible est l'unique distribution invariante de la chaîne et comment comment construire une chaîne ayant cette distribution comme distribution invariante.

Il est important de noter à ce stade que dans la mise en oeuvre pratique des méthodes MCMC le noyau de transition n'est pas utilisé sous une forme explicite. Il peut être reconstruit à partir des équations des algorithmes MCMC pour prouver la convergence des algorithmes, mais n'est en fait pas utilisé pour construire la chaîne de Markov.

2.4.2.4 Les équations de Chapman-Kolmogorov

Les équations de Chapman-Kolmogorov régissent les probabilités de transition d'un état à un autre en plusieurs étapes et permettent d'établir la convergence de la n^{e} itérée des probabilités de transition $\Pi^{(n)}$ vers la distribution cible, Π^* , au sens suivant :
Pour n assez grand et quel que soit $\boldsymbol{\theta}$ à l'étape initiale,

$$\Pi^{(n)}(\boldsymbol{\theta}, A) = \Pi^{*}(A) = \int \pi^{*}(t)\,\mathrm{d}t \tag{2.67}$$

$$\text{quel que soit } \boldsymbol{\theta} \text{ à l'étape initiale} \tag{2.68}$$

qui signifie que pour n assez grand, la distribution des états de la chaîne de Markov est indépendante des conditions initiales. Cette propriété est dite *propriété ergodique* et est vérifiée sous certaines hypothèses sur la chaîne rappelées dans le paragraphe suivant.

Les équations de Chapman-Kolmogorov sont obtenues par récurrence à partir de la formule suivante où $\Pi^{(2)}$ est la probabilité de transition de l'état $\boldsymbol{\theta}$ à un état de A en **2** itérations

$$\Pi^{(2)}(\boldsymbol{\theta}, A) = \int_{t\in\Theta}\int_{\boldsymbol{\theta}'\in A} k(\boldsymbol{\theta}, \boldsymbol{t})\, k\left(\boldsymbol{t}, \boldsymbol{\theta}'\right)\mathrm{d}\boldsymbol{t}\mathrm{d}\boldsymbol{\theta}' \tag{2.69}$$

$$= \int_{\boldsymbol{\theta}'\in A} k^{(2)}\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right)\mathrm{d}\boldsymbol{\theta}' \tag{2.70}$$

$$= \int_{t\in\Theta} k(\boldsymbol{\theta}, \boldsymbol{t})\,\Pi^{(1)}(\boldsymbol{t}, A)\,\mathrm{d}\boldsymbol{t} \tag{2.71}$$

où t est un point intermédiaire de Θ représentant toutes les trajectoires possibles en deux itérations pour passer de $\boldsymbol{\theta}$ à un point quelconque $\boldsymbol{\theta}'$ de A et $\Pi^{(1)}(\boldsymbol{t}, A) = K(\boldsymbol{t}, A)$.

A l'itération $n+m$ on obtient par récurrence

$$\Pi^{(n+m)}(\boldsymbol{\theta}, A) = \int_{t\in\Theta} k^{(n)}(\boldsymbol{\theta}, \boldsymbol{t})\,\Pi^{(m)}(\boldsymbol{t}, A)\,\mathrm{d}\boldsymbol{t} \tag{2.72}$$

où $k^{(n)}(\boldsymbol{\theta}, \boldsymbol{t}) = \int_{\boldsymbol{\theta}'\in\Theta} k\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) k^{(n-1)}\left(\boldsymbol{\theta}', \boldsymbol{t}\right)\mathrm{d}\boldsymbol{\theta}'$ et $Pi^{(m)}(\boldsymbol{t}, A) = \int_{\boldsymbol{\theta}'\in A} k^{(n)}\left(\boldsymbol{t}, \boldsymbol{\theta}'\right)\mathrm{d}\boldsymbol{\theta}'$.

2.4.2.5 La convergence vers la distribution cible

Une chaîne est π-*irréductible* si il y a une probabilité strictement positive de retour en un nombre fini d'étapes dans l'ensemble A tel que $\pi(A)$ est strictement positive.

Théorème 1
Théorème général de convergence ergodique

Si une chaîne de Markov est π-irréductible et qu'elle possède une distribution invariante π, alors elle est π-récurrente positive et π est l'unique distribution invariante.

Si la chaîne est apériodique et Harris-récurrente on a de plus convergence en variation totale

$$\lim_{n\to\infty}\left(\sup\left|\Pi^{(n)}(\boldsymbol{\theta}, A) - \pi(A)\right|\right) = 0 \tag{2.73}$$

Le théorème ergodique signifie que la chaîne oublie son état initial et que la répartition des états possibles après n itérations ressemble directement de plus en plus à la distribution de probabilité limite π [Parent et Bernier, 2007].

2.4.2.6 Moyenne le long d'une trajectoire

Les réalisations des chaînes de Markov obtenues dans les conditions du théorème d'ergodicité peuvent être directement utilisées pour l'inférence des paramètres ou de fonctions des paramètres. C'est cette propriété qui rend les chaînes de Markov si attractives.

Précisément on peut chercher l'espérance de $\boldsymbol{\theta}$, $E(\boldsymbol{\theta})$, ou encore sa variance (moment d'ordre 2) $E(\boldsymbol{\theta}^2) - [E(\boldsymbol{\theta})]^2$ ou toute autre fonction h des paramètres π-mesurable.

On cherche donc à évaluer l'espérance $E^{\pi}(h(\boldsymbol{\theta}))$ de forme intégrale

$$E^{\pi}(h(\boldsymbol{\theta})) = \int_{\Theta} h(\boldsymbol{\theta})\, \pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{2.74}$$

à partir de la moyenne correspondante $S_n(h)$ sur une trajectoire quelconque de la chaîne de Markov

$$S_n(h) = \frac{1}{n} \sum_{r=n_0+1}^{n_0+n} h(\boldsymbol{\theta}_r) \tag{2.75}$$

où n_0 est le temps de chauffe des chaînes.

Sous les hypothèses du théorème ergodique (c.f. [Robert et Casella, 2004]), pour n_0 assez grand, la moyenne $S_n(h)$ le long de la trajectoire converge presque sûrement (p.s.) vers l'espérance que l'on cherche à estimer

$$S_n(h) \xrightarrow[n \to +\infty]{p.s.} E^{\pi}(h(\boldsymbol{\theta})) \tag{2.76}$$

quel que soit le point de départ $\boldsymbol{\theta}$ dans le support de π.

Ainsi cette espérance est estimée par la moyenne des réalisations de la chaîne de Markov sur les paramètres d'intérêt.

2.4.2.7 La réversibilité

La notion centrale sur laquelle s'appuie la théorie des chaînes de Markov est la notion de distribution invariante. En effet les propriétés de convergence supposent toutes que la chaîne admet la bonne distribution invariante. Dans cette section on donne une condition suffisante pour qu'une distribution soit invariante pour une chaîne : la condition de *réversibilité*.

Définition 2.4.1 ***Réversibilité*** *[Robert et Casella, 2004] : une chaîne de Markov* $\left(\boldsymbol{\theta}^{(t)}\right)$ *est réversible si la distribution de* $\left(\boldsymbol{\theta}^{(t+1)}\right)$ *sachant* $\left(\boldsymbol{\theta}^{(t+2)} = \boldsymbol{\theta}'\right)$ *est la même que la distribution de* $\left(\boldsymbol{\theta}^{(t+1)}\right)$ *sachant* $\left(\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}'\right)$

La réversibilité signifie que le temps n'a pas d'importance dans la construction de la chaîne.

La réversibilité est liée à la condition d'équilibre des chaînes de Markov définie ci-dessous

Définition 2.4.2 ***Condition d'équilibre*** *[Robert et Casella, 2004] : une chaîne de Markov de noyau de transition K satisfait la condition d'équilibre si il existe une fonction f telle que*

$$f(\boldsymbol{\theta}) K(\boldsymbol{\theta}, \boldsymbol{\theta}') = f(\boldsymbol{\theta}') K(\boldsymbol{\theta}', \boldsymbol{\theta}) \tag{2.77}$$

pour tout $(\boldsymbol{\theta}, \boldsymbol{\theta}')$.

La condition d'équilibre signifie que la probabilité d'être en $\boldsymbol{\theta}$ et d'aller en $\boldsymbol{\theta}'$ est la même que la probabilité d'être en $\boldsymbol{\theta}'$ et de retourner en $\boldsymbol{\theta}$.

Théorème 2
Réversibilité et distribution invariante *[Robert et Casella, 2004] : Si une chaîne de Markov de noyau de transition K vérifie la condition d'équilibre avec π une densité de probabilité alors (i) la densité π est la densité invariante de la chaîne*
(ii) la chaîne est réversible

En pratique il suffit donc de vérifier la condition d'équilibre sur la distribution *a posteriori* .

2.4.3 L'algorithme de Gibbs

L'algorithme de Gibbs est central en statistique bayésienne car il permet de réduire un problème complexe de simulation, typiquement la simulation selon la distribution jointe *a posteriori* des paramètres, en une suite d'étapes simples à simuler.

Pour cette raison, l'algorithme de Gibbs est aussi connu sous le nom d'échantillonneur de Gibbs (en anglais, Gibbs sampler).

D'un point de vue historique, l'algorithme de Gibbs tire son nom d'un physicien et mathématicien américain du 19^{e} siècle Josiah Willard Gibbs, considéré comme l'un des fondateurs de la thermodynamique moderne et de la mécanique statistique. Josiah Willard Gibbs a donné son nom aux distributions de Gibbs utilisées en traitement d'image. L'association entre le nom et l'algorithme a été réalisée dans l'article de [Geman et Geman, 1984] où cette méthode a été développée pour l'étude bayésienne des champs de Gibbs en traitement d'image, et a perduré depuis.

2.4.3.1 Construction

Dans la suite on se place dans les "bons cas" où la distribution jointe des grandeurs inconnues est entièrement spécifiée par la donnée de toutes les distributions conditionnelles de ces grandeurs et où on peut simuler à partir de toutes les distributions conditionnelles *a posteriori* .

Précisément, l'algorithme de Gibbs permet de simuler les distributions marginales *a posteriori* de toutes les grandeurs inconnues du modèle à partir des distributions conditionnelles *a posteriori* de toutes ces grandeurs.

En d'autres termes l'algorithme de Gibbs simule des échantillons issus de la loi jointe *a posteriori* $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ à partir de simulations dans les distributions conditionnelles *a posteriori* $\pi_i(\theta_i|\boldsymbol{y}, \theta_j,\, j \neq i)$.

L'algorithme de Gibbs, sous sa formulation générale, s'écrit

Algorithme 2.4.1 ***L'algorithme de Gibbs.***
Initialisation : $\theta_2^{(0)}, \ldots, \theta_K^{(0)}$
A l'itération t $(t = 1, \ldots, T)$
- ***Etape 1 :*** *Générer* $\theta_1^{(t)}$ *selon* $\pi_1\left(\theta_1|\theta_2^{(t-1)}, \ldots, \theta_K^{(t-1)}, \boldsymbol{y}\right)$
- ***Etape 2 :*** *Générer* $\theta_2^{(t)}$ *selon* $\pi_2\left(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \ldots, \theta_K^{(t-1)}, \boldsymbol{y}\right)$
- ...
- ***Etape K :*** *Générer* $\theta_K^{(t)}$ *selon* $\pi_K\left(\theta_K|\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_{K-1}^{(t)}, \boldsymbol{y}\right)$

2.4.3.2 Un algorithme markovien

La convergence de l'algorithme de Gibbs repose sur les propriétés markoviennes des échantillons ainsi construits.

En effet les réalisations successives de la loi jointe forment une chaîne de Markov dont la distribution stationnaire est $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. En revanche les sous-chaînes ainsi construites ne sont plus des chaînes de Markov (sauf dans le cas où $k = 2$).

Remarque : Lorsqu'il est difficile de simuler selon l'une des distributions conditionnelles $\pi_i(\theta_i|\theta_j,\, j \neq i)$ on peut remplacer cette étape par une étape de Métropolis-Hastings dont la distribution cible est $\pi_i(\theta_i|\theta_j,\, j \neq i)$. D'après [Robert, 2001], l'introduction d'une étape de Métropolis-Hastings ne modifie pas la distribution stationnaire de la chaîne et est donc valide.

2.4.4 L'algorithme de Metropolis-Hastings

2.4.4.1 Introduction

L'algorithme de Metropolis-Hastings peut être vu comme une alternative à l'algorithme de Gibbs dans le cas où on ne peut pas simuler facilement dans les distributions conditionnelles *a posteriori* , par exemple lorsque les distributions ne sont pas conjuguées ou lorsque les expressions ne peuvent pas s'exprimer sous forme analytique.

L'algorithme de Metropolis-Hastings est un algorithme d'acceptation/rejet. L'idée de l'algorithme est de simuler selon une autre distribution, plus simple à simuler, appelée la *loi de proposition*, et d'accepter la valeur simulée avec une certaine probabilité d'être effectivement un tirage selon la loi cible.

Sous les conditions de convergence rappelées à la section 2.4, les valeurs successivement acceptées de la loi de proposition forment une chaine de Markov convergeant vers la distribution cible.

La différence avec un algorithme acceptation/rejet classique est que tant qu'une nouvelle valeur n'est pas acceptée l'algorithme retourne la dernière valeur acceptée comme nouvelle valeur de la chaîne.

2.4.4.2 Construction

Cette section a pour but d'amener à la construction de l'algorithme de Metropolis-Hastings en reprenant la présentation de l'algorithme dans l'article pédagogique de [Chib et Greenberg, 1990].

Pour construire une chaîne de Markov, la loi de proposition prend en argument la valeur courante de la chaîne. En notant q la loi de proposition, $q(x,y)$ est la fonction permettant à la chaîne de changer d'état en passant de l'état x à l'état y. Par exemple y peut être une réalisation d'une loi normale centrée en x.

D'après la section 2.4 la loi de proposition q doit satisfaire la condition de réversibilité à savoir $\pi(x)\,q(x,y) = \pi(y)\,q(y,x)$. Si cette condition est vérifiée alors le noyau de transition recherché est q.

Au contraire si on peut trouver x et y tels que $\pi(x)\,q(x,y) > \pi(y)\,q(y,x)$ alors la condition de réversibilité n'est plus vérifiée, c'est-à-dire que le processus favorise le passage de x à y au détriment du passage de y à x.

Pour rétablir la situation de réversibilité on introduit une probabilité $\alpha(x,y) < 1$ dans le membre de gauche et une probabilité $\alpha(y,x) = 1$ dans le membre de droite appelée *probabilité de mouvement*.

Puis on détermine $\alpha(x,y)$ en posant la condition de réversibilité

$$\pi(x)\,q(x,y)\,\alpha(x,y) = \pi(y)\,q(y,x)\,\alpha(y,x) \tag{2.78}$$

d'où $\alpha(x,y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}$.

De la même façon, pour les couples (x,y) qui vérifient l'inégalité dans l'autre sens, à savoir $\pi(x)\,q(x,y) < \pi(y)\,q(y,x)$, on pose $\alpha(x,y) = 1$ et on obtient l'expression symétrique pour $\alpha(y,x) = \frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}$.

Pour tout x et tout y on peut donc définir la probabilité de mouvement

$$\alpha(x,y) = \min\left[\frac{\pi(y)\,q(y,x)}{\pi(x)\,q(x,y)}, 1\right] \tag{2.79}$$

$$= 1 \text{ sinon} \tag{2.80}$$

qui vérifie bien la condition de réversibilité

$$\pi(x)\,q(x,y)\,\alpha(x,y) = \pi(y)\,q(y,x)\,\alpha(y,x) \tag{2.81}$$

Algorithme 2.4.2 ***L'algorithme de Metropolis-Hastings .*** *Initialisation :*
A l'itération t $(t = 1, \ldots, T)$

- *Générer* y *selon* $q\left(x^{(t-1)}, .\right)$
- *Générer* u *selon* $Unif\,(0,1)$
- *Si* $u \leq \alpha\left(x^{(t-1)}, y\right)$ *accepter* y *comme réalisation de la loi cible et prendre* $x^{(t)} = y$
- *Sinon rejeter* y *et prendre* $x^{(t)} = x^{(t-1)}$

Fil rouge 4 ***Ecriture de l'algorithme de Metropolis***

La loi de proposition q *en deux dimensions est la marche aléatoire définie par*

$$q\,(s,r) = (s,r) + \beta\,(\xi_1, \xi_2) \tag{2.82}$$

où β *est le déplacement maximum autorisé et* ξ_1 *et* ξ_2 *sont des nombres aléatoires entre* -1 *et* 1*. Le point généré se situe donc dans un carré de côté* 2β *centré en le dernier point accepté.*

La probabilité de mouvement $\alpha\,(s,r)$ *d'un état d'énergie* s *à un état d'énergie* r *dépend de la différence d'énergie* $\Delta E = E_r - E_s$ *entre les deux états. La probabilité de mouvement est la probabilité de cette différence d'énergie*

$$\alpha\,(s,r) = \exp -\frac{\Delta E}{k_B T} \tag{2.83}$$

Si $\Delta E < 0$ *alors* $\alpha\,(s,r) = 1$ *car* α *est une probabilité donc admet 1 comme borne supérieure. Ainsi la probabilité de mouvement vers un état de moindre énergie est 1.*

Si $\Delta E > 0$ *alors* $\alpha\,(s,r) < 1$*. Ainsi la probabilité de mouvement vers un état d'énergie supérieur est* $\alpha\,(s,r)$*.*

En résumé quels que soient les états d'énergie s *et* r *la probabilité de mouvement est définie par*

$$\alpha\,(s,r) = \min\left\{\exp -\frac{\Delta E}{k_B T}, 1\right\} \tag{2.84}$$

$$= \min\left\{\frac{\pi\,(r)}{\pi\,(s)}, 1\right\} \tag{2.85}$$

où $\pi\,(r) = p_r$ *et* $\pi\,(s) = p_s$ *sont les probabilités des énergies* r *et* s *respectivement définies à la section 2.1.5.*

L'algorithme de Metropolis est donc bien un cas particulier de l'algorithme de Metropolis-Hastings dans le cas où la loi de proposition est symétrique.

L'algorithme de Metropolis
A l'itération t $(t = 1, \ldots, T)$
- *Générer* ξ_1 *et* ξ_2 *selon* $Unif(-1, 1)$
- *Générer* y *selon* $x^{(t-1)} + \beta(\xi_1, \xi_2)$
- *Générer* u *selon* $Unif(0, 1)$
- *Si* $u \leq min\left\{\exp -\frac{E_y - E_{x^{(t-1)}}}{k_B T}, 1\right\}$ *accepter* y *comme réalisation de la loi cible et prendre* $x^{(t)} = y$

- *Sinon rejeter* y *et prendre* $x^{(t)} = x^{(t-1)}$

2.4.4.3 Expression du noyau de transition et convergence

D'après la condition de réversibilité de l'équation 2.81 la quantité $k(x, y) = q(x, y)\,\alpha(x, y)$ est la partie continue du noyau de transition.

Il reste à prendre en compte la probabilité $r(x)$ de rester en x qui s'exprime par

$$r(x) = 1 - \int q(x, y)\,\alpha(x, y)\,\mathrm{d}y \tag{2.86}$$

La forme complète du noyau de transition exprimée sous forme intégrale est donc

$$K(x, A) = \int_A k(x, y)\,\mathrm{d}y + r(x)\,\delta_A(x) \tag{2.87}$$

Il reste à vérifier que π est bien la distribution invariante de la chaîne de Markov ainsi construite. Pour cela il faut montrer que la formule 2.65 est vérifiée. On a

$$\begin{aligned}
\int K(x, A)\,\pi(x)\,\mathrm{d}x &= \int\int_A k(x, y)\,\mathrm{d}y\pi(x)\,\mathrm{d}x + \int r(x)\,\delta_A(x)\,\pi(x)\,\mathrm{d}x && (2.88)\\
&= \int \delta_A(x)\,\pi(x)\,\mathrm{d}x && (2.89)\\
&= \pi(A) && (2.90)
\end{aligned}$$

ce qui prouve que π est la distribution invariante pour la chaîne (d'après le théorème 2.4.2.7).

2.5 L'expansion paramétrique

Les méthodes d'augmentation des données puis d'expansion paramétrique ont d'abord été appliquées avec succès aux algorithmes EM. On renvoie ainsi aux articles de [Dempster *et al.*, 1977] pour l'implémentation de l'augmentation des données au sein de l'algorithme EM et de [Liu *et al.*, 1998] pour l'implémentation de l'augmentation paramétrique sous la forme de l'algorithme PX-EM. L'article (à paraître) [Lewandowski *et al.*, 2011] reprend l'interprétation de l'algorithme PX-EM en termes d'efficacité d'estimation mesurée par la réduction induite de biais.

Ces méthodes ont été développées pour accélérer la convergence de l'algorithme EM en introduisant des variables latentes et des paramètres non identifiables et ont prouvé leur efficacité. Ces méthodes permettent également de surmonter des problèmes d'identifiabilité et c'est la raison qui a motivé notre utilisation de l'expansion paramétrique.

L'article séminal de [Tanner et Wong, 1987] marque la volonté d'appliquer ces méthodes au calcul des distributions *a posteriori* .

Ainsi l'algorithme d'augmentation des données peut être vu comme l'analogue bayésien de l'algorithme EM [Liu *et al.*, 1994] et son implémentation en fait un cas particulier de l'algorithme de Gibbs.

Plus généralement l'algorithme de Gibbs peut être vu comme une généralisation stochastique/bayésienne des algorithmes de type EM, qui pourrait de la même manière bénéficier des améliorations apportées à l'algorithme EM par les méthodes basées sur l'augmentation des données. On renvoie aux articles de [Liu et Wu, 1999] et [van Dyk et Meng, 2001] pour la mise en oeuvre de l'expansion paramétrique dans les méthodes MCMC, sous la forme d'algorithmes PX-Gibbs.

L'implémentation de l'expansion paramétrique mise en oeuvre au chapitre 3 interprète en termes bayésiens les étapes de l'algorithme PX-EM décrit dans [Liu *et al.*, 1998] et ainsi diffère de l'implémentation classique de l'expansion paramétrique pour les algorithmes MCMC.

Dans la première partie de cette section on rappelle l'algorithme d'augmentation des données de [Tanner et Wong, 1987] sur lequel s'appuient les implémentations des algorithmes PX-EM et PX-Gibbs.

Dans la deuxième partie de cette section on présente la mise en oeuvre classique de l'expansion paramétrique pour l'algorithme de Gibbs.

On renvoie au chapitre 3 pour une illustration de la mise en oeuvre bayésienne de l'algorithme PX-EM sous la forme d'un algorithme PX-Gibbs dont l'implémentation est en fait celle du schéma 2.1 de [Liu et Wu, 1999] présenté à la section 2.5.2.3, à ceci près que nous utilisons un prior informatif au lieu d'un prior de Haar pour modéliser le paramètre d'expansion. Dans ces conditions les propriétés d'amélioration de convergence ne sont pas établies mais la convergence est de toute façon assurée [Lawrence *et al.*, 2008].

On notera avant de poursuivre que deux des auteurs mentionnées portent fortuitement le même nom à savoir Jun S Liu ([Liu et Wu, 1999]) et Chuanhai Liu ([Liu *et al.*, 1998]) car le référencement ne permet pas de les distinguer.

2.5.1 L'augmentation des données [Tanner et Wong, 1987]

L'augmentation des données a pour but de simplifier l'estimation de la distribution *a posteriori* des paramètres $\pi(\boldsymbol{\theta}|y)$ en introduisant des données non observées z telles que

- sachant la totalité des données (observées et non observées) la distribution complète *a posteriori* de $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z})$, est connue,
- la distribution *a posteriori* des données non observées conditionnellement aux données observées $\pi(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})$ est connue.

L'augmentation des données fournit donc deux modélisations des mêmes données :

– le modèle des données observées $f(\boldsymbol{y}|\boldsymbol{\theta})$, appelé modèle (O),
– le modèle des données complètes $f(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta})$, appelé modèle (C).

Puisque les variables $\boldsymbol{z}$ ne sont pas observées, leur introduction doit préserver le modèle des données observées. Cela signifie qu'une augmentation de données est valide si on retrouve le modèle (0) en intégrant le modèle (C) sur les données non observées c'est-à-dire que l'augmentation de données doit vérifier

$$\int f(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{z} = f(\boldsymbol{y}|\boldsymbol{\theta}) \tag{2.91}$$

La justification de l'approche repose sur les points suivants [Tanner et Wong, 1987]

– si, en plus de connaître $\pi(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z})$, on connaît $\pi(\boldsymbol{z}|\boldsymbol{y})$ alors on peut simuler selon la distribution *a posteriori* recherchée $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ en moyennant sur les données latentes prédites :

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z})\,\pi(\boldsymbol{z}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{z} \tag{2.92}$$

– si, en plus de connaître $\pi(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{y})$, on connaît $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ alors on peut simuler selon $\pi(\boldsymbol{z}|\boldsymbol{y})$ en moyennant sur les valeurs de $\boldsymbol{\theta}$ *a posteriori* :

$$\pi(\boldsymbol{z}|\boldsymbol{y}) = \int \pi\left(\boldsymbol{z}|\boldsymbol{\theta}',\boldsymbol{y}\right)\pi\left(\boldsymbol{\theta}'|\boldsymbol{y}\right)\mathrm{d}\boldsymbol{\theta}' \tag{2.93}$$

Ces deux conditions semblent s'éliminer mutuellement car elles supposent la connaissance simultanée de $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ et $\pi(\boldsymbol{z}|\boldsymbol{y})$ pour pouvoir estimer $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ et $\pi(\boldsymbol{z}|\boldsymbol{y})$!

Au contraire, ces deux distributions sont obtenues à l'équilibre des deux équations 2.92 et 2.93 après substitutions successives dans les deux équations.

Précisément la distribution *a posteriori* vérifie l'équation

$$\pi\left(\boldsymbol{\theta}'|\boldsymbol{y}\right) = \int k\left(\boldsymbol{\theta},\boldsymbol{\theta}'\right)\pi(\boldsymbol{\theta}|\boldsymbol{y})\,\mathrm{d}\boldsymbol{\theta} \tag{2.94}$$

où $k\left(\boldsymbol{\theta},\boldsymbol{\theta}'\right) = \int \pi\left(\boldsymbol{\theta}'|\boldsymbol{y},\boldsymbol{z}\right)\pi(\boldsymbol{z}|\boldsymbol{\theta},\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}$.

On reconnaît ainsi la problématique des chaînes de Markov, à savoir atteindre une situation d'équilibre.

L'algorithme d'augmentation des données consiste à itérer les deux étapes correspondant aux équations 2.93 et 2.92, et peut être vu comme un algorithme de Gibbs à deux niveaux.

Le premier niveau (Etape 1) consiste à simuler un échantillon $\boldsymbol{z}_1,\ldots,\boldsymbol{z}_m$ des variables latentes dans leur distribution prédictive $p_t(\boldsymbol{z}|\boldsymbol{y})$.

Le deuxième niveau (Etape 2) consiste à approcher la distribution *a posteriori* de $\boldsymbol{\theta}$ par le mélange des distributions complètes *a posteriori* :

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \approx \frac{1}{m}\sum_{i=1}^{m}\pi(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{z}_i) \tag{2.95}$$

Le déroulement de l'algorithme d'augmentation des données est formalisé ci-dessous.

Algorithme 2.5.1 ***L'algorithme d'augmentation des données***

[Tanner et Wong, 1987]

A l'itération t :

- ***Etape 1 : "imputation des variables latentes" :*** *générer un échantillon* $\boldsymbol{z}_1^{(t+1)}, \ldots, \boldsymbol{z}_m^{(t+1)}$ *de l'approximation courante de la distribution prédictive* $p_t(\boldsymbol{z}|\boldsymbol{y})$*, pour cela :*
  - ***Etape 1.1*** *générer* $\boldsymbol{\theta}$ *selon* $\pi_t(\boldsymbol{\theta}|\boldsymbol{y})$
  - ***Etape 1.2*** *générer* $\boldsymbol{z}_1^{(t+1)}, \ldots, \boldsymbol{z}_m^{(t+1)}$ *selon* $\pi(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{y})$
- ***Etape 2 : "tirage dans la distribution a posteriori " :*** *mise à jour de l'approximation courante de la distribution a posteriori des paramètres à partir des imputations des variables latentes*

$$\pi_{t+1}(\boldsymbol{\theta}|\boldsymbol{y}) \approx \frac{1}{m}\sum_{i=1}^{m} \pi_t\left(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}_i^{(t+1)}\right) \tag{2.96}$$

Pour $m = 1$ on retrouve l'algorithme de Gibbs.

La preuve de la convergence vers la distribution *a posteriori* est donnée dans [Tanner et Wong, 1987] qui montre en particulier que k est le noyau de transition de la chaîne de Markov ainsi construite. Dans ces conditions l'équation 2.94 signifie que la distribution *a posteriori* $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ est la distribution invariante de la chaîne.

L'exemple suivant montre la construction du modèle des données complètes (modèle C) à partir du modèle des données observées (modèle O) dans le cas du modèle probit univarié.

Exemple 2.5.1 ***Le modèle probit univarié***

Soit $\boldsymbol{y} = (y_1, \ldots, y_n)$ *un échantillon de variables binaires 0/1 dont la probabilité de succès est à rapprocher des caractéristiques des individus, regroupées dans un vecteur de prédicteurs* $\boldsymbol{x}_i$ *pour l'individu i.*

La probabilité de succès π_1 *est alors reliée aux prédicteurs par un lien probit défini par le modèle suivant*

$$\pi_1 = \Pr(y_i = 1) = \Phi\left(\boldsymbol{x}_i^t\boldsymbol{\theta}\right) \tag{2.97}$$

où $\boldsymbol{\theta}$ *est le vecteur des coefficients de régression et* Φ *est la fonction de répartition de la loi normale centrée réduite.*

Le modèle des données observées (modèle O) s'écrit encore

Modèle (0)

$$y_i|\boldsymbol{\theta} \sim \text{Bernoulli}\left(\Phi\left(\boldsymbol{x}_i^t\boldsymbol{\theta}\right)\right) \tag{2.98}$$

Suivant [Albert et Chib, 1993] le modèle (O) est équivalent à un modèle de régression linéaire dans lequel la réponse binaire y_i est quantifiée en un réel z_i selon une procédure garantissant l'équivalence.

La procédure d'augmentation des données consiste donc à associer à chaque donnée observée y_i une donnée non observée z_i issue d'un tirage aléatoire dans une loi normale tronquée dont le seuil est directement lié à la proportion de succès dans l'échantillon. Ainsi chaque réalisation z_i est caractérisée par sa position par rapport au seuil, sans perte d'information.

D'après 2.97 on peut choisir comme variable latente la loi normale standard et caractériser chaque réalisation par rapport au seuil $\boldsymbol{x_i^t\theta}$ auquel cas le seuil est le quantile de la proportion de succès.

Cependant en pratique, il est plus maniable d'étudier les positions de la variable latente par rapport à un seuil nul.

En remarquant que

$$\Phi\left(\boldsymbol{x}_i^t\boldsymbol{\theta}\right) = P\left(N\left(0,1\right) \leq \boldsymbol{x}_i^t\boldsymbol{\theta}\right) = P\left(N\left(\boldsymbol{x}_i^t\boldsymbol{\theta},1\right) \geq 0\right), \tag{2.99}$$

on choisit donc d'étudier les positions des réalisations de la loi $N\left(\boldsymbol{x}_i^t\boldsymbol{\theta},1\right)$ par rapport à 0. On prend donc $z_i \sim N\left(\boldsymbol{x}_i^t\boldsymbol{\theta},1\right)$.

L'équivalence entre ces deux choix est illustrée dans le schéma de la figure 2.1.

Le modèle des données complètes s'écrit donc

Modèle (C)

$$y_i|z_i = I_{\{z_i>0\}} \tag{2.100}$$

$$z_i|\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{x_i^t\theta},1\right) \tag{2.101}$$

Les deux modèles présentent les mêmes ensembles de paramètres. Cependant l'estimation du modèle (C) est plus simple car elle revient en fait à l'estimation d'un modèle de régression linéaire.

Remarque : variable latente et identifiabilité : *On a vu qu'il existe en réalité une infinité de variables latentes qui satisfont le modèle (C) à variance fixée égale à 1, obtenues en jouant sur le seuil. On peut également décider de libérer l'échelle de la variable latente en ne l'imposant plus égale à 1. A seuil fixé à 0, si on augmente l'échelle de la variable latente on doit augmenter la moyenne de la variable latente en proportion car cette moyenne a le rôle d'un seuil. En effet, d'après 2.99,*

$$P\left(N\left(0,\alpha\right) \leq \boldsymbol{x}_i^t\boldsymbol{\theta}\sqrt{\alpha}\right) = P\left(N\left(\boldsymbol{x}_i^t\boldsymbol{\theta}\sqrt{\alpha},\alpha\right) \geq 0\right) \tag{2.102}$$

Ainsi introduire une variance α implique de multiplier la moyenne (vue comme un seuil) par $\sqrt{\alpha}$.

Cependant les données n'apportent pas d'information sur la variance de la variable latente, l'information n'agit que que sur la moyenne (vue comme un seuil). A variance

inconnue, le coefficient de régression est donc également inconnu, précisément connu à un coefficient près (l'écart-type de la variable latente). Il faut donc fixer la variance de la variable latente à une valeur de notre choix, et ne pas oublier de post-multiplier la valeur du coefficient de régression par l'écart-type correspondant.

Libérer la variance revient donc à rajouter un paramètre non identifiable au modèle. L'idée peut paraître curieuse cependant elle est à la base du procédé d'amélioration de convergence qu'est l'expansion paramétrique.

2.5.2 L'augmentation des données pour l'expansion paramétrique [Liu et Wu, 1999]

2.5.2.1 Modèle complet et modèle étendu

L'expansion paramétrique consiste à construire un modèle étendu basé sur le modèle des données complètes pour lequel le modèle des données complètes correspond à une valeur précise du paramètre rajouté dit *paramètre d'expansion*.

On peut créer ainsi une infinité de modèles étendus indicés par la valeur du paramètre d'expansion. [Liu et Wu, 1999] démontrent même que l'on crée de cette façon une classe d'équivalence de modèles dont le représentant est le modèle des données complètes.

En appelant α le paramètre d'expansion, on définit t_α la transformation qui permet le passage du modèle $\mathcal{M}_\alpha$ au modèle $\mathcal{M}_{\alpha_0}$ des données observées correspondant à $\alpha = \alpha_0$ (supposé équivalent au modèle des données observées d'après 2.91).

L'expansion paramétrique se traduit souvent par une transformation, également indicée par α, des données latentes z introduites par l'algorithme d'augmentation des données.

Ainsi en notant w les variables latentes du modèle étendu (après expansion paramétrique), le passage du modèle $\mathcal{M}_\alpha$ au modèle $\mathcal{M}_{\alpha_0}$ se traduit dans ce cas par la transformation suivante au niveau des variables latentes

$$z = t_\alpha(w) \tag{2.103}$$

Le modèle étendu est donc paramétrisé par $\boldsymbol{\theta}$ et α et les données non observées correspondant à l'augmentation des données paramétrisée par α sont désormais w.

