université PARIS-SACLAY

Tests des composantes de la variance dans les modèles à effets mixtes pour des petits échantillons. Application à l'étude de la variabilité génotypique chez *Arabidopsis thaliana*

Variance components testing in mixed effects models with small sample sizes. Application to the study of genotypic variability in Arabidopsis thaliana

Thèse de doctorat de l'université Paris-Saclay

École doctorale de Mathématiques Hadamard (EDMH) n° 574
Spécialité de doctorat : Mathématiques aux interfaces
Graduate School : Mathématiques. Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **MaIAGE** (Université Paris-Saclay, INRAE), sous la direction de **Estelle KUHN**, directrice de recherche (MaIAGE, INRAE) et le co-encadrement de **Charlotte BAEY**, maîtresse de conférence (Laboratoire Paul Painlevé, Université de Lille).

Thèse soutenue à Paris-Saclay, le 5 décembre 2024, par

Tom GUÉDON

Composition du jury (par ordre alphabétique)

Membres du jury avec voix délibérative

Cécile DUROT Professeure des universités, Université Paris Nanterre	Rapporteure & Examinatrice
Sébastien GADAT Professeur des universités, Université Toulouse 1 Capitole	Examinateur
Anne GEGOUT-PETIT Professeure des universités, Université de Lorraine	Examinatrice
Jean-Michel MARIN Professeur des universités, Université de Montpellier	Rapporteur & Examinateur
Tristan MARY HUARD Directeur de recherche, Université Paris-Saclay INRAE	Examinateur

THESE DE DOCTORAT
NNT : 2020UPASA001

université PARIS-SACLAY
ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Thèse préparée au sein de l'école doctorale de Mathématiques Hadamard (EDMH) n° 574.

Title : Variance components testing in mixed effects models with small sample sizes. Application to the study of genotypic variability in *Arabidopsis thaliana*.

Keywords : mixed effects models, variance components testing, parametric Bootstrap, Monte Carlo methods, stochastic approximation, mechanistic models.

Abstract : Mixed-effects models allow for the analysis of data with a hierarchical structure, such as longitudinal data, which are measurements collected on the same individual over time. These models take into account variability within measurements for each individual (intra-individual variability) and between different individuals (inter-individual variability) through two types of effects : fixed effects, which are common to all individuals in the population, and random effects, which vary from one individual to another.

The latter are modeled by unobserved latent variables that account for the inter-individual variability in the population. Identifying the model parameters that are the source of this variability is an important issue, especially for studying genotypic variability in plant breeding. This objective can be formulated as a statistical test of the nullity of the random effects variances. The likelihood ratio statistic is considered.

In this context, however, the test presents several challenges. Theoretically, the nullity of the variances is problematic because these variances are on the boundary of the parameter space. Classical results of methods based on maximum likelihood are not valid anymore. Moreover, the Fisher information matrix is singular in this context. Additionally, if non-tested variances are null, asymptotic tests are no longer applicable.

In this thesis work, a likelihood ratio test procedure for the nullity of variance components in mixed-effects models is proposed, based on parametric Bootstrap. The consistency of this test procedure under the null hypothesis is demonstrated, provided that a judicious choice is made for the parameter used for simulating the Bootstrap samples. A shrinkage step for this parameter is proposed to address boundary issues and the singularity of the Fisher Information matrix.

From a computational perspective, since mixed-effects models are latent variable models, their likelihood is often not explicit, making the estimation of the likelihood ratio statistic challenging. Estimating this statistic amounts to estimating a ratio of normalization constants of probability densities. A new estimation procedure for this ratio, based on stochastic approximation, is proposed. It involves finding the zero of a function defined by an expectation. This new estimator is consistent and asymptotically Gaussian. It demonstrates very good theoretical performance and can be integrated into a parameter estimation procedure in a latent variable model.

Finally, motivated by the analysis of genotypic variability for plant breeding, a study of 48 genotypes of *Arabidopsis thaliana* is presented. A complex mechanistic model describing carbon and nitrogen exchanges between the plant and its environment is integrated into a mixed-effects model to identify which biological parameters are responsible for the observed genotypic variability.

Titre : Tests des composantes de la variance dans les modèles à effets mixtes pour des petits échantillons. Application à l'étude de la variabilité génotypique chez *Arabidopsis thaliana*.

Mots clés : modèles à effets mixtes, test des composantes de la variance, Bootstrap paramétrique, méthodes de Monte-Carlo, approximation stochastique, ratio de constantes de normalisation, modèles mécanistes, variabilité génotypique.

Résumé : Les modèles à effets mixtes permettent d'analyser des données présentant une structure hiérarchique, telles que les données longitudinales, qui sont des mesures collectées sur un même individu au cours du temps. Ces modèles prennent en compte la variabilité au sein des mesures effectuées sur chaque individu (variabilité intra-individuelle) et entre différents individus (variabilité inter-individuelle), grâce à deux types d'effets : les effets fixes, communs à tous les individus de la population, et les effets aléatoires, variables d'un individu à l'autre.

Ces derniers sont modélisés par des variables latentes non observées, qui portent la variabilité inter-individuelle de la population. Identifier les paramètres du modèle à la source de cette variabilité est une question importante, en particulier pour l'étude de la variabilité génotypique en amélioration des plantes. Cet objectif peut être formulé comme un test statistique de nullité des variances des effets aléatoires. La statistique du rapport de vraisemblances est considérée.

Dans ce contexte, ce test présente cependant plusieurs défis. Théoriquement, la nullité des variances pose problème car ces variances se trouvent à la frontière de l'espace des paramètres. Les résultats classiques des méthodes basées sur le maximum de vraisemblance ne sont pas valides. Par ailleurs, la matrice d'information de Fisher est singulière dans ce cadre. De plus, si des variances non testées sont nulles, les tests asymptotiques ne sont plus applicables.

Dans ce travail de thèse, une procédure de test du rapport de vraisemblances de nullité des composantes de la variance dans les modèles à effets mixtes est proposée, basée sur le Bootstrap paramétrique. La consistance de cette procédure de test sous l'hypothèse nulle est démontrée, pour un choix judicieux du paramètre utilisé pour simuler les échantillons Bootstrap. Une étape de seuillage de ce paramètre est proposée pour pallier les problèmes de frontière et de singularité de la matrice d'Information de Fisher.