Suivant le même raisonnement qui a mené à l'égalité des vraisemblances 2.91, le modèle étendu doit préserver le modèle des données observées. Le procédé d'expansion paramétrique couplé à l'augmentation des données doit donc vérifier la suite d'égalités suivante

$$\int p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{\theta}, \alpha)\, \mathrm{d}\boldsymbol{w} = \int f(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{z} = f(\boldsymbol{y}|\boldsymbol{\theta}) \tag{2.104}$$

2.5.2.2 Forme générale de l'algorithme

L'algorithme d'augmentation des données 2.5.1 s'applique directement au cadre de l'expansion paramétrique (voir [Liu et Wu, 1999]) en supposant que les distributions suivantes sont connues ou facilement simulables **dans le modèle étendu**

- la distribution *a posteriori* de tous les paramètres sachant les données complètes : $\pi(\boldsymbol{\theta}, \alpha|\boldsymbol{y}, \boldsymbol{w})$

– la distribution *a posteriori* des variables latentes w : $\pi(\boldsymbol{w}|\alpha, \boldsymbol{\theta}, \boldsymbol{y})$

De manière similaire à la section 2.5.1, l'algorithme d'augmentation des données consiste alors à alterner les simulations dans les distributions *a posteriori* de $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, et prédictive de $\boldsymbol{w}$, $\pi(\boldsymbol{w}|\boldsymbol{y})$ où

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \int\int \pi(\boldsymbol{\theta}, \alpha|\boldsymbol{y}, \boldsymbol{w})\, \pi(\boldsymbol{w}|\boldsymbol{y})\, \mathrm{d}\alpha \mathrm{d}\boldsymbol{w} \tag{2.105}$$

$$\pi(\boldsymbol{w}|\boldsymbol{y}) = \int\int \pi(\boldsymbol{w}|\alpha, \boldsymbol{\theta}, \boldsymbol{y})\, \pi(\alpha|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}|\boldsymbol{y})\, \mathrm{d}\alpha \mathrm{d}\boldsymbol{\theta} \tag{2.106}$$

avec $\pi(\alpha|\boldsymbol{\theta})$ la distribution *a priori* de α.

L'algorithme, dont on reprend le nom donné par les auteurs [Liu et Wu, 1999], à savoir PX-DA pour Parameter EXpansion - Data Augmentation, s'écrit donc sous la forme générale suivante

Algorithme 2.5.2 ***L'algorithme PX-DA***
A l'itération t :
– ***Etape 1 : "imputation des variables latentes dans le modèle étendu" :***
  – ***Etape 1.1*** *générer* α *selon* $\pi\left(\alpha|\boldsymbol{\theta}^{(t-1)}\right)$
  – ***Etape 1.2*** *générer* $\boldsymbol{w}^{(t)}$ *selon* $\pi\left(\boldsymbol{w}|\alpha, \boldsymbol{\theta}^{(t-1)}, \boldsymbol{y}\right)$
– ***Etape 2 : "tirage dans la distribution jointe a posteriori " des paramètres originaux*** $\boldsymbol{\theta}$ ***et du paramètre d'expansion*** α ***:*** *générer* $\boldsymbol{\theta}^{(t)}, \alpha$ *selon* $\pi\left(\boldsymbol{\theta}, \alpha|\boldsymbol{y}, \boldsymbol{w}^{(t)}\right)$

L'algorithme peut être spécifié dans le cas d'hypothèses sur le passage entre les modèles complet et étendu, ce qui est l'objet du paragraphe suivant.

2.5.2.3 Cas particulier où l'expansion paramétrique se traduit par une transformation des données

On se place dans le cadre où il existe une relation du type de celle décrite à la section 2.5.2.1, à savoir que le paramètre d'expansion α indice en fait une transformation du type $\boldsymbol{z} = t_\alpha(\boldsymbol{w})$ où t_α est un C^1 difféomorphisme entre les variables latentes du modèle complet $\boldsymbol{z}$ et les variables latentes du modèle étendu $\boldsymbol{w}$. On suppose de plus que la distribution *a priori* du paramètre d'expansion est indépendante des paramètres du modèle complet, à savoir $p(\alpha|\boldsymbol{\theta}) = p_0(\alpha)$.

Les propriétés de convergence de l'algorithme PX-DA peuvent être améliorées si la distribution *a priori* du paramètre d'expansion est un prior de Haar.

Par exemple le prior de Haar d'un paramètre de variance α^2 est $p_0\left(\alpha^2\right) \propto \frac{1}{\alpha}$. La transformation associée est $t_\alpha(\boldsymbol{z}) = \alpha\boldsymbol{z}$. Si α est une matrice $k \times k$ non singulière le prior de Haar de α est $p_0(\alpha) \propto \frac{1}{|\alpha|^k}$. La transformation associée est $t_\alpha(\boldsymbol{z}) = \alpha^{1/2}\boldsymbol{z}$.

Sous ces conditions l'algorithme PX-DA devient

Algorithme 2.5.3 ***L'algorithme PX-DA : le schéma 2.1 de [Liu et Wu, 1999].***
- ***Etape 1 :*** *Générer* $\boldsymbol{z} \sim f(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{y})$
- ***Etape 2 :*** *Générer* $\alpha \sim p(\alpha|\boldsymbol{y}, \boldsymbol{z})$
- ***Etape 3 :*** *Calculer* $\boldsymbol{z}' = t_\alpha(\boldsymbol{z})$
- ***Etape 4 :*** *Générer* $\boldsymbol{\theta} \sim f(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{z}')$

Les propriétés de convergence sont étudiées dans l'article [Liu et Wu, 1999] et plus généralement dans l'ouvrage [Liu, 2001] qui a suivi l'écriture de l'article.

Des études empiriques y sont proposées, illustrant en particulier sur le modèle probit univarié, les propriétés d'amélioration de la convergence attendues suite à l'utilisation des méthodes d'expansion paramétriques. Dans ce cas l'utilisation de l'expansion paramétrique permet de diminuer les autocorrélations dans les chaînes et d'accélérer la convergence, comme illustré dans l'exemple suivant.

Exemple 2.5.2 ***Le modèle probit univarié, suite de l'exemple 2.5.1***

Dans la remarque de la première partie de l'exemple on a introduit α *la variance résiduelle de la variable latente sous-jacente* z*.*

Dans la deuxième partie de l'exemple on peut maintenant définir les nouvelles variables latentes sous-jacentes $\boldsymbol{w} = \alpha^{1/2}\boldsymbol{z}$ *dans le modèle étendu grâce à cette expansion paramétrique.*

Le modèle des données complètes est ainsi étendu dans le modèle (X) défini par

Modèle (X) :

$$y_i|w_i = I_{\{w_i>0\}} \tag{2.107}$$

$$w_i|\theta, \alpha \sim N\left(\boldsymbol{x_i^t}\boldsymbol{\theta}\alpha^{\frac{1}{2}}, \alpha\right) \tag{2.108}$$

$$w_i = z_i\alpha^{\frac{1}{2}} \tag{2.109}$$

On illustre l'efficacité de l'expansion paramétrique dans le cas du modèle probit univarié avec les graphiques d'autocorrélations des chaînes extraits de [Liu et Wu, 1999] résultant de l'implémentation des algorithmes d'augmentation des données DA (algorithme 2.5.1) et d'expansion paramétrique PX-DA (algorithme 2.5.3) dans des cas simulés.

Le prior de Haar utilisé dans l'implémentation de l'algorithme d'expansion paramétrique est

$$\pi(\alpha) \propto \frac{1}{\sqrt{\alpha}} \tag{2.110}$$

En prenant $n = 100$, $\boldsymbol{x_i} = (1x_i)$ avec $x_i \sim N(0, 1)$, les réponses y_i ont été générées selon le modèle $y_i \sim bernoulli(\Phi(\theta_0 + \theta_1 x_i))$ avec la valeur $\theta_0 = 0$ et avec les quatre valeurs suivantes des coefficients de régression $\theta_1 = 1, 2, 4, 8$.

Les résultats, présentés à la figure 2.2, montrent une forte diminution des autocorrélations dans l'estimation de θ_1, avec une diminution d'autant plus importante que la valeur de θ_1 est élevée.

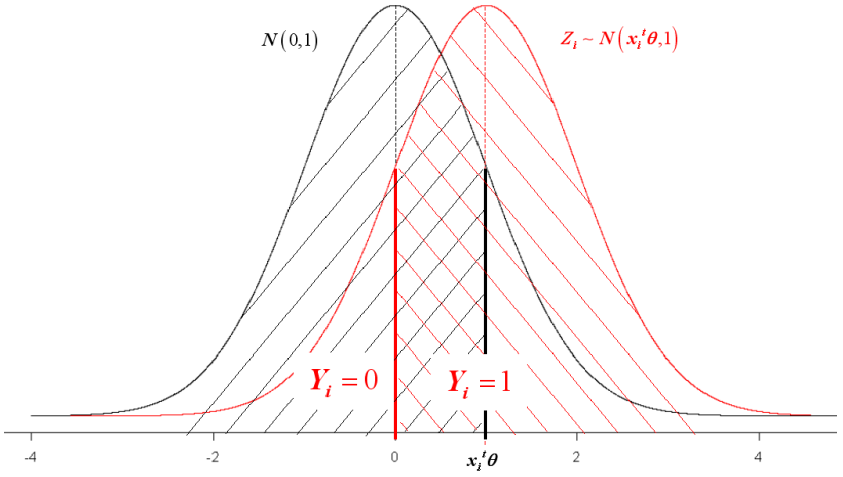


FIGURE 2.1 – .

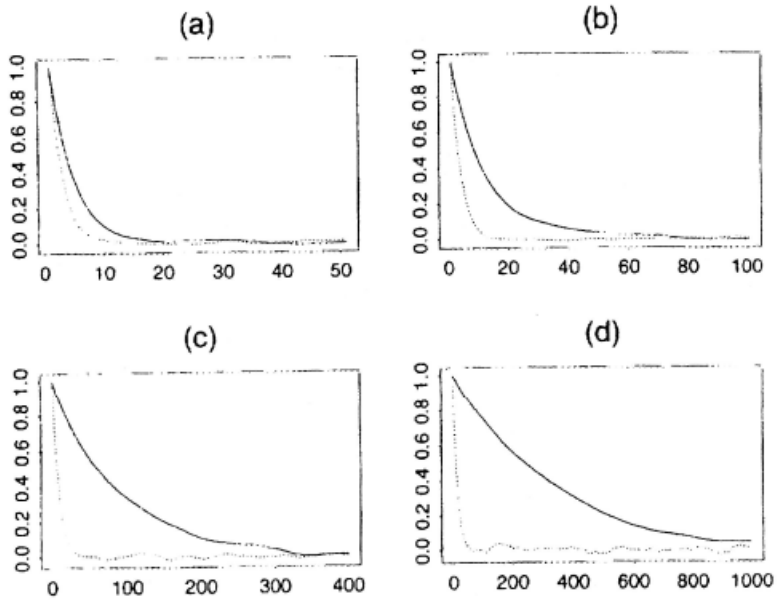


Figure 2. Autocorrelation Functions for DA (Solid Lines) and PX-DA (Dashed Lines) With Various Values of β_1: (a) $\beta_1 = 1$; (b) $\beta_1 = 2$; (c) $\beta_1 = 4$; (d) $\beta_1 = 8$.

FIGURE 2.2 – Comparaisons des autocorrélations résultant de l'algorithme DA (algorithme 2.5.1, courbe pleine) par rapport à l'algorithme PX-DA (algorithme 2.5.3, courbe en pointillés) pour différentes valeurs de θ_1
: (a) $\theta_1 = 1$; (b) $\theta_1 = 2$; (c) $\theta_1 = 4$; (d) $\theta_1 = 8$. Schéma extrait de [Liu et Wu, 1999].

Chapitre 3

Approche bayésienne des modèles à équations structurelles utilisant l'expansion paramétrique

3.1 Variables latentes et modèles à équations structurelles

3.1.1 Nature ambivalente des variables latentes

L'intérêt pour les variables latentes trouve son origine dans la nature ambivalente de ces variables, qui, puisqu'elles ne sont pas observées, peuvent soit servir d'auxiliaire de calcul, soit représenter un concept latent porteur de sens.

L'utilisation en tant qu'auxiliaire de calcul a été abordée dans le chapitre précédent, de telle sorte que la curiosité est désormais aiguisée pour imaginer des variables latentes porteuses de sens. Un indice est qu'une variable non observée porteuse de sens est forcément rattachée à des variables observées par rapport auxquelles définir sa signification.

Ainsi l'utilisation des variables latentes est très répandue dans de nombreux domaines : psychologie, sociologie, éducation, marketing, médecine où on cherche à quantifier chez des individus des caractéristiques ou des comportements tels que respectivement l'*estime de soi*, l'*aliénation*, la *capacité en arithmétique*, la *fidélité* et la *dépression* à partir de variables observées sur les individus.

Un objectif est alors d'identifier les causes de la dépression, les conditions favorisant l'aptitude en arithmétique ou encore les leviers de la fidélité parmi les variables observées associées à la variable latente correspondante.

3.1.2 Les modèles à équations structurelles à variables latentes (SEM)

Plus généralement un ensemble de variables latentes peut être utilisé pour extraire une structure de données observées contribuant à la réduction de la dimension des données pourvu que les données soient corrélées. La matrice de corrélation des variables observées est alors "résumée" en la matrice de corrélation des variables latentes. Le modèle à variables

latentes ainsi caractérisé est un modèle à équations structurelles à variables latentes dit SEM (de l'anglais Structural Equation Modelling).

En termes statistiques, les modèles SEM sont des modèles à variables latentes multivariés utilisés pour représenter des structures de causalité dans les données. Les variables observées sont associées à des variables latentes dans le modèle externe, dit aussi modèle de mesure et des relations de causalité entre les variables latentes sont représentées dans le modèle interne dit aussi modèle structurel.

Le modèle de la figure 3.1 représente un exemple de structure de causalité construite à partir de dires d'experts sur les relations entre trois concepts propres à un processus de mesure : la préparation de la mesure, la mesure et le contrôle qualité, estimée à partir de variables observées du processus de mesure.

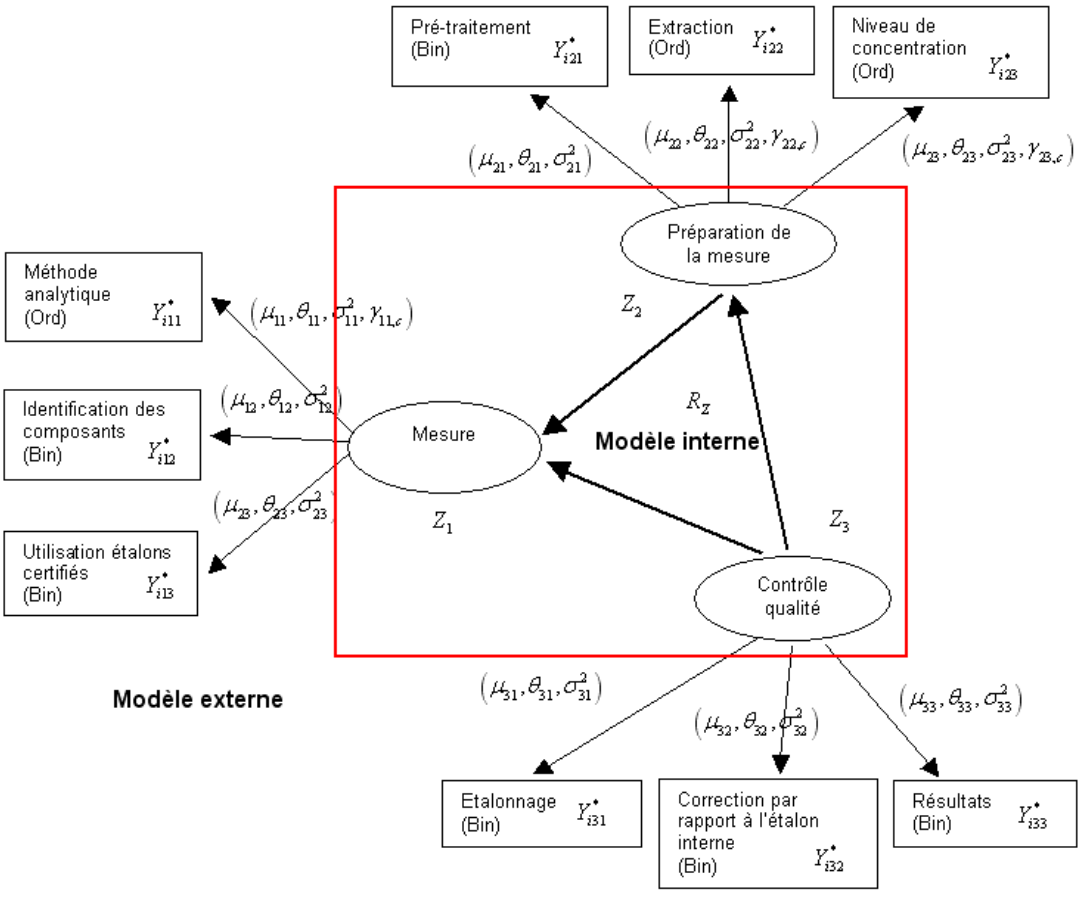


FIGURE 3.1 – Exemple de modélisation SEM d'un processus de mesure

3.1.3 Variables latentes et identifiabilité

Au-delà des perspectives alléchantes offertes par les variables latentes, il faut garder à l'esprit que le fait de rajouter des variables latentes à un modèle ne rajoute pas d'information et une donnée non observée reste une donnée conditionnelle à une donnée observée et cet ajout a un prix à payer en terme de contraintes d'identifiabilité : l'ajout de paramètres qui s'ensuit n'est pas un ajout de paramètres identifiables ! Le recours aux variables latentes doit donc se faire sous contrôle des contraintes d'identifiabilité.

3.1.4 Intérêt des modèles SEM : exemple en étude de la satisfaction

En marketing l'indice de satisfaction des consommateurs européens (ECSI) est un indicateur économique qui mesure la satisfaction des consommateurs.

Un modèle SEM a été établi pour calculer cet indicateur reposant sur les relations entre la satisfaction et d'autres concepts soit à l'origine de la satisfaction (l'image, la qualité perçue, la valeur perçue, l'attente du consommateur) soit découlant de la satisfaction (la fidélité et les déboires).

Il apparaît que ces concepts ont également des relations directes de causalité entre eux représentées par des flèches à la figure 3.2, empruntée à [Tenenhaus *et al.*, 2005], reflétant une structure complexe dans les données observées.

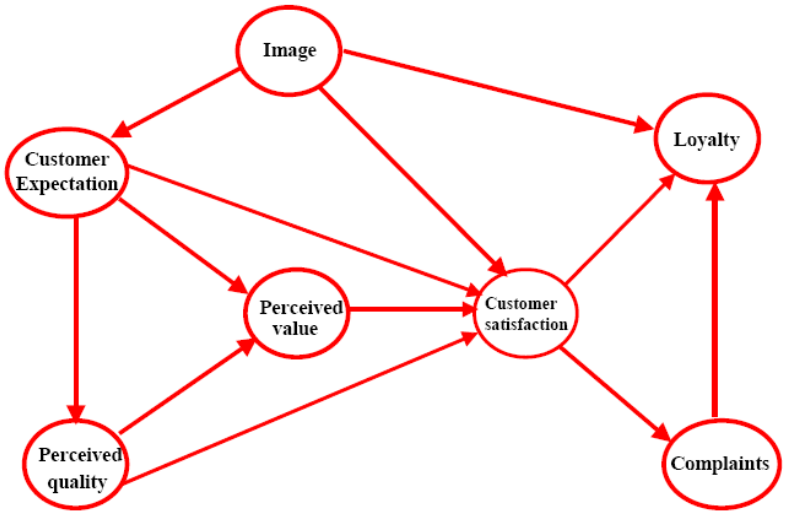


FIGURE 3.2 – Modèle ECSI décrivant les causes et les conséquences de la satisfaction des consommateurs

Chaque concept dont la satisfaction est mesuré sur plusieurs variables observées non représentées.

Les variables observées de la satisfaction déterminent des dimensions de la satisfaction ([Tenenhaus *et al.*, 2005])

- la satisfaction globale (*Overall satisfaction 1-10*)
- la satisfaction des attentes (*Fulfillment of expectations 1-10*)
- la satisfaction par rapport à un idéal (*How well do you think "your mobile phone provider" compares with your ideal mobile phone provider ? 1-10*)

L'intérêt est alors d'identifier les variables agissant sur les différentes dimensions de la satisfaction et d'identifier la(les) dimension(s) de la satisfaction qui influence(nt) le plus la fidélité.

De telles informations sont cruciales pour améliorer la politique marketing d'une entreprise autour d'un produit. Les résultats s'interprètent directement en termes d'indicateurs de la qualité de la politique marketing globale donnant les leviers d'action. Ces informations sont à confronter aux résultats du produit pour décider du suivi ou bien d'une action d'amélioration dont les axes sont définis par les indicateurs.

3.1.5 Typologie des modèles SEM

3.1.5.1 Analyse exploratoire ou confirmatoire ?

Les modèles SEM sont des constructions reposant sur des hypothèses structurelles. Ces hypothèses peuvent être inférées au cours d'une analyse dite *exploratoire* des données visant à identifier des blocs de variables corrélées et à décrire la dimension qui les résume. Par exemple, en médecine, des questions relatives à des symptômes dont la probabilité d'être observés ensemble est grande sont des indications qu'ils sont causés par le même facteur.

Les hypothèses peuvent au contraire reposer sur des fondements théoriques par exemple provenant d'un consensus d'experts dans un but *confirmatoire* de la théorie à partir des données. C'est le cas du modèle ECSI présenté à la section précédente où l'intérêt réside dans la quantification de relations de causalité données.

3.1.5.2 Relations réflexives, formatives ou mixtes ?

Les relations entre les variables latentes et les variables observées peuvent également être de nature différente. Dans les cas abordés jusqu'à présent, les variables observées d'un bloc reflètent leur variable latente. Les relations du modèle externe ainsi construites sont dites *réflexives*.

Au contraire la variable latente peut être construite à partir des variables observées de son bloc. Les relations du modèle externe ainsi construites sont dites *formatives*.

Un modèle SEM peut contenir des blocs réflexifs et formatifs (modèle de mesure *MIMIC*).

3.1.5.3 Choix de l'analyse et des relations

L'estimation bayésienne proposée dans ce chapitre concerne les modèles SEM confirmatoires dont la structure est fixée.

Les variables latentes sont des concepts nommés, porteurs d'un sens intrinsèque, reflétés par les variables observées, dans des relations réflectives.

Une caractéristique supplémentaire des modèles SEM considérés dans ce chapitre est qu'ils combinent des données observées mixtes continues, binaires et catégorielles.

3.1.6 Présentation de l'approche bayésienne des modèles SEM proposée, dite hybride

L'approche bayésienne que nous proposons se distingue de l'approche bayésienne classique des modèles SEM confirmatoires rappelée à la section 3.3.4 en permettant l'estimation *a posteriori* de toutes les relations de causalité d'un modèle SEM prédéfini.

L'analyse que nous proposons est donc un hybride entre l'analyse confirmatoire, où les blocs sont donnés et seules les relations *a priori* entre les variables latentes sont estimées, et l'analyse exploratoire où à la fois les blocs et les relations entre les blocs sont déterminés *a posteriori* .

En d'autres termes nous proposons une approche pleinement bayésienne des modèles SEM reposant sur la structure en blocs des données observées où toutes les relations entre les blocs sont estimées *a posteriori* , prenant ainsi en compte toute l'information portée par des données.

L'approche permet en particulier de retrouver les relations "marquées" du modèle SEM graphique et permet le cas échéant de revoir la théorie qui a mené à la construction du modèle graphique si les estimations des relations "marquées" sont moins significatives que des relations non "marquées", mises en évidence par l'estimation bayésienne hybride.

L'approche hybride que nous proposons est rendue possible par un travail direct au niveau de la matrice de corrélation des variables latentes, dans le cadre d'une expansion paramétrique.

3.2 Spécification du modèle SEM

3.2.1 Ecriture du modèle SEM

3.2.1.1 Le modèle externe

Soit $\boldsymbol{Y_i}$ le vecteur ligne des réalisations de l'individu i sur les p variables observées mixtes continues, binaires et catégorielles ordonnées, réparties dans q blocs de variables disjoints indicés par $k = 1, \ldots, q$ et n_k le nombre de variables observées dans le bloc k. Chaque bloc de variables est supposé refléter un concept unidimensionnel, résumé par une unique variable latente continue.

On appelle $\boldsymbol{Z_i}$ le vecteur ligne des scores de l'individu i sur les q variables latentes.

Le traitement simultané de données observées mixtes motive la spécification du modèle dans le cadre des modèles linéaires généralisés où les variables binaires et catégorielles ordonnées sont quantifiées en réponses latentes par l'utilisation de fonctions de lien probit et de valeurs de seuils suivant [Albert et Chib, 1993].

Soit $\boldsymbol{Y_i^*} = \left\{ Y_{ikj}^*,\ k = 1 \ldots q,\ j = 1 \ldots n_k \right\}$ le vecteur ligne des réponses latentes définies dans les expressions 3.3 et 3.5.

Le modèle de mesure relie chaque réponse latente Y^*_{ikj} à la variable latente structurelle Z_k qui lui est associée dans un modèle réflexif où on suppose que les réponses latentes sont conditionnellement indépendantes sachant les variables latentes structurelles.

L'écriture matricielle du modèle de mesure relie le vecteur des réponses latentes au vecteur des variables latentes structurelles pour l'individu i

$$\boldsymbol{Y_i^*} = \boldsymbol{\mu} + \boldsymbol{Z_i\theta} + \boldsymbol{E_i},\ 1 \leq i \leq n \tag{3.1}$$

où $\boldsymbol{E_i}$ est l'erreur de mesure distribuée $\boldsymbol{E_i} \sim \mathcal{N}(0, \boldsymbol{\Sigma_\varepsilon})$ avec $\boldsymbol{\Sigma_\varepsilon}$ diagonale de dimension $p \times p$ et $\boldsymbol{\theta}$ est la matrice $q \times p$ des coefficients de régressions.

Afin d'illustrer la formule matricielle on exprime la matrice $\boldsymbol{\theta}$ correspondant au modèle externe de la figure 3.1

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_{21} & \theta_{22} & \theta_{23} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta_{31} & \theta_{32} & \theta_{33} \end{pmatrix} \tag{3.2}$$

Les fonctions de liens utilisées dépendent de la nature des variables observées. Le lien identité est utilisé pour les variables observées continues et des liens probit sont préférés dans le cas de variables binaires et catégorielles ordonnées pour maintenir la conjugaison avec la vraisemblance gaussienne.

Dans le cas de variables binaires on modélise donc la probabilité de succès $p(Y_{ikj} = 1) = \Phi(\mu_{kj} + \theta_{kj} Z_{ik})$.

Le modèle probit univarié pour une observation binaire s'écrit donc

$$\begin{aligned} Y_{ikj} &= 1_{\{Y^*_{ikj} \geq 0\}} \\ Y^*_{ikj} &\sim \mathrm{N}\left(\mu_{kj} + \theta_{kj} Z_{ik}, \sigma^2_{kj}\right) \end{aligned} \tag{3.3}$$

où σ^2_{kj} n'est pas identifiable.

Dans le cas de variables catégorielles ordonnées on modélise les probabilités cumulées

$$p(Y_{ikj} \leq c) = \Phi(\gamma_{kj,c} + \theta_{kj} Z_{ik}) \tag{3.4}$$

où la position relative des seuils et des modalités est représentée à la figure 3.4.

Le modèle probit univarié pour une observation discrète ordonnée s'écrit

$$\begin{aligned} Y_{ikj} = c &\Longleftrightarrow \gamma_{kj,c-1} < Y^*_{ikj} \leq \gamma_{kj,c} \\ Y^*_{ikj} &\sim \mathrm{N}\left(\mu_{kj} + \theta_{kj} Z_{ik}, \sigma^2_{kj}\right) \end{aligned} \tag{3.5}$$

où σ^2_{kj} n'est pas identifiable, et avec $\gamma_{kj,0} = -\infty$, $\gamma_{kj,1} = 0$ et $\gamma_{kj,n_{kj}} = \infty$ pour assurer l'identifiabilité des seuils.

Si $\boldsymbol{Y_i^*}$ et $\boldsymbol{Z_i}$ étaient observées le modèle de mesure serait un modèle de régression linéaire.

Remarques : L'utilisation des liens probit rend le modèle externe invariant par inversion de l'ordre des modalités des variables à catégories ordonnées (binaires et catégorielles ordonnées) d'après la propriété d'*invariance palindromique* [McCullagh, 1980]. Dans ce cas l'inversion des modalités est compensée par l'inversion du signe du coefficient de régression associé. On en déduit, avec anticipation sur la suite, que l'estimation des variables latentes ne dépend pas du sens de l'intensité des modalités des variables observées, conditionnellement au modèle externe.

3.2.1.2 Le modèle interne

Le modèle interne résume la structure des données observées qui est transférée dans la structure de corrélation des variables latentes structurelles. Ainsi la donnée de la distribution des variables latentes structurelles est un résumé de l'information portée par les variables observées.

On décompose le vecteur des variables latentes en $\boldsymbol{Z_i} = (\boldsymbol{H_i}, \boldsymbol{\Xi_i})$, où $\boldsymbol{H_i}$ est le vecteur des q_1 variables latentes endogènes (expliquées par d'autres variables latentes du système) et $\boldsymbol{\Xi_i}$ est le vecteur des q_2 variables latentes exogènes (expliquées par aucune autre variable latente) (en particulier $q = q_1 + q_2$).

Dans l'exemple de la figure 3.1, les variables latentes *Préparation de la mesure* et *Mesure* sont endogènes car elles sont expliquées par la variable *Contrôle qualité*. En revanche la variable *Contrôle qualité* est exogène.

Le modèle des variables latentes structurelles est un modèle d'équations simultanées qui peut s'écrire sous la forme suivante

$$\boldsymbol{H_i \Pi_0} = \boldsymbol{\Xi_i \Lambda} + \boldsymbol{\Delta_i} \qquad \boldsymbol{\Pi_0} = \boldsymbol{Id} - \boldsymbol{\Pi} \tag{3.6}$$

où $\boldsymbol{\Pi}$ est la matrice $q_1 \times q_1$ des coefficients de régression entre variables latentes endogènes et $\boldsymbol{\Gamma}$ est la matrice $q_2 \times q_1$ des coefficients de régressions entre les variables endogènes et les variables exogènes, $\boldsymbol{\Delta_i}$ est le terme d'erreur distribué $\boldsymbol{\Delta_i} \sim \mathcal{N}(0, \boldsymbol{\Sigma_\delta})$ avec $\boldsymbol{\Sigma_\delta}$ diagonale, indépendant de $\boldsymbol{\Xi_i}$.

Dans l'exemple de la figure 3.1, le modèle interne s'écrit

$$(Z_{i1}\, Z_{i2}) \begin{pmatrix} 1 & 0 \\ -\pi_{12} & 1 \end{pmatrix} = Z_{i3}\ (\lambda_1\, \lambda_2) + (\delta_{i1}\, \delta_{i2}) \tag{3.7}$$

avec $\boldsymbol{\Pi_0} = \begin{pmatrix} 1 & 0 \\ -\pi_{12} & 1 \end{pmatrix}$, $\boldsymbol{\Lambda} = (\lambda_1\, \lambda_2)$, $\boldsymbol{\Delta_i} = (\delta_{i1}\, \delta_{i2})$.

Les variables exogènes ont leur propre modèle : $\boldsymbol{\Xi_i}$ est distribué selon $\mathcal{N}(0, \boldsymbol{\Phi})$ où $\boldsymbol{\Phi}$ est une matrice de corrélation.

A partir de l'équation 3.6 on peut exprimer la matrice de covariance des variables latentes en fonction des paramètres internes

$$\mathbf{\Sigma_Z} = \begin{pmatrix} \left(\mathbf{\Pi_0^t}\right)^{-1}\left(\mathbf{\Gamma^t \Phi \Gamma} + \mathbf{\Sigma_\delta}\right)\mathbf{\Pi_0^{-1}} & \left(\mathbf{\Pi_0^t}\right)^{-1}\mathbf{\Gamma^t \Phi} \\ \mathbf{\Phi \Gamma \Pi_0^{-1}} & \mathbf{\Phi} \end{pmatrix} \tag{3.8}$$

3.2.2 Deux types de contraintes d'identifiabilité : ancrage et réduction des variables latentes

Les problèmes d'identifiabilité apparaissant lors de la quantification des réponses latentes dans les modèles probit sont communément résolus en fixant la variance résiduelle de la variable latente sous-jacente à l'unité. On propose de mettre en oeuvre une expansion paramétrique pour bénéficier des propriétés d'amélioration de convergence présentées au chapitre 2 dans le cas des modèles probit univariés.

Les problèmes d'identifiabilité qui sont l'objet de cette section sont ceux portés par les variables latentes structurelles dont la variance n'est pas identifiable.

L'indétermination de l'échelle des variables latentes peut être surmontée en mettant chaque variable latente à l'échelle d'une des variables de son bloc en fixant le coefficient de régression associé à l'unité. Ce procédé s'appelle l'ancrage (en anglais, *anchoring*). Cette solution a été adoptée dans GLAMM et dans l'estimation des modèles SEM proposée par [Lee, 2007] et [Palomo *et al.*, 2007].

Suivant [Skrondal et Rabe-Hesketh, 2004], il est équivalent de travailler sur la matrice de corrélation des variables latentes, avec tous les coefficients de régression libres et identifiés (en anglais, *scaling*).

Dans cette section nous montrons l'équivalence entre les deux sous-ensembles correspondants de paramètres identifiés sous les deux ensembles de contraintes. Chaque ensemble de contraintes d'identifiabilité conduit à une paramétrisation particulière du modèle.

Dans cette section nous considérons une paramétrisation générale de la matrice de covariance des variables latentes structurelles et non pas la paramétrisation 3.6. On note, localement dans cette section, Σ_Z la matrice de covariance de terme général $\psi_{kk'}$, $1 \leq k, k' \leq q$.

Les deux sous ensembles de paramètres identifiés sont

$$\mathcal{V} = \left\{\theta_{kj}, \psi_{kk'},\, 1 \leq j \leq n_k,\, 1 \leq k < k' \leq q,\, \right\} \quad \text{sous la réduction } (\{\psi_{kk} = 1,\, 1 \leq k \leq q\}) \tag{3.9}$$

et

$$\tilde{\mathcal{V}} = \left\{\tilde{\theta}_{kj}, \tilde{\psi}_{kk}, \tilde{\psi}_{kk'},\, 2 \leq j \leq n_k,\, 1 \leq k < k' \leq q\right\} \quad \text{sous l'ancrage } \left(\left\{\tilde{\theta}_{k1} = 1,\, 1 \leq k \leq q\right\}\right) \tag{3.10}$$

Pour passer d'une paramétrisation à l'autre, les formules de transition entre les deux

sous-ensembles sont obtenues en comparant les éléments des matrices de covariance marginales des réponses latentes Y^*_{ikj} sous les deux paramétrisations.

Les éléments des matrices de covariance des Y^*_{ikj} sont reportées dans la table 3.1 et les formules de transition sont reportées dans la table 3.2.