D'un point de vue computationnel, les modèles à effets mixtes étant des modèles à variables latentes, leur vraisemblance est la plupart du temps non explicite, ce qui rend difficile l'estimation de la statistique du rapport de vraisemblances. Estimer cette statistique revient à estimer un ratio de constantes de normalisation de densités de probabilité. Une nouvelle procédure d'estimation de ce ratio, basée sur une approximation stochastique, est proposée. Elle consiste à trouver le zéro d'une fonction définie par une espérance. Ce nouvel estimateur est consistant et asymptotiquement gaussien. Il présente de très bonnes performances théoriques et peut s'intégrer dans une procédure d'estimation de paramètres dans un modèle à variables latentes.

Pour finir, motivé par l'analyse de la variabilité génotypique pour l'amélioration des plantes, une étude de 48 génotypes d'*Arabidopsis thaliana* est présentée. Un modèle mécaniste complexe décrivant les échanges de carbone et d'azote entre la plante et son environnement est intégré à un modèle à effets mixtes afin d'établir quels sont les paramètres biologiques responsables de la variabilité génotypique observée.

Remerciements

Je tiens tout d'abord à remercier l'ensemble des membres du jury pour avoir accepté de participer à l'examen de cette thèse. Je remercie tout particulièrement Madame Durot et Monsieur Marin pour avoir consacré leur temps à rapporter ce travail de manière si complète et rigoureuse.

Je souhaite exprimer ma profonde gratitude à Charlotte et Estelle, mes directrices de thèse, sans qui ce travail n'aurait pas vu le jour. Merci pour votre présence constante tout au long de ces années, ainsi que pour votre bienveillance, qui a permis de maintenir un environnement de travail fondé sur la confiance. Sur le plan scientifique, je vous suis reconnaissant pour la rigueur et les précieux conseils que vous m'avez transmis. J'espère sincèrement que la fin de ce doctorat marquera le début d'une nouvelle phase de notre collaboration.

Je tiens également à remercier l'ensemble de mon comité de thèse pour la richesse des échanges que nous avons pu mener et pour la bienveillance dont vous avez toujours fait preuve.

Merci à tous les membres du projet Stat4Plant, pour les échanges enrichissants et l'ouverture scientifique que les journées du projet ont apportées. Je remercie particulièrement Jean-Benoist Leger pour ses conseils et son enseignement en programmation, ainsi que Céline Richard-Molard pour sa collaboration et pour avoir continué à nourrir mon intérêt pour l'application des statistiques. J'espère que ces collaborations se poursuivront dans l'avenir.

Je remercie également l'ensemble de l'unité MaIAGE pour leur accueil chaleureux, en particulier ceux du bâtiment 210. Merci pour vos conseils et pour les moments conviviaux partagés. Merci à Gildas pour tes conseils avisés et pour nos échanges scientifiques, ainsi que pour cette agréable semaine passée à Varsovie en ta compagnie. Un grand merci à Christian pour ton aide précieuse face à mes péripéties informatiques ! Merci également à Béatrice, Kascia, Laurent, Olivier, Mathieu, Auguste, Thibaut, Jeanne, Lorenzo, Vincent, Mahendra, Maud, Marc, Alain et tous ceux que je n'ai pas mentionnés, pour avoir contribué à rendre ces trois années si plaisantes.

Je tiens bien sûr à remercier mes collègues stagiaires et doctorants, avec qui j'ai partagé tant de bons moments. Je pense notamment à nos stagiaires : Julie, Pablo, Hugo, Inès, Thomas, Rita, Nicolas, Paul, Jeanne, Viviana, Ambre et tous les autres. Et bien sûr, un immense merci à l'ensemble des doctorants de MaIAGE. Marion, merci et désolé pour toutes les fois où tu as dû penser pour deux ! Je suis très heureux d'avoir partagé cette période avec toi. Merci à Madeleine pour ta gentillesse et ta bienveillance, à Marie, ma "presque co-bureau" pour reprendre tes mots, pour ta bonne humeur contagieuse (tu nous as manqué cette dernière année !), à Henri pour ta sympathie, à Antoine, mon "petit frère" de thèse, pour nos échanges et nos moments partagés au travail comme en dehors (courage pour cette dernière année où tu bénéficieras de toute l'attention de notre chère directrice !). Merci aussi à Arthur, pour ta bonne humeur, même depuis un étage différent ; à Katia, pour ton énergie et ta motivation. Ensuite merci à Andréa, même si je t'avais dit ne pas aimer cet exercice, je suis quand même obligé de te dire que je suis très heureux de t'avoir rencontré, et je suis sûr que malgré

la fin de notre covoiturage, ce ne sera pas la fin de notre amitié ! Enfin, merci à la nouvelle génération de doctorantes : Julie, ma nouvelle voisine de bureau, Anastasia, la stéphanoise (courage pour ces années restantes chez les Parigots !), et Eleonora, bon courage à toutes pour la suite.

Un grand merci à tous mes amis : la Bégomoly, Bagel, les Lazos, les copains de l'ENSAE, et ceux de Bologne. Même si je n'écris pas de messages individuels, sachez que le cœur y est. Mention spéciale pour les événements marquants de cette période : félicitations à Ju – cela ne nous rajeunit pas ! En tant que "petit frère de cœur", je suis ravi de devenir "tonton de cœur". Félicitations également à Mou et Gui, bon courage pour les préparatifs ! Enfin, félicitations à Cam et Pierrot pour votre appartement. Pierrot, un immense merci pour ton amitié indéfectible et pour avoir toujours été là.

Je ne peux pas ne pas réserver un paragraphe spécial à mes colocataires, qui ont été un soutien constant durant ces trois ans. Merci pour nos riches échanges, qui ont grandement attisé ma curiosité scientifique.

Enfin, je souhaite remercier ma famille. À la famille stéphanoise : Mamie Marinette, mes cousins et tous les petits (peut-être liront-ils un jour ce manuscrit ?), Brigitte, Élisabeth et bien sûr Thierry, à qui je dédie une pensée émue. À la famille lyonnaise de coeur : merci à Yvette Co pour sa présence constante depuis toujours, et merci à Lélé, pour ta bonne humeur, et merci de m'avoir accueilli pendant les derniers moments de la rédaction de ce travail ! Je tiens à faire un remerciement tout particulier à ma Tatie Yvette : pour la relecture de ce manuscrit à court terme, et à long terme pour avoir nourri mon goût pour les mathématiques. Tu occupes une place toute spéciale dans ces lignes.

Je termine avec mes parents, pour qui aucun mot ne suffira à exprimer ma gratitude. J'ai pleinement conscience de la chance que j'ai eue d'être si bien guidé tout au long de ma vie. Merci pour tout ; j'espère que vous savez à quel point je vous suis reconnaissant.

Un merci tout particulier à Clara et Missile : je suis si heureux de vous avoir à mes côtés et j'ai hâte de continuer à vous accompagner dans vos propres recherches.