	$\mathcal{V}$ (réduction)	$\tilde{\mathcal{V}}$ (ancrage)
$cov\left(Y^*_{ikj}, Y^*_{ikj}\right)$	$\theta^2_{kj} + 1$	$\tilde{\theta}^2_{kj}\tilde{\psi}_{kk} + 1$
$cov\left(Y^*_{ikl}, Y^*_{ikj}\right)$	$\theta_{kl}\theta_{kj}$	$\tilde{\psi}_{kk}\tilde{\theta}_{kl}\tilde{\theta}_{kj}$
$cov\left(Y^*_{ikl}, Y^*_{ik'j}\right)$	$\psi_{kk'}\theta_{kl}\theta_{k'j}$	$\tilde{\psi}_{kk'}\tilde{\theta}_{kl}\tilde{\theta}_{k'j}$

TABLE 3.1 – Eléments des matrices de covariance marginales sous les deux contraintes

$\mathcal{V}$ (réduction)	$\tilde{\mathcal{V}}$ (ancrage)
$\theta_{k1} = \sqrt{\tilde{\psi}_{kk}}$	$\tilde{\theta}_{k1} = 1$
$\theta_{kj} = \sqrt{\tilde{\psi}_{kk}}\tilde{\theta}_{kj}$	$\tilde{\theta}_{kj} = \frac{\theta_{kj}}{\theta_{k1}}$
$\psi_{kk} = 1$	$\tilde{\psi}_{kk} = \theta^2_{k1}$
$\psi_{kk'} = \frac{\tilde{\psi}_{kk'}}{\sqrt{\tilde{\psi}_{kk}}\sqrt{\tilde{\psi}_{k'k'}}}$	$\tilde{\psi}_{kk'} = \theta_{k1}\theta_{k'1}\psi_{kk'}$

TABLE 3.2 – Formules de transition entre les deux sous ensembles sous les deux contraintes

L'équivalence entre les deux sous-ensembles vient à la fois de l'interprétation de θ^2_{k1} en tant que la variance de la variable latente associée sous l'ancrage et de la conservation de la variance expliquée sous les deux paramétrisations.

Il s'en suit que l'ancrage et la réduction ont pour même effet de fixer les variances des variables latentes et la transition entre les deux sous-ensembles est opérée en multipliant ou en divisant les paramètres par les écarts-type appropriés issus de la table 3.2.

Le choix des contraintes d'identifiabilité doit donc être motivé par l'efficacité de la mise en oeuvre de l'algorithme de Gibbs.

3.2.3 Sens des variables latentes (*label switching*)

En vue de leur usage ultérieur décrit aux chapitres 1 et 4, les variables latentes sont supposées refléter une intensité, du moins bon au meilleur laboratoire, image de la graduation représentée par l'ordre des modalités des variables catégorielles ou des réalisations des variables continues.

Or, seul le produit de chaque variable latente avec son coefficient de régression externe est identifiable, ce qui implique que le signe de la variable latente peut se caler à l'opposé

de l'intensité voulue et le produit sera inchangé car le signe du coefficient de régression sera également inversé.

Ce phénomène, connu sous le nom de *label switching*, apparaît lors des simulations MCMC et a été largement décrit dans le cadre des modèles de mélange à propos de la non identifiabilité des étiquettes (*labels*) des variables latentes (voir par exemple [Jasra *et al.*, 2005]) lorsque la distribution *a posteriori* des paramètres est invariante sous la permutation des étiquettes.

La description de ce phénomène pour les modèles SEM est réalisée dans l'ouvrage [Congdon, 2003] et correspond à la bascule de signe de la variable latente et du coefficient de régression associé qui peut intervenir suite à l'échantillonnage des paramètres et des variables latentes. Ce phénomène n'est par exemple pas observé dans le cas d'une estimation par maximum de vraisemblance.

Une conséquence de ce phénomène est que les chaînes de Markov vont converger soit vers la vraie valeur des coefficients de régression externes soit vers leur inverse selon les chaînes.

Une solution au problème de *label switching* dans les modèles SEM est de définir chaque variable latente comme un reflet direct d'une des réponses latentes de son bloc qui devient la variable de référence pour le signe de la variable latente associée.

Pour cela on peut fixer le signe d'au moins un coefficient de régression dans chaque bloc, par exemple en simulant dans une loi tronquée.

On remarque que lorsque l'ancrage est retenu comme contrainte d'identifiabilité des paramètres (voir section 3.2.2) alors les labels (les signes) des variables latentes sont automatiquement identifiables.

Le choix de la réduction des variables latentes structurelles s'accompagne donc de contraintes supplémentaires visant à identifier le signe des variables latentes.

Le choix de ces contraintes dans le cadre de l'expansion paramétrique est discuté à la section 3.4.6.2.

3.3 Méthodes courantes d'estimation, méthodes bayésiennes et implémentations logicielles

3.3.1 Méthodes courantes d'estimation

Les modèles SEM sont couramment estimés par les méthodes PLS (Partial Least Squares) ou par maximum de vraisemblance.

L'approche PLS des modèles SEM est connue sous la terminologie PLS Path Modelling (PLS PM) car elle exploite directement les relations de causalité (paths) représentées dans le modèle graphique sous peu d'hypothèses contraignantes (pas d'hypothèses distributionnelles). Cette approche est décrite dans [Tenenhaus *et al.*, 2005] et résumée à la section 3.3.3.

Parmi les méthodes d'estimation par maximum de vraisemblance on peut distinguer la méthode LISREL d'analyse de la covariance des données observées qui s'applique sous

l'hypothèse de multinormalité des données observées et des variables latentes, et la méthode GLLAMM (Generalized Linear Latent And Mixed Models) qui permet de lever ces hypothèses grâce à l'implémentation propre à Stata de l'algorithme de Newton-Raphson.

Outre les hypothèses distributionnelles du modèle, le choix de la méthode d'estimation est guidé par l'intérêt porté aux variables latentes en elles-mêmes. En effet dans l'approche PLS les scores des variables latentes sont obtenus automatiquement à la convergence de l'algorithme alors que dans les deux approches par maximum de vraisemblance les scores ne sont pas directement explicités et leur calcul doit être réalisé après convergence de l'algorithme. Les scores de la variable latente d'un bloc à partir de LISREL peuvent être obtenus comme combinaison linéaire des variables observées dans ce bloc dont les poids sont les *loadings* à la convergence (méthode de Tenenhaus, [Jakobowicz, 2007]). La méthode GLLAMM repose sur l'utilisation d'intégrations numériques basées sur des quadratures adaptatives.

Les méthodes PLS et LISREL ont été en particulier comparées dans les travaux de thèse de [Jakobowicz, 2007].

Le traitement des données catégorielles a fait l'objet de travaux récents dans le cadre des méthodes PLS. Ainsi [Jakobowicz et Derquenne, 2007] ont développé une méthode combinant les modèles linéaires généralisés à l'approche PLS classique appelée approche PML (Partial Maximum Likelihood) suivis par les travaux de thèse de [Russolillo, 2009] reposant sur la quantification optimale (*optimal scaling*) des variables catégorielles.

3.3.2 Implémentations logicielles (liste non exhaustive)

La méthode PLS PM est implémentée dans le logiciel XLSTAT.

La méthode LISREL est implémentée dans le logiciel du même nom (http ://www.ssicentral.com/lisrel/).

La proc TCALIS de la version 9.2 de SAS implémente une variété de méthodes d'estimations basée sur le maximum de vraisemblance mais pas la méthode PLS PM ni de méthode bayésienne.

GLAMM est implémenté dans Stata.

Le package sem de R propose une implémentation libre de la méthode glamm.

3.3.3 Présentation de la méthode PLS-PM pour l'estimation du modèle SEM sur données continues

La méthode PLS-PM sur données observées continues repose sur une approche directe du modèle graphique en alternant l'estimation des variables latentes à partir du modèle externe et l'estimation des variables latentes à partir de leurs relations structurelles dans le modèle interne. Les poids du modèle externe sont obtenus par régressions des variables manifestes sur la variable latente associée. Les poids du modèle interne sont calculés à la convergence de l'algorithme par régression multiple ou PLS entre variables latentes selon les relations du modèle.

Les variables latentes sont réduites à chaque itération de l'algorithme et le signe de

chaque variable latente est choisi comme celui qui rend la variable latente corrélée posivement à la majorité des variables de son bloc.

Les étapes de l'algorithme sont représentées en reprenant les notations de [Tenenhaus *et al.*, 2005] locales à cette section :

- la variable latente j est notée $\boldsymbol{\xi_j}$ et a pour moyenne m_j
- l'estimation externe de $\boldsymbol{\xi_j}$ est notée $\boldsymbol{y_j}$
- l'estimation interne de $\boldsymbol{\xi_j}$ est notée $\boldsymbol{z_j}$
- le symbole $\propto$ signifie réduction du membre de droite

Algorithme 3.3.1 ***Algorithme PLS-PM, [Tenenhaus* et al.*, 2005]***
A l'itération $t+1$, $t = 1, \ldots, T$
1- Estimation externe $\boldsymbol{y_j}$ ***de la version réduite de*** $\boldsymbol{\xi_j} - m_j$

$$\boldsymbol{y_j}^{t+1} \propto \pm \left[w_{jh}^t \left(\boldsymbol{x_{jh}} - \bar{\boldsymbol{x}}_{\boldsymbol{jh}} \right) \right] \tag{3.11}$$

où $\propto$ *signifie que* $\boldsymbol{y_j}^{t+1}$ *est la version réduite du membre de droite, le signe* $\pm$ *marque le choix du signe de la variable latente.*
2- Estimation interne z_j ***de la version réduite de*** $\boldsymbol{\xi}_j - m_j$

$$\boldsymbol{z_j}^{t+1} \propto \sum_{j' : \boldsymbol{\xi_{j'}} \text{ est connecté avec } \boldsymbol{\xi_j}} e_{jj'} \boldsymbol{y_{j'}}^{t+1} \tag{3.12}$$

où $\propto$ *signifie que* $\boldsymbol{z_j}^{t+1}$ *est la version réduite du membre de droite,* $e_{jj'}$ *est égal au signe de la corrélation entre* $\boldsymbol{y_j}$ *et l'ensemble des* $\boldsymbol{y_{j'}}$ *connectés à* $\boldsymbol{y_j}$ *(cas du schéma centroïde).*
3- Estimation des poids du modèle externe (modèle réflexif)

$$w_{jh}^{t+1} = cov\left(\boldsymbol{x_{jh}}, \boldsymbol{z_j}^{t+1} \right) \tag{3.13}$$

où $\boldsymbol{z_j}^{t+1}$ *est réduite.*
A la convergence,
la variable latente réduite s'écrit finalement

$$\hat{\boldsymbol{\xi}}_{\boldsymbol{j}} = \sum \tilde{w}_{jh}^T \boldsymbol{x_{jh}} = \boldsymbol{y_j}^T + \hat{m}_j \tag{3.14}$$

où $\tilde{w}_{jh}^T$ *est tel que* $\boldsymbol{y_j}^T = \sum \tilde{w}_{jh}^T \left(\boldsymbol{x_{jh}} - \bar{\boldsymbol{x}}_{\boldsymbol{jh}} \right)$, $\hat{m}_j = \sum \tilde{w}_{jh}^T \bar{\boldsymbol{x}}_{\boldsymbol{jh}}$
4- Estimation des poids du modèle interne
Par régression multiple ou régression PLS en cas de corrélations fortes entre variables latentes.

Remarque : La convergence de l'algorithme PLS-PM n'est démontrée que dans le cas de deux blocs.

3.3.4 Méthode bayésienne existante pour l'estimation du modèle SEM sur données mixtes

Les méthodes bayésiennes ne sont pas encore couramment utilisées dans l'estimation des modèles SEM, certainement car elles ne sont pas encore représentées dans les principaux logiciels, ce qui force l'utilisateur à une implémentation propre.

L'estimation des modèles SEM dans un cadre bayésien a été formalisée dans l'ouvrage [Lee, 2007] sous les hypothèses de multinormalité des variables observées et des variables latentes, et également dans un cadre plus large qui n'est pas l'objet de cette étude. L'ouvrage a d'ailleurs été primé *"Winner of the 2008 Ziegel Prize for outstanding new book of the year"* d'après le site internet de l'éditeur.

Les principales caractéristiques de l'algorithme de Lee sur données observées mixtes continues, binaires et catégorielles ordonnées sont résumées dans les points suivants

- les contraintes d'identifiabilité consistent à fixer un coefficient de régression dans chaque bloc à 1 (ancrage)
- les contraintes sur les seuils des variables catégorielles ordonnées sont fixées à partir des données, ainsi

$$\alpha_{kj,0} = -\infty,\ \alpha_{kj,1} = \Phi^{-1}\left(f_{kj,1}\right),\ \alpha_{kj,n_{kj}-1} = \Phi^{-1}\left(f_{kj,n_{kj}-1}\right),\ \alpha_{kj,n_{kj}} = +\infty \quad (3.15)$$

  où Φ est la fonction de répartition de la loi normale centrée réduite, $f_{kj,1}$ est la proportion d'observations dans la catégorie 1 et $f_{kj,n_{kj}-1}$ est la proportion d'observations dans la catégorie n_{kj}.
- la variance résiduelle des réponses latentes est identifiable (car le seuil $\gamma_{kj,n_{kj}-1}$ est fixé)
- les paramètres internes sont simulés à chaque itération
- la matrice de covariance des variables latentes est calculée à partir de son expression analytique de la formule 3.8 à chaque itération en fonction des paramètres internes

La structure de l'algorithme de Gibbs implémenté par Lee et donnée à la figure 3.3 est à rapprocher de l'algorithme PX-Gibbs décrit à la fin de ce chapitre.

Remarques :

- les paramètres de régression du modèle externe sont estimés de la même façon que dans ce chapitre
- les paramètres de régression du modèle interne sont calculés de la même façon mais l'algorithme de ce chapitre reporte leur estimation hors du corps de l'algorithme de Gibbs
- une étape de Metropolis est introduite pour l'estimation des seuils des variables catégorielles ordonnées.

3.3.5 Motivation d'une approche alternative par expansion paramétrique sous la contrainte de réduction des variables latentes

D'après l'algorithme de Gibbs décrit à la figure 3.3, les paramètres du modèle externe sont estimés conditionnellement aux variables latentes structurelles et les variables latentes

Algorithme 3.3.2 ***Algorithme de Gibbs, [Lee, 2007]***

A l'itération $t+1$, $t = 1, \ldots, T$

1- Calculer la matrice de covariance Σ_Z ***des variables latentes en fonction des paramètres internes*** $\mathbf{\Pi}, \mathbf{\Gamma}, \mathbf{\Sigma_\delta}, \mathbf{\Phi}$
2- Générer les variables latentes en fonction de Σ_Z ***et des paramètres externes***

$$\boldsymbol{Z_i} \sim p\left(\boldsymbol{Z_i}|\boldsymbol{Y_i^*}, \boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{\Sigma_\varepsilon}, \mathbf{\Sigma_Z}\right) \tag{3.16}$$

3- Générer les paramètres du modèle externe

$$\mu_{kj}, \theta_{kj}, \sigma^2_{kj} \sim p\left(\mu_{kj}, \theta_{kj}, \sigma^2_{kj}|\boldsymbol{Y^*}, \boldsymbol{Z_k}\right) \tag{3.17}$$

4- Générer les paramètres du modèle interne

$$\mathbf{\Lambda}, \mathbf{\Sigma_\delta} \sim p\left(\mathbf{\Lambda}, \mathbf{\Sigma_\delta}|\boldsymbol{Z_i}\right) \tag{3.18}$$
$$\mathbf{\Phi} \sim p\left(\mathbf{\Phi}|\Xi\right) \tag{3.19}$$

5- Générer les seuils et les réponses latentes

$$\boldsymbol{\gamma_{kj}}, \boldsymbol{Y^*_{kj}} \sim p\left(\boldsymbol{\gamma_{kj}}, \boldsymbol{Y^*_{kj}}|\mu_{kj}, \theta_{kj}, \sigma^2_{kj}, \boldsymbol{Z_k}, \boldsymbol{Y_{kj}}\right) \tag{3.20}$$

FIGURE 3.3 – Algorithme de Gibbs pour estimer les modèles SEM sur données observées mixtes continues, binaires et catégorielles ordonnées

structurelles sont imputées *a posteriori* conditionnellement à leur matrice de covariance et non directement conditionnellement aux paramètres internes.

On en déduit que l'estimation des paramètres internes peut être repoussée à l'extérieur de l'échantillonneur de Gibbs, pourvu que les imputations des variables latentes correspondant aux différentes itérations aient été stockées.

Plus généralement, puisque l'approche bayésienne amène à travailler sur la distribution jointe des variables latentes et que cette distribution est multinormale conditionnellement à la matrice de corrélation des variables latentes, il est équivalent de modéliser la loi jointe des variables latentes par leur matrice de corrélation de telle sorte que le modèle interne de l'expression 3.6 est remplacé dans le reste de cette étude par

$$\boldsymbol{Z_i}|\boldsymbol{R_Z} \sim N\left(\mathbf{0}, \boldsymbol{R_Z}\right) \tag{3.21}$$

où $\boldsymbol{R_Z}$ est une matrice de corrélation.

Cette approche est également justifiée par le fait que la matrice de précision $\boldsymbol{R_Z^{-1}}$ contient les paramètres de régression de toutes les régressions possibles entre les variables

latentes.

Le choix de la contrainte d'identifiabilité consistant à ce que les variables latentes soient réduites (réduction) nécessite d'affecter une distribution *a priori* à la matrice de corrélation des variables latentes.

Or affecter directement une distribution *a priori* à la matrice de corrélation est complexe. Dans ces conditions, il apparaît d'autant plus judicieux de faire appel à l'expansion paramétrique. En effet, en plus des améliorations de convergence qu'elle apporte et qui sont à l'origine de son utilisation dans la plupart des problèmes, l'expansion paramétrique permet d'induire une distribution *a priori* dans le modèle des données complètes.

[Ghosh et Dunson, 2009] ont utilisé l'expansion paramétrique à cet effet dans le cas de modèles factoriels, sur des paramètres scalaires (coefficients de régression et variances résiduelles sous la contrainte que la matrice de covariance des variables latentes est l'identité) et ont montré l'amélioration de la convergence sur des données simulées.

Dans ce chapitre, nous utilisons l'expansion paramétrique pour induire une distribution *a priori* sur la matrice de corrélation des variables latentes dans le modèle des données complètes, et définir un cadre pour l'amélioration de la convergence. L'étude par simulation des propriétés de convergence est l'objet de perspectives.

3.4 Estimation bayésienne des modèles structurels utilisant l'expansion paramétrique sur données observées mixtes continues, binaires et catégorielles ordonnées

3.4.1 Mise en oeuvre de l'expansion paramétrique

A notre connaissance l'implémentation de l'expansion paramétrique n'a pas été encore été formalisée dans la littérature bayésienne autour des modèles à équations structurelles mais a été mentionnée comme étant une voie d'amélioration de la convergence de l'algorithme de Gibbs [Palomo *et al.*, 2007] présenté à la section 3.3.4.

Notre choix est d'implémenter l'expansion paramétrique pour les deux niveaux de variables latentes. Ainsi l'expansion paramétrique dans le modèle externe permet de travailler avec des réponses latentes de variance résiduelle unité dont la mise en oeuvre est en fait la mise en oeuvre de l'expansion paramétrique de [Liu et Wu, 1999] dans un modèle probit présentée en exemple dans la section 2.5.2.3. On pourra se référer à [Meza *et al.*, 2009] pour la mise en oeuvre.

Ainsi l'objectif de cette section est de présenter la mise en oeuvre de l'expansion paramétrique pour générer la matrice de corrélation des variables latentes structurelles.

Partant de l'écriture du modèle des données complètes 3.21

Modèle (C) :

$$\boldsymbol{Z_i}|\boldsymbol{R_Z} \sim N(\mathbf{0}, \boldsymbol{R_Z}) \tag{3.22}$$

où $\boldsymbol{R_Z}$ est une matrice de corrélation, on introduit les variances des variables latentes, $\alpha_1, \ldots, \alpha_q$, et on définit le paramètre d'expansion sous la forme de la matrice diagonale

$\boldsymbol{\alpha} = diag\,(\alpha_1, \ldots, \alpha_q)$.

Les variables latentes dans le modèle des données augmentées sont obtenues par la transformation $\boldsymbol{W} = \boldsymbol{\alpha}^{\frac{1}{2}}\boldsymbol{Z}$ indicée par $\boldsymbol{\alpha}$ où $\boldsymbol{\alpha}^{\frac{1}{2}} = diag\left(\sqrt{\alpha_1}, \ldots, \sqrt{\alpha_q}\right)$.

Le modèle des données augmentées par l'expansion paramétrique s'écrit donc

Modèle (X) :

$$\boldsymbol{W_i}|\boldsymbol{R_Z}, \boldsymbol{\alpha} \sim N\,(0, \boldsymbol{\Sigma_W}) \tag{3.23}$$

$$\boldsymbol{W_i} = \boldsymbol{\alpha}^{\frac{1}{2}}\boldsymbol{Z_i} \tag{3.24}$$

où $\boldsymbol{\Sigma_W} = \boldsymbol{\alpha}^{\frac{1}{2}}\boldsymbol{R_Z}\boldsymbol{\alpha}^{\frac{1}{2}}$.

Pour obtenir des variables latentes réduites dans le modèle (C) il suffit de générer la matrice de covariance $\boldsymbol{\Sigma_W}$ des variables $\boldsymbol{W}$ dans le modèle (X) et de la réduire par la transformation

$$\boldsymbol{R_Z} = \boldsymbol{\alpha}^{-\frac{1}{2}}\boldsymbol{\Sigma_W}\boldsymbol{\alpha}^{-\frac{1}{2}}, \tag{3.25}$$

les variables latentes réduites $\boldsymbol{Z}$ sont alors obtenues par la transformation

$$\boldsymbol{Z_i} = \boldsymbol{\alpha}^{-\frac{1}{2}}\boldsymbol{W_i} \tag{3.26}$$

3.4.2 Imputation des variables latentes

3.4.2.1 Imputation des réponses latentes

Cas des variables manifestes binaires Les variables observées binaires sont quantifiées selon la méthode proposée par Albert et Chib (1993) :

$$Y^*_{ikj}|\mu_{kj}, \theta_{kj}, Z_{ik}, Y_{ikj} \sim \mathrm{NT}\,(\mu_{kj} + \theta_{kj}Z_{ik}, 1; 0, \infty) \text{ si } Y_{ikj} = 1 \tag{3.27}$$

$$Y^*_{ikj}|\mu_{kj}, \theta, Z_{ik}, Y_{ikj} \sim \mathrm{NT}\,(\mu_{kj} + \theta_{kj}Z_{ik}, 1; -\infty, 0)) \text{ si } Y_{ikj} = 0 \tag{3.28}$$

où $NT\,(\mu, 1; a, b)$ est la distribution normale $N\,(\mu, 1)$ tronquée à gauche au point a et à droite au point b.

Cas des variables manifestes ordinales Les variables manifestes ordinales sont également quantifiées selon la méthode proposée par Albert et Chib (1993) pour les modèles probit univariés sur variables catégorielles ordonnées. Le modèle s'écrit

$$Y_{ikj} = c \Longleftrightarrow \gamma_{kj,c-1} < Y^*_{ikj} \leq \gamma_{kj,c} \tag{3.29}$$

$$Y^*_{ikj} \sim \mathrm{N}\,(\mu_{kj} + \theta_{kj}Z_{ik}, 1) \tag{3.30}$$

Pour l'identifiabilité des seuils on pose : $\gamma_{kj,0} = -\infty$, $\gamma_{kj,1} = 0$ et $\gamma_{kj,n_{kj}} = \infty$ où n_{kj} est le nombre de catégories de la question kj.

Les variables observées catégorielles ordonnées sont alors quantifiées selon

$$Y^*_{ikj}|, \theta, Z_{ik}, Y_{ikj} \sim \text{NT}\left(\mu_{kj} + \theta_{kj} Z_{ik}, 1; \gamma_{kj,Y_{ikj}-1}, \gamma_{kj,Y_{ikj}}\right) \tag{3.31}$$

3.4.2.2 Imputation des variables latentes structurelles

Les variables latentes sont imputées à partir des versions quantifiées centrées des variables manifestes et des valeurs courantes des paramètres $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\Sigma_\varepsilon}, \boldsymbol{R_Z}\}$ vérifiant les contraintes d'identifiabilité.

L'application de la formule de Bayes pour l'observation i s'écrit

$$\begin{aligned} \left[\boldsymbol{W_i}|\boldsymbol{Y_i^*}, \boldsymbol{\Theta}\right] &\propto \left[\boldsymbol{Y_i^*}|\boldsymbol{Z_i}, \boldsymbol{\Theta}\right] \left[\boldsymbol{Z_i}|\boldsymbol{\Theta}\right] && (3.32)\\ &\propto \left[\boldsymbol{Y_i^*}|\boldsymbol{Z_i}, \boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\Sigma_\varepsilon}\right] \left[\boldsymbol{Z_i}|\boldsymbol{R_Z}\right] && (3.33)\\ & && (3.34) \end{aligned}$$

où $\boldsymbol{Y_i^*}|\boldsymbol{Z_i}, \boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\Sigma_\varepsilon} \sim N\left(\boldsymbol{\mu} + \boldsymbol{\theta Z_i}, \boldsymbol{\Sigma_\varepsilon}\right)$ est la vraisemblance de l'observation i calculée à partir du modèle de mesure 3.1 et $\boldsymbol{Z_i}|\boldsymbol{R_Z} \sim N\left(\boldsymbol{0}, \boldsymbol{R_Z}\right)$ est la distribution jointe des variables latentes structurelles issue du modèle structurel 3.21.

La combinaison de ces deux distributions gaussiennes est immédiate et donne

$$\boldsymbol{W_i}|\boldsymbol{Y_i^*}, \boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\Sigma_\varepsilon}, \boldsymbol{R_Z} \sim N\left(\boldsymbol{\Sigma_W \theta \Sigma_\varepsilon^{-1}}\left(\boldsymbol{Y_i^*} - \boldsymbol{\mu}\right), \boldsymbol{\Sigma_W}\right) \tag{3.35}$$

où $\boldsymbol{\Sigma_W^{-1}} = \boldsymbol{\theta \Sigma_\varepsilon^{-1} \theta^t} + \boldsymbol{R_Z^{-1}}$.

La matrice $\boldsymbol{\Sigma_W}$ est la matrice de covariance des variables latentes dans le modèle des données augmentées indicé par la diagonale de $\boldsymbol{\Sigma_W}$.

Le terme $\boldsymbol{R_Z^{-1}}$ prend en compte les corrélations entre les variables latentes.

Le terme $\boldsymbol{\theta \Sigma_\varepsilon^{-1} \theta}$ prend en compte les corrélations entre les variables manifestes et les variables latentes dans le modèle des données complètes et permet le passage dans le modèle étendu

L'espérance *a posteriori* de $\boldsymbol{W_i}$ est donc une moyenne pondérée des réponses quantifiées aux questions, prenant en compte la force des relations entre les questions et les variables latentes et la force des relations entre les variables latentes.

3.4.3 Simulation de la matrice de covariance en tant que paramètre d'expansion

La matrice de covariance est simulée dans le modèle des données augmentées sous la distribution conjuguée dont la forme a été déterminée à la section 2.3.3.2

$$\boldsymbol{\Sigma_W} \sim \text{Inverse-Wishart}_{\nu_0}\left(\left(\nu_0 \boldsymbol{S_0}\right)^{-1}\right) \tag{3.36}$$

où ν_0 est le nombre de degrés de liberté, $\boldsymbol{S_0}$ est notre croyance *a priori* sur la matrice de covariance $\boldsymbol{\Sigma_W}$.

Une distribution *a priori* faiblement informative est obtenue pour $\nu_0 = q$ ou $q+1$.

La distribution conditionnelle *a posteriori* de $\boldsymbol{\Sigma_W}$ est

$$\boldsymbol{\Sigma_W}|\boldsymbol{W} \sim \text{Inverse-Wishart}_{\nu_0+n}\left(\left(\nu_0\boldsymbol{S_0} + \boldsymbol{W^t W}\right)^{-1}\right) \tag{3.37}$$

3.4.4 Estimation des paramètres du modèle externe

3.4.4.1 Estimation des coefficients de régression du modèle externe

La distribution conditionnelle *a posteriori* de $\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\Sigma_\varepsilon}$ sachant les variables latentes structurelles réduites et les réponses latentes s'écrit

$$[\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\Sigma_\varepsilon}|\boldsymbol{Y^*}, \boldsymbol{Z}] = [\boldsymbol{Y^*}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\Sigma_\varepsilon}]\,[\boldsymbol{\mu}, \boldsymbol{\theta}, \boldsymbol{\Sigma_\varepsilon}] \tag{3.38}$$

$$= \prod_{k=1}^{K}\prod_{j=1}^{n_k}\left[Y^*_{ikj}|Z_{ik}, \mu_{kj}, \theta_{kj}, \sigma^2_{kj}\right]\left[\mu_{kj}, \theta_{kj}, \sigma^2_{kj}\right] \tag{3.39}$$

La factorisation de la vraisemblance à la dernière égalité vient de l'indépendance conditionnelle des réponses latentes sachant les variables latentes structurelles.

D'après la section 2.3.2.1 du chapitre 2 la distribution conjuguée *a priori* se factorise en le produit

$$\left[\mu_{kj}, \theta_{kj}, \sigma^2_{kj}\right] = \left[\mu_{kj}, \theta_{kj}|\sigma^2_{kj}\right]\left[\sigma^2_{kj}\right] \tag{3.40}$$

sous les distributions *a priori* conjuguées suivantes

$$\mu_{kj}, \theta_{kj}|\sigma^2_{kj} \sim N\left(\mu_{0kj}, \theta_{0kj}, \sigma^2_{kj}\boldsymbol{H_0^{-1}}\right) \tag{3.41}$$

$$\sigma^2_{kj} \sim \text{Inverse-Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}s_0^2\right) \tag{3.42}$$

où $\boldsymbol{H_0^{-1}}$ est la matrice de covariance *a priori* 2×2 de (μ_{kj}, θ_{kj}), s_0^2 représentant notre croyance *a priori* sur la variance résiduelle. Une distribution *a priori* faiblement informative est donnée par $\nu_0 = 1$ or 2.

Dans le cas des modèles probit la croyance *a priori* sur les variances résiduelles est 1 ce qui amène à la distribution *a priori* suivante

$$\sigma^2_{kj} \sim \text{Inverse-Gamma}\left(\frac{1}{2}, \frac{1}{2}\right) \tag{3.43}$$

En posant $\boldsymbol{X_k} = (\boldsymbol{1}, \boldsymbol{Z_k})$, $\boldsymbol{\beta_{kj}} = (\mu_{kj}, \theta_{kj})$, $\boldsymbol{\beta_{0kj}} = (\mu_{0kj}, \theta_{okj})$ la distribution conditionnelle *a posteriori* est

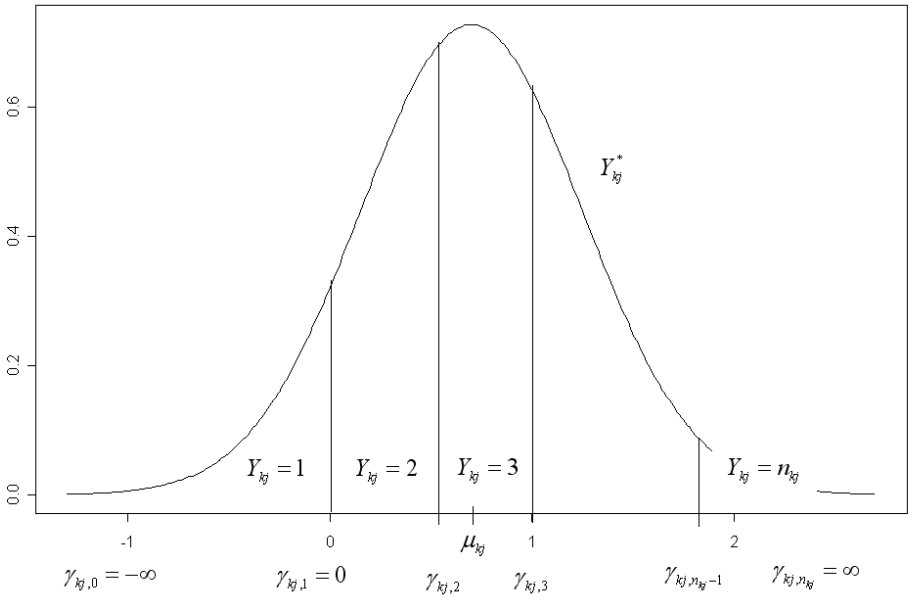


FIGURE 3.4 – Illustration des variables latentes sous-jacentes aux variables catégorielles : positionnement des seuils et des modalités.

$$\boldsymbol{\beta_{kj}}|\sigma^2_{kj}, \boldsymbol{Y^*_{kj}}, \boldsymbol{Z_k} \sim N\left(\left(\boldsymbol{X_k^t X_k} + \boldsymbol{H_0}\right)^{-1}\left(\boldsymbol{X_k^t Y^*_{kj}} + \boldsymbol{H_0\beta_0}\right), \sigma^2_{kj}\left(\boldsymbol{X_k^t X_k} + \boldsymbol{H_0}\right)^{-1}\right) \tag{3.44}$$

$$\sigma^2_{kj}|\boldsymbol{Y^*_{kj}}, \boldsymbol{Z^*_k} \sim \text{Inverse-Gamma}\left(\frac{1}{2} + \frac{n}{2}, \frac{1}{2} + \frac{n}{2}\tilde{s}^2_{kj}\right) \tag{3.45}$$

$$\text{avec } n\tilde{s}^2_{kj} = \left(\boldsymbol{Y^*_{kj}} - \boldsymbol{X_k\beta_{kj}}\right)^t\left(\boldsymbol{Y^*_{kj}} - \boldsymbol{X_k\beta_{kj}}\right) + \left(\boldsymbol{\beta_{kj}} - \boldsymbol{\beta_{0kj}}\right)^t H_0\left(\boldsymbol{\beta_{kj}} - \boldsymbol{\beta_{0kj}}\right)$$

3.4.4.2 Distribution conditionnelle *a posteriori* des seuils des variables catégorielles

Par convention dans ce document, le seuil $\gamma_{kj,c}$ sépare les modalités c et $c+1$, comme illustré à la figure 3.4.

Sous les distributions non informatives $[\gamma_{kj,c}] \propto 1$ la distribution *a posteriori* du seuil $\gamma_{kj,c}$ pour $2 \leq c \leq n_{kj} - 1$ s'écrit

$$\gamma_{kj,c}|\boldsymbol{Y_{kj}}, \boldsymbol{Y^*_{kj}}, \left\{\gamma_{kj,c'},\, c' \neq c\right\} \sim \text{Unif}\left(\max\left\{Y^*_{kj} : Y_{kj} = c\right\}, \min\left\{Y^*_{kj} : Y_{kj} = c+1\right\}\right) \tag{3.46}$$

où l'on reconnaît un estimateur EM stochastique des valeurs des seuils.

Une proposition alternative pourrait être de supposer, comme dans [Foulley et Jaffrézic, 2010], que les différences $\Delta_{kj,c} = \gamma_{kj,c} - \gamma_{kj,c-1}$ sont uniformément distribuées sur l'intervalle $[0, \delta]$.

3.4.5 Algorithme PX-Gibbs pour modèle SEM sur données observées mixtes

L'algorithme PX-Gibbs pour le modèle SEM sur données mixtes enchaîne deux étapes d'expansion paramétrique pour prendre en compte les problèmes d'identifiabilité dus aux deux niveaux de variables latentes.

L'algorithme compte ainsi trois étapes, homologues des étapes implémentées dans l'algorithme PX-EM comme annocé à la section 2.5.2 du chapitre 2, décrites comme suit

- **Etape 1 :** Mise en oeuvre de l'expansion paramétrique dans les modèles probit pour générer les réponses latentes sous la contrainte de variance résiduelle unité, conditionnellement aux variables latentes structurelles réduites et aux paramètres du modèle des données complètes.
- **Etape 2 :** Mise en oeuvre de l'expansion paramétrique dans le modèle structurel pour générer les variables latentes structurelles sous la contrainte qu'elles soient réduites, conditionnellement aux réponses latentes de variance résiduelle unité et aux paramètres du modèle des données complètes.
- **Etape 3 :** Tirage des paramètres du modèle externe dans le modèle des données complètes sous les deux contraintes.

3.4.6 Mise en oeuvre pratique de l'algorithme PX-Gibbs

3.4.6.1 Choix des paramètres *a priori*

Il est recommandé de choisir des distributions *a priori* informatives avec une variance raisonnable (ni trop grande ni trop petite) en raison de la complexité du modèle SEM.

Puisque les modèles SEM sont confirmatoires, dans le but de valider une théorie, il est raisonnable d'afficher une bonne croyance dans la structure du modèle, qui se traduit par des coefficients de corrélations *a priori* élevés et de variance autour de 1 (au maximum) entre variables latentes et entre variables observées et variables latentes.

3.4.6.2 Prise en compte des contraintes d'identifiabilité des signes des variables latentes

Garantir la consistance des signes des variables latentes lors de l'échantillonnage des chaînes de Markov consiste à conserver l'interprétation de l'intensité de la variable latente dans le sens de l'intensité de la variable observée à laquelle elle est raccordée à chaque itération de la chaîne.

Cette contrainte peut être prise en compte directement à l'étape 3 de l'algorithme PX-Gibbs présenté dans ce chapitre en tronquant la distribution d'un des coefficients de régression dans chaque bloc sur les valeurs positives. Cette solution est proposée par [Congdon, 2003] dans le cadre général des algorithmes MCMC (sans implémentation spécifique dans le cas de l'expansion paramétrique).

Algorithme 3.4.1 ***Algorithme PX-Gibbs pour estimer les modèles SEMs sur données observées mixtes continues, binaires et catégorielles ordonnées***
A l'itération $t+1$, $t = 1, \ldots, T$

1- Expansion paramétrique dans le modèle externe
- *Générer les réponses latentes dans le modèle (X)*

$Y_{ikj}^{*(t+1)} = Y_{ikj}$ *if* Y_{ikj} *is continuous*

$Y_{ikj}^{*(t+1)} \sim f\left(Y_{ikj}^{*} | \mu_{kj}^{(t)}, \theta_{kj}^{(t)}, \sigma_{kj}^{2} = 1, Z_{ik}^{(t)}, Y_{ikj}\right)$ *if* Y_{ikj} *is binary*

$$Y_{ikj}^{*(t+1)} \sim f\left(Y_{ikj}^{*} | \mu_{kj}^{(t)}, \theta_{kj}^{(t)}, \sigma_{kj}^{2} = 1, \gamma_{kj,Y_{ikj}}^{(t)}, \gamma_{kj,Y_{ikj}+1}^{(t)}, Z_{ik}^{(t)}, Y_{ikj}\right)$$
if Y_{ikj} *is ordered categorical*

- *Générer les paramètres d'expansion des modèles probit (variances résiduelles)*

$$\sigma_{kj}^{2} \sim f\left(\sigma_{kj}^{2} | \mu_{kj}^{(t)}, \theta_{kj}^{(t)}, Y_{ikj}^{*(t+1)}, Z_{ik}^{(t)}\right)$$

- *Calculer les réponses latentes dans le modèle (C)*

$$Y_{ikj}^{*(t+1)} \leftarrow Y_{ikj}^{*(t+1)} / \sqrt{\sigma_{kj}^{2}}$$

2-Expansion paramétrique dans le modèle structurel
- *Générer les réponses latentes dans le modèle (X)*

$$\boldsymbol{W_i^{(t+1)}} \sim f\left(\boldsymbol{W_i} | \boldsymbol{\mu^{(t)}}, \boldsymbol{\theta^{(t)}}, \boldsymbol{\Sigma_\varepsilon = I_q}, \boldsymbol{Y^{*(t)}}, \boldsymbol{R_Z^{(t)}}\right)$$

- *Générer les paramètres d'expansion du modèle structurel (matrice de corrélation des VL)*

$$\boldsymbol{\Sigma_Z} \sim f\left(\boldsymbol{\Sigma_Z} | \boldsymbol{W}\right)$$

- *Calculer les VL dans le modèle (C)*

$$\boldsymbol{R_Z^{(t+1)}} = \left[diag\left(\boldsymbol{\Sigma_Z}\right)\right]^{-\frac{1}{2}} \boldsymbol{\Sigma_Z} \left[diag\left(\boldsymbol{\Sigma_Z}\right)\right]^{-\frac{1}{2}}$$
$$\boldsymbol{Z^{(t+1)}} = \boldsymbol{W^{(t+1)}} \left[diag\left(\boldsymbol{\Sigma_Z}\right)\right]^{-\frac{1}{2}}$$

3- Distributions conditionnelles a posteriori des paramètres dans le modèle (C)

$$\mu_{kj}^{(t+1)}, \theta_{kj}^{(t+1)} \sim f\left(\mu_{kj}, \theta_{kj} | \boldsymbol{Y_{kj}^{*(t)}}, \boldsymbol{Z_k^{(t)}}\right)$$
$$\gamma_{kj,c}^{(t+1)} \sim f\left(\gamma_{kj,c} | \gamma_{kj,c-1}^{(t)}, \gamma_{kj,c+1}^{(t)}, \boldsymbol{Y_{kj}^{*}}^{(t)}, \boldsymbol{Y_{kj}}^{(t)}\right)$$

FIGURE 3.5 – Algorithme PX-Gibbs pour estimer les modèles SEMs sur données observées mixtes continues, binaires et catégorielles ordonnées

En choisissant une distribution *a priori* déjà décalée vers les valeurs positives (par exemple $\theta_{kj} \sim N(1,1)$) la distribution *a posteriori* du coefficient de régression θ_{kj} conditionnelle à l'intercepte μ_{kj} est obtenue à partir de l'expression du vecteur gaussien 3.44.

Pour simplifier l'expression de la distribution conditionnelle on écrit

$$\begin{pmatrix} \mu_{kj} \\ \theta_{kj} \end{pmatrix} \sim N\left(\begin{pmatrix} m_1 \\ m_2 \end{pmatrix}, \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} \right) \tag{3.47}$$

où $\begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = \left(\boldsymbol{X_k^t X_k} + \boldsymbol{H_0}\right)^{-1} \left(\boldsymbol{X_k^t Y_{kj}^*} + \boldsymbol{H_0 \beta_0}\right)$

et

$R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix} = \sigma_{kj}^2 \left(\boldsymbol{X_k^t X_k} + \boldsymbol{H_0}\right)^{-1}$ d'où

$$\theta_{kj} | \mu_{kj}, \sigma_{kj}^2, \boldsymbol{Y_{kj}^*}, \boldsymbol{Z_k} \sim N\left(m_1 + R_{12} R_{22}^{-1} (\mu_{kj} - m_2), R_{11} - R_{12} R_{22}^{-1} R_{12}^t\right) \tag{3.48}$$

Cette écriture permet la prise en compte immédiate de la constrainte d'identifiabilité sur les signes des variables latentes $\boldsymbol{Z_k}$ en générant chaque θ_{k1} dans la loi normale tronquée de moyenne et de variance définies à la l'expression précédente, à savoir

$$\theta_{k1} | \mu_{k1}, \sigma_{k1}^2, \boldsymbol{Y_{k1}^*}, \boldsymbol{Z_k} \sim NT\left(m_1 + R_{12} R_{22}^{-1} (\mu_{kj} - m_2), R_{11} - R_{12} R_{22}^{-1} R_{12}^t; 0, +\infty\right) \tag{3.49}$$

La mise en oeuvre de cette contrainte conserve donc l'algorithme PX-Gibbs tel qu'il a été implémenté dans ce chapitre.

On a relevé dans la littérature une autre façon de prendre en compte cette contrainte d'identifiabilité sur les signes, qui consiste à réaliser l'estimation des paramètres externes dans le modèle des données augmentées puis à appliquer des formules de reconstitution pour retrouver les paramètres dans le modèle des données complètes. Cette façon revient à rajouter l'estimation des paramètres externes dans la partie 2 de l'algorithme PX-Gibbs et à affecter la partie 3 à la reconstitution des paramètres. Cette méthode est mise en oeuvre par [Ghosh et Dunson, 2009] dans l'analyse bayésienne des modèles factoriels utilisant l'expansion paramétrique.

3.4.7 Validation du modèle SEM par posterior predictive p-values

La validation repose sur l'utilisation des Posterior Predictive p-values (PP p-values) développées dans [Gelman *et al.*, 1996]. Les PP p-values sont calculées à partir des distributions prédictives *a posteriori* , intégrées sur les paramètres et les variables latentes.

Soit $\boldsymbol{Y_{rep}}$ des données simulées sous le même modèle que celui qui a généré les données observées $\boldsymbol{Y}$, disons H_0, et les mêmes paramètres $\boldsymbol{\Theta}$ et de variables latentes $\boldsymbol{Z}$.

La distribution prédictive *a posteriori* des données répliquées y_{rep} est alors définie par

$$\begin{aligned}\mathbb{P}^{H_0}\left(\boldsymbol{Y_{rep}}|\boldsymbol{y}\right) &= \int \mathbb{P}^{H_0}\left(\boldsymbol{Y_{rep}}, \boldsymbol{\Theta}, \boldsymbol{Z}|\boldsymbol{Y}\right) \mathrm{d}\boldsymbol{\Theta}\mathrm{d}\boldsymbol{Z} \\ &= \int \mathbb{P}^{H_0}\left(\boldsymbol{Y_{rep}}|\boldsymbol{\Theta}, \boldsymbol{Z}\right) \left[\boldsymbol{\Theta}, \boldsymbol{Z}|\boldsymbol{y}\right] \mathrm{d}\boldsymbol{\Theta}\mathrm{d}\boldsymbol{Z}\end{aligned} \tag{3.50}$$

La PP p-value est définie comme la probabilité de queue d'une certaine fonction de discrépance D (analogue aux satistiques utilisées pour calculer les p-values classiques) sous la distribution prédictive *a posteriori*

$$\begin{aligned}PPp\left(\boldsymbol{Y}\right) &= \mathbb{P}^{H_0}\left(D\left(\boldsymbol{Y_{rep}}, \boldsymbol{\Theta}, \boldsymbol{Z}\right) \geq D\left(\boldsymbol{Y}, \boldsymbol{\Theta}, \boldsymbol{Z}\right)|\boldsymbol{Y}\right) \\ &= \int \mathbb{P}^{H_0}\left(D\left(\boldsymbol{Y_{rep}}, \boldsymbol{\Theta}, \boldsymbol{Z}\right) \geq D\left(\boldsymbol{Y}, \boldsymbol{\Theta}, \boldsymbol{Z}\right)\right) \left[\boldsymbol{\Theta}, \boldsymbol{Z}|\boldsymbol{Y}\right] \mathrm{d}\boldsymbol{\Theta}\mathrm{d}\boldsymbol{Z} \qquad (3.51) \\ &= \int\int \mathbf{1}_{D(\boldsymbol{Y_{rep}}, \boldsymbol{\Theta}) \geq D(\boldsymbol{Y}, \boldsymbol{\Theta})} p^{H_0}\left(\boldsymbol{Y_{rep}}|\boldsymbol{\Theta}\right) p\left(\boldsymbol{\Theta}|\boldsymbol{Y}\right) \mathrm{d}\boldsymbol{Y_{rep}}\mathrm{d}\boldsymbol{\Theta} \qquad (3.52)\end{aligned}$$

Lee (2007) a choisi la fonction de discrépance du Chi2

$$D\left(\boldsymbol{Y_{rep}}, \boldsymbol{\Theta}, \boldsymbol{Z}\right) = \sum_{i=1}^{n} \left(\boldsymbol{Y_{rep_i}} - \boldsymbol{Z^i}\boldsymbol{\theta^i}\right)^t \boldsymbol{\Sigma_\varepsilon^i} \left(\boldsymbol{Y_{rep_i}} - \boldsymbol{Z^i}\boldsymbol{\theta^i}\right) \tag{3.53}$$

où $\boldsymbol{\theta^i}$ et $\boldsymbol{\Sigma_\varepsilon^i}$ sont les valeurs courantes de l'algorithme de Gibbs.

La PP p-value est alors calculée comme la proportion de tirages pour lesquels $D\left(\boldsymbol{Y_{rep}}, \boldsymbol{\Theta^i}, \boldsymbol{Z^i}\right)$ est supérieure ou égale à $D\left(\boldsymbol{Y}, \boldsymbol{\Theta^i}, \boldsymbol{Z^i}\right)$. H_0 est rejetée si la PP p-value est proche de 0.5.

3.5 Remarques conclusives

Originellement, l'expansion paramétrique a été développée pour accélérer la convergence des algorithmes EM puis de l'échantillonneur de Gibbs en libérant les contraintes sur les paramètres ce qui permet d'améliorer le mélange des chaînes et donc d'améliorer la convergence (voir les références de la section 2.5.2).

Une pratique plus récente consiste à utiliser l'expansion paramétrique pour prendre en compte des contraintes d'identifiabilité difficiles à prendre en compte autrement [Lawrence *et al.*, 2008].

L'approche développée dans ce chapitre utilise essentiellement l'expansion paramétrique pour induire une distribution *a priori* sur les paramètres non identifiables, sous les contraintes choisies (réduction).

A la différence, l'approche développée dans le cadre des modèles factoriels par [Ghosh et Dunson, 2009] a pour but explicite d'accélérer la convergence de l'échantillonneur de Gibbs sur ces modèles, de la même façon que l'approche par expansion paramétrique a permis d'accélérer la convergence des algorithmes EM dans ces mêmes modèles factoriels (voir [Liu *et al.*, 1998]). Heuristiquement, la liberté donnée aux réalisations des paramètres

dans le modèle des données étendues améliore la convergence des paramètres reconstruits dans le modèle des données observées.

Nous avons donc établi, dans ce chapitre, un cadre théorique permettant aux modèles structurels de bénéficier des propriétés de convergence améliorées grâce à l'expansion paramétrique. Ces améliorations ont été observées sur les modèles factoriels et bien d'autres modèles en pratique (voir entre autre [Ghosh et Dunson, 2009], [Liu et Wu, 1999], [Lawrence *et al.*, 2008]).

Néanmoins l'étude, par simulation, de l'amélioration des propriétés de convergence de l'algorithme PX-Gibbs par rapport à l'algorithme de Lee [Lee, 2007] pour les modèles SEM est nécessaire. Il est également intéressant de compléter cette étude par l'étude de l'influence du choix de la contrainte d'identifiabilité sur le signe des variables latentes.

Chapitre 4

Approche bayésienne pour assigner la valeur de consensus et son incertitude associée dans les comparaisons interlaboratoires en présence d'information auxiliaire

Ce chapitre a pour objectif de proposer une démarche statistique pour l'amélioration de la qualité des mesures des laboratoires de routine et le diagnostic des biais de mesure des laboratoires.

Cette démarche s'appuie sur une nouvelle modélisation et un nouveau traitement des données de comparaisons interlaboratoires.

4.1 Fondements de la démarche : des méthodes actuelles à une démarche statistique globale

L'approche consiste à valoriser l'information à disposition ou facilement accessible sur les comparaisons interlaboratoires (réplicats, incertitudes,...) et à s'appuyer sur les méthodes existantes selon le type de la comparaison d'après la classification réalisée au chapitre 1.

On propose également une démarche complète permettant de valoriser l'utilisation de méthodes de référence par les laboratoires nationaux de métrologie dans le traitement des données de comparaisons interlaboratoires.

La stratégie choisie, qui consiste à modéliser les biais de mesure des laboratoires et introduite à la fin du chapitre 1, permet de compléter les méthodes existantes en proposant, pour la première fois lors du traitement des données de comparaisons interlaboratoires, la prise en compte explicite des sources de biais et le calcul des contributions au biais.

4.1.1 Présentation de la démarche statistique globale

Nous proposons une méthodologie afin de valoriser à terme la participation d'un laboratoire national de métrologie dans un test d'aptitude, dans le calcul de futures valeurs de consensus et de leurs incertitudes associées. Notre approche s'appuie sur une modélisation prédictive des biais de mesure des laboratoires reposant sur l'expertise des sources de biais. La nature prédictive de l'approche nécessite de connaître la confiance des laboratoires dans leurs résultats, c'est-à-dire les incertitudes de mesures de chaque laboratoire et les réplicats de mesures de chaque laboratoire.

Nous proposons une approche alternative lorsqu'il n'existe pas encore de méthode de référence, où la connaissance apportée par la méthode de référence est exclusivement remplacée par de l'expertise sur les sources de biais, tirant ainsi profit d'une information jusqu'à présent laissée de côté dans le calcul de la valeur de consensus (cf. chapitre 1). L'approche est déclinée selon la disponibilité (ou l'existence) des réplicats de mesure et des incertitudes des laboratoires.

Il est intéressant de noter qu'à l'heure actuelle, l'utilisation de valeurs de référence (selon la nomenclature 3 des test d'aptitude [1], cf. introduction) à la place de valeurs de consensus en chimie (biologie médicale, alimentaire, environnement) concerne entre 10 et 50% des tests d'aptitude. Il est également plus simple d'utiliser les valeurs de référence en environnement où les composés sont bien identifiés avec un effet matrice réduit. En revanche l'utilisation de valeurs de référence est plus difficile en biologie médicale et en alimentation où les contaminants non naturels comme la mélamine peuvent être évalués par valeur de référence mais plus difficilement les contaminants endogènes [2].

D'après ces chiffres on peut donc évaluer que la démarche statistique globale (i.e. sous l'existence ou non de la méthode de référence) présentée dans ce chapitre est susceptible de concerner entre 50 et 90% des tests d'aptitude en biologie médicale, en alimentation et en environnement.

4.1.2 Au coeur de la démarche : la modélisation des sources de biais de mesure

Le coeur de ces deux nouvelles approches est la modélisation des sources de biais de mesure suivant un consensus d'experts, présentée à la fin du chapitre 1.

Dans les deux cas la modélisation des biais de mesure par un modèle à équations structurelles permet d'étudier les relations entre les différentes sources de biais (cf. chapitre 3).

L'utilisation couplée de la méthode de référence et de la nouvelle approche permet de quantifier les biais de mesure par rapport à la valeur vraie, pourvu que l'on connaisse les incertitudes des résultats des laboratoires et d'utiliser ces quantifications dans un but prédictif lors de nouveaux tests d'aptitude réalisés dans les mêmes conditions pour améliorer

1. La valeur de référence est déterminée par analyse, mesurage ou comparaison de l'entité soumise à l'essai d'aptitude avec un matériau de référence ou un étalon traçable à un étalon national ou international [ISO-CASCO, 2010].

2. Informations recueillies par Michèle Désenfant, LNE auprès de Dan Tholen, convenior de la norme ISO/IEC 17043.

la fiabilité de la valeur de consensus et de son incertitude associée.

En revanche, si l'on ne dispose pas de méthode de référence alors seule la quantification des biais de mesure par rapport à la valeur de consensus est possible. Dans ces conditions l'information contenue dans la modélisation des biais de mesure est la seule information extérieure pour améliorer le calcul de la valeur de consensus et de son incertitude associée.

4.1.3 Cadre bayésien du calcul de la valeur de consensus, de son incertitude associée et des biais de mesure

La valeur de consensus, son incertitude associée et les biais de mesure sont calculés conditionnellement à l'information auxiliaire sur les sources de biais de mesure.

L'information auxiliaire a le rôle d'une information *a priori* sur les résultats de mesure, obtenue indépendamment des résultats de mesure, combinée aux résultats de mesure dans le but de mettre à jour la connaissance de la valeur de consensus, de son incertitude associée et des biais de mesure.

Le cadre bayésien d'estimation du modèle à équations structurelles fournit un cadre bayésien au calcul de la valeur de consensus, de son incertitude associée et des biais de mesure.

La combinaison du modèle des résultats des laboratoires et du modèle des biais de mesure repose sur l'hypothèse que les résultats des laboratoires sont conditionnellement indépendants sachant l'information sur les biais

$$[\boldsymbol{x_1} \dots \boldsymbol{x_n}|\boldsymbol{Y}] = \prod_{i=1}^{n} [\boldsymbol{x_i}|\boldsymbol{Y}] \tag{4.1}$$

4.1.4 A la base de la démarche : l'utilisation des méthodes existantes

La démarche que nous proposons permet de compléter les méthodes présentées dans la revue de littérature du chapitre 1 sur le point crucial de l'étude des contributions au biais de mesure et de la prise en compte d'information auxiliaire sur les biais de mesure.

Ainsi dès que l'on dispose des incertitudes de mesure ou des réplicats de mesure des laboratoires nous recommandons d'effectuer le calcul de la valeur de consensus, de son incertitude associée et des biais de mesure à partir des méthodes existantes. L'apport de la thèse se situe alors au niveau du calcul des contributions au biais, voir les sections 4.4.1 et 4.4.2.

En l'absence d'incertitude et des réplicats nous recommandons une première application des méthodes robustes à l'échantillon des résultats de mesure afin de traiter les observations aberrantes. L'apport de la thèse se situe alors au niveau de la prise en compte d'information auxiliaire sur le biais dans le traitement des données de comparaisons interlaboratoires, voir section 4.4.3.

4.2 De l'étude des relations entre les sources de biais de mesure au diagnostic des biais de mesure

En amont de l'étude des contributions au biais, l'estimation du modèle structurel des biais de mesure apporte une information importante sur les relations entre les sources de biais. La structure des biais s'en trouve absolument décortiquée.

La décomposition du biais en composantes latentes porteuses de sens permet de comprendre les mécanismes qui engendrent les biais de mesure et mène au diagnostic des biais de mesure via l'étude des contributions au biais.

La traduction d'une source de biais en terme de qualité de la pratique, à la base de la construction du modèle structurel des biais, donne non seulement du sens à la démarche statistique mais rend également les estimations du modèle directement interprétables par les praticiens des laboratoires, augmentant de fait l'impact d'une telle approche auprès des laboratoires.

4.3 Méthodologie proposée pour valoriser l'utilisation de la valeur de référence dans les tests d'aptitude

L'existence d'une méthode de référence est en soi une avancée majeure dans la mesure d'une grandeur. Or la mise en oeuvre d'une méthode de référence est coûteuse en temps et en compétences requises et de fait ne peut pas être utilisée lors de chaque comparaison interlaboratoires.

L'idée est alors de proposer une méthode pour relier la calcul d'une valeur de consensus à la valeur de référence de cette grandeur obtenue lors d'une précédente comparaison réalisée dans les mêmes conditions. Cela est possible par l'intermédiaire des biais de mesure des laboratoires, sous la donnée des incertitudes de mesure des laboratoires.

Supposons que l'on ait accès à la valeur de référence de la grandeur mesurée lors de la comparaison, soit grâce à la participation d'un laboratoire de métrologie à la comparaison, soit parce que l'on connaît par certificat d'étalonnage la valeur et l'incertitude de ce qui est mesuré.

Sous la connaissance de la valeur de référence nous pouvons modéliser la moyenne et la variance du biais de mesure, au lieu de simplement modéliser sa variance. Notre propos est que l'information sur les biais, en moyenne et en variance, inférée grâce à la valeur de référence apporte une information *a priori* sur les biais utilisable dans des comparaisons suivantes réalisées dans les mêmes conditions.

Nous illustrons notre propos en proposant une démarche dans le cas simplifié où les biais de mesure sont constants en moyenne égaux à b_i^0 et que l'incertitude du résultat de mesure u_i^2 est identique pour un mesurande donné et un laboratoire donné et se décompose de la même manière en la somme $u_i^2 = \tau_i^2 + \sigma_i^2$ sur les comparaisons considérées.

On appelle comparaison de référence, la comparaison utilisée pour inférer la distribution *a priori* des biais à partir de la valeur de référence.

Pour chaque comparaison qui suit la comparaison de référence la distribution *a priori*

du biais du laboratoire i est $b_i \sim N\left(b_i^0, \tau_i^2\right)$.

Il s'agit alors de mettre à jour le biais des laboratoires lors de la nouvelle comparaison. Pour cela il faut d'abord calculer la valeur de consensus sachant les biais *a priori* , c'est-à-dire sachant les données de la comparaison de référence.

Pour une comparaison vérifiant les hypothèses dont les données et les paramètres ont l'exposant (1) la démarche que nous proposons consiste en les deux étapes suivantes

1. **Etape 1 :** calculer la distribution *a posteriori* de la valeur de consensus sachant la distribution *a priori* des biais, les nouvelles mesures et la variance résiduelle

$$\mu^{(1)}|x_{ij}^{(1)}, b_i, \sigma_i^2 \sim N\left(\frac{1}{\sum \frac{1}{\sigma_i^2/n_i}} \sum \frac{1}{\sigma_i^2/n_i}\left(\bar{x}_{i.}^{(1)} - b_i\right), \frac{1}{\sum \frac{1}{\sigma_i^2/n_i}}\right) \tag{4.2}$$

2. **Etape 2 :** mettre à jour la moyenne et la variance *a posteriori* des biais de chaque laboratoire d'après la formule 1.35 de [Chunovkina *et al.*, 2008]

$$b_i|x_{ij}^{(1)} = \frac{\tau_i^2}{\tau_i^2 + \frac{\sigma_i^2}{n_i}}\left(\bar{x}_{i.}^{(1)} - \mu^{(1)}\right) \tag{4.3}$$

$$u^2\left(b_i\right)|x_{ij}^{(1)} = \tau_i^2 \frac{\frac{\sigma_i^2}{n_i} + \tau_i^2 \frac{u^2\left(\mu^1\right)}{\tau_i^2 + \frac{\sigma_i^2}{n_i}}}{\tau_i^2 + \frac{\sigma_i^2}{n_i}} \tag{4.4}$$

Le calcul des contributions au biais peut alors s'effectuer par la méthode de la section 4.4.1.

La méthode permet une mise à jour séquentielle des biais de mesure des laboratoires en utilisant comme distribution *a priori* des biais la distribution *a posteriori* obtenue sur une précédente comparaison et peut être généralisée pour prendre en compte d'autres hypothèses sur les biais.

4.4 Méthodologie proposée indépendamment de l'existence de la valeur de référence

La modélisation par équations structurelles permet de quantifier les sources de biais pour chaque laboratoire par rapport à la valeur de consensus, si l'on dispose également des incertitudes des laboratoires. Puisque la valeur de consensus dépend des résultats des laboratoires, la modélisation des biais ne peut pas être utilisée dans un but prédictif lors d'autres comparaisons. Au contraire ces contributions sont calculées lors de chaque traitement de données de comparaisons interlaboratoires.

Si l'on ne dispose pas des incertitudes des laboratoires (cas fréquent en pratique) mais si l'on dispose de réplicats de mesure alors la variance du terme de biais (la variance interlaboratoire), commune à tous les laboratoires, est estimable et permet d'estimer des contributions moyennes au biais sur tous les individus.

Si l'on ne dispose ni des incertitudes ni des réplicats de mesure (cas également fréquent en pratique), les données ne permettent pas d'estimer les contributions aux biais.

Ces éléments ont mené à l'élaboration de méthodologies spécifiques permettant de tirer profit des incertitudes de mesures et des réplicats de mesures en l'absence des incertitudes au moment du calcul des contributions aux biais. Le calcul des contributions au biais est nécessairement réalisé après le calcul de la valeur de consensus et de son incertitude où l'on préconise, en l'état actuel des recherches, d'utiliser des méthodes existantes pour leur calcul (respectivement [Toman, 2007] ou [Chunovkina *et al.*, 2008] et [AMC, 1989]).

En l'absence d'incertitude de mesure, on propose une méthodologie complète dans laquelle l'information sur les biais de mesure est valorisée dès le calcul de la valeur de consensus et de son incertitude associée et utilisée pour la mise à jour des biais de mesure. L'approche repose sur l'utilisation de l'algorithme robuste de la norme NF ISO 13528

4.4.1 Valorisation des incertitudes de mesure et des réplicats de mesure dans le calcul des contributions au biais

En présence des incertitudes de mesure de tous les laboratoires et des réplicats de mesure de tous les laboratoires, nous recommandons de modéliser le biais de mesure sous la forme d'un effet aléatoire.

La démarche que nous proposons s'appuie sur la démarche de [Toman, 2007] et [Chunovkina *et al.*, 2008] présentée à la section 1.4.4.2 du chapitre 1 où la connaissance de l'écart-type τ_i du terme de biais du laboratoire i nous permet de calculer les contributions au biais de chaque laboratoire.

On rappelle l'écriture du modèle à effets aléatoires de la section 1.4.4.2

$$x_{ij} = \mu + b_i + \varepsilon_{ij} \tag{4.5}$$

où μ est la valeur vraie de la comparaison, $b_i \sim \mathcal{N}\left(0, \tau_i^2\right)$ et $\varepsilon_{ij} \sim \mathcal{N}\left(0, \sigma_i^2\right)$.

Nous illustrons notre propos dans le cas simple où une seule composante latente résume les variables de biais. Dans ce cas les composantes de biais sont conditionnellement indépendantes sachant la composante latente, le modèle interne est réduit à cette seule composante latente et les corrélations entre variables de biais sont directement modélisées dans la matrice de covariance des variables de biais.

Le calcul des contributions au biais doit prendre en compte la structure de corrélation des variables de biais. Pour cela nous nous aidons d'une propriété de l'algorithme de Gibbs, à savoir qu'à partir de la convergence de l'algorithme de Gibbs les tirages dans les distributions conditionnelles *a posteriori* sont des réalisations des distributions marginales *a posteriori* .

Ainsi l'utilisation des réalisations de l'algorithme de Gibbs utilisé pour l'estimation du modèle des biais permet de travailler avec les distributions marginales *a posteriori* des variables latentes structurelles et des réponses latentes (dans le cas de variables catégorielles) où les imputations successives sont (entre autres) intégrées sur les corrélations avec les autres variables latentes et les autres réponses latentes (respectivement).

D'après [Chunovkina *et al.*, 2008] (voir section 1.4.4.2), les biais des laboratoires sont *a posteriori* estimés par

$$\hat{b_i} = \frac{\tau_i^2}{\tau_i^2 + \frac{\sigma_i^2}{n_i}} (\bar{x}_{i.} - \hat{\mu}) \tag{4.6}$$

$$u\left(\hat{b_i}\right)^2 = \tau_i^2 \frac{\frac{\sigma_i^2}{n_i} + \tau_i^2 \frac{u(\hat{\mu})^2}{\tau_i^2 + \frac{\sigma_i^2}{n_i}}}{\tau_i^2 + \frac{\sigma_i^2}{n_i}} \tag{4.7}$$

La contribution au biais s'exprime donc *a posteriori* en pourcentage de la variance *a posteriori* du biais du laboratoire i, $u\left(b_i\right)^2$.

La régression de la forme réduite du biais *a posteriori* $\tilde{b_i} = \frac{\hat{b_i}}{\sqrt{Var(\hat{b})}}$ sur la composante latente Z (centrée et réduite) permet d'obtenir la contribution marginale de la composante latente au biais, sous la forme d'un pourcentage indépendant de l'échelle des biais

$$\tilde{b_i} = \alpha Z_i + \delta_i \tag{4.8}$$

avec une contribution moyenne au biais de

$$contrib\left(Z\right) = \frac{\alpha^2}{\alpha^2 + \sigma_\delta^2} \tag{4.9}$$

La contribution au biais du laboratoire i s'obtient en rapportant le pourcentage de biais moyen expliqué à l'échelle du biais du laboratoire i

$$contrib_i\left(Z\right) = \frac{\alpha^2}{\alpha^2 + \sigma_\delta^2} u\left(\hat{b}_i\right)^2 \tag{4.10}$$

De la même manière nous pouvons également exprimer les contributions marginales moyennes au biais, de chaque variable de biais observée à partir de la régression suivante, visant à étudier la contribution de la variable Y_k à partir de sa réponse latente Y_k^* imputée à chaque itération de l'algorithme de Gibbs

$$\tilde{b}_i = \beta_k Y_{ik}^* + \tau_{ik} \tag{4.11}$$

On obtient de cette manière la contribution marginale moyenne de la variable Y_k^* au biais sous la forme

$$contrib\left(Y_k^*\right) = \frac{\beta_k^2}{\beta^2 + \sigma_{\tau_k}^2} \tag{4.12}$$

De la même manière que précédemment on obtient la contribution de la variable Y_k au biais du laboratoire i

$$contrib_i\left(Y_k\right) = \frac{\beta_k^2}{\beta^2 + \sigma_{\tau_k}^2} u\left(\hat{b}_i\right)^2 \tag{4.13}$$

Dans le cas général où le modèle des biais de mesure est un modèle à équations structurelles il faut en plus prendre en compte la structure des variables latentes, ce qui est automatique si on réutilise les imputations jointes des variables latentes structurelles au cours de l'algorithme de Gibbs.

On remarque que le calcul des contributions à chaque itération après convergence permet d'obtenir la distribution des contributions au biais prenant en compte la variabilité dans l'échantillonnage des variables latentes.

4.4.2 Valorisation de l'information auxiliaire sur les biais de mesure en l'absence des incertitudes des laboratoires

On a vu au chapitre 1 que la modélisation du biais de mesure sous la forme d'un effet aléatoire en présence de réplicats de mesure permet l'estimation de la variance interlaboratoire soit par l'algorithme de Mandel-Paule soit par l'algorithme robuste de l'Analytical Methods Committee.

La donnée des réplicats de mesure permet ainsi l'estimation des contributions moyennes au biais des variables du modèle structurel.

La contribution au biais s'exprime dans ce cas en pourcentage de la variance interlaboratoire.

Afin de prendre en compte les corrélations entre les variables latentes (problématique décrite à la section précédente), nous travaillons également avec les distributions marginales *a posteriori* des variables latentes dont les réalisations sont fournies au cours de l'algorithme de Gibbs utilisé pour l'estimation du modèle des biais.

Nous illustrons également notre propos dans le cas simple où une seule composante latente résume les variables de biais.

On note σ_L^2 l'estimation de la variance interlaboratoire par l'une des deux méthodes décrite à la section 1.4.4.1.