Enfin, je souhaite remercier mon frère Pablo, qui a grandement façonné la personne que je suis aujourd'hui. Je suis très fier d'être ton petit frère, et je te remercie d'assumer si bien ton rôle de grand frère. Je vous souhaite à Léa et à toi, de belles prochaines années dans votre nouvelle vie !

Productions scientifiques

Publications

- *Guédon, T., Baey, C., & Kuhn, E. (2024). Bootstrap test procedure for variance components in nonlinear mixed effects models in the presence of nuisance parameters and a singular Fisher information matrix. Biometrika, asae025.*
- *Guédon, T., Baey, C., & Kuhn, E. (2024). Estimation of ratios of normalizing constants using stochastic approximation : the SARIS algorithm. arXiv preprint arXiv :2408.13022.*

Poster

- *Guédon, T., Baey, C., & Kuhn, E. Bootstrap test for variance components in nonlinear mixed effects models for small sample size in presence of nuisance parameters and singular Fisher Information Matrix. In European Meeting of Statisticians 2023 Warsaw 3–7 July, 2023 Book of Abstracts (p. 112).*

Codes informatiques

- https://github.com/tguedon

Présentations orales

- Rencontres des jeunes statisticien.ne.s 2022 (Porquerolles) (Du 03 au 07 avril 2022). Présentation orale
- StatMathAppli 2022, du 29 août au 02 septembre 2022 (Frejus). Présentation orale
- CMStatistics 2022, du 17 au 19 décembre 2022 (Londres). Présentation orale à distance
- European Meeting of Statisticians (EMS) 2023, du 03 au 07 Juillet 2023 (Varsovie). Présentation d'un poster
- StatMathAppli 2023, du 18 au 22 septembre 2023 (Frejus). Présentation d'un poster
- Séminaire MIA Paris Saclay, le 23 novembre 2023. Présentation orale
- Journées des statistiques (JDS), du 27 au 31 mai 2024 (Bordeaux). Présentation orale
- Séminaire LMAC (UTC), le 18 Juin 2023 (Compiègne). Présentation orale

Table des matières

Chapitre 1

Introduction

1.1 . Motivations biologiques et cadre statistique

1.1.1 . Étude de la variabilité génotypique

Le changement climatique, ainsi que la réduction de l'usage des intrants (fertilisants, pesticides) dans le souci de préserver l'environnement posent de nouveaux défis, auxquels l'agriculture moderne doit s'adapter. Dans ce contexte, comprendre l'interaction complexe entre les plantes et leur environnement est essentiel. En décryptant les mécanismes biologiques qui régissent la croissance et l'adaptation des plantes, il est possible de sélectionner des variétés plus performantes, ou plus résilientes, mieux adaptées aux nouveaux contextes environnementaux.

Le projet interdisciplinaire ANR *Stat4Plant*, dans lequel ce travail de doctorat s'inscrit, vise à développer de nouvelles méthodologies statistiques permettant de modéliser et d'analyser les interactions entre la plante et son environnement, en s'appuyant sur des collaborations étroites entre biologistes et statisticiens. Ce projet est centré autour de différents axes de recherche, dans le but de développer des méthodes statistiques pertinentes aux problématiques biologiques considérées. Cette thèse est associée à l'axe de recherche dédié à l'étude de la variabilité génotypique des plantes, en réponse à l'environnement.

De manière plus concrète, au sein d'une même espèce, une grande variabilité phénotypique est observée. La variabilité phénotypique fait référence aux différences observables dans les traits physiques et fonctionnels (les phénotypes) des individus au sein d'une population. Ces traits peuvent inclure des caractéristiques telles que la taille, la forme, la couleur, la résistance aux maladies, la production de fruits, le rendement, la surface foliaire et bien d'autres. Les deux facteurs principaux à la source de cette variabilité sont l'environnement (E) et la génétique (G). A condition environnementale fixée, la variabilité phénotypique observée a donc pour source principale la variabilité génétique de la population, dite variabilité génotypique. Afin d'améliorer ou de sélectionner les variétés (ou génotypes) les plus performantes, il est nécessaire d'identifier les caractéristiques biologiques à l'origine de la variabilité observée.

À cette fin, la modélisation mathématique mécaniste est un outil efficace car elle permet de modéliser explicitement les différents processus biologiques en jeu, de les intégrer pour prendre en compte leurs interactions et simuler dynamiquement le phénotype final en fonction de l'environnement considéré. Les modèles mécanistes sont construits à partir de connaissances expertes sur les processus décrits, et permettent mathématiquement de faire le lien entre des traits phénotypiques mesurables (variables de sortie du modèle) et différents paramètres biologiques sous tendant les relations mathématiques impliquées dans le modèle. Ces paramètres sont supposés indépendants du temps et de l'environnement, contrairement aux traits modélisés. Ainsi identifier les paramètres à la source de la variabilité observée permettrait

d'identifier les leviers d'action pertinents pour la sélection variétale. Pour cela, l'approche classique consiste à estimer pour chaque génotype les différents paramètres du modèle, puis à les comparer afin d'identifier ceux soumis à variation entre génotypes. Cependant cette approche peut souffrir du manque de données. En effet dans ce type d'étude les données sont difficiles et longues à obtenir et sont donc souvent en petits effectifs. Une approche individuelle génotype par génotype est alors souvent peu précise car basée sur un petit nombre de mesures. L'approche populationnelle permet de limiter cet effet car l'inférence des paramètres se fonde sur l'ensemble des individus disponibles. En particulier, quelques mauvaises mesures individuelles peuvent donc être compensées par celles du reste de la population.

La modélisation à effets mixtes est un outil statistique puissant permettant de considérer différents niveaux de variabilité au sein d'une population d'étude. Dans le cadre de la modélisation dynamique des processus de croissance de plantes, les données utilisées sont des données longitudinales, c'est à dire des mesures répétées au cours du temps sur chaque génotype considéré. Les modèles à effets mixtes permettent à la fois de considérer la variabilité intra-individuelle existant au sein de chaque individu statistique (i.e. au sein de chaque génotype), et la variabilité inter-individuelle existant entre les différents individus statistiques. Ces modèles impliquent deux types d'effets, d'une part les effets fixes communs à tous les individus, et d'autre part les effets aléatoires, qui varient d'un individu à l'autre. La capacité à distinguer parmi tous les effets du modèle mécaniste ceux qui peuvent être modélisés comme des effets fixes permettrait d'identifier plus clairement les processus à l'origine de cette variabilité génotypique observée.