D'après [AMC, 1989] les biais des laboratoires sont estimés par

$$b_i = \bar{x}_i - \mu \tag{4.14}$$

On travaille ensuite avec $\tilde{b}_i = \frac{b_i}{\sigma_L}$ le biais désormais sans unité.

La régression du biais observé réduit $\tilde{B}_i$ sur la composante latente Z (centrée et réduite) permet d'obtenir la contribution de la composante latente au biais moyen à chaque itération de l'itération de l'algorithme de Gibbs

$$\tilde{B}_i = \alpha Z_i + \delta_i \tag{4.15}$$

où α est le coefficient de régression et δ_i le terme d'erreur distribué $\delta_i \sim N\left(0, \sigma_\delta^2\right)$

On en déduit la contribution moyenne au biais de la composante latente Z

$$contrib\,(Z) = \frac{\alpha^2}{\alpha^2 + \sigma_\delta^2} \tag{4.16}$$

De la même manière nous pouvons également exprimer les contributions moyennes au biais de chaque variable de biais observée.

$$\tilde{B}_i = \beta_k Y_{ik}^* + \tau_{ik} \tag{4.17}$$

On obtient de cette manière la *contribution marginale moyenne* de la variable Y_k^* au biais sous la forme

$$contrib\,(Y_k^*) = \frac{\beta_k^2}{\beta^2 + \sigma_{\tau_k}^2} \tag{4.18}$$

Pour obtenir une contribution à l'échelle de la variance interlaboratoire on multiplie les contributions obtenues par σ_L^2.

Dans le cas général où le modèle des biais de mesure est un modèle à équations structurelles il faut en plus prendre en compte la structure des variables latentes structurelles, ce qui est immédiat d'après la justification de la section précédente.

4.4.3 Valorisation de l'information auxiliaire sur les biais en l'absence des incertitudes des laboratoires et des réplicats de mesures

La modélisation des biais par un modèle à équations structurelles fournit des composantes latentes de biais, reflétant la qualité de la pratique des laboratoires sur chacune de ces composantes. En l'absence d'une échelle adaptée aux données (typiquement fixée par une composante d'incertitude) ces composantes sans échelle ne permettent pas d'estimer les contributions aux biais.

En revanche l'information sur la qualité relative des pratiques des laboratoires demeure, et constitue une information extérieure fiable à combiner aux résultats de mesure des laboratoires pour améliorer la fiabilité de la valeur de consensus en diminuant sa dépendance aux résultats.

4.4.3.1 Combinaison du modèle des mesures et du modèle des biais : construction des poids des laboratoires

Nous proposons de transformer les composantes latentes des biais en des poids reflétant la bonne pratique des laboratoires à combiner aux résultats de mesures des laboratoires.

Si les corrélations entre les composantes de biais sont de l'ordre des coefficients de régression entre les composantes alors on peut raisonnablement construire un poids w_i pour le laboratoire i à partir de la somme s_i des composantes de biais du laboratoire i obtenue par

$$s_i = \sum_{k=1}^{K} Z_{ik} \tag{4.19}$$

Il reste à fixer l'échelle des poids, inconnue, c'est-à-dire les écarts de poids entre laboratoires. Or dans le cadre de cette section aucune donnée ne peut renseigner ce point.

La solution que l'on propose consiste à définir un écart de poids proportionnel à la différence induite par la fonction de répartition empirique sur l'échantillon des s_i définie par

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{s_i \leq x} \tag{4.20}$$

Dans le cas où l'échantillon est ordonné, en notant $s_{(i)}$, $i = 1, \ldots, n$ la i$^{\text{e}}$ statistique d'ordre de l'échantillon des s_i on a

$$F_n(x) = \frac{i}{n} \tag{4.21}$$

A titre d'illustration la figure 4.1 propose cinq représentations graphiques de différentes interpolations à partir des points de coordonnées

$$\left(s_{(i)}, \frac{i}{n}\right) \tag{4.22}$$

dans cinq cas artificiels avec $n = 18$ laboratoires correspondant à différentes positions relatives des statistiques d'ordre $s_{(i)}$, $i = 1, \ldots, n$ de l'échantillon des s_i.

Par construction les variables latentes prennent des valeurs réelles centrée autour de 0 de variance unité.

La représentation sigmoïde des deux courbes en pointillés noirs est adaptée pour des scores dont la médiane est nulle, avec une tendance centrale plus ou moins forte qui peut être ajustée par le coefficient λ de la fonction sigmoïde

$$f(x) = \frac{\exp(\lambda x)}{1 + \exp(\lambda x)} \tag{4.23}$$

En revanche la fonction sigmoïde n'est pas adaptée si la répartition des scores est dissymétrique par rapport à 0, par exemple dans le cas d'un petit nombre de mauvais laboratoires ou d'un petit nombre de bons laboratoires.

Ceci est par exemple le cas des configurations représentées par les courbes bleue et verte. La configuration bleue représente un petit nombre de laboratoires dont le score est positif et un grand nombre dont le score est négatif, tels que l'écart entre deux scores consécutifs augmente entre deux statistiques d'ordres consécutives. La médiane de l'échantillon (ici strictement négative) est affectée du poids $\frac{1}{2}$. La courbe verte représente la configuration inverse.

La droite noire représente une configuration dans laquelle la fonction de répartition empirique est proportionnelle aux scores.

Ce qu'il faut retenir de cet exemple est qu'en l'absence d'échelle fournie par les données, rien ne sert d'appliquer des transformations "au hasard" aux scores pour former des poids au risque de perdre l'information présente dans les différences de scores.

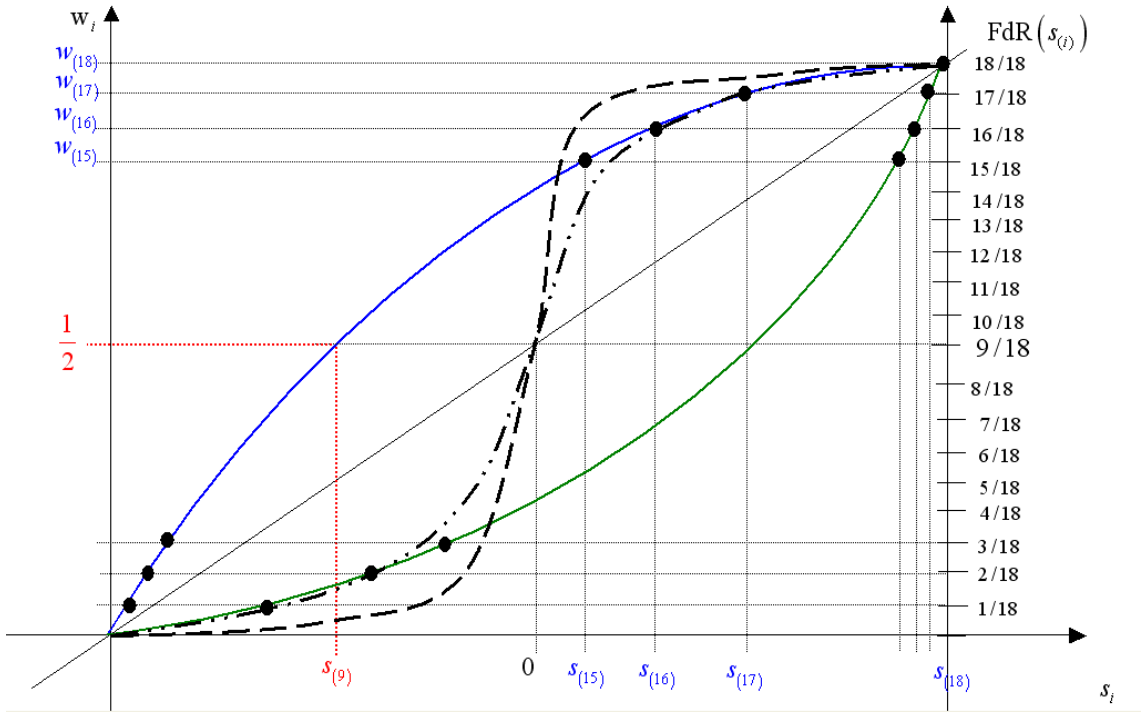


FIGURE 4.1 – , voir l'interprétation du schéma dans le corps du texte.

Le recours à la fonction de répartition empirique des statistiques d'ordre fournit une réponse indépendante de l'échelle inconnue des scores.

De cette manière chaque itération t de l'algorithme d'estimation du modèle des biais de mesure fournit un poids $w_i^{(t)}$ égal à

$$w_i^{(t)} = F_n\left(s_i^{(t)}\right) \tag{4.24}$$

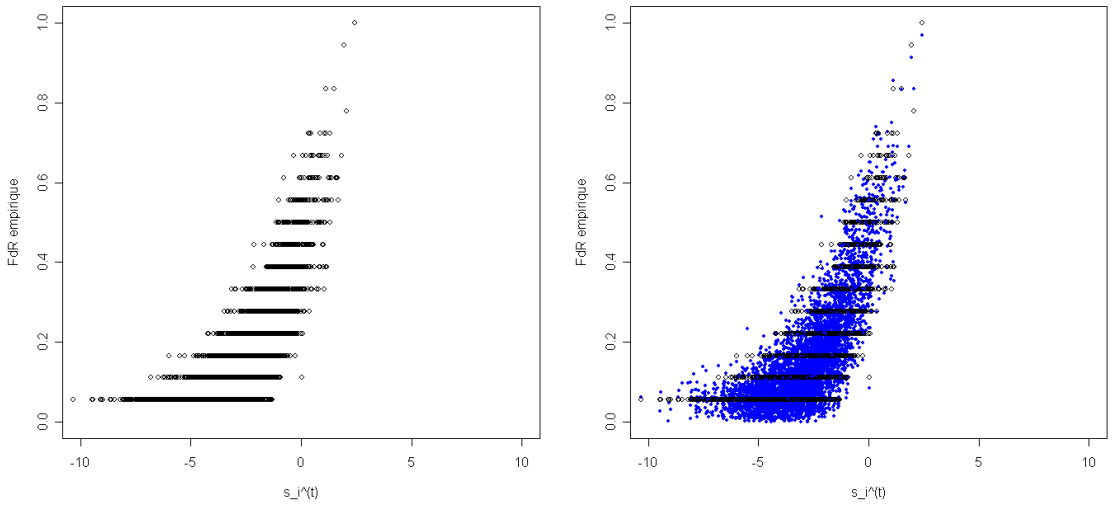
où F_n est la fonction de répartition empirique définie à l'expression 4.20.

L'avantage de cette méthode est qu'elle est exclusivement basée sur les sorties d'algorithme à disposition.

L'inconvénient de cette méthode est que l'échelle discrète des poids ne permet pas de reproduire la variabilité des sorties d'algorithme.

La représentation graphique des poids obtenus à chaque itération donne l'intuition de la réponse à apporter. A cet effet on emprunte la figure 4.2a au chapitre 5.

L'intuition suscitée par la figure 4.2a est d'introduire une variabilité artificielle au moment du calcul des poids. Une façon simple de procéder est d'introduire un bruit gaussien (tronqué sur l'intervalle de définition de la fonction de répartition empirique $[0, 1]$) au moment du calcul des poids où l'écart-type ς du bruit dépend du nombre de laboratoires par la formule suivante



(a) Poids égal à la FdR empirique

(b) Poids bruité (en bleu)

FIGURE 4.2 – Représentation graphique des poids des laboratoires obtenus à chaque itération t de l'algorithme. Dans la figure de gauche ces poids sont égaux à la FdR empirique appliquée à la somme $s_i^{(t)}$. Dans la figure de droite on applique à ces poids un bruit gaussien d'écart-type égal à 0.028 (en bleu).

$$\varsigma = \frac{1}{2}\frac{1}{n} \tag{4.25}$$

où $\frac{1}{n}$ est en fait la différence entre deux valeurs consécutives de la fonction de répartition empirique. L'écart-type du bruit est donc le même pour tous les laboratoires d'une même comparaison.

En conclusion, nous proposons d'utiliser les poids suivants à chaque itération t de l'algorithme et pour tous les laboratoires, indicés par i

$$w_i^{(t)} \sim NT\left(F_n\left(s_i^{(t)}\right), \varsigma; 0, 1\right) \tag{4.26}$$

Les poids seront normalisés à chaque itération.

4.4.3.2 Combinaison du modèle des mesures et du modèle des biais : calcul de la valeur de consensus comme moyenne pondérée

A l'itération t, les valeurs prises par les laboratoires sur les variables latentes permettent d'obtenir les poids $w_i^{(t)}$ et d'en déduire les valeurs courantes de la moyenne pondérée $x_p^{(t)}$ et de sa variance $V\left(x_p^{(t)}\right)$ à partir des formules appliquées à l'échantillon winsorizé $x_1^R, \ldots, x_n^R$

$$x_w^{(t)} = \sum_{i=1}^{n} w_i^{(t)} x_i^R \tag{4.27}$$

$$V\left(x_w^{(t)}\right) = \sum_{i=1}^{n} w_i^{2(t)} V\left(x_i^R\right) + C \tag{4.28}$$

où $C = \sum_{1 \leq i < j \leq n} w_i w_j Cov\left(x_i^R, x_j^R\right)$ prend en compte les covariances entre les résultats winsorizés.

En effet la transformation d'un échantillon en un échantillon winsorizé entraîne la perte de l'indépendance conditionnelle définie à l'expression 4.1 et induit des corrélations dans l'échantillon winsorizé.

Chaque résultat winsorizé $x_i^{(R)}$ dépend des autres résultats via l'ensemble des transformations appliquées au résultat de mesure observé x_i définies par la relation

$$x_i^{(R)} = \prod_{r=1}^{R} \varpi_i^{(r)} x_i \tag{4.29}$$

où $\varpi_i^{(r+1)}, r = 0, \ldots, R-1$ est défini par

$$\varpi_i^{(r+1)} = \begin{cases} 1 \text{ si } x_i^{(r+1)} = x_i^{(r)} \\ \dfrac{c\sigma}{|x_i^{(r)} - \hat{\mu}|} \text{ sinon} \end{cases} \tag{4.30}$$

d'après la section 1.4.1 du chapitre 1.

Les corrélations induites sont complexes et difficilement calculables car il faut prendre en compte le parcours de chaque résultat de mesure à travers l'algorithme robuste.

De plus en l'absence de réplicats de mesure le terme des covariances C n'est pas estimable à partir des résultats par la formule de covariance empirique. En l'état actuel des travaux nous proposons donc de le considérer comme nul en l'absence de données pour l'estimer.

Les chaînes ainsi construites convergent vers la distribution *a posteriori* de la moyenne pondérée et de sa variance associée.

Nous proposons de calculer la valeur de consensus sous la forme de la moyenne pondérée 4.28. Le théorème ergodique assure alors que les éléments $x_w^{(t)}$ sont des réalisations de la distribution *a posteriori* de la valeur de consensus et que les éléments $V\left(x_w^{(t)}\right)$ sont des réalisations de la distribution *a posteriori* de la variance de la valeur de consensus.

4.4.3.3 Distribution marginale *a posteriori* de la valeur de consensus

La distribution marginale *a posteriori* de la valeur de consensus prend en compte la variabilité de la moyenne pondérée et de sa variance.

Partant d'une formulation ponctuelle au niveau 1, la modélisation hiérarchique décrit ces deux sources de variabilité dans les niveaux 2 et 3 respectivement

$$x_c \sim N\left(\mu_w, \sigma_w^2\right) \tag{4.31}$$
$$\mu_w \sim N\left(\mu_{w_0}, \sigma_{w_0}^2\right) \tag{4.32}$$
$$\sigma_w^2 \sim \text{Inverse-Gamma}\left(\frac{\alpha_0}{2}, \frac{\alpha_0}{2} S_0\right) \tag{4.33}$$

où μ_{w_0} et $\sigma_{w_0}^2$ sont respectivement la moyenne et la variance de la distribution *a posteriori* de la valeur de consensus, et α_0 et S_0 sont des hyperparamètres connus et identifiés à partir de la moyenne et de la variance de la distribution *a posteriori* de la variance de la moyenne pondérée.

En pratique l'intégration sur les deux niveaux de variabilité est réalisée par simulations de Monte-Carlo.

On peut montrer après calculs que la distribution marginale *a posteriori* de la valeur de consensus suit une loi de Student. La preuve est réalisée ci-après, preuve 1.

Preuve 1 ***La distribution marginale a posteriori de la valeur de consensus suit une loi de Student***

La combinaison des deux premières lignes donne l'expression de la valeur de consensus en fonction de la variance de la moyenne pondérée, intégrée sur les valeurs possibles de la moyenne pondérée

$$x_c \sim N\left(\mu_{w_0}, \sigma_w^2 + \sigma_{w_0}^2\right) \tag{4.34}$$

Cette première étape d'intégration donne l'expression analytique de la distribution de la valeur de consensus où on observe que

- *la valeur de consensus reste centrée sur le maximum a posteriori de la moyenne pondérée μ_{w_0},*
- *la variance de la valeur de consensus est la variance a posteriori de la moyenne pondérée σ_w^2 augmentée de la variabilité de l'échantillonnage de la moyenne pondérée $\sigma_{w_0}^2$.*

Il reste à intégrer la distribution de la valeur de consensus sur toutes les valeurs possibles de la variance de la moyenne pondérée σ_w^2 à partir du modèle suivant

$$x_c \sim N\left(\mu_{w_0}, \sigma_w^2 + \sigma_{w_0}^2\right) \tag{4.35}$$
$$\sigma_w^2 \sim \textit{Inverse-Gamma}\left(\frac{\alpha_0}{2}, \frac{\alpha_0}{2} S_0\right) \tag{4.36}$$

En notant $\tau^2 = \sigma_w^2 + \sigma_{w_0}^2$ il reste à exprimer la distribution de τ^2 en fonction de α_0 et S_0. En introduisant les paramètres β_0 et $\tilde{S}_0$ de la distribution de τ^2 le modèle devient

$$x_c|\tau^2 \sim N\left(\mu_{w_0}, \tau^2\right) \tag{4.37}$$

$$\tau^2 \sim \textit{Inverse-Gamma}\left(\frac{\beta_0}{2}, \frac{\beta_0}{2}\tilde{S}_0\right) \tag{4.38}$$

où il reste à exprimer β_0 et $\tilde{S}_0$ en fonction de α_0 et S_0.

Les formules de l'espérance et de la variance de lois inverse Gamma permettent de relier ces paramètres L'espérance de τ^2 est $E\left(\tau^2\right) = \frac{\beta_0\tilde{S}_0}{\beta_0-2}$ et s'exprime en fonction de α_0 et S_0 comme suit

$$E\left(\tau^2\right) = \sigma^2_{w_0} + E\left(\sigma^2_w\right) \tag{4.39}$$

$$= \sigma^2_{w_0} + \frac{\alpha_0 S_0}{\alpha_0 - 2} = \frac{\beta_0\tilde{S}_0}{\beta_0 - 2} \tag{4.40}$$

De même on exprime la variance de τ^2, $V\left(\tau^2\right) = \frac{\left(\beta_0\tilde{S}_0\right)^2}{(\beta_0-2)^3}$ en fonction de α_0 et S_0

$$V\left(\tau^2\right) = V\left(\sigma^2_w\right) \tag{4.41}$$

$$= \frac{\left(\alpha_0 S_0\right)^2}{\left(\alpha_0 - 2\right)^3} = \frac{\left(\beta_0\tilde{S}_0\right)^2}{\left(\beta_0 - 2\right)^3} \tag{4.42}$$

On déduit des égalités 4.40 et 4.42

$$\beta_0 - 2 = \frac{\left(\alpha_0 - 2\right)^3}{\left(\alpha_0 S_0\right)^2}\left[\sigma^2_{w_0} + \frac{\alpha_0 S_0}{\alpha_0 - 2}\right]^2 \tag{4.43}$$

$$\tilde{S}_0 = \frac{\beta_0 - 2}{\beta_0}\left(\sigma^2_{w_0} + \frac{\alpha_0 S_0}{\alpha_0 - 2}\right) \tag{4.44}$$

La distribution marginale a posteriori de la valeur de consensus est donc l'intégrale par rapport à la distribution de τ^2 dont les paramètres sont identifiés à partir de l'identification des paramètres de σ^2_w

$$f_{x_c}\left(x_c\right) \propto \int \frac{1}{\sqrt{\tau^2}} \exp\left\{-\frac{1}{2}\frac{\left(x_c - \mu_{w_0}\right)^2}{\tau^2}\right\} \frac{1}{\left(\tau^2\right)^{\frac{\beta_0}{2}+1}} \exp\left\{-\frac{1}{2}\frac{\beta_0\tilde{S}_0}{\tau^2}\right\} \mathrm{d}\tau^2 \tag{4.45}$$

$$\propto \int \frac{1}{\left(\tau^2\right)^{\frac{\beta_0+1}{2}}} \exp -\frac{1}{2\tau^2}\left[\left(x_c - \mu_{w_0}\right)^2 + \beta_0\tilde{S}_0\right] \mathrm{d}\tau^2 \tag{4.46}$$

$$\propto \left[\left(x_c - \mu_{w_0}\right)^2 + \beta_0\tilde{S}_0\right]^{-\frac{\beta_0+1}{2}} \tag{4.47}$$

$$\propto \left[1 + \frac{\left(x_c - \mu_{w_0}\right)^2}{\beta_0\tilde{S}_0}\right]^{-\frac{\beta_0+1}{2}} \tag{4.48}$$

La distribution marginale a posteriori de la valeur de consensus est donc la loi de Student suivante

$$x_c \sim t_{\beta_0}\left(\mu_{w_0}, \tilde{S}_0\right) \tag{4.49}$$

où $\beta_0 = 2 + \frac{(\alpha_0-2)^3}{(\alpha_0 S_0)^2}\left[\sigma^2_{w_0} + \frac{\alpha_0 S_0}{\alpha_0-2}\right]^2$ *et* $\tilde{S}_0 = \frac{\beta_0-2}{\beta_0}\left(\sigma^2_{w_0} + \frac{\alpha_0 S_0}{\alpha_0-2}\right)$.

Ainsi, a posteriori , la valeur de consensus est centrée en $E(x_c) = \mu_{w_0}$, *de variance* $V(x_c) = \frac{\beta_0}{\beta_0-2}\tilde{S}_0 = \sigma^2_{w_0} + \frac{\alpha_0 S_0}{\alpha_0-2}$.

4.4.3.4 Moyenne *a posteriori* des biais de mesure des laboratoires

En l'absence d'incertitude et de réplicats de mesure il impossible de calculer la variance des biais de mesure. Seule l'estimation en moyenne est possible.

Deux options s'offrent à nous pour calculer le biais moyen d'un laboratoire participant :

- à chaque itération, calculer l'écart du résultat du laboratoire à la moyenne pondérée (qui a l'avantage d'être déjà calculée pour le calcul de la valeur de consensus) ou,
- à chaque itération calculer directement le biais du laboratoire à partir des poids et des résultats winsorizés des laboratoires

La *première option* est immédiate et consiste à calculer à chaque itération la quantité

$$B_i^{(t)} = x_i - x_w^{(t)} \tag{4.50}$$

La *deuxième option* nécessite d'exprimer $x_w^{(t)}$ en fonction des poids $w_i^{(t)}$

$$x_w^{(t)} = \sum_{i=1}^{n} w_i^{(t)} \bar{x}_i^{(R)} \tag{4.51}$$

Le biais de mesure du laboratoire i peut donc être réécrit en fonction des résultats winsorizés de tous les laboratoires à l'itération t

$$B_i^{(t)} = -w_1^{(t)}x_1^{(R)} - w_2^{(t)}x_2^{(R)} \ldots - w_{i-1}^{(t)}x_{i-1}^{(R)} + x_i - w_i^{(t)}x_i^{(R)} - w_{i+1}^{(t)}x_{i+1}^{(R)} \ldots w_n^{(t)}x_n^{(R)} \tag{4.52}$$

A la convergence de l'algorithme A, de dernière itération R, l'observation 'brute' x_i est soit une *valeur non aberrante* auquel cas $x_i^{(R)} = x_i$ soit une *valeur aberrante* auquel cas $x_i^{(R)} = \prod_{r=1}^{R} \varpi_i^{(r)} x_i$ d'après la section 1.4.1 du chapitre 1 où on a montré que, à l'itération $r+1$, $r = 0, \ldots, R-1$,

$$\varpi_i^{(r+1)} = \begin{cases} 1 \text{ si } x_i^{(r+1)} = x_i^{(r)} \\ \dfrac{c\sigma}{|x_i^{(r)} - \hat{\mu}|} \text{ sinon} \end{cases} \tag{4.53}$$

Le biais d'un laboratoire non aberrant (i.e. tel que $x_i^{(R)} = x_i$) peut donc s'exprimer directement en fonction des poids et des résultats winsorizés sous la forme

$$B_i^{(t)} = \left(1 - w_i^{(t)}\right) x_i - \sum_{j \neq i} w_j^{(t)} x_j^{(R)} \tag{4.54}$$

ou encore sous la forme

$$B_i^{(t)} = \sum_{j \neq i} w_j^{(t)} \left(x_i - x_j^{(R)}\right) =: B_i^{(t)}(x_i) \tag{4.55}$$

en remarquant que $(1 - w_i) = \sum_{j \neq i} w_j$

En notant $d_{ij}^{(R)} = x_i - x_j^{(R)}$ le "degré d'équivalence" (cette appellation peut être un abus de langage) entre les résultats winsorizés des laboratoires, le biais du laboratoire i s'interprète comme la somme pondérée des degrés d'équivalence entre le laboratoire i et les résultats winsorizés des autres laboratoires, où la pondération est issue du traitement de l'information auxiliaire sur les biais de mesure.

Le biais d'un laboratoire i aberrant vaut à l'itération t

$$\tilde{B}_i^{(t)} = x_i - \sum_{j=1}^{n} w_j x_j^{(R)} \tag{4.56}$$

$$= x_i - x_i^{(R)} + \sum_{j=1}^{n} w_j \left(x_i^{(R)} - x_j^{(R)}\right) \tag{4.57}$$

$$= x_i - x_i^{(R)} + \sum_{j \neq i} w_j \left(x_i^{(R)} - x_j^{(R)}\right) \tag{4.58}$$

$$= x_i - x_i^{(R)} + B_i^{(t)}\left(x_i^{(R)}\right) \tag{4.59}$$

La distribution des biais des laboratoires obtenue à partir des itérations des chaînes de Markov prend en compte la variabilité d'échantillonnage dans les chaîne, mais l'écart-type de cette distribution n'est pas l'écart-type du biais.

Chapitre 5

Applications

La méthodologie permettant d'intégrer de l'information auxiliaire sur les biais de mesure dans le traitement des données de comparaisons interlaboratoires est illustrée dans ce chapitre sur des données de tests d'aptitude en environnement.

Les applications en santé dans le cadre du projet européen présenté en introduction de la thèse seront illustrées ultérieurement (dans des articles) sur la mesure de la concentration en créatinine, un marqueur de l'insuffisance rénale, dès réception des données, non disponibles au moment de la rédaction des travaux de thèse.

5.1 Démarche de l'étude

Un travail collaboratif a été instauré avec le deuxième organisateur européen de tests d'aptitude, le BIPEA, Bureau Interprofessionnel d'Etudes Analytiques dès 2008, sur la base du volontariat des laboratoires participant aux tests d'aptitude.

L'étude a nécessité la mise en oeuvre des étapes suivantes détaillées ci-après

1. choix d'un circuit de tests d'aptitude et d'analytes, conjointement entre le LNE et le BIPEA
2. constitution d'un ensemble d'experts sur la mesure des analytes du circuit sélectionné
3. sollicitation des laboratoires du circuit sélectionné
4. réalisation de mesures de références des analytes sélectionnés
5. recueil de l'information auxiliaire sur les biais de mesure
6. application de la méthode statistique permettant de combiner les résultats de mesure à l'information auxiliaire dans le cadre du circuit choisi.

5.1.1 Choix d'un circuit de tests d'aptitude et d'analytes

Parmi les nombreux circuits de tests d'aptitude mis en oeuvre par le BIPEA, le choix, motivé par des experts du LNE après discussion avec le BIPEA, s'est orienté vers le circuit de mesure des micropolluants organiques de l'eau douce, le circuit 37.

Parmi les micropolluants organiques proposés par le BIPEA dans le cadre du circuit 37, cinq analytes de la famille des triazines ont été sélectionnés.

Cette famille d'analytes est mesurée trois fois par an (i.e. lors de trois campagnes de tests d'aptitude) à des niveaux de concentration bas, moyen et haut.

Dans cette étude nous nous intéressons à la campagne de février 2009 pour laquelle nous disposons exceptionnellement des valeurs de références (voir section 5.1.5) des cinq analytes.

5.1.2 Constitution d'un ensemble d'experts sur la mesure des analytes sélectionnés

Le projet réunit des experts du LNE, du BIPEA et des experts provenant des laboratoires suivants

- LDAR (Laboratoire Départemental d'Analyses et de Recherche de Périgueux)
- LA DROME Laboratoires
- LCDI
- CRECEP (Centre de recherche d'expertise et de contrôle des eaux de Paris)
- ESPCI (Laboratoire Environnement et Chimie analytique de l'Ecole Supérieure de Physique et Chimie Industrielle de la ville de Paris)

L'ensemble des experts a été réuni en octobre 2008 afin d'identifier et de classer les variables susceptibles d'expliquer les biais de mesure des laboratoires lors de la mesure des triazines.

5.1.3 Sollicitation des laboratoires participant à l'étude

La sollicitation de la quarantaine de laboratoires du circuit 37 a été effectuée dans un courrier en février 2009 (voir annexe 2) s'appuyant sur les bénéfices attendus pour les laboratoires de la méthode statistique proposée.

Dèjà, les bénéfices attendus étaient exprimés au niveau du suivi de l'évolution des biais de mesure, en proposant une estimation plus fiable de la valeur de consensus et de son incertitude associée, et au niveau du contrôle des biais de mesure, en permettant pour la première fois une étude statistique des biais de mesure à partir de variables de biais identifiées.

La sollicitation des laboratoires a consisté en pratique à une invitation à participer à l'enquête correspondant au recueil de l'information auxiliaire sur les biais de mesure.

5.1.4 Réalisation de mesures de référence

Une information cruciale consiste en la connaissance de la valeur de référence et de son incertitude associée.

Les mesures de référence ont été réalisées, à titre exceptionnel, dans le cadre du projet européen où le LNE a réalisé la mesure des triazines lors de la campagne de mars 2009.

L'utilisation de valeurs de référence dans ce travail a pour but de positionner les ré-

sultats de différentes méthodes par rapport aux valeurs de référence, où les valeurs de référence ont pleinement le rôle de *référence*.

En l'absence des incertitudes de mesure des laboratoires et des réplicats de mesure (voir section 5.1.6) l'information apportée par les valeurs de référence ne peut être pleinement exploitée, par exemple dans le cadre défini à la section 4.3 du chapitre 4.

5.1.5 Recueil de l'information auxiliaire sur les biais de mesure

Le recueil de l'information auxiliaire sur les biais de mesure a été réalisé par un questionnaire en ligne sur internet développé grâce à un soutien technique de l'équipe informatique du LNE qui a également permis un accès sécurisé et privé au questionnaire.

De nombreuses relances téléphoniques ont permis d'atteindre le nombre de 18 laboratoires répondant pour les triazines.

5.1.6 Application de la nouvelle modélisation et du nouveau traitement des données des tests d'aptitude aux données du circuit 37

D'après le chapitre 4 le choix de la méthode à mettre en oeuvre dans le cadre de la nouvelle approche dépend de la nature des résultats de la comparaison considérée.

Les résultats des comparaisons organisées dans le cadre du circuit 37 considéré dans ce chapitre consistent en **un seul résultat de mesure sans incertitude associée et sans réplicats de mesure**.

D'après la section 4.4 ces éléments ne permettent pas de calculer les contributions au biais de mesure de chaque laboratoire ni les contributions moyennes au biais des laboratoires.

D'après la section 4.4.3 la démarche que nous proposons consiste dans le cas de cette application à

1. estimer le modèle structurel des biais de mesure
2. transformer les estimations de chaque laboratoire sur les composantes latentes du biais de mesure en poids par laboratoire
3. appliquer l'algorithme A à l'échantillon formé par les résultats de mesure afin d'obtenir des résultats winsorizés
4. combiner les poids aux résultats de mesures winsorizés

5.1.7 Remarques générales sur le modèle des biais de mesure

On insiste sur le fait que dans cette étude un modèle des biais de mesure est supposé propre à une famille d'analytes, c'est-à-dire que les poids calculés pour chaque laboratoire à partir du modèle des biais sont propres à la mesure d'une famille complète d'analytes.

En particulier le modèle des biais de mesure des triazine diffère par exemple du modèle des biais de mesure des BTEX afin de prendre en compte le caractère volatil des BTEX.

Cependant au sein d'une même famille, la mesure de certains analytes peut être plus ou moins bien maîtrisée. C'est une idée similaire qui a finalement orienté le choix des analytes

de la famille des triazines : nous avons choisi les cinq analytes parmi les huit proposés par le BIPEA pour lesquels la mesure de référence existait au LNE en mars 2009 car une méthode de référence est propre à la mesure d'un analyte.

Ce fait connu des praticiens se retrouve dans les données que nous avons recueillies où on observe que pour un même laboratoire certaines mesures peuvent être dans les bornes de confiance de la valeur de référence et d'autres en dehors.

Dans ces conditions il serait légitime de proposer un modèle des biais de mesure pour chaque analyte. Nous pensons que cela sera possible dès que les procédés seront bien maîtrisés, c'est-à-dire dès que les sources de biais propres à chaque analyte auront été identifiées.

En attendant, les sources de biais prises en compte dans chacun des deux modèles des biais que nous proposons sont communes à la mesure de tous les analytes de telle sorte que les sources de biais restantes caractérisent les analytes. D'un point de vue statistique le modèle complet des biais revient à une modélisation hiérarchique des sources de biais.

En d'autres termes la part de variabilité des biais de mesure expliquée par les modèles de ce chapitre est en fait la part de la variance des biais non expliquée par les variables de biais propres à chaque analyte, inconnues en l'état actuel de nos connaissances.

5.2 Mise en oeuvre de la nouvelle approche sur la mesure des Triazine lors des tests d'aptitude du BIPEA

5.2.1 Résultats du LNE

Le tableau 5.1 résume les résultats de mesure du LNE sur les cinq analytes obtenues avec la méthode de référence pour chaque analyte.

	Analyte 1	Analyte 2	Analyte 3	Analyte 4	Analyte 5
valeur de réf	29.21	107.27	98.41	44.67	33
incertitude	1.61	2.25	2.01	1.34	1.73
intervalle 95%	[25.99,32.43]	[102.77,111.77]	[94.39,102.43]	[41.99,47.35]	[29.54,36.46]

TABLE 5.1 – Résultats du LNE (en ng /L) sur la mesure des 5 analytes de la famille des triazine : valeur de référence, incertitude-type et intervalle de confiance à 95 %.

5.2.2 Résultats de mesures des laboratoires

5.2.2.1 Données

Le tableau 5.2 regroupe les résultats de mesures brut et winsorizés des laboratoires sur les cinq analytes. Le grisé met en évidence les résultats qui ont été winsorizés lors de l'application de l'algorithme A, pour chaque analyte et chaque laboratoire.

Suivant une remarque de la section 5.1.7 on distingue pour chaque analyte et chaque laboratoire les résultats tombant dans l'intervalle de confiance correspondant du LNE (en vert) et les résultats tombant en dehors (en rouge).

	Id_labo	Analyte 1	rob_1	Analyte 2	rob_2	Analyte 3	rob_3	Analyte 4	rob_4	Analyte 5	rob_5
1	11014	50.00	48.68	115.00	115.00	93.00	93.00	50.00	50.00	50.00	40.65
2	10493	32.70	32.70	111.10	111.10	93.90	93.90	44.40	44.40	32.60	32.60
3	10490	80.00	48.68	100.30	100.30	93.10	93.10	50.00	50.00	50.00	40.65
4	11626	24.70	24.70	59.30	91.15	64.70	81.65	40.70	40.70	20.00	27.15
5	11157	20.00	20.03	89.00	91.15	73.00	81.65	38.00	38.00	25.00	27.15
6	11456	32.00	32.00	88.00	91.15	90.00	90.00	50.00	50.00	31.00	31.00
7	10437	26.00	26.00	95.00	95.00	79.00	81.65	42.00	42.00	40.00	40.00
8	11082	24.90	24.90	98.00	98.00	89.00	89.00	39.30	39.30	28.10	28.10
9	11323	29.00	29.00	101.00	101.00	124.00	105.35	44.00	44.00	36.00	36.00
10	11080	45.00	45.00	116.00	116.00	114.00	105.35	46.00	46.00	36.00	36.00
11	11079	48.60	48.60	105.40	105.40	134.70	105.35	52.70	52.40	57.50	40.65
12	11088	37.00	37.00	109.00	109.00	91.00	91.00	37.00	38.00	30.00	30.00
13	11753	26.90	26.90	80.70	91.15	98.30	98.30	40.30	40.30	34.50	34.50
14	10415	36.00	36.00	113.00	113.00	106.00	105.35	47.00	47.00	32.00	32.00
15	11885	61.00	48.68	103.90	103.90	82.50	82.50	58.80	52.40	42.40	40.65
16	10199	45.00	45.00	126.00	118.15	98.00	98.00	35.00	38.00	31.00	31.00
17	11095	26.00	26.00	112.00	112.00	116.00	105.35	50.00	50.00	39.00	39.00
18	10872	68.70	48.68	127.80	118.15	94.40	94.40	48.40	48.40	33.30	33.30

TABLE 5.2 – Données brutes et winsorizées (colonne rob) pour les 5 analytes mesurés de la famille des Triazine. En vert, les résultats dans l'intervalle de confiance du LNE, en rouge les résultats en dehors.

5.2.2.2 Commentaires

En l'absence d'incertitude associée aux résultats de mesure la position du résultat par rapport à l'intervalle de confiance du LNE ne permet pas de conclure qu'un laboratoire est bon ou mauvais sur la mesure d'un analyte donné.

Prenons l'exemple du laboratoire 11079 : de sa seule mesure de l'analyte 2 on conclurait que le laboratoire est bon dans la mesure de cet analyte, cependant comme il semble être très mauvais sur la mesure des autres analytes, on ne peut décidément rien conclure.

Il existe un moyen de confirmer ou d'infirmer l'idée que l'on se fait d'un laboratoire : le recours à une information extérieure sur le laboratoire.

Cette information extérieure peut être la façon dont le laboratoire met en oeuvre le contrôle de la qualité. D'après le tableau 5.3 le laboratoire 11079 met partiellement en oeuvre le contrôle qualité en n'ayant qu'une seule bonne pratique sur les trois variables définissant le contrôle qualité dans le cadre de cette application. Cette information extérieure dite auxiliaire ne permet pas d'accorder *a priori* une confiance élevée dans les résulats de ce laboratoire.

Toujours d'après le tableau 5.3 ce laboratoire met en oeuvre la meilleure des 5 techniques relevées, ce qui influence désormais positivement notre jugement sur la qualité des mesures réalisées. Mais on sait par expertise que la mise en oeuvre du contrôle qualité influence la mise en oeuvre de la technique. Dans ces conditions, que conclure sur la qualité de ce laboratoire ?

C'est là qu'intervient la puissance de la modélisation des relations entre les sources de biais que nous proposons. En effet il est clair sur cet exemple que la qualité d'un laboratoire ne peut être notée isolément des autres laboratoires : c'est la pratique relative des laboratoires qui permet d'évaluer leur qualité relative. Puisque les relations entre les variables peuvent être complexes, seule une analyse statistique peut permettre d'estimer

ces relations.

L'approche bayésienne que nous avons développée au chapitre 3 permet à cet effet de travailler conjointement sur toutes les relations possibles en dégageant le degré de confiance (sous forme de distribution de probabilité) que les données du modèle des biais apportent à chacune des relations du modèle.

5.2.3 Réponses au questionnaire et structure des biais

En raison du faible nombre de laboratoires seules les variables les plus pertinentes, parmi toutes celles susceptibles d'expliquer le biais de mesure, ont été retenues pour modéliser la structure des biais de mesure.

Ainsi la structure des biais repose sur la structure de corrélation entre les six variables suivantes réparties dans les trois blocs annoncés dans le chapitre 1 dont le nom figure sur la ligne précédé de la nature de la variable.

- Technique ordinale à 5 modalités, bloc *Mesure*
- Extraction, ordinale à 4 modalités, bloc *Préparation de la mesure*
- Niveau de concentration, ordinale à 3 modalités, bloc *Préparation de la mesure*
- Nature de l'étalonnage, binaire, bloc *Contrôle qualité*
- Utilisation d'étalons certifiés, binaire, bloc *Contrôle qualité*
- Correction par rapport à l'étalon interne, binaire, bloc *Contrôle qualité*

Les réponses des 18 laboratoires sur ces 6 variables explicatives du biais sont représentées dans le tableau 5.3.

Id_labo	Technique	Extraction	Niveau concentration	Nature étalonnage	Utilisation d'étalons certifiés	Correction vs étalon interne
11014	1	2	1	0	0	0
10493	1	3	1	0	1	0
10490	1	4	2	0	0	0
11626	4	4	1	0	0	0
11157	5	3	2	0	1	0
11456	1	2	2	0	1	0
10437	5	3	1	0	1	0
11082	5	2	2	0	1	0
11323	2	3	2	1	1	1
11080	5	1	2	1	0	1
11079	5	3	2	0	1	0
11088	5	2	2	1	1	1
11753	5	4	2	1	0	0
10415	5	4	2	0	1	1
11885	2	3	3	1	1	1
10199	5	4	3	0	1	0
11095	5	3	3	1	1	1
10872	3	4	3	1	1	1

TABLE 5.3 – Réponses au questionnaires : information auxiliaire sur les biais de mesure.

La structure des biais est donnée par le modèle structurel de la figure 5.1.

Pour chaque bloc, la variable observée qui détermine les sens de la variable latente,

afin de surmonter le phénomène de *label switching* défini au chapitre, est signifiée avec une flèche en pointillés.

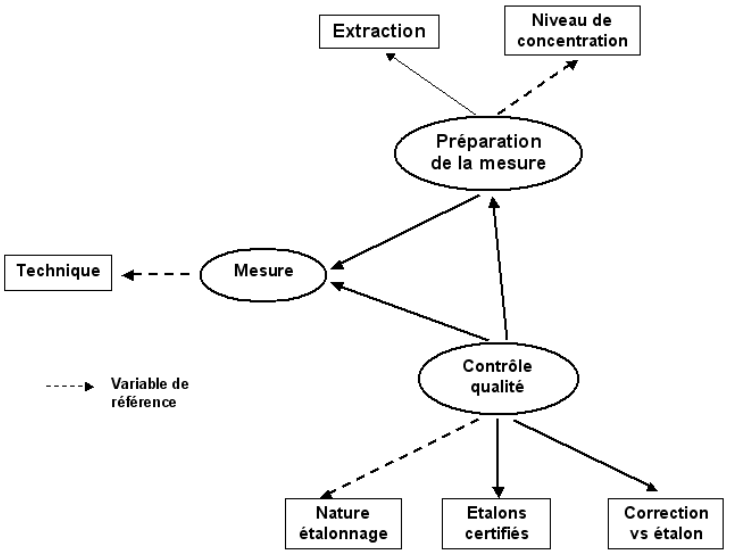


FIGURE 5.1 – Modèle structurel représentant la structure des biais de mesure lors de la mesure des triazine.

On rappelle que le schéma structurel est commun à tous les analytes de la famille des triazine.

5.2.4 Estimation du modèle structurel des biais

5.2.4.1 Choix des paramètres *a priori*

On choisit des distributions *a priori* faiblement informatives pour les variances résiduelles et la matrice de covariance des variables latentes

$$\mu_{kj}, \theta_{kj} | \sigma^2_{kj} \sim N\left(\begin{pmatrix} 0 \\ 0.5 \end{pmatrix}, \sigma^2_{kj} I_2 \right) \tag{5.1}$$

$$\sigma^2_{kj} \sim \text{Inverse-Gamma}\left(\frac{1}{2}, \frac{1}{2}\right) \tag{5.2}$$

$$\Sigma_W \sim \text{Inverse-Wishart}_3\left(I_3\right) \tag{5.3}$$

où I_2 et I_3 sont respectivement les matrices identités 2×2 et 3×3.

5.2.4.2 Choix des variables de référence pour le signe des variables latentes structurelles

Puisque la variable latente *Mesure* est définie par la seule variable *Technique*, on choisit de fixer le coefficient de régression entre ces deux variables à 1.

Les variables observées de référence pour le signe des deux autres variables latentes (en pointillés) sont choisies arbitrairement.

5.2.4.3 Résultats de l'estimation du modèle structurel des biais de mesure

Le modèle structurel des biais est estimé par l'algorithme PX-Gibbs présenté au chapitre 3. On rappelle que l'inférence n'est pas réalisée conditionnellement aux flèches du modèle interne. Les coefficients de régression internes sont estimés en dehors de l'algorithme de Gibbs.

Les principaux résultats de l'estimation du modèle structurel sont donnés à la figure 5.2.

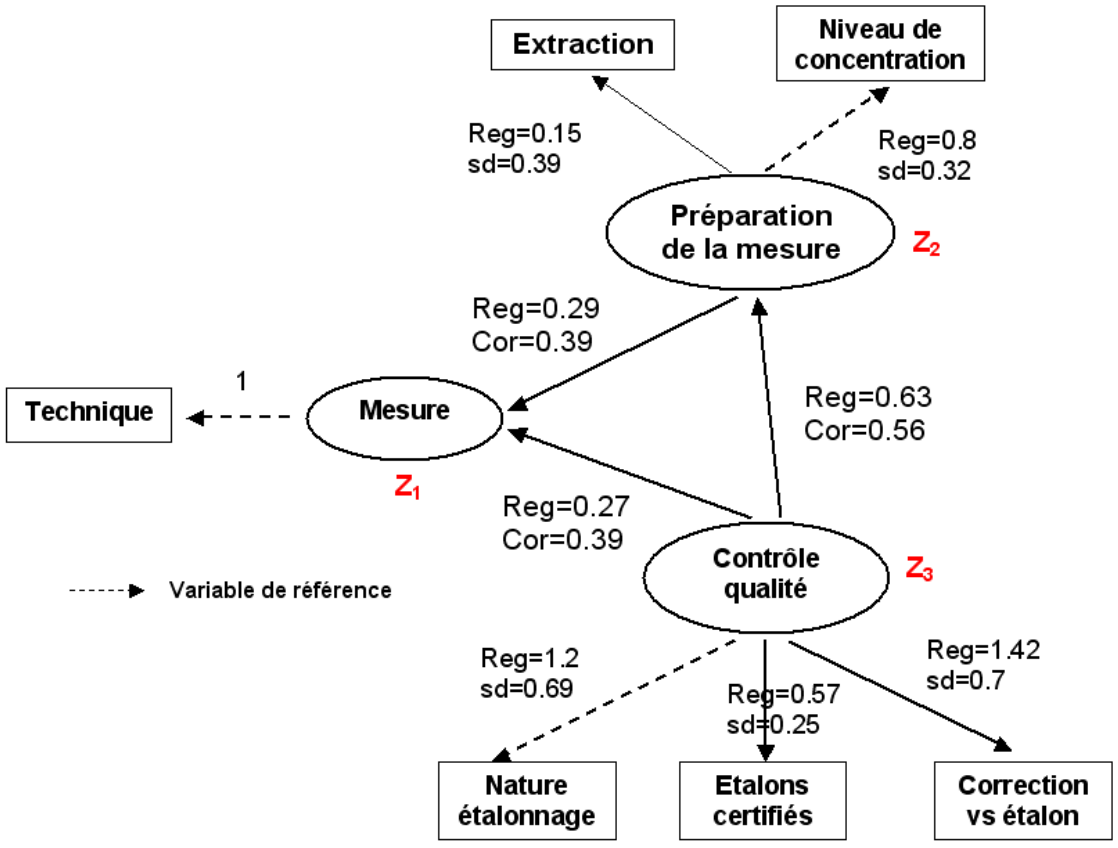


FIGURE 5.2 – Estimation du modèle structurel. *Reg* est l'estimation du coefficient de régression, *sd* est l'écart-type associé et *cor* est l'estimation du coefficient de corrélation.

Les variances résiduelles résultant de l'estimation du modèle intérieur sont une indication de la qualité du modèle intérieur. Les densités des deux variances résiduelles du modèle intérieur sont représentées à la figure 5.3 où on observe que les deux variances sont raisonnables, indiquant que le modèle intérieur est satisfaisant.

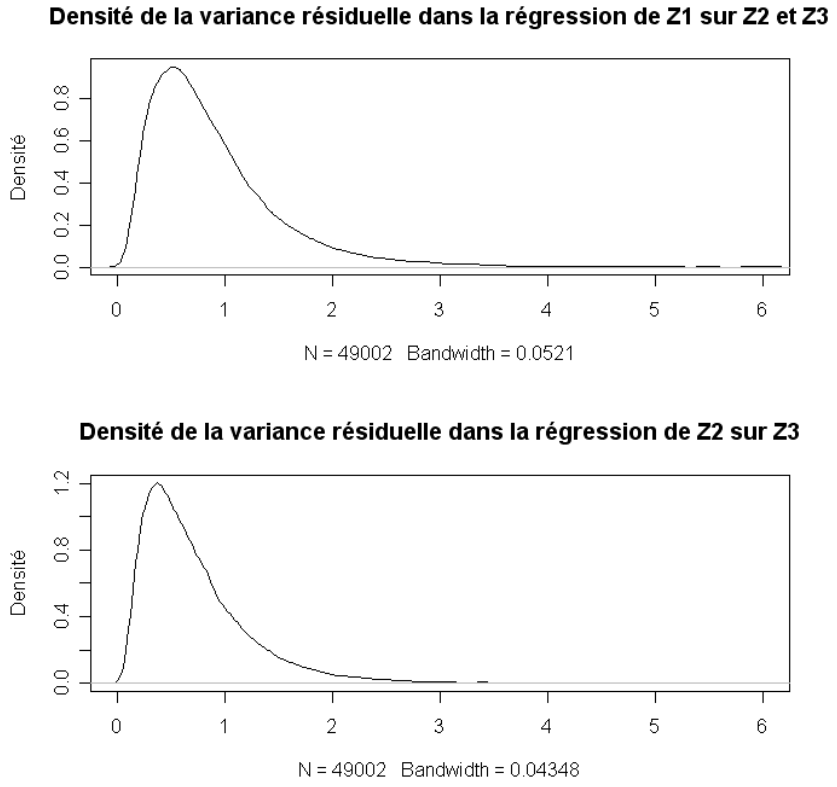


FIGURE 5.3 – Densités des variances résiduelles dans l'estimation du modèle intérieur.

5.2.4.4 Profil des laboratoires sur les composantes de biais

L'imputation des variables latentes à chaque itération de l'algorithme PX-Gibbs pour chaque laboratoire permet d'établir le **profil *a posteriori* de chaque laboratoire sur chacune des composantes de biais**.

L'estimation des densités des composantes de biais de six laboratoires est représentée à la figure 5.4.

Une densité centrée en 0 s'interprète comme une pratique moyenne du laboratoire vis à vis de la composante de biais, une densité significativement décalée à droite (du côté positif) caractérise une bonne pratique du laboratoire sur la composante et une densité significativement décalée à gauche (du côté des négatifs) signifie une mauvaise pratique du laboratoire sur la composante de biais.

5.2.4.5 Corrélations entre les variables latentes internes

Puisque la distribution *a priori* sur la matrice de corrélation des variables latentes est faiblement informative, les corrélations *a posteriori* des variables latentes sont évaluées (presque) exclusivement à partir des données du modèle des biais.

La représentation de la densité des corrélations sur les itérations des chaînes de Markov, à la figure 5.5, permet de représenter le niveau de confiance (l'incertitude) dans les corrélations conditionnellement aux données et au modèle.

La queue lourde à gauche des deux premières distributions, qui contient 0, diminue la significativité de ces deux corrélations, centrées en 0.4.

Seule la dernière corrélation, entre le contrôle qualité et la mesure représentée ici par la technique, est significativement différente de 0, centrée en 0.56.

Remarque : A la vue de ces distributions il semble évident de prendre en compte la distribution des corrélations dans le traitement des résultats de comparaisons interlaboratoires afin de prendre en compte l'incertitude sur ces corrélations dans l'incertitude de la valeur de consensus.

Or on perd cette information en travaillant sur la somme des variables latentes comme base de calcul des poids des laboratoires.

La piste pour surmonter ce point est d'établir un modèle des poids en fonction des variables latentes selon la méthodologie de la section 4.3 du chapitre 4.

5.2.5 Poids des laboratoires et construction de profils unidimensionnels

5.2.5.1 Calcul des poids des laboratoires

D'après la section 4.4.3.1 on choisit comme poids de laboratoire la fonction de répartition empirique de l'échantillon formé par les sommes des variables latentes pour chaque laboratoire.

Ainsi, à l'itération t, on rappelle le poids du laboratoire i donné à la formule 4.24

$$\tilde{w}_i^{(t)} = F_n\left(s_i^{(t)}\right) \tag{5.4}$$

Le poids est bruité selon la formule 4.25 qui devient ici (avec $\varsigma = \frac{1}{2}\frac{1}{18} = 0.028$

$$\tilde{w}_i^{(t)} \sim NT\left(F_n\left(s_i^{(t)}\right), 0.028; 0, 1\right) \tag{5.5}$$

Le poids est ensuite normalisé selon la formule

$$w_i^{(t)} = \frac{\tilde{w}_i^{(t)}}{\sum_{i=1}^{n} \tilde{w}_i^{(t)}} \tag{5.6}$$

5.2.5.2 Représentation graphique de profils unidimensionnels des laboratoires

Nous proposons de représenter graphiquement, pour chaque laboratoire, sa position relative (donnée par la valeur de la fonction de répartition empirique) par rapport aux autres laboratoires dans l'échantillon formé par la somme des variables latentes, à chaque itération de l'algorithme PX-Gibbs.

Ces représentations nous donnent les profils moyens sur les trois composantes de biais des laboratoires, à confronter avec les représentations graphiques des scores sur chacune des composantes de la figure 5.4.

Le schéma pour le laboratoire 1 montre que le score (sommé sur les trois composantes) du laboratoire est le plus souvent en tête de l'échantillon ordonné, signifiant que sur toutes les itérations de l'algorithme, le laboratoire 1 est le plus souvent celui qui a la moins bonne pratique *a posteriori* .

Au contraire les laboratoires 17 et 18 montrent qu'ils sont le plus souvent meilleurs que tous les autres laboratoires, le laboratoire 17 étant même plus souvent mieux classé que le laboratoire 18.

Les laboratoires 5 et 11 ont des profils moyens. Le laboratoire 10 a tendance a être un bon laboratoire.

On retrouve ces informations dans la représentation graphique des densités des poids normalisés des laboratoires à la figure 5.7. Chaque densité peut s'interpréter comme le profil unidimensionnel du laboratoire en terme de qualité de la pratique. Un poids faible caractérise une mauvaise pratique et un poids élevé caractérise une bonne pratique. Un poids autour de $1/18 = 0.05$ est un laboratoire de pratique moyenne.

5.2.6 Estimation de la valeur de consensus, de son incertitude associée et des biais de mesure

5.2.6.1 Résultats pour la valeur de consensus et son incertitude associée

La distribution *a posteriori* de la valeur de consensus est obtenue par simulation de Monte-Carlo du modèle de la section 4.4.3.3 dont les paramètres sont évalués pour chaque analyte.

Nous proposons de comparer sur cette application les résultats obtenus par le BIPEA, par l'algorithme A et par l'approche que nous proposons menant à une valeur de consensus robustifée.

On précise que les résultats du BIPEA dont nous disposons pour cette comparaison ont été calculés à partir des résultats du "meilleur" sous-ensemble de laboratoires de l'ensemble des laboratoires participant et non pas à partir des "meilleurs" laboratoires parmi les 18 laboratoires ayant répondu à l'enquête, où la qualification en meilleur laboratoire est laissée à l'interprétation du BIPEA.

Les résultats sont regroupés par méthode pour tous les analytes dans le tableau 5.8

Une information supplémentaire est fournie par l'approche bayésienne qui fournit la distribution complète de la valeur de consensus intégrée sur toutes les sources de variabilité.

Ainsi pour chaque analyte nous proposons de plus la représentation graphique de la

distribution *a posteriori* de la valeur de consensus à la figure 5.9 où les intervalles de confiance à 95 % des trois méthodes et du LNE sont également représentés.

5.2.6.2 Résultats pour les biais de mesure des laboratoires

Les distributions *a posteriori* des biais de mesure de chaque laboratoire sur la mesure des cinq analytes sont estimées à partir des réalisations des biais lors de chaque itération de l'algorithme PX-Gibbs à partir des formules 4.54 et 4.59. On rappelle que la variabilité des biais provient de l'échantillonnage des chaînes de Markov utilisées dans le calcul des biais.

Les schémas des figures 5.10 et 5.11 représentent les biais des laboratoires 10 et 18 respectivement dans la mesure des cinq analytes.

On fait figurer sur chaque schéma l'estimation ponctuelle des biais par la méthode de référence (triangle noir), par la méthode du BIPEA (point vert) et par la méthode de l'algorithme A (point rouge).

5.2.6.3 Interprétation des résultats

Le fait que les intervalles de confiance suivant les trois méthodes se recouvrent montre que les trois méthodes statistiques sont consistantes entre elles dans l'estimation de la valeur de consensus et de son incertitude associée.

En revanche aucune des méthodes n'est consistante avec les résultats du LNE pour l'analyte 1 car les intervalles de confiance des trois méthodes ne recouvrent pas l'intervalle de confiance du LNE.

On remarque également que les trois méthodes sous-estiment la valeur de consensus dans la mesure de l'analyte 3 et les trois méthodes surestiment la valeur de consensus dans de la mesure de l'analyte 5.

En l'absence des incertitudes des biais, l'interprétation des positions relatives des estimations ponctuelles des biais n'a pas de sens. On rappelle que l'incertitude apparente des biais par la méthode que nous proposons ne reflète que la variabilité d'échantillonnage sur les chaînes de Markov.

En revanche les positions relatives des biais sur les différents analytes nous permet de conclure que les biais ne sont pas constants sur la mesure des différents analytes. Ce qui confirme que la pratique des laboratoire doit être différenciée pour chaque analyte.

5.2.7 Conclusion de l'application

L'application que l'on vient de développer de façon exhaustive a permis d'illustrer la réponse statistique à un problème existant dont la mise en oeuvre est nécessaire en raison de la fréquence de la situation en pratique.

L'intérêt de l'approche réside cependant davantage dans la mise en évidence des limitations de toute approche statistique en l'absence des "bonnes données". En effet aucun artifice statistique ne pourra remplacer la connaissance de l'incertitude des résultats de mesure.

La mise en oeuvre de l'approche, en l'absence des "bonnes données" permet tout de même d'entrevoir la richesse et les possibilités quasi infinies de l'approche bayésienne que nous proposons du traitement de données de comparaisons interaboratoires :

- la construction de profils de laboratoires
- la connaissance de la distribution complète de la valeur de consensus sachant toutes les sources de biais évaluée sur tous les laboratoires
- la connaissance de la distribution complète des biais de chaque laboratoire pour un suivi amélioré prenant en compte l'historique des biais et toutes les sources de biais

Plus généralement la flexibilité de l'approche développée permet de prendre en compte facilement tout type de données. En particulier, l'approche permet la prise en compte de l'historique des mesures de chaque laboratoire, la prise en compte d'un réponse manquante lors d'une comparaison ou d'un résultat tronqué ou censuré du type inférieur à la limite de détection. Une approche plus globale peut consister en la prise en compte de données par pays, par région en introduisant une partie hiérarchique.

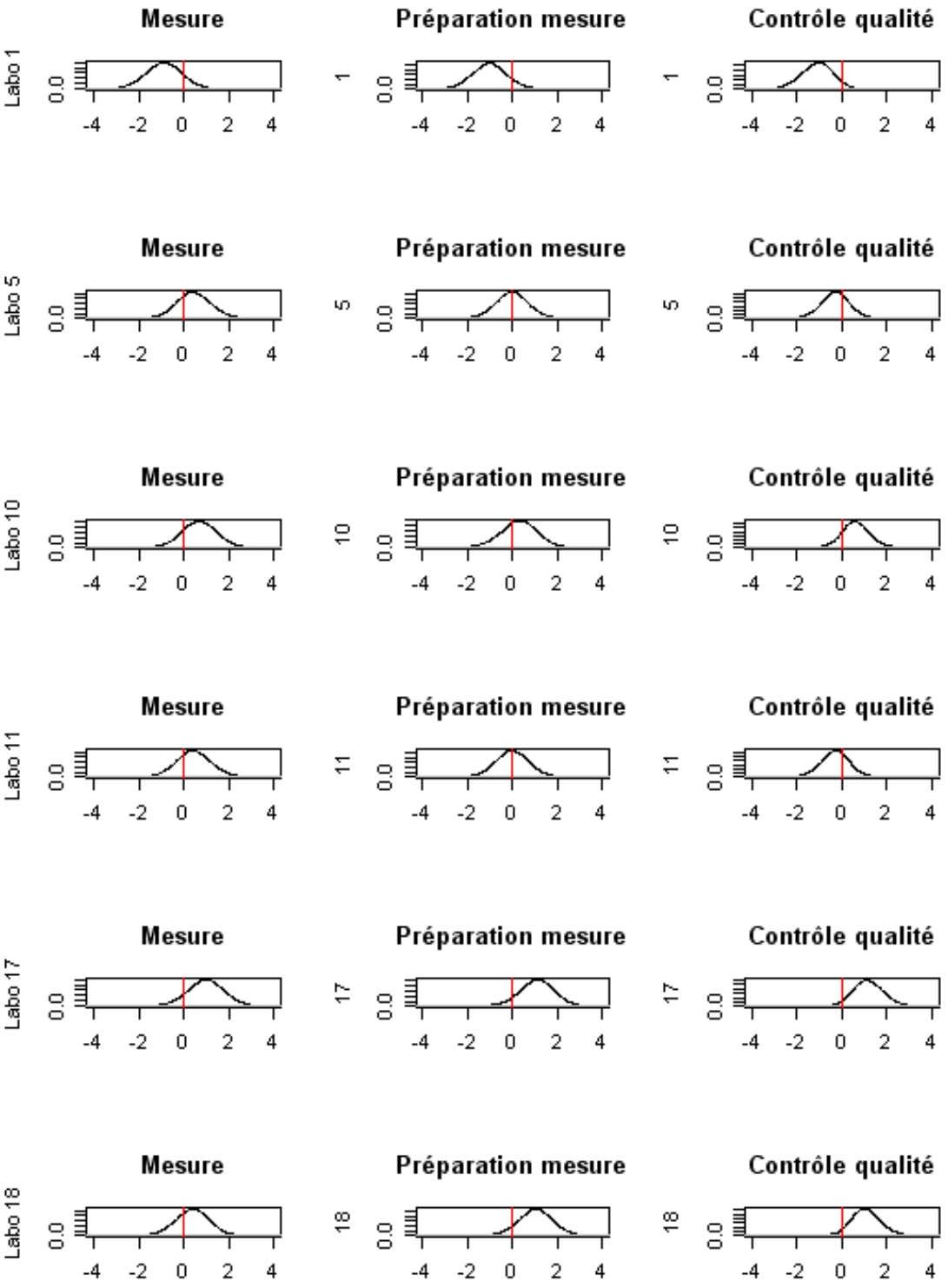


FIGURE 5.4 – Scores des laboratoires 1,2,5,17 et 18 sur les trois variables latentes. En rouge, le score 'nul' de référence.

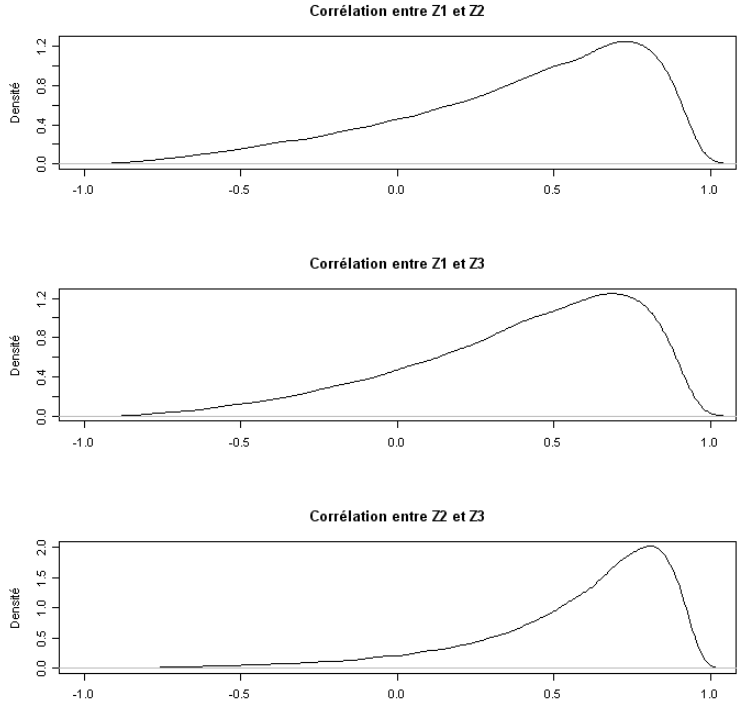


FIGURE 5.5 – Corrélations entre les variables latentes structurelles

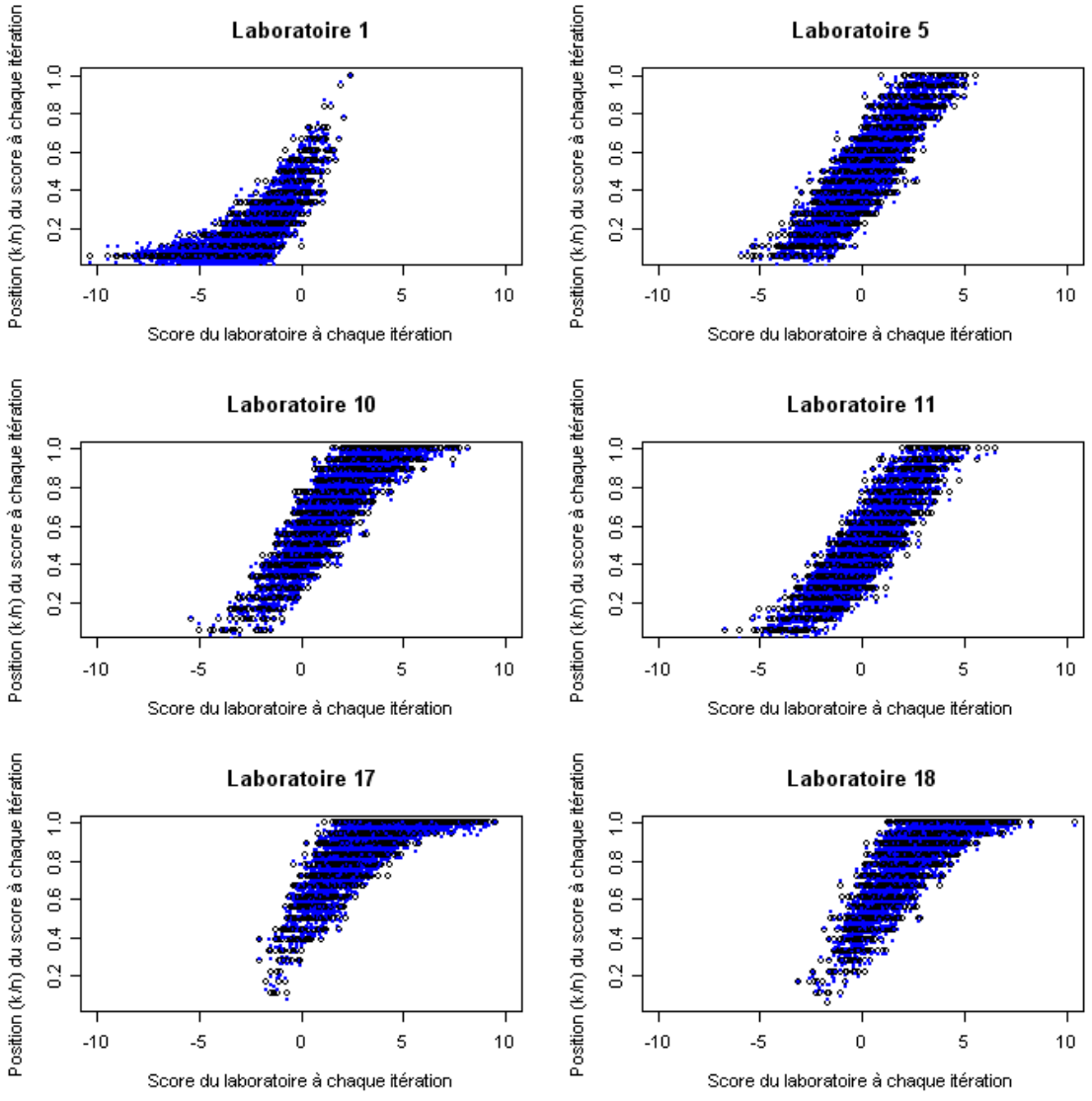


FIGURE 5.6 – Répartition de la position relative des scores, **sommés sur les trois variables latentes**, de six laboratoires sur les itérations de l'algorithme PX-Gibbs. Le nuage bleu représente les valeurs bruitées de la fonction de répartition empirique.