Dans le cadre du projet *Stat4Plant*, ce travail de thèse a bénéficié de la collaboration de Céline Richard-Molard, écophysiologiste végétale (UMR EcoSys, INRAE Saclay) qui étudie l'adaptation d'*Arabidopsis thaliana* à un faible niveau de nutrition azotée. Afin d'identifier les processus à l'origine de la variabilité génotypique observée chez *Arabidopsis thaliana*, le modèle mécaniste complexe ARNICA, a été implémenté (Richard-Molard et al., 2007). Ce modèle vise à simuler le comportement de différents génotypes d'Arabidopsis thaliana en réponse à des nutritions azotées contrastées. Il simule dynamiquement les flux de carbone et d'azote dans la plante entière au cours de la phase végétative, les croissances en biomasse des parties aériennes et racinaires, et la surface foliaire.
Ce modèle intègre et quantifie certaines caractéristiques phénotypiques de la plante au cours de sa croissance, à l'aide d'équations dynamiques dépendant de différents paramètres biologiques. Identifier les paramètres du modèle ARNICA présentant une forte variabilité permettrait de comprendre la variabilité génotypique de l'adaptation de *Arabidopsis thaliana* à une faible alimentation azotée. Pour cela, un jeu de données construit sur 48 génotypes de *Arabidopsis*

thaliana a été mis à notre disposition par Céline Richard-Molard afin de mener une inférence statistique des paramètres du modèle ARNICA.

Ainsi l'objectif global de cette thèse, est de développer des méthodologies statistiques et computationnelles permettant d'identifier les paramètres présentant une variabilité inter-individuelle dans des modèles à effets mixtes, et d'appliquer ces outils à l'étude de la variabilité génotypique chez *Arabidopsis thaliana*.

1.1.2 . Les modèles à effets mixtes

Introduits par Laird and Ware (1982), les modèles à effets mixtes permettent d'analyser des données présentant une structure hiérarchique. Ils sont particulièrement utiles pour traiter les données longitudinales. L'application considérée dans la section 1.1.1 comporte ce type de données : les mesures répétées sont les différents traits phénotypiques d'un même génotype, et les individus sont les 48 génotypes d'*Arabidopsis thaliana*. Comme expliqué dans la section précédente ce type de données présente différentes sources de variabilité, prises en compte par la modélisation à effets mixtes qui considèrent à la fois des effets fixes, communs à tous les individus, et des effets aléatoires, dont les réalisations seront spécifiques à chaque individu statistique.

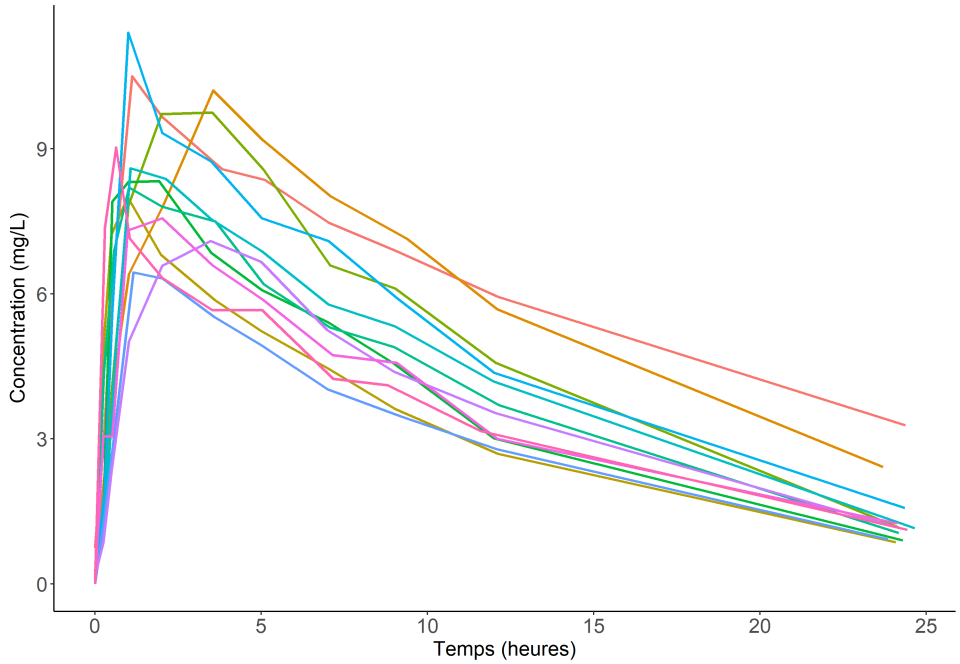


Figure 1.1 – Concentration de théophylline mesurée au cours du temps chez 12 sujets d'une étude pharmacologique

Un exemple couramment utilisé dans la littérature (Pinheiro and Bates, 2006; Davidian and Giltinan, 2003, 2017) pour illustrer les données sur lesquelles les modèles à effets mixtes peuvent être utilisés, est celui de l'étude pharmacocinétique de Boeckmann et al. (1994). Le jeu de données correspondant contient des mesures de concentrations de théophylline dans le sang (en mg/L) de plusieurs sujets, à différents moments après l'administration orale d'une dose unique de théophylline. Les données sont représentées graphiquement figure (1.1). Chaque courbe de couleur représente la concentration en théophylline mesurée chez un individu au cours du temps. On observe sur ce graphique une tendance générale, correspondant à la variabilité intra-individuelle, avec cependant une variabilité inter-individuelle marquée, par exemple sur les valeurs maximales atteintes.
De manière plus formelle, supposons que l'on observe N individus, chacun mesuré J_i fois ($i = 1, ..., N$). Soit $y_{ij} \in \mathbb{R}$ la j-ème mesure de l'individu i, on considère le modèle à effets mixtes suivant (Pinheiro and Bates, 2006; Baey et al., 2019) :

$$\begin{cases} y_{ij} &= g(x_{ij}, \varphi_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \varphi_i &= A_i\beta + B_i b_i, \quad b_i \sim \mathcal{N}(0, \Gamma) \end{cases} \tag{1.1}$$

Ce modèle est défini en deux couches hiérarchiques.
La première couche modélise le comportement individuel où :

- g est une fonction connue
- x_{ij} regroupe toutes les covariables connues de l'individu i au temps j. Ces covariables peuvent être par exemple des caractéristiques de l'individu ou des caractéristiques environnementales relatives à la j-ème mesure. Ces covariables sont connues
- φ_i est un paramètre individuel inconnu
- ε_{ij} est un terme aléatoire d'erreur centré, de variance σ^2 supposé Gaussien.

La seconde couche modélise le paramètre individuel φ_{ij}, où :

- β est un vecteur d'effets fixes de taille q, commun à tous les individus
- b_i est un vecteur de taille p d'effets aléatoires indépendant des résidus, associé à l'individu i, normalement distribué, de matrice de covariance Γ inconnue
- A_{ij} et B_{ij} sont des matrices de covariables connues.