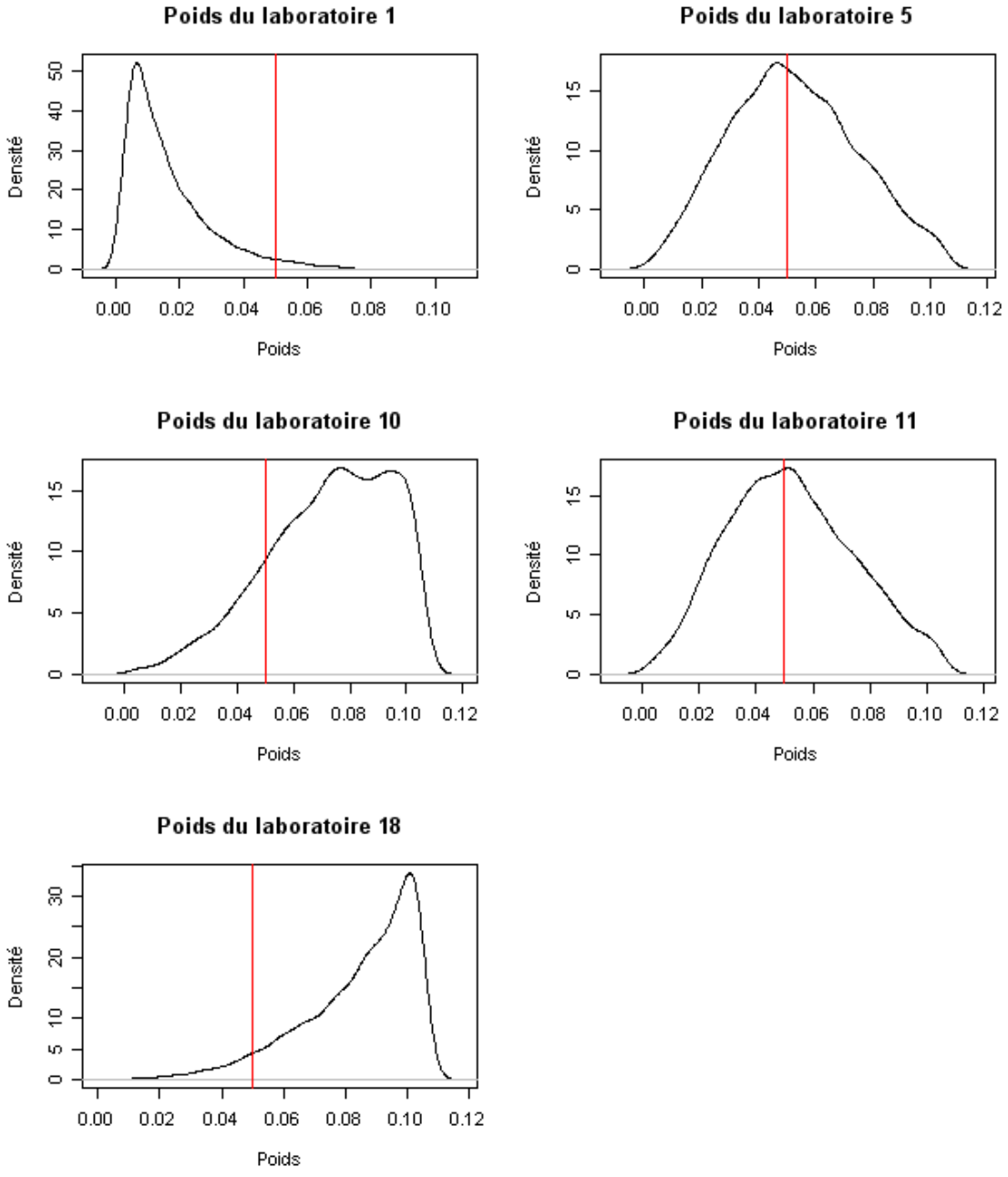


FIGURE 5.7 – Densité des poids de 6 laboratoires et position par rapport au poids uniforme de 1/18=0.05.

LNE	Analyte 1	Analyte 2	Analyte 3	Analyte 4	Analyte 5
valeur de réf	29.21	107.27	98.41	44.67	33
incertitude	1.61	2.25	2.01	1.34	1.73
intervalle 95%	[25.99,32.43]	[102.77,111.77]	[94.39,102.43]	[41.99,47.35]	[29.54,36.46]

BIPEA	Analyte 1	Analyte 2	Analyte 3	Analyte 4	Analyte 5
valeur de cons.	39.6	107.5	96.3	43.6	34
incertitude	4.2	3.7	3.8	1.6	1.6
intervalle 95%	[31.2,48]	[100.1,114.9]	[88.7,103.9]	[40.4,46.8]	[30.8,37.2]

Appr. Bay.	Analyte 1	Analyte 2	Analyte 3	Analyte 4	Analyte 5
valeur de cons.	36.37	106	95.51	45.01	34.49
incertitude	3.27	3.14	2.78	1.63	1.55
intervalle 95%	[29.83,42.91]	[99.72,112.28]	[89.95,101.07]	[41.75,48.27]	[31.39,37.59]

Algorithme A	Analyte 1	Analyte 2	Analyte 3	Analyte 4	Analyte 5
valeur de cons.	36.03	104.48	94.16	45.05	34.47
incertitude	3.46	3.32	2.94	1.73	1.64
intervalle 95%	[29.11,42.95]	[97.84,111.12]	[88.28,100.04]	[41.59,48.51]	[31.19,37.75]

FIGURE 5.8 – Résultats pour les cinq analytes sous la forme des intervalles de confiance à 95 % pour les quatre méthodes comparées.

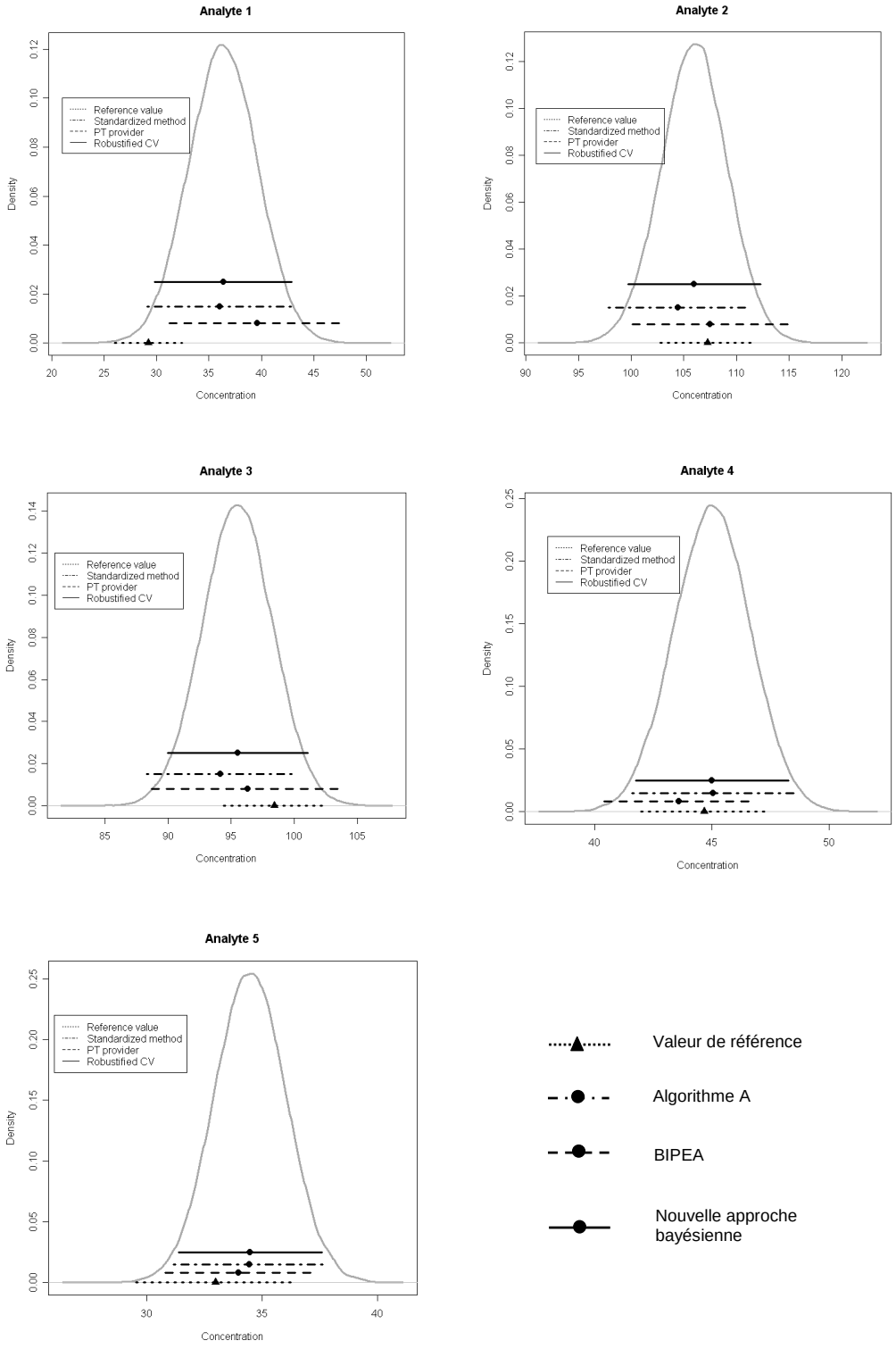


FIGURE 5.9 – Représentation graphique des résultats pour les cinq analytes sous forme des intervalles de confiance à 95 % pour les quatre méthodes comparées et en plus la distribution *a posteriori* de la valeur de consensus intégrée sur sa variance.

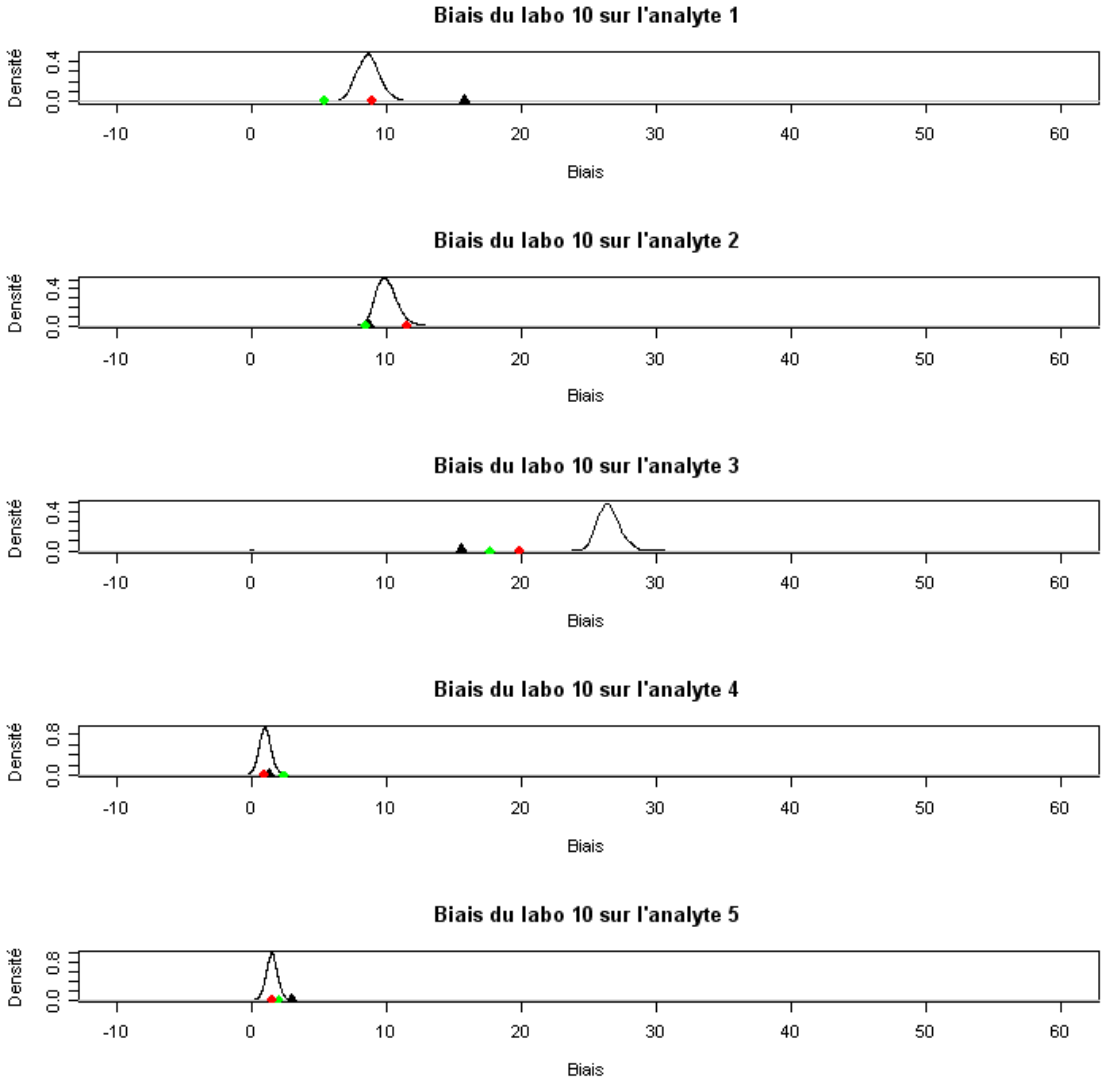


FIGURE 5.10 – Représentation graphique de la distribution *a posteriori* des biais de mesure du laboratoire 10 sur les cinq analytes.

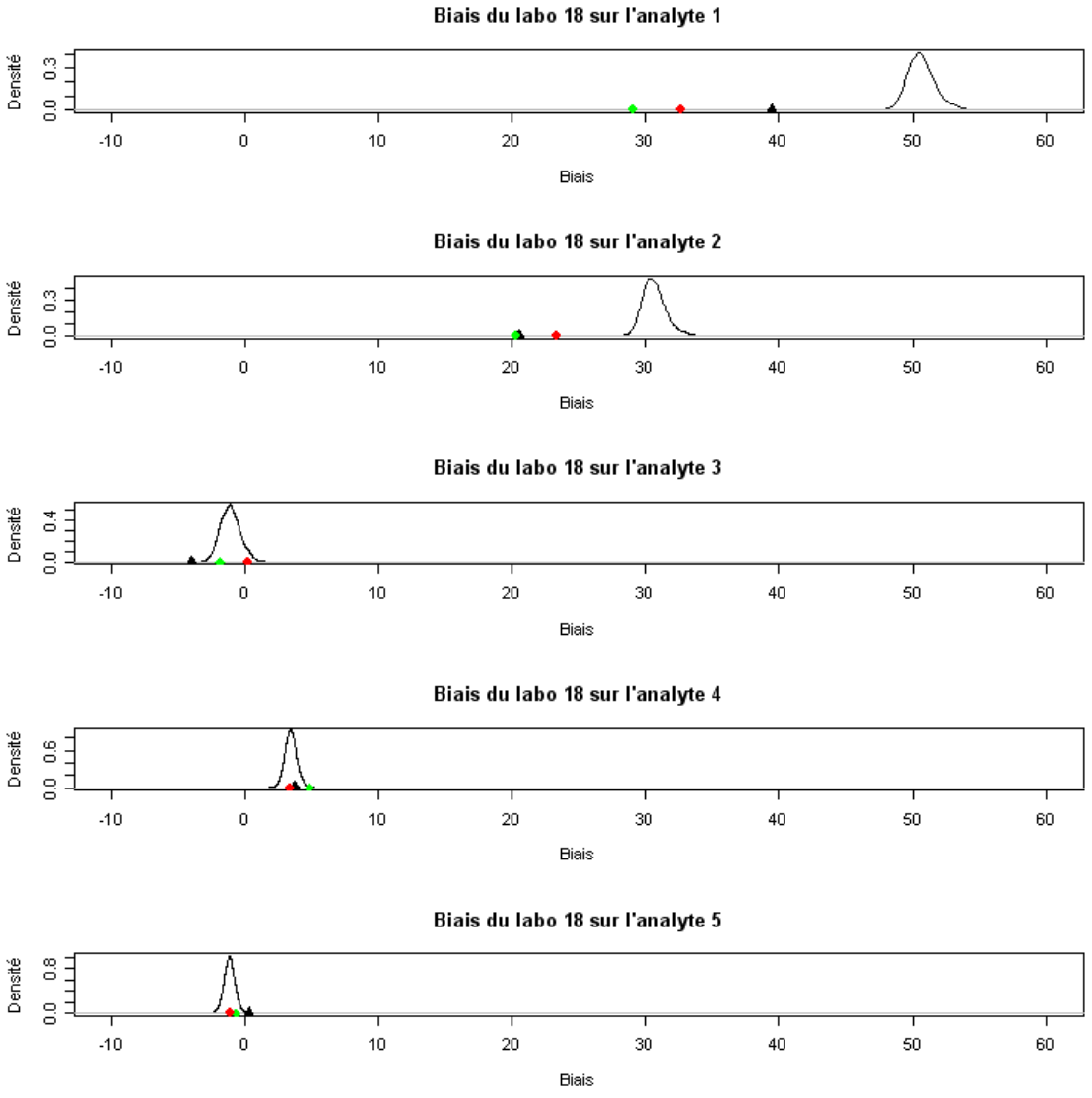


FIGURE 5.11 – Représentation graphique de la distribution *a posteriori* des biais de mesure du laboratoire 18 sur les cinq analytes.

Conclusions et perspectives

Partant du sujet ambitieux de l'étude de l'apport de la statistique bayésienne aux comparaisons interlaboratoires, les travaux de thèse ont contribué au développement d'une nouvelle approche bayésienne du traitement des données de comparaisons interlaboratoires dont le but est d'améliorer à terme la qualité des mesures des laboratoires.

C'est l'étude des méthodes existantes de traitement des données de comparaisons interlaboratoires qui nous a permis de dégager le fil conducteur des travaux présentés, à savoir la volonté de modéliser les biais de mesure des laboratoires qui s'est traduite dans nos travaux par une modélisation effective des biais de mesure à partir de variables identifiées.

Le travail de modélisation des biais de mesure que nous avons mis en oeuvre a nécessité de faire appel à des modèles permettant de prendre en compte des structures possiblement complexes dans les sources de biais de mesure.

C'est ainsi que nous avons introduit le cadre conceptuel général des modèles à équations structurelles (SEM) qui permet une modélisation flexible de la structure des biais de mesure. Ces modèles permettent en outre une meilleure compréhension de la structure des biais en permettant d'introduire dans l'analyse des composantes latentes des biais de mesure issues de connaissances d'experts. Le modèle des biais se traduit ainsi en une modélisation graphique des relations entre ces composantes.

La force de la nouvelle approche bayésienne que nous proposons est de prendre en compte la structure des biais de mesure dans le traitement des données de comparaisons interlaboratoires en travaillant sur les distributions conjointes de toutes les grandeurs en jeu, en nous reposant sur l'hypothèse que les résultats de mesure des laboratoires sont conditionnellement indépendants sachant les sources de biais.

Afin de proposer une meilleure étude des relations entre les biais de mesure nous avons développé une nouvelle approche bayésienne d'estimation des modèles SEM centrée sur l'estimation des corrélations entre toutes les composantes de biais c'est-à-dire sur la matrice de corrélation de toutes les variables latentes.

Par rapport une approche bayésienne traditionnelle, la méthode que nous proposons permet de prendre en compte une information *a priori* sur la matrice de corrélation et d'améliorer les propriétés de convergence de l'algorithme d'estimation.

De plus le fait de raisonner sur un modèle des biais saturé permet de tester et valider des structures particulières des biais *a posteriori* et non seulement d'estimer la structure *a priori* des biais.

Le cadre bayésien de l'approche permet d'étendre le modèle global développé à des données plus complexes, par exemple en présence de données manquantes, censurées ou hiérarchiques en prenant en compte dans chaque cas l'incertitude introduite par ces données particulières. Le cadre bayésien est également un cadre rigoureux pour la comparaison et la validation de modèles.

Ce qu'il faut retenir de la contribution de la nouvelle approche à l'amélioration de la qualité des mesures est qu'elle permet de quantifier les contributions des différentes sources de biais au biais des laboratoires si l'on dispose des incertitudes des mesures et des réplicats de mesure, dont la conséquence directe est l'amélioration du processus de mesure au niveau de tous les laboratoires.

En l'absence d'incertitude et sans les réplicats de mesure ces contributions ne sont pas

estimables et nous avons développé une approche permettant de prendre en compte les sources de biais dans le calcul de la valeur de consensus, de son incertitude associée et des biais de mesure.

Nous avons mis en oeuvre cette approche dans le cas de tests d'aptitude en environnement où nous avons montré que nos résultats sont cohérents à la fois avec les résultats fournis par l'organisateur de la comparaison et avec les résultats obtenus par la méthode de référence, réalisés de manière exceptionnelle.

Dans les conditions de la comparaison, qui peuvent s'étendre à un grand nombre d'autres comparaisons en santé, en environnement et dans l'alimentation, l'avantage de notre méthode est qu'elle permet la prise en compte d'expertise de façon transparente et entièrement décrite (par exemple par rapport à un choix des meilleurs laboratoires par l'organisateur du test d'aptitude). C'est la prise en compte de cette expertise dans le calcul de la valeur de consensus et de son incertitude associée qui permet d'améliorer la fiabilité de la valeur de consensus et de son incertitude associée en limitant leur dépendance aux données. En conséquence le suivi des biais de mesure est amélioré. Un avantage considérable demeure également dans l'étude des relations entre les sources de biais.

Plus généralement nous avons proposé une méthodologie permettant de valoriser la participation d'un laboratoire national de métrologie à un test d'aptitude dans le traitement de futures données de tests d'aptitude réalisés dans les mêmes conditions. Cette approche permet la modélisation prédictive des biais de mesure des laboratoires afin de calculer la valeur de consensus et son incertitude associée à partir des résultats corrigés. La mise en oeuvre de cette approche nécessite les compétences de tous les acteurs de la mesure.

Ce qu'il faut retenir sur la mise en oeuvre d'une étude statistique pour des études futures sur ce même sujet est que l'on ne peut pas tirer des données plus d'informations qu'elles n'en contiennent. Ainsi un résultat de mesure seul ne contient aucune information sur le biais du laboratoire mais son incertitude est une information précieuse si on considère que le biais est pris en compte dans une composante de l'incertitude. La donnée des réplicats de mesure est également importante pour quantifier la répétabilité du laboratoire. De notre côté nous sommes partis à la recherche d'informations sur les biais de mesure. Précisons que la donnée de l'incertitude d'un résultat (qui prend en compte toutes les sources de biais connus) n'est pas incompatible avec l'étude des sources de biais. En effet la contribution d'une source de biais évaluée à partir de tous les laboratoires est une information différente de la contribution de cette même variable dans le calcul d'incertitude d'un laboratoire.

En tant que statisticiens nous sommes conscients que l'amélioration des mesures passe d'abord par le réalisation du calcul d'incertitude par tous les laboratoires et ensuite par un traitement statistique d'autant enrichi. En l'absence d'incertitude de mesure la statistique fournit une méthode nécessaire mais qui ne résout pas les problèmes de fond et dont l'existence doit au contraire motiver l'évolution vers le calcul d'incertitude. En cela nos conclusions vont le sens de l'évolution des normes et des règlementations qui visent à demander l'incertitude du résultat de mesure.

Les bénéfices attendus de la méthodologie proposée dans ces travaux pour l'amélioration de la qualité des mesures sont donc à long terme et nous sommes persuadés qu'il est nécessaire de continuer à travailler à la modélisation des biais de mesure conjointement avec les laboratoires et les organisateurs de comparaisons interlaboratoires. Alors, en sui-

vant l'évolution des normes et des règlementations et en accompagnant le développant de nouvelles méthodes de référence par les laboratoires nationaux de métrologie, l'objectif sera atteint.

Nous espérons avoir apporté notre pierre à l'édifice.

Bibliographie

[Albert et Chib, 1993] Albert, J. H. et Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

[Amarouche et Désenfant, 2004] Amarouche, S. et Désenfant, M. (2004). Formulaire pour le calcul d'une valeur de référence dans les comparaisons clés internationales (KCRV). Rapport technique, LNE.

[AMC, 1989] AMC (1989). Robust statistics-How not to reject outliers. part 2. inter-laboratory trials. *Analyst*, 114:1699–1702.

[Beichl et Sullivan, 2000] Beichl, I. et Sullivan, F. (2000). The Metropolis algorithm. *Computing in Science and Engineering*, 2(1):65–69.

[BIPM *et al.*, 2008] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IOPAP et OIML (2008). Evaluation des données de mesure – Supplément 1 du 'Guide pour l'expression de l'incertitude de mesure' – propagation de distributions par une méthode de monte carlo. Joint Committee for Guides in Metrology, Bureau International des Poids et Mesures, JCGM 101 :2008.

[BIPM *et al.*, 1995] BIPM, IEC, IFCC, ISO, IUPAC, IOPAP et OIML (1995). Guide pour l'expression de l'incertitude de mesure. ISO/IEC Guide 98 :1995. International Organization for Standardization (ISO), Geneva, Switzerland.

[Box, 1953] Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, 40:318–335.

[Chib et Greenberg, 1990] Chib, S. et Greenberg, E. (1990). Understanding the Metropolis-Hastings algorithm. *Journal of the American Statistical Association*, 85(412): 972–985.

[Chunovkina *et al.*, 2008] Chunovkina, A. G., Elster, C., Lira, I. et Wöger, W. (2008). Analysis of key comparison data and laboratory biases. *Metrologia*, 45(2):211–216.

[Congdon, 2003] Congdon, P. (2003). *Applied Bayesian modelling*. Wiley (Wiley Series in Probability and Statistics).

[Cox, 2007] Cox, M. G. (2007). The evaluation of key comparison data : determining the largest consistent subset. *Metrologia*, 44(3):187–200.

[Dempster *et al.*, 1977] Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm "with discussion". *Journal of the Royal Statistical Society, Serie B*, 39(1):1–38.

[Eckhardt, 1987] Eckhardt, R. (1987). Stan Ulam, John von Neumann, and the Monte Carlo method. *Los Alamos Science*, (15):125–130.

[Elster *et al.*, 2007] Elster, C., Wöger, W. et Cox, M. G. (2007). Draft GUM Supplement 1 and bayesian analysis. *Metrologia*, 44(3):L31–L32.

[Foulley et Jaffrézic, 2010] Foulley, J.-L. et Jaffrézic, F. (2010). Modelling and estimating heterogeneous variances in threshold models for ordinal discrete data via winbugs/openbugs. *Computer Methods and Programs in Biomedicine*, 97(1):19–27.

[Gelman *et al.*, 2004] Gelman, A., Carlin, J. B., Stern, H. S. et Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC.

[Gelman et Hill, 2006] Gelman, A. et Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

[Gelman *et al.*, 1996] Gelman, A., Meng, X.-L. et Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807.

[Geman et Geman, 1984] Geman, S. et Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 6, pages 721–741.

[Ghosh et Dunson, 2009] Ghosh, J. et Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320.

[Graybill et Deal, 1959] Graybill, F. A. et Deal, R. B. (1959). Combining unbiased estimators. *Biometrics*, 15:543–550.

[Hastings, 1970] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

[Huber et Ronchetti, 2009] Huber, P. et Ronchetti, E. (2009). *Robust Statistics*. Wiley.

[ISO-CASCO, 2005] ISO-CASCO (2005). Exigences générales concernant la compétence des laboratoires d'étalonnage et d'essais. ISO/IEC 17025 :2005. International Organization for Standardization (ISO), Geneva, Switzerland.

[ISO-CASCO, 2010] ISO-CASCO (2010). Evaluation de la conformité – exigences générales concernant les essais d'aptitude. ISO/IEC 17043 :2010. International Organization for Standardization (ISO), Geneva, Switzerland.

[ISO/TC69, 2005] ISO/TC69 (2005). Méthodes statistiques utilisées dans les essais d'aptitude par comparaisons interlaboratoires. ISO 13528 :2005. International Organization for Standardization (ISO), Geneva, Switzerland.

[Jakobowicz, 2007] Jakobowicz, E. (2007). *Contributions aux modèles d'équations structurelles à variables latentes*. Thèse de doctorat, Conservatoire National des Arts et Métiers, Paris.

[Jakobowicz et Derquenne, 2007] Jakobowicz, E. et Derquenne, C. (2007). A modified PLS Path Modeling algorithm handling reflective categorical variables and a new model building strategy. *Computational Statistics and Data Analysis*, 51(7):3666–3678.

[Jasra *et al.*, 2005] Jasra, A., Holmes, C. C. et Stephens, D. A. (2005). MCMC and the label switching problem in Bayesian mixture models. *Statistical Science*, 20:50–67.

[JCGM, 2008] JCGM (2008). Vocabulaire international de métrologie – concepts fondamentaux et généraux et termes associés. ISO/IEC Guide 99 :2007. International Organization for Standardization (ISO), Geneva, Switzerland.

[Kafadar, 2003] Kafadar, K. (2003). John tukey and robustness. *Statistical Science*, 18(3):319–331.

[Lawrence *et al.*, 2008] Lawrence, E., Bingham, D., Liu, C. et Nair, V. N. (2008). Bayesian inference for multivariate ordinal data using parameter expansion. *Technometrics*, 50(2):182–191.

[Lee, 2007] Lee, S.-Y. (2007). *Structural Equation Modelling : A Bayesian Approach*. Wiley (Wiley Series in Probability and Statistics).

[Lewandowski *et al.*, 2011] Lewandowski, A., Liu, C. et Wiel, S. V. (2011). Parameter expansion and efficient inference. *Statistical Science, to appear*.

[Lindley et Smith, 1972] Lindley, D. V. et Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(1):1–41.

[Liu *et al.*, 1998] Liu, C., Rubin, D. B. et Wu, Y. N. (1998). Parameter expansion to accelerate EM : The PX-EM algorithm. *Biometrika*, 85:755–770.

[Liu, 2001] Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*. Springer-Verlag, New York.

[Liu *et al.*, 1994] Liu, J. S., Wong, W. H. et Kong, A. (1994). Covariance structure of the gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40.

[Liu et Wu, 1999] Liu, J. S. et Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274.

[Mandel et Paule, 1970] Mandel, J. et Paule, R. C. (1970). Interlaboratory evaluation of a material with unequal numbers of replicates. *Analytical chemistry*, 42(11).

[Marin et Robert, 2007] Marin, J.-M. et Robert, C. P. (2007). *Bayesian Core : a practical approach to computational Bayesian statistics*. Springer.

[McCullagh, 1980] McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142.

[Metropolis, 1987] Metropolis, N. (1987). The beginning of the Monte Carlo method. *Los Alamos Science*, (15):125–130.

[Metropolis *et al.*, 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. et Teller, E. (1953). Equations of state calculations by fast computing machines. *J.Chem.Phys*, 21:1087–1092.

[Metropolis et Ulam, 1949] Metropolis, N. et Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341.

[Meza *et al.*, 2009] Meza, C., Jaffrézic, F. et Foulley, J.-L. (2009). Estimation in the probit normal model for binary outcomes using the SAEM algorithm. *Computational Statistics and Data Analysis*, 53(4):1350–1360.

[Müller, 2000] Müller, J. W. (2000). Possible advantages of a robust evaluation of comparisons. *J. Res. Natl. Inst. Stand. Technol.*, 105(4):551–555.

[Palomo *et al.*, 2007] Palomo, J., Dunson, D. B. et Bollen, K. (2007). *Bayesian structural equation modeling*, chapitre 8, pages 163–188. Elsevier.

[Parent et Bernier, 2007] Parent, E. et Bernier, J. (2007). *Le raisonnement bayésien. Modélisation et inférence.* Springer-Verlag, France.

[Robert, 2001] Robert, C. P. (2001). *The Bayesian choice.* Springer-Verlag, New York, second edition.

[Robert et Casella, 2004] Robert, C. P. et Casella, G. (2004). *Monte Carlo statistical methods.* Springer-Verlag, New York, second edition.

[Rosenthal, 2004] Rosenthal, J. (2004). W.K. Hastings, statistician and developer of the Metropolis-Hastings algorithm. *http ://www.probability.ca/hastings/.*

[Rousseeuw, 1991] Rousseeuw, P. J. (1991). Tutorial to robust statistics. *Journal of Chemometrics*, 5:1–20.

[Rukhin, 2009] Rukhin, A. L. (2009). Weighted means statistics in interlaboratory studies. *Metrologia*, 46(3):323–331.

[Russolillo, 2009] Russolillo, G. (2009). *Partial Least Squares methods for non-metric data.* Thèse de doctorat, Università degli Studi di Napoli Federico II.

[Saporta, 2006] Saporta, G. (2006). *Probabilités, analyse des données et statistique.* Technip, 2ème édition révisée et augmentée.

[Skrondal et Rabe-Hesketh, 2004] Skrondal, A. et Rabe-Hesketh, S. (2004). *Generalized latent variable modeling.* Chapman Hall CRC.

[Tanner et Wong, 1987] Tanner, M. A. et Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

[Tenenhaus *et al.*, 2005] Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M. et Lauro, C. (2005). PLS Path Modeling. *Computational Statistics and Data Analysis*, 48:159–205.

[Toman, 2007] Toman, B. (2007). Bayesian approaches to calculating a reference value in key comparison experiments. *Technometrics*, 49(1):81–87.

[Toman et Possolo, 2009] Toman, B. et Possolo, A. (2009). Laboratory effects models for interlaboratory comparisons. *Accred Qual Assur*, 14:553–563.

[Tukey, 1960a] Tukey, J. W. (1960a). Discussion of Anscombe and Daniel papers. *Technometrics*, 2:159–163.

[Tukey, 1960b] Tukey, J. W. (1960b). *A survey of sampling from contaminated distributions.* Stanford Univ Press.

[van Dyk et Meng, 2001] van Dyk, D. A. et Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.

[Weise et Wöger, 1993] Weise, K. et Wöger, W. (1993). A Bayesian theory of measurement uncertainty. *Meas. Sci. Technol*, 4(1):1–11.

[Weise et Wöger, 1994] Weise, K. et Wöger, W. (1994). Comparison of two measurement results using the Bayesian theory of measurement uncertainty. *Meas. Sci. Technol*, 5(3):879–882.

[Zhang, 2006] Zhang, N. F. (2006). The uncertainty associated with the weighted mean of measurement data. *Metrologia*, 43(3):195–204.

Annexes

Annexe A

Distributions utilisées

On reprend les notations de l'ouvrage de Gelman [Gelman *et al.*, 2004].

A.1 Modélisation des paramètres d'espérance

A.1.1 Distribution de Student

La distribution de Student apparaît lorsque l'on estime la moyenne d'un échantillon de petite taille et de variance inconnue. Elle est paramétrisée par une fonction de la taille de l'échantillon (le nombre de degrés de liberté) et on peut montrer que la distribution de la moyenne converge en probabilité vers la loi normale lorsque la taille de l'échantillon augmente.

En notant ν le nombre de degrés de liberté de l'échantillon, μ sa moyenne, connue, et σ^2 sa variance inconnue, la densité de probabilité de la moyenne, notée $\boldsymbol{\theta}$, est la distribution de Student de paramètres ν, μ et σ^2 notée $t_\nu\left(\mu, \sigma^2\right)$ définie par

$$p\left(\boldsymbol{\theta}\right) \propto \left[1 + \frac{1}{\nu}\left(\frac{\theta - \mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}} \tag{A.1}$$

L'espérance de la moyenne est conservée par rapport à la loi normale correspondante et la variance est augmentée, ce qui signifie que la loi de Student est plus robuste pour modéliser des événements rares, éloignés de la moyenne.

$$E\left(\boldsymbol{\theta}\right) = \mu \tag{A.2}$$

$$V\left(\boldsymbol{\theta}\right) = \frac{\nu}{\nu - 2}\sigma^2,\ \nu > 2 \tag{A.3}$$

A.1.2 Distribution de Student multivariée

La distribution de Student admet une généralisation multivariée lorsque les observations sont multidimensionnelles.

En notant $\boldsymbol{\theta}$ le vecteur des observations de taille K, de moyenne le vecteur $\boldsymbol{\mu}$ et de matrice de covariance Σ et de nombre de degrés de liberté ν, la densité de probabilité de la loi de Student multivariée notée $t_\nu(\boldsymbol{\mu}, \Sigma)$ s'écrit

$$p(\boldsymbol{\theta}) \propto \left[1 + \frac{1}{\nu}(\boldsymbol{\theta} - \boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu})\right]^{-\frac{\nu+K}{2}} \tag{A.4}$$

Les formules correspondantes pour l'espérance et la variance sont

$$E(\boldsymbol{\theta}) = \boldsymbol{\mu} \tag{A.5}$$

$$V(\boldsymbol{\theta}) = \frac{\nu}{\nu - 2}\Sigma,\ \nu > 2 \tag{A.6}$$

A.2 Modélisation des paramètres de variance

A.2.1 La distribution Gamma

La distribution Gamma est la distribution *a priori* conjuguée de l'inverse de la variance (la précision), notée α^{-1} d'une distribution normale univariée.

En notant a le paramètre de forme et b le paramètre de précision (inverse scale) de la loi Gamma alors $\alpha^{-1} \sim \text{Gamma}(a, b)$ admet pour densité

$$p\left(\alpha^{-1}\right) \propto \frac{1}{\alpha^{a-1}} \exp\left(-\frac{b}{\alpha}\right),\ \alpha > 0 \tag{A.7}$$

A.2.2 La distribution Inverse Gamma

Réciproquement, la distribution Inverse-Gamma est la distribution *a priori* conjuguée de la variance α d'une distribution normale univariée.

En reprenant les notations du paragraphe précédent, si $\alpha^{-1} \sim \text{Gamma}(a, b)$ alors $\alpha \sim \text{Inv} - \text{Gamma}\left(a, \frac{1}{b}\right)$.

En notant a le paramètre de forme et b le paramètre d'échelle de la loi Inverse-Gamma alors $\alpha \sim \text{Inv} - \text{Gamma}(a, b)$ admet pour densité

$$p(\alpha) \propto \frac{1}{\alpha^{a+1}} b^a \exp\left(-\frac{b}{\alpha}\right),\ \alpha > 0 \tag{A.8}$$

A.2.3 Expression de la distribution *a posteriori* du paramètre de variance

Soit X_i, $1 \leq i \leq n$ un échantillon de n variables gaussiennes identiquement distribuées de moyenne μ connue et de variance α inconnue, $X_i \sim \mathcal{N}(\mu, \alpha)$.

La distribution *a priori* de Jeffreys est $p(\alpha) \propto \frac{1}{\alpha}$

La distribution *a posteriori* sous le prior de Jeffreys s'écrit

$$p(\alpha|X,\mu) \propto \frac{1}{\alpha}\frac{1}{\sqrt{\alpha}^n}\exp -\frac{1}{2}\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\alpha} \tag{A.9}$$

$$\propto \frac{1}{\alpha^{1+\frac{n}{2}}}\exp -\frac{1}{\alpha}\left(\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^2\right) \tag{A.10}$$

On reconnaît l'expression de la loi $\text{Inv}-\text{Gamma}\left(\frac{n}{2},\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^2\right)$.

Soit le prior conjugué informatif $p(\alpha) \sim \text{Inv}-\text{Gamma}(a_0,b_0)$. La distribution *a posteriori* de α sous ce prior informatif s'écrit

$$p(\alpha|X,\mu) \propto \frac{1}{\alpha^{a_0+1}}\exp -\frac{b_0}{\alpha}\frac{1}{\sqrt{\alpha}^n}\exp -\frac{1}{2}\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\alpha} \tag{A.11}$$

$$\propto \frac{1}{\alpha^{1+a0+\frac{n}{2}}}\exp -\frac{1}{\alpha}\left(b0+\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^2\right) \tag{A.12}$$

On reconnaît l'expression de la loi $\text{Inv}-\text{Gamma}\left(a0+\frac{n}{2},b0+\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^2\right)$.

A.2.4 La distribution Khi2

Soit $\alpha^{-1} \sim \chi^2(\nu)$, alors α^{-1} admet pour densité

$$p\left(\alpha^{-1}\right) \propto \frac{1}{\alpha^{\frac{\nu}{2}-1}}\exp\left(-\frac{1}{2}\frac{1}{\alpha}\right),\ \alpha>0 \tag{A.13}$$

On reconnaît la distribution $\text{Gamma}\left(\frac{\nu}{2},\frac{1}{2}\right)$. On a donc la relation suivante :

$$\alpha^{-1} \sim \chi^2(\nu) \tag{A.14}$$

$$\sim \text{Gamma}\left(\frac{\nu}{2},\frac{1}{2}\right) \tag{A.15}$$

A.2.5 La distribution inverse Khi2

Soit $\alpha \sim \text{Inv}{-}\chi^2(\nu)$, alors α admet pour densité

$$p(\alpha) \propto \frac{1}{\alpha^{\frac{\nu}{2}+1}}\exp\left(-\frac{1}{2}\frac{1}{\alpha}\right),\ \alpha>0 \tag{A.16}$$

On reconnaît la distribution $\text{Inv}-\text{Gamma}\left(\frac{\nu}{2},\frac{1}{2}\right)$. On a donc la relation suivante :

$$\alpha \sim \text{Inv}{-}\chi^2(\nu) \tag{A.17}$$

$$\sim \text{Inv} - \text{gamma}\left(\frac{\nu}{2}, \frac{1}{2}\right) \tag{A.18}$$

Plus généralement, si α admet pour densité

$$p(\alpha) \propto S^{\frac{1}{2}} \frac{1}{\alpha^{\frac{\nu}{2}+1}} \exp\left(-\frac{1}{2}\frac{S}{\alpha}\right), \alpha > 0 \tag{A.19}$$

alors $\alpha \sim \frac{S}{\chi^2_\nu}$.

Cette distribution est un cas particulier de la distribution scaled inverse chi square de degré de liberté ν et d'échelle S notée $Inv - \chi^2(\nu, S)$ obtenue en posant $\nu = 1$. Ainsi, $\alpha \sim Inv - \chi^2(1, S)$.

Soit $\beta \sim \text{Inv} - \chi^2(\nu, S)$ alors β admet pour densité

$$p(\beta) \propto S^{\frac{\nu}{2}} \frac{1}{\beta^{\frac{\nu}{2}+1}} \exp\left(-\frac{1}{2}\frac{\nu S}{\beta}\right), \beta > 0 \tag{A.20}$$

On reconnaît la densité de la loi Inverse Gamma de paramètres $a = \frac{\nu}{2}$ et $b = \frac{\nu}{2}S$.

En résumé,

$$\alpha \sim \frac{S}{\chi^2_\nu} \tag{A.21}$$

$$\sim \text{Inv}{-}\chi^2(\nu, S) \tag{A.22}$$

$$\sim \text{Inv} - \text{gamma}\left(\frac{\nu}{2}, \frac{S}{2}\right) \tag{A.23}$$

A.2.6 La distribution inverse Wishart

A.2.6.1 Définition

La distribution inverse Wishart est la distribution *a priori* conjuguée de la matrice de variance covariance d'une distribution normale multivariée. Soit $X = (X_1, \ldots, X_p)$, p vecteurs gaussiens centrés indépendants et identiquement distribués de distribution $X_i \sim \mathcal{N}(0, W)$. La distribution *a priori* conjuguée de W est la distribution inverse Wishart. Cette distribution est une généralisation multivariée d'une distribution gamma.

La densité est toujours finie et l'intégrale est toujours finie.

Il existe plusieurs paramétrisations de la loi inverse Wishart. La paramétrisation choisie est la paramétrisation adoptée dans l'ouvrage de Gelman et al.

Soit W une matrice $q \times q$ définie positive. W est distribuée selon la loi inverse Wishart de paramètres ν et S^{-1}. La densité de W est proportionnelle à

$$p(W) \propto |W|^{-\frac{\nu+q+1}{2}} \exp -\frac{1}{2}\text{tr}SW^{-1} \tag{A.24}$$

On note $W \sim \text{Inv} - \text{Wishart}_\nu\left(S^{-1}\right)$ où ν est le nombre de degrés de liberté et S^{-1} est la matrice de précision (inverse scale) de dimension $q \times q$.

Lien avec la distribution de Wishart : si $W \sim \text{Inv} - \text{Wishart}_\nu\left(S^{-1}\right)$ alors $W^{-1} \sim \text{Wishart}_\nu(S)$.

Le générateur de distributions IW sous R riwish se trouve dans le package MCMCpack et prend pour argument la matrice scale : riwish (ν, S).

A.2.6.2 Expression de la distribution *a posteriori* de la matrice de covariance

La distribution *a priori* de Jeffreys est $p(W) \propto |W|^{-\frac{1}{2}(q+1)}$

La distribution *a posteriori* de W sous le prior de Jeffreys est

$$p(W|X) \propto |W|^{-\frac{q+1}{2}}|W|^{-\frac{n}{2}} \exp -\frac{1}{2}\text{tr}X^tXW^{-1} \tag{A.25}$$

$$\propto |W|^{-\frac{n+q+1}{2}} \exp -\frac{1}{2}\text{tr}X^tXW^{-1} \tag{A.26}$$

où la matrice X^tX est n fois la matrice de covariance empirique de X. En identifiant avec l'expression de la densité A.24, $W|X \sim \text{Inv} - \text{Wishart}_n\left(\left(X^tX\right)^{-1}\right)$

La distribution *a posteriori* de W sous le prior informatif conjugué $p(W) \sim \text{Inv} - \text{Wishart}_{\nu_0}\left(S_0^{-1}\right)$ s'écrit

$$p(W|X) \propto |W|^{-\frac{\nu_0+q+1}{2}} \exp -\frac{1}{2}\text{tr}S_0W^{-1}|W|^{-\frac{n}{2}} \exp -\frac{1}{2}\text{tr}X^tXW^{-1} \tag{A.27}$$

$$\propto |W|^{-\frac{n+\nu_0+q+1}{2}} \exp -\frac{1}{2}\text{tr}\left(S_0 + X^tX\right)W^{-1} \tag{A.28}$$

En identifiant avec l'expression de la densité A.24, $W|X \sim \text{Inv} - \text{Wishart}_{n+\nu_0}\left(\left(S_0 + X^tX\right)^{-1}\right)$

Annexe B

Lettre de sollicitation des laboratoires

Lettre de sollicitation des laboratoires participant au circuit 37

Le but de ce courrier est de vous présenter le projet de recherche TRACEBIOACTIVITY dans le domaine des micropolluants organiques de l'eau et les modalités de participation du BIPEA et des laboratoires dans cette étude.

Le projet TRACEBIOACTIVITY

Cadre de l'étude

Dans le cadre de la **thèse en statistique appliquée** de Séverine Demeyer (2008-2011), doctorante au LNE (Laboratoire national de métrologie et d'essais) et au CNAM (Conservatoire National des Arts et Métiers), le service Mathématiques et Statistiques du LNE collabore avec le BIPEA au développement de nouvelles approches statistiques dans le traitement des comparaisons interlaboratoires.

Cette thèse s'inscrit dans le cadre d'un projet du programme de métrologie européen intitulé TRACEBIOACTIVITY dont le but est **d'améliorer la traçabilité métrologique des mesures** dans des domaines relatifs à la santé publique.

Le LNE et le BIPEA collaborent depuis de nombreuses années sur différents sujets de recherche et cette étude s'inscrit dans le cadre de cette collaboration. Nous avons alors choisi de développer ces nouvelles approches dans le cadre du circuit 37 du BIPEA. Le choix d'un travail sur les micropolluants organiques de l'eau doit donc permettre d'illustrer et de valider les modèles statistiques développés.

Motivation de l'étude

Actuellement la valeur de consensus d'un test d'aptitude est calculée comme la moyenne robuste des valeurs données par les laboratoires participants. Dans certains cas ce choix de la moyenne robuste peut apparaître discutable. En effet, les démarches initiées notamment dans le laboratoire de la Drôme par Félix Massat ont mis en évidence des **limitations de cette méthode pour les polluants de l'eau** (circuit 37 du BIPEA).

Les inconvénients de la moyenne robuste proviennent du fait qu'elle ne prend en compte que les résultats des laboratoires et que ceux-ci sont en trop petit nombre d'un point de vue statistique. Plus précisément :

- on ne dispose que d'un seul résultat par laboratoire ce qui ne permet pas d'estimer la bonne répétabilité du laboratoire
- les résultats des laboratoires sont traités comme s'ils provenaient d'un laboratoire unique: on ne prend pas en compte la **disparité des laboratoires** alors qu'elle existe. Un seul résultat par laboratoire ne permet pas non plus d'estimer cette disparité.

Objectif

Partant de ce constat, nos travaux de recherche ont pour but de développer et de proposer des modèles alternatifs de calcul de la valeur consensuelle et de l'incertitude associée prenant en compte de l'information auxiliaire propre au laboratoire.

L'originalité de l'approche réside dans l'utilisation de cette information auxiliaire. Pour cela il est nécessaire de :

- recueillir cette information auxiliaire à partir d'un questionnaire envoyé aux laboratoires participant aux comparaisons du circuit 37

- combiner cette information aux résultats des laboratoires afin de proposer une valeur de consensus plus représentative de la réalité, c'est à dire de la valeur vraie.

Etapes du projet

1- Choix des micropolluants organiques d'intérêt pour le projet

Le choix des familles de l'étude est : BTEX, PCB et Triazines. Au sein de chaque famille tous les analytes du circuit ont été sélectionnés.

2- Elaboration d'un questionnaire à destination des laboratoires du circuit 37

Un collège de 5 experts dans le domaine des micropolluants organiques a été constitué afin de finaliser et de valider le questionnaire préalablement établi par le LNE et le BIPEA (cf. Partenaires du projet).

Le questionnaire comprend des questions d'ordre général sur le laboratoire et des questions spécifiques à la pratique du laboratoire dans la mesure des analytes des trois familles de micropolluants organiques sélectionnés.

3- Développement d'un outil statistique pour identifier les facteurs qui contribuent aux biais des laboratoires

Les facteurs pertinents seront identifiés à partir des réponses des laboratoires au questionnaire.

Pour cela l'outil statistique en cours de développement s'appuie sur des groupes de questions établis par les experts.

Diffusion et intérêt pour les laboratoires

La nouvelle méthode de calcul de la valeur de consensus permettra aux laboratoires de mieux **suivre l'évolution de leur biais**.

L'identification des facteurs et leurs contributions aux biais, permettra aux laboratoires de **mieux contrôler leur biais**.

Pour cela, nous vous communiquerons les résultats de l'étude dès leur validation.

Partenaires du projet

Le projet réunit le LNE, le BIPEA ainsi qu'un représentant de chacun des laboratoires suivants qui constituent notre collège d'experts:

- LDAR (Laboratoire Départemental d'Analyses et de Recherche de Périgueux)
- LA DROME Laboratoires
- LCDI
- CRECEP (Centre de recherche d'expertise et de contrôle des eaux de Paris)
- ESPCI (Laboratoire Environnement et Chimie analytique de l'Ecole Supérieure de Physique et Chimie Industrielle de la ville de Paris)

Participation des laboratoires du BIPEA

Nous vous sollicitons afin de renseigner le questionnaire, établi spécifiquement à destination des laboratoires participant au circuit 37 du BIPEA, intitulé :

« Pratique du laboratoire dans la mesure des micropolluants organiques »

Votre participation à cette étude est indispensable pour la validation de la méthodologie développée. Elle s'inscrit dans le cadre d'un programme de recherche et n'a donc aucun but commercial. Elle permettra de développer et de valider des outils statistiques utiles dans le cadre du projet européen.

Cette participation est strictement confidentielle. Elle ne revêt pas de caractère officiel et ne saurait en aucun cas faire l'objet d'un contrôle par les organismes d'accréditation ou les commissions d'agrément.

Une large participation des laboratoires est vivement souhaitée et nécessaire au développement et à la validation de nos travaux de recherche.

Nous vous invitons à y répondre d'ici le 31 mars 2009.

Comparaisons sélectionnées pour ce questionnaire

Trois comparaisons par niveau de concentration ont été sélectionnées. On considère donc 9 participations à des tests d'aptitude par famille. Au total ce sont 18 comparaisons qui sont couvertes par ce questionnaire.

Certaines questions dans le questionnaire nécessitent de se reporter au rapport correspondant à la comparaison pour renseigner la question.

Dans le questionnaire les comparaisons sont systématiquement repérées par les chiffres 1 à 9 pour chaque famille.

Le tableau ci-dessous donne **le numéro du rapport** correspondant au numéro de la comparaison dans le questionnaire et à la famille considérée.

Par exemple, la première comparaison considérée pour la famille BTEX correspond au rapport n°53 de l'essai de février 2009.

N° de la comparaison dans le questionnaire	BTEX	PCB	Triazines
1	53 (02/2009)	53 (02/2009)	54 (03/2009)
2	52 (11/2008)	51 (09/2008)	52 (11/2008)

3	51 (09/2008)	49 (04/2008)	50 (06/2008)
4	49 (04/2008)	47 (02/2008)	48 (03/2008)
5	48 (03/2008)	45 (09/2007)	46 (11/2007)
6	47 (02/2008)	43 (04/2007)	44 (06/2007)
7	46 (11/2007)	41 (02/2007)	42 (03/2007)
8	45 (09/2007)	39 (09/2006)	40 (11/2006)
9	43 (04/2007)	37 (04/2006)	38 (06/2006)

Confidentialité

Nous nous engageons à respecter la plus stricte confidentialité sur les informations qui nous seront fournies à travers le questionnaire.

Seul le BIPEA aura accès à l'identité des laboratoires et codera les laboratoires.

Seuls le BIPEA et le LNE auront accès aux données du questionnaire.

Les informations ne seront pas communiquées aux partenaires du projet européen.

Le questionnaire est hébergé sur le site internet du LNE et la confidentialité des informations fournies par les laboratoires est garantie par le LNE au moment où vous cliquerez sur le lien qui vous est affecté vers le questionnaire.

Ce lien vous sera envoyé individuellement début mars par Séverine Demeyer dans un e-mail intitulé : *LNE/BIPEA : Invitation à participer au questionnaire 'Pratique du laboratoire dans la mesure des micropolluants organiques'*.

Contacts

Vos contacts au LNE :

Séverine Demeyer
Doctorante
01 30 69 32 47
severine.demeyer@lne.fr

Nicolas Fischer
Responsable du Service Mathématiques et Statistiques
01 30 69 10 79
nicolas.fischer@lne.fr

Philippe Charlet
Coordinateur R&D
01 30 69 21 95
philippe.charlet@lne.fr

LNE, 29 avenue Roger Hennequin 78197 Trappes Cedex

Votre contact au BIPEA :

Marie-Philippe Seiller
Responsable qualité
01 47 33 96 39
mpseiller@bipea.org

Annexe C

Aperçus du questionnaire en ligne

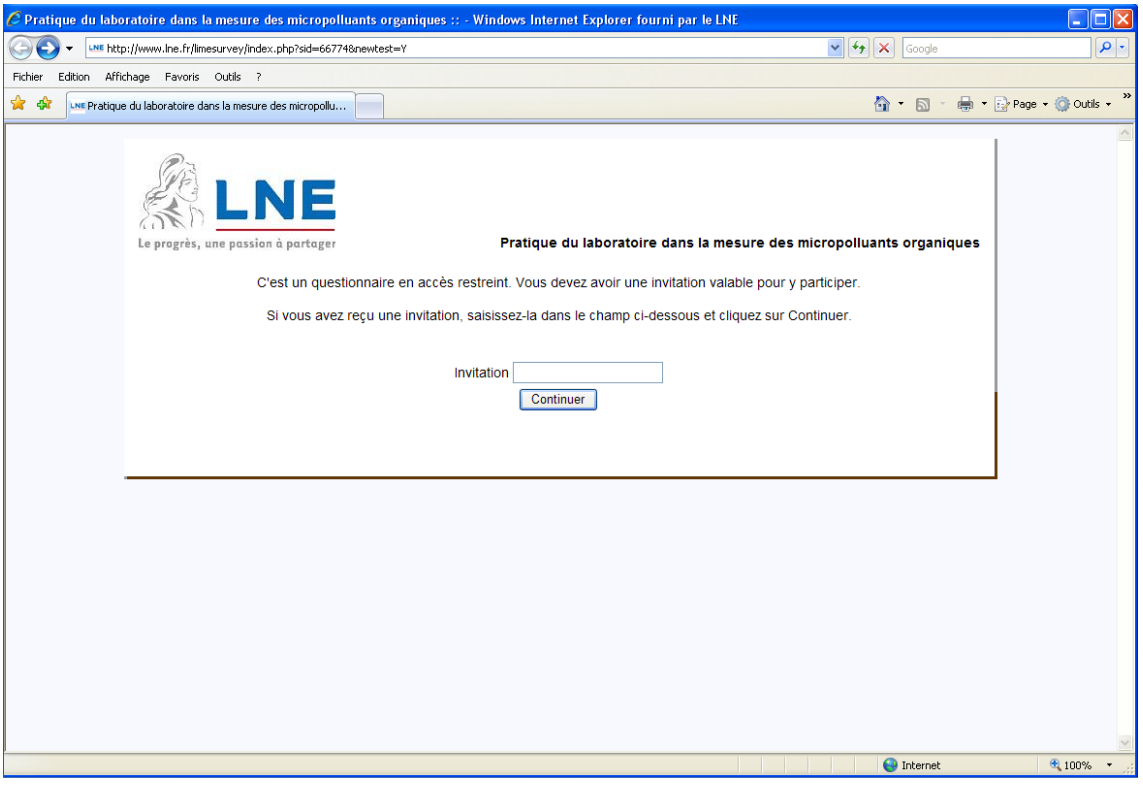


TABLE C.1 – Accès au questionnaire

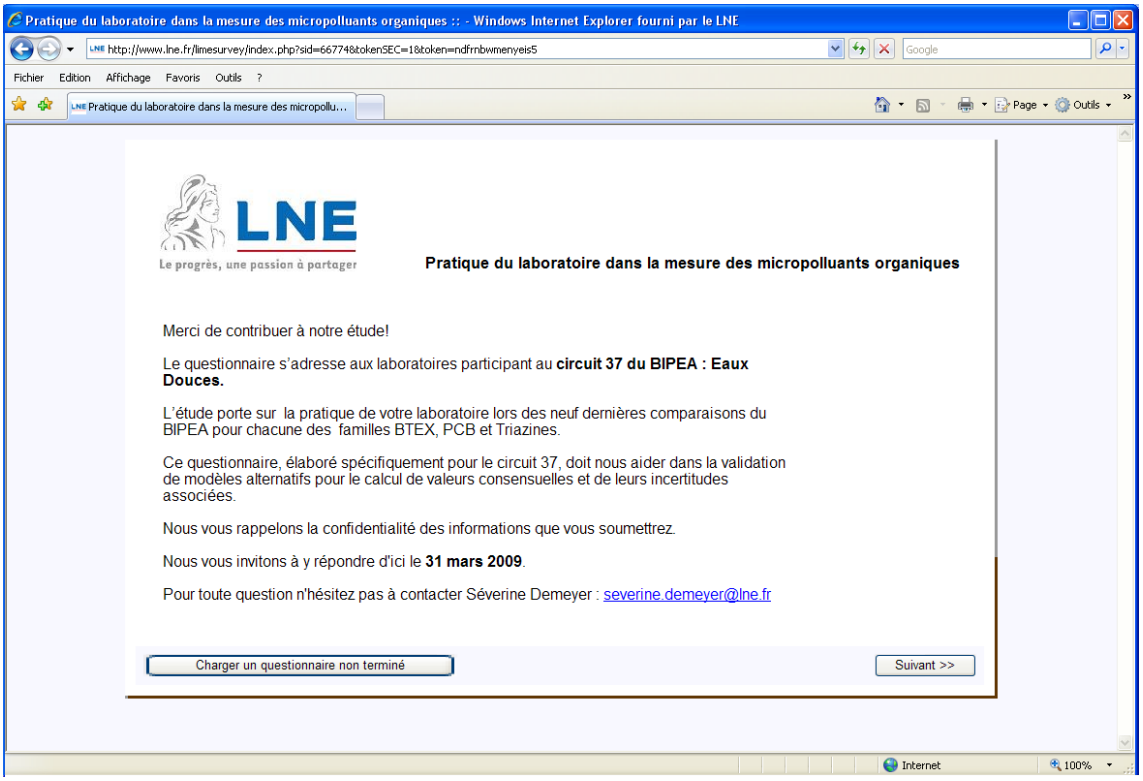


TABLE C.2 – Accueil du questionnaire

APERÇU DU QUESTIONNAIRE EN LIGNE

LNE
Le progrès, une passion à partager

Pratique du laboratoire dans la mesure des micropolluants organiques

0% 100%

QUESTIONS GENERALES - ENVIRONNEMENT DU LABORATOIRE

Le laboratoire est-il accrédité ISO 17025 ?

- (●) Oui
- ( ) Non

Est-il accrédité pour le programme 100-1 (physico-chimie), pour tout ou partie de ces familles ?

	Oui	Non
BTEX	( )	( )
PCB	( )	( )
Triazines	( )	( )

?

La portée d'accréditation est-elle la même pour les trois familles?

- ( ) Oui
- ( ) Non

Le laboratoire utilise-t-il les méthodes normalisées suivantes ?

	Oui	Non
BTEX NF EN 11423-1	( )	( )
PCB NF EN ISO 6468	( )	( )
Triazines NF EN ISO11369	( )	( )

Nombre de participations par an aux tests d'aptitude toute famille, tout organisateur :

Seuls les chiffres sont autorisés pour ce champ

Nombre total d'échantillons traités par an en micropolluants organiques :

Seuls les chiffres sont autorisés pour ce champ

TABLE C.3 – Déroulement du questionnaire

APERÇU DU QUESTIONNAIRE EN LIGNE

Nombre total d'échantillons traités par an :

Seuls les chiffres sont autorisés pour ce champ

Nombre de personnes de l'entite micropolluants organiques ?

Seuls les chiffres sont autorisés pour ce champ

Estimez-vous l'incertitude de mesure par :

Cochez la ou les réponses

- [ ] décomposition des sources (5M)
- [ ] carte de controle
- [ ] estimation globale
- [ ] essais interlaboratoires
- [ ] plans d'experiences
- [ ] Autre:

Rendez-vous l'incertitude à vos clients ?

- ( ) Oui
- ( ) Non

Reprendre plus tard << Précédent Suivant >>

TABLE C.4 – Déroulement du questionnaire

Annexe D

Architecture des programmes d'estimation

D.1 Programmes d'estimation du modèle à équations structurelles

Le programme `estime_SEM.r` regroupe les programmes suivants utilisés pour l'estimation du modèle à équations structurelles décrits ci-dessous

- `val_init.r` : programme qui génère les valeurs initiales pour les variables latentes structurelles et qui comprend les valeurs initiales des paramètres
- `rep_latentes.r` : programme qui impute les réponses latentes associées aux variables catégorielles dans leur distribution conditionnelle *a posteriori* sous la contrainte d'identifiabilité les concernant
- `VL_struct.r` : programme qui impute les variables latentes structurelles dans leur distribution conditionnelle *a posteriori* sous la contrainte d'identifiabilité les concernant
- `seuils.r` : programme qui tire les seuils associés aux réponses latentes des variables catégorielles dans leur distribution conditionnelle *a posteriori*
- `mu_theta.r` : programme qui tire les coefficients de régression des modèles de régression du modèle extérieur dans leur distribution conditionnelle *a posteriori*
- `stockage.r` : programme qui stocke les réalisations de l'algorithme de Gibbs pour tous les paramètres et toutes les variables latentes à chaque itération

Les relations entre les programmes sont représentées à la figure D.1.

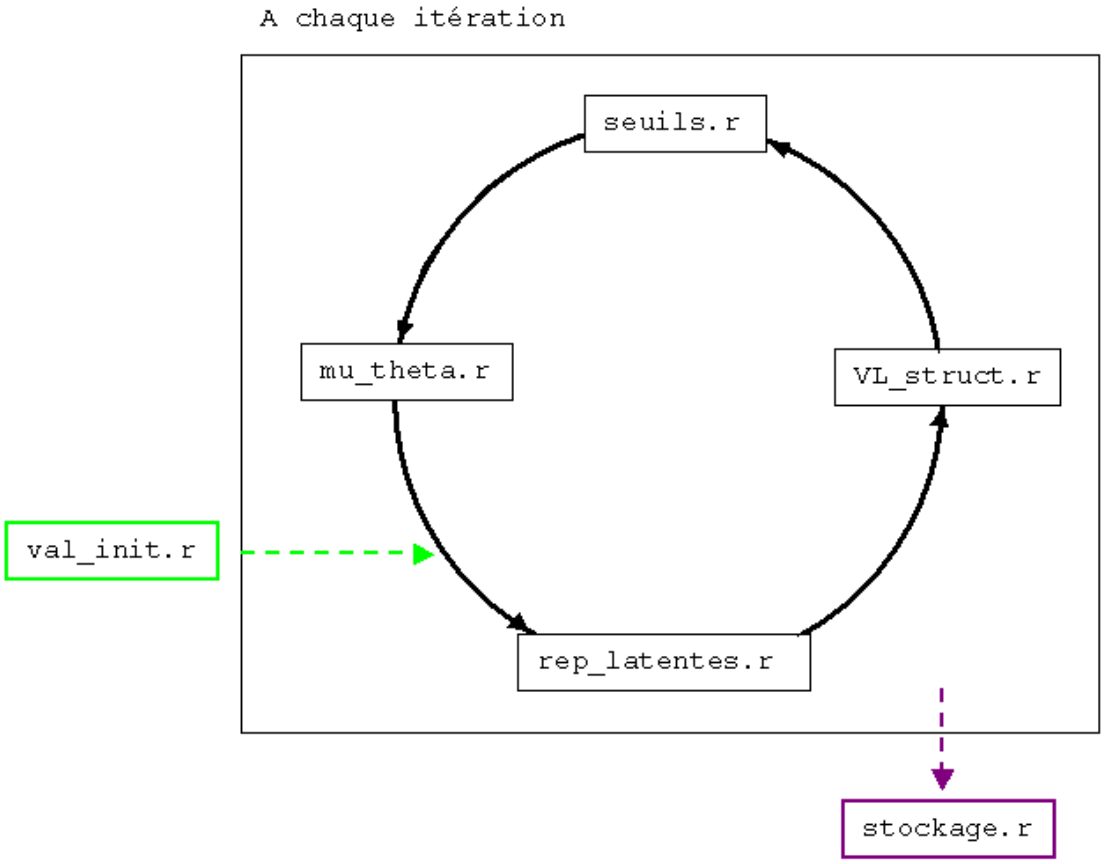


TABLE D.1 – Relations entre les programmes d'estimation du modèle structurel.

D.2 Programmes d'estimation de la valeur de consensus et des biais de mesure

L'estimation de la valeur de consensus, de son incertitude associée et des biais de mesure est réalisée à partir des valeurs stockées appelées `stockage` des variables latentes lors de l'estimation du modèle à équations structurelles.

L'estimation fait appel à différents programmes décrits ci-dessous

- `scores.r` : programme qui calcule les scores des laboratoires à partir des valeurs stockées des variables latentes structurelles
- `poids.r` : programme qui calcule les poids des laboratoires à partir des scores
- `algorithme_A.r` : programme qui implémente la mise en oeuvre de l'algorithme A de la norme NF ISO 13528
- `val_cons.r` : programme qui calcule la distribution marginale *a posteriori* de la valeur de consensus
- `biais.r` : programme qui calcule la distribution *a posteriori* des biais de mesure des

laboratoires

Les relations entre les programmes sont représentées à la figure D.2.

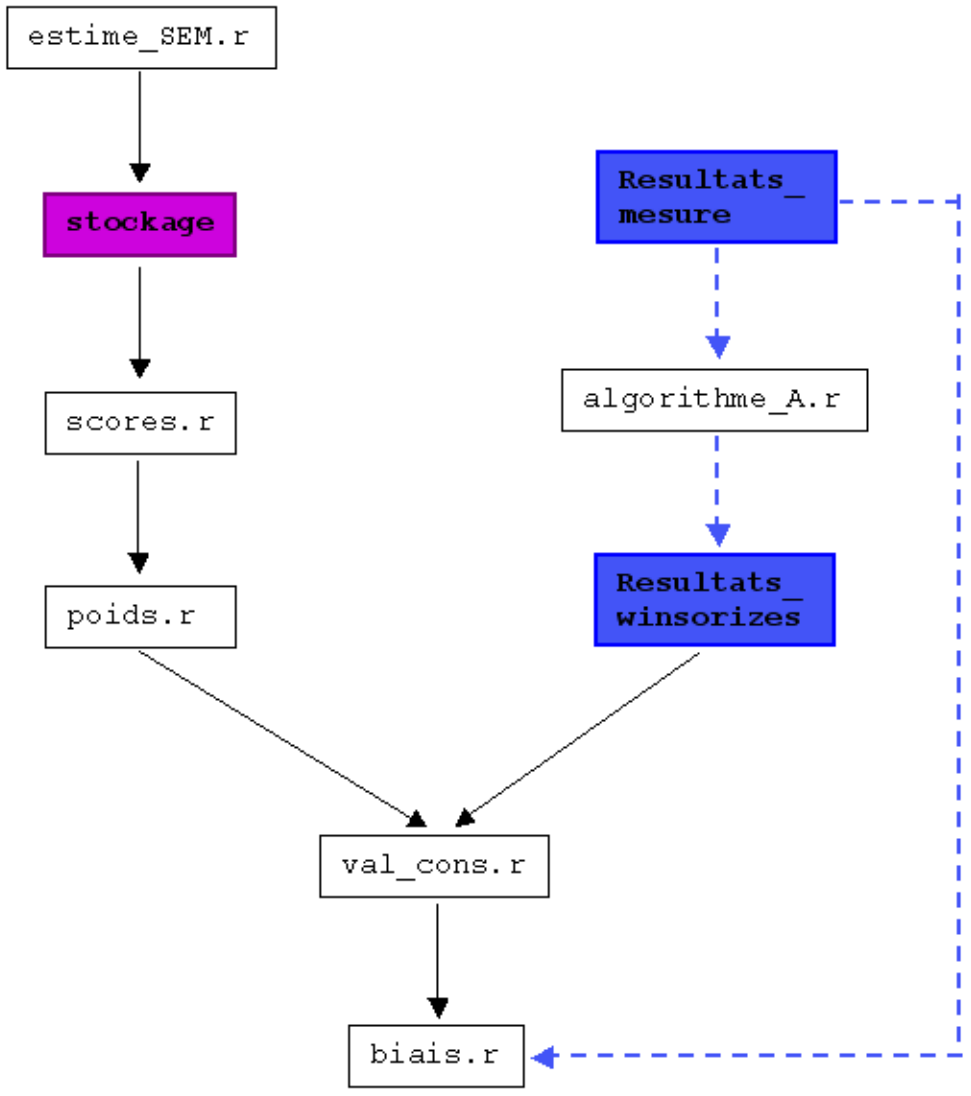


TABLE D.2 – Relations entre les programmes d'estimation menant au calcul de la valeur de consensus et des biais de mesure.