Les résidus $(\varepsilon_{ij})_{ij}$ et les effets aléatoires $(b_i)_i$ sont supposés mutuellement indépendants.

Remarque 1.1. *Le modèle considéré ici est un modèle à uniquement deux couches hiérarchiques. Il est possible de le généraliser à plus de couches (Pinheiro and Bates, 2006).*

Le paramètre à estimer dans le modèle défini dans (1.1) est $\theta = (\beta, \Gamma, \sigma^2)$. Une approche possible pour l'inférence dans les modèles à effets mixtes est la méthode du maximum de vraisemblance. La vraisemblance relative aux variables aléatoires indépendantes $(y_i, b_i)_{i=1,...,N}$, est supposée appartenir à une famille paramétrique $\{f_i(y_i, b_i; \theta); \theta \in \Theta\}$ où :

$$\Theta = \{(\beta, \Gamma, \sigma^2); \beta \in \mathbb{R}^q, \Gamma \in \mathbb{S}_p^+, \sigma^2 \in \mathbb{R}_*^+\}$$

avec $\mathbb{S}_k^+$ l'ensemble des matrices symétriques semi-définies positives de taille k et $\mathbb{S}_k^{+*}$ l'ensemble des matrices symétriques définies positives de taille k.

Le modèle présenté ici est un modèle à variables latentes. En effet, les effets aléatoires $(b_i)_{i=1,\dots,N}$ ne sont pas observés. Pour l'inférence par maximum de vraisemblance, il est donc nécessaire de considérer la vraisemblance relative aux variables observées $y_{1:N}$:

$$L(\theta; y_{1:N}) = \prod_{i=1}^{N} \int f_i(y_i, b_i; \theta) db_i \tag{1.2}$$

Cette vraisemblance marginale correspond à la vraisemblance complète intégrée sur toutes les valeurs possibles des variables latentes. La vraisemblance complète $f_i(y_i, b_i; \theta)$ est connue, et calculable à partir de la structure hiérarchique du modèle, comme le produit de la densité de y_i conditionnellement à l'effet aléatoire b_i et de la densité de b_i.

On définit l'estimateur du maximum de vraisemblance par :

$$\hat{\theta}_N = \arg\max_{\theta \in \Theta} \quad L(\theta; y_{1:N})$$

Dans le cadre linéaire cette vraisemblance marginale est explicite. Cependant dans le cadre non linéaire, l'intégrale définie en (1.2) n'a généralement pas de forme explicite et doit être approchée par des méthodes numériques. La partie (1.4) introduit cette problématique. Les deux exemples suivants illustrent ce point.

Exemple 1.1. *Un exemple spécifique mais cependant très commun, et très étudié est le modèle linéaire à effets mixtes, défini par exemple dans Pinheiro and Bates (2006) p. 58 :*

$$\begin{cases} y_i = X_i\beta + Z_i b_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{J_i}) \\ b_i \sim \mathcal{N}_p(0, \Gamma) \end{cases} \tag{1.3}$$

Dans cet exemple, le modèle peut se réécrire comme un modèle gaussien, non hiérarchique. Par indépendance des effets aléatoires et des résidus le modèle est équivalent à :

$$y_i = X_i\beta + \tilde{\varepsilon}_i, \quad \tilde{\varepsilon}_i \sim \mathcal{N}(0, Z_i \Gamma Z_i^T + \sigma^2 I_{J_i})$$

où A^T correspond à la transposée d'une matrice quelconque A et I_k est la matrice identité de taille $k \times k$.

Exemple 1.2. *Un exemple courant de modèle non linéaire à effets mixtes est le modèle de croissance logistique, par exemple présenté dans Pinheiro and Bates (2006) page 274. Ce modèle est défini de la*

manière suivante, pour $i = 1, ..., N$ et $j = 1, ..., J_i$:

$$\begin{cases} y_{ij} = \frac{\varphi_{i1}}{1+\exp\left\{-\frac{(t_j-\varphi_{i2})}{\varphi_{i3}}\right\}} + \varepsilon_{ij}, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I) \\ \\ \varphi_i = (\varphi_{i1}, \varphi_{i2}, \varphi_{i3})^T = \beta + b_i, \quad b_i \sim \mathcal{N}(0, \Gamma) \end{cases} \tag{1.4}$$

Ce modèle étant non linéaire en φ_{i2} et φ_{i3}, la vraisemblance marginale telle que définie en (1.2)*, n'est pas calculable explicitement. Ce modèle est utilisé dans Pinheiro and Bates (2006) pour modéliser un jeu de données qui contient des mesures longitudinales de la circonférence des troncs de cinq orangers, collectées à différents âges. De manière similaire à la modélisation mécaniste présentée en section* (1.1.1)*, ici les paramètres $(\varphi_i)_i$ ont un sens physique. Pour chaque arbre i, φ_{i1} représente la circonférence maximale atteignable par le tronc, φ_{i2} définit l'âge auquel le tronc atteint la moitié de sa circonférence maximale et φ_{i3} est un taux de croissance de la circonférence du tronc.*

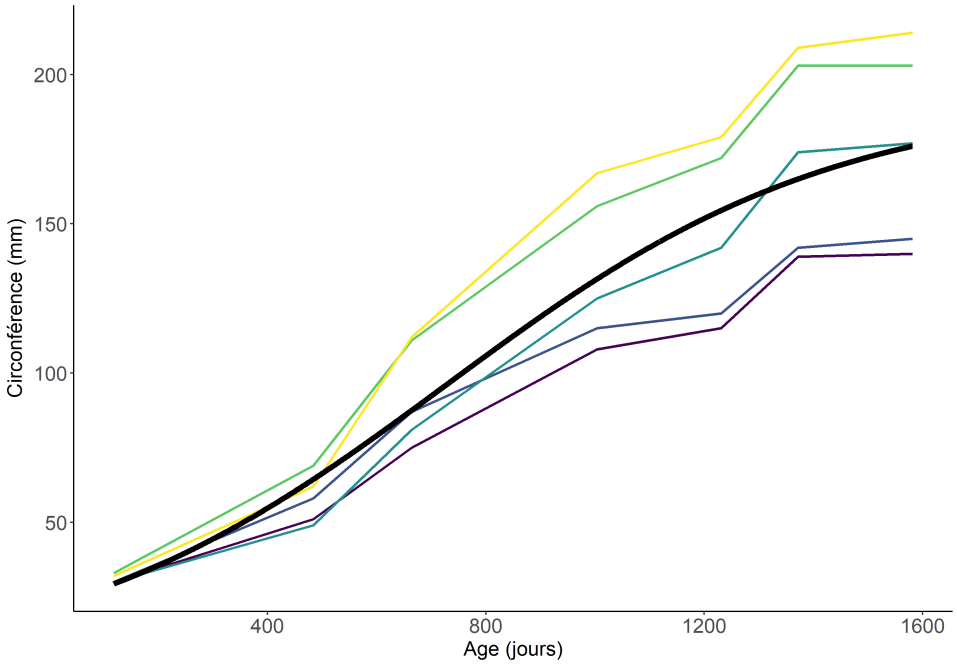


Figure 1.2 – Evolution au cours du temps (en jours) de la circonférence (en mm) du tronc de cinq orangers. La courbe noire épaisse correspond à une courbe de croissance logistique ajustée sur les données.

La figure (1.2) *représente graphiquement ces données. Chaque ligne de couleur correspond à un oranger. La courbe noire plus épaisse est une courbe de croissance logistique ajustée sur les données.*

1.2 . Tests des composantes de la variance dans les modèles à effets mixtes :

définitions et état de l'art

Cette section présente le test d'hypothèses portant sur les composantes de la variance dans les modèles à effets mixtes qui motive cette thèse. La section 1.2.1 présente le test et les statistiques de test basées sur la vraisemblance. La section 1.2.2 présente la problématique théorique de l'inférence sur la frontière de l'espace des paramètres, et établit une revue de la littérature des méthodes asymptotiques proposées pour tester la nullité de composantes de la variance dans les modèles à effets mixtes. La section 1.2.3 présente les limites de ces approches.

1.2.1 . Tests des composantes de la variance dans les modèles à effets mixtes

La section 1.1.1 a introduit la problématique qui a motivé ce travail de thèse : identifier parmi les différents paramètres du modèle ceux qui peuvent être considérés comme fixes. Pour illustrer graphiquement cette question sur la variabilité, reprenons l'exemple du modèle de croissance logistique (1.4). La figure 1.3 représente des données simulées selon le modèle de croissance logistique décrit dans (1.4), avec :

$$\varphi_i = \begin{pmatrix} \varphi_{i1} \\ \varphi_{i2} \\ \varphi_{i3} \end{pmatrix} \sim \mathcal{N}_3 \left(\beta, \begin{pmatrix} \gamma_1 & 0 & 0 \\ 0 & \gamma_2 & 0 \\ 0 & 0 & \gamma_3 \end{pmatrix} \right)$$

Comme expliqué dans l'exemple (1.2) sur l'étude de croissance des orangers, les paramètres individuels φ_{ik} ($k = 1, 2, 3$) ont un sens physique précis dans la modélisation d'un phénomène de croissance :

- φ_{i1} est la valeur asymptotique de la réponse y_ij lorsque j croît
- φ_{i2} correspond au temps auquel la réponse atteint la moitié de sa valeur asymptotique
- φ_{i3} est un paramètre d'échelle agissant comme un taux de croissance autour du point d'inflexion

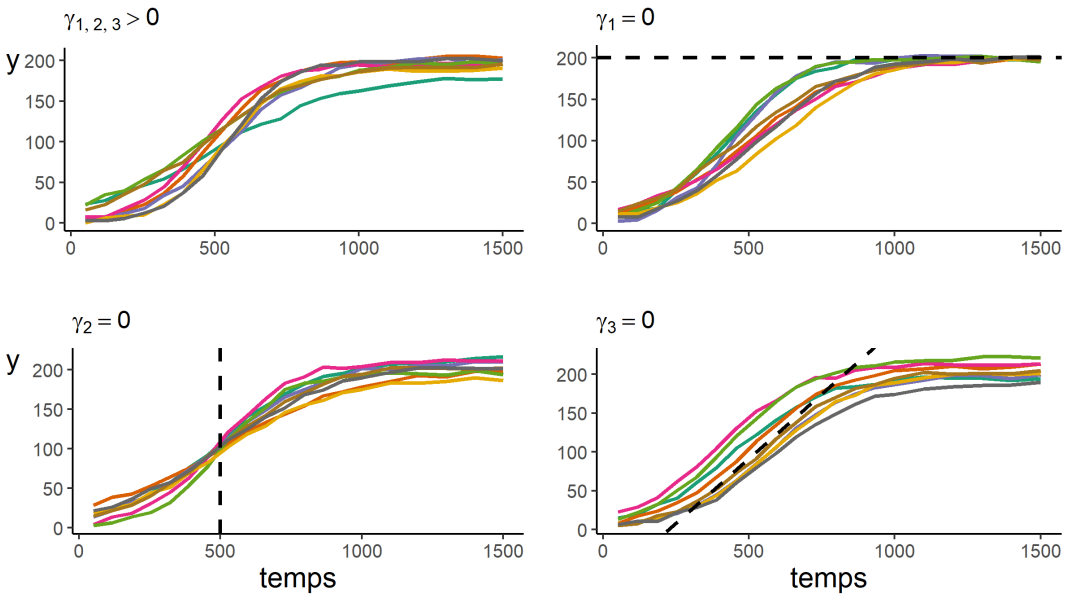


Figure 1.3 – Données simulées selon un modèle de croissance logistique, illustrant l'effet de l'absence de variabilité individuelle sur la valeur asymptotique φ_{i1} ($\gamma_1 = 0$), le temps de demi-vie φ_{i2} ($\gamma_2 = 0$) et le paramètre d'échelle φ_{i3} ($\gamma_3 = 0$)

La figure (1.3) représente quatre exemples de simulations de données longitudinale avec un modèle de croissance logistique. Cette figure illustre l'impact de la variabilité portée par chacun des paramètres du modèle. En haut à gauche, les données ont été simulées avec chacun des trois paramètres variables, puis les trois autres graphiques présentent des données simulées avec à chaque fois un paramètre non variable. Ces graphiques décrivent donc des exemples de données longitudinales avec :

- une valeur asymptotique commune à tous les individus ($\gamma_1 = 0$)
- un temps de demi-vie commun ($\gamma_2 = 0$) qui correspond à un point d'inflexion identique pour chacune des courbes
- un taux de croissance commun à chacun des individus ($\gamma_3 = 0$) qui conduit au parallélisme des tangentes au temps de demi-vie.

D'un point de vue statistique, la question d'identification des paramètres ne présentant pas de variabilité peut se poser sous la forme d'un test d'hypothèses sur les variances des effets aléatoires.

Le modèle considéré est celui défini en (1.1). Supposons que l'objectif soit de tester la nullité de r variances des effets aléatoires. Quitte à permuter les lignes, on peut supposer qu'il s'agit

des dernières r variances. Ainsi on souhaite tester si la matrice Γ est de la forme :

$$\Gamma = \left(\begin{array}{c|c} \Gamma_1 & \Gamma_{12} \\ \hline \Gamma_{12}^t & \Gamma_2 \end{array}\right) = \left(\begin{array}{c|c} \Gamma_1 & 0_{(p-r)\times r} \\ \hline 0_{r\times(p-r)} & 0_{r\times r} \end{array}\right)$$

où Γ_1 est une matrice symétrique semi-définie positive.

Le paramètre du modèle est $\theta = (\beta, \Gamma, \sigma^2)$. On chercher à tester :

$$H_0 \;\; \theta \in \Theta_0 \quad \text{contre} \quad H_1 \;\; \theta \in \Theta$$

où :

$$\Theta_0 = \{\beta \in \mathbb{R}^q, \Gamma_1 \in \mathbb{S}_{p-r}^+, \Gamma_2 = 0_{r\times r}, \Gamma_{12} = 0_{(p-r)\times r}, \sigma^2 \in \mathbb{R}_*^+\}$$

$$\Theta = \{\beta \in \mathbb{R}^q, \Gamma \in \mathbb{S}_p^+, \sigma^2 \in \mathbb{R}_*^+\}$$

Comme mentionné dans la section précédente, nous considérons l'estimation par maximum de vraisemblance. De plus, le test considéré est un test qui compare des modèles emboîtés ($\Theta_0 \subset \Theta$), les tests basés sur la vraisemblance tels que le test du rapport de vraisemblances, le test du score et le test de Wald sont donc des choix naturels vers lesquels s'orienter. Dans un premier temps nous rappelons la définition de ces statistiques de test.

Soit $y_{1:N} = (y_i)_{i=1,..,N}$ les observations, et $l(\theta; y_{1:N}) = \sum_{i=1}^N \log L(\theta; y_i)$ la log-vraisemblance d'un paramètre $\theta \in \Theta$. Supposons que $\theta \mapsto l(\theta; y_{1:N})$ soit 3 fois différentiable par rapport à θ et que les dérivées première et seconde soient majorées par des fonctions indépendantes de θ intégrables par rapport à la loi de $y_{1:N}$.

On définit tout d'abord les quantités nécessaires à l'étude de l'inférence par maximum de vraisemblance :

- le maximum de vraisemblance sous l'alternative, également appelé non restreint :

$$\hat{\theta}_N = \underset{\theta\in\Theta}{\arg\sup} \quad l(\theta; y_{1:N})$$

- le maximum de vraisemblance sous l'hypothèse nulle, également appelé restreint :

$$\tilde{\theta}_N = \underset{\theta\in\Theta_0}{\arg\sup} \quad l(\theta; y_{1:N})$$

- le score d'un paramètre $\theta \in \Theta$:

$$s_N(\theta) = \nabla_\theta l(\theta; y_{1:N})$$

- la matrice d'information de Fisher observée d'un paramètre $\theta \in \Theta$:

$$I_N(\theta) = -\nabla^2_\theta l(\theta; y_{1:N})$$

où ∇_θ correspond au gradient et ∇^2_θ à la hessienne, par rapport à θ.

Pour la présentation des différentes statistiques de test, on supposera que les quantités à inverser sont inversibles. La statistique du rapport de vraisemblance est définie comme :

$$\text{lrt}(y_{1:N}) = -2\left(l(\tilde{\theta}_N; y_{1:N}) - l(\hat{\theta}_N; y_{1:N})\right) \tag{1.5}$$

La statistique du score est définie comme :

$$S_N(y_{1:N}) = s_N(\tilde{\theta}_N)^T I_N(\tilde{\theta}_N)^{-1} s_N(\tilde{\theta}_N) \tag{1.6}$$

La statistique de Wald est définie comme :

$$W_N(y_{1:N}) = R\hat{\theta}_N \left[RI_N(\hat{\theta}_N)R^T\right]^{-1} (R\hat{\theta}_N)^T \tag{1.7}$$

où la matrice R est une matrice d'indicatrice telle que $R\theta = 0$ sous H_0.

Sous des hypothèses de régularité, les trois tests basés sur ces statistiques sont asymptotiquement équivalents, car ces trois statistiques ont la même loi asymptotique sous H_0. Dans la suite de ce manuscrit nous nous consacrerons à l'étude de la statistique du maximum de vraisemblance car c'est la seule qui ne requiert pas le calcul de l'information de Fisher. Ce point est expliquée en détail dans la section 1.2.3 et encore plus précisément dans le chapitre 2.

Le théorème de Wilks (Wilks, 1938) décrit la distribution asymptotique de la statistique du rapport de vraisemblance (et donc a fortiori la distribution asymptotique de la statistique de Wald et celle du score). Nous introduisons les hypothèses suivantes :

1. La vraie valeur du paramètre, que l'on note $\theta_0 \in \Theta_0$, appartient à l'intérieur de Θ
2. Le modèle est identifiable
3. $\theta \mapsto l(\theta; y_{1:N})$ est trois fois différentiable sur Θ et les trois premières dérivées sont dominées par une fonction dont l'espérance existe
4. La matrice d'information de Fisher $I(\theta_0) = \lim_{N\to+\infty} \frac{1}{N}\mathbb{E}_{\theta_0}\left[I_N(\theta_0)\right]$ est une matrice symétrique définie positive (et donc non singulière)

où $\mathbb{E}_{\theta_0}\left[h(y_{1:N})\right] = \int h(y_{1:N})L(\theta_0; y_{1:N})dy_{1:N}$ pour une fonction h mesurable

Sous ces hypothèses, $\text{lrt}(y_{1:N})$ converge en distribution vers une variable aléatoire qui suit une loi du Chi-deux, dont le nombre de degrés de liberté correspond à la dimension de l'espace complémentaire à Θ_0 dans Θ. Cependant dans le contexte de tests de composantes de la variance dans les modèles à effets mixtes, la première hypothèse n'est pas vérifiée. La section suivante présente la théorie qui s'applique dans ce cadre d'inférence sous contrainte.

1.2.2 . Test du rapport de vraisemblance asymptotique sous conditions non standards

Cette section vise à présenter la théorie asymptotique du rapport de vraisemblance lorsque la vraie valeur du paramètre est sur la frontière de l'espace. Pour plus de détail sur cette théorie, le lecteur est renvoyé à Chernoff (1954); Self and Liang (1987); Andrews (1999); Silvapulle and Sen (2011).

Sous les hypothèses de régularités requises par le théorème de Wilks, les résultats asymptotiques usuels en estimation par maximum de vraisemblance s'obtiennent par une approximation quadratique en θ de la vraisemblance autour de la vraie valeur du paramètre $\theta_0 \in \Theta_0$:

$$\begin{aligned} l(\theta; y_{1:N}) = l(\theta_0; y_{1:N}) + (\theta - \theta_0)^T s_N(\theta_0) \\ - \frac{1}{2}(\theta - \theta_0)^T I_N(\theta_0)(\theta - \theta_0) + R_N(\theta) \end{aligned} \tag{1.8}$$

où $R_N(\theta)$ est un reste qui tend vers 0 en probabilité lorsque $N \to +\infty$.
Si θ_0 est sur la frontière de l'espace des paramètres, la propriété de normalité asymptotique n'est plus vérifiée. Par exemple dans le cas du test $H_0 : \theta_0 = 0$ contre $H_1 : \theta_0 > 0$, la statistique $\sqrt{n}\hat{\theta} \geq 0$ est une variable aléatoire presque sûrement positive, et n'est donc pas asymptotiquement normale. Cependant, sous certaines conditions, le comportement de l'estimateur du maximum de vraisemblance et celui du rapport de vraisemblance peuvent être obtenus. Sous les hypothèses de régularité du théorème de Wilks, le score $\frac{1}{\sqrt{N}} s_N(\theta_0)$, étant une somme de variables aléatoires indépendantes et centrées, converge en distribution vers une loi normale centrée de variance asymptotique $I(\theta_0)^{-1}$. Une hypothèse supplémentaire cruciale porte sur la forme de l'espace des paramètres au voisinage de θ_0. Il est requis que, près de θ_0, l'espace des paramètres se comporte comme un cône. Une définition formelle de cette propriété a été donnée par Chernoff en 1954 dans Chernoff (1954) :

Un ensemble $\Theta \subset \mathbb{R}^p$ est dit Chernoff régulier s'il peut être approximé en θ_0 par C_{θ_0}, un cône ayant pour sommet θ_0, si :

$$\inf_{x \in C_{\theta_0}} \|x - y\| = o_{y \to \theta_0}(\|y - \theta_0\|), \qquad y \in \Theta$$

$$\inf_{y \in \Theta} \|x - y\| = o_{x \to \theta_0}(\|x - \theta_0\|), \qquad x \in C_{\theta_0}$$

Tout d'abord, sous les conditions énoncées ci-dessus, l'estimateur du maximum de vraisemblance $\hat{\theta}_N$ est $\sqrt{N}$-consistant, et est asymptotiquement équivalent au maximum sur $\theta \in C_{\theta_0}$ de la forme quadratique suivante :

$$\left(\frac{1}{\sqrt{N}} I^{-1}(\theta_0) s_N(\theta_0) - \sqrt{N}(\theta - \theta_0)\right)^T I(\theta_0) \left(\frac{1}{\sqrt{N}} I^{-1}(\theta_0) s_N(\theta_0) - \sqrt{N}(\theta - \theta_0)\right) \tag{1.9}$$

En remarquant que $\frac{1}{\sqrt{N}} I^{-1}(\theta_0) s_N(\theta_0)$ a pour distribution asymptotique $\mathcal{N}\left(0, I^{-1}(\theta_0)\right)$, la loi limite de $\sqrt{N}(\theta - \theta_0)$ est la même que celle d'un estimateur du maximum de vraisemblance basé sur une seule observation d'une variable aléatoire distribuée selon une $\mathcal{N}\left(0, I^{-1}(\theta_0)\right)$ restreinte à se situer dans $(C_{\theta_0} - \theta_0)$.

Grâce à cette forme explicite, nous comprenons que la distribution asymptotique est étroitement liée à la forme du cône d'approximation de l'espace des paramètres.
Donnons maintenant deux exemples simples pour illustrer ces résultats (pour plus de détails le lecteur est renvoyé à Self and Liang (1987) :
Si $\theta_0 = (\theta_{01}, ..., \theta_{0p})^T$ est un point intérieur de $\Theta \subset \mathbb{R}^p$, alors le cône d'approximation à Θ en θ_0 est l'espace $\mathbb{R}^p$ tout entier, et nous retrouvons la normalité asymptotique de l'estimateur du maximum de vraisemblance.
Maintenant, si nous notons $Z = (Z_1, ..., Z_p)^T$ une variable aléatoire gaussienne centrée avec une matrice de covariance $I(\theta_0)^{-1}$, et que nous avons $\Theta = [0, +\infty[\times \mathbb{R}^{p-1}$ avec $\theta_{01} = 0$, alors :

$$\lim_{N \to +\infty} \sqrt{N}(\hat{\theta}_N - \theta_0) = Z \times \mathbb{1}(Z_1 > 0) + \text{proj}_{E_1}(Z) \times \mathbb{1}(Z_1 \leq 0)$$

où proj_{E_1} est le projeté orthogonal sur l'espace $\{(x_1, ..., x_p) \in \mathbb{R}^p | x_1 \leq 0\}$. Dans ce cas, nous observons que $\lim_{N \to +\infty} \sqrt{N}(\hat{\theta}_N - \theta_0)$ a une distribution normale tronquée. Notons qu'ici un seul paramètre est sur la frontière. Si un deuxième l'était aussi, alors il y aurait quatre cas à considérer : $Z_1, Z_2 > 0$, $Z_1 > 0 \ \& \ Z_2 \leq 0$, $Z_1, Z_2 \leq 0$, et $Z_1 \leq 0 \ \& \ Z_2 > 0$. Plus il y a de paramètres sur la frontière, plus la distribution asymptotique est complexe, car le nombre de termes dans l'équation précédente croît de manière exponentielle.
Cette expression permet également de déterminer la distribution asymptotique de la statistique du rapport de vraisemblance. Notons $\tilde{C}_{\theta_0}$ le cône d'approximation de Θ_0 en θ_0. Alors la statistique du rapport de vraisemblance a la forme asymptotique suivante :

$$\begin{aligned} \text{lrt}(y_{1:N}) \underset{N \to \infty}{\overset{d}{\to}} & \inf_{\theta \in \tilde{C}_{\theta_0}} \|Z - (\theta - \theta_0)\|_{I(\theta_0)} \\ & - \inf_{\theta \in C_{\theta_0}} \|Z - (\theta - \theta_0)\|_{I(\theta_0)} \end{aligned} \tag{1.10}$$

où $Z = (Z_1, ..., Z_p)^T \sim \mathcal{N}_p(0, I^{-1}(\theta_0))$ et $\|x\|_A = \sqrt{x^T A x}$, pour une matrice symétrique définie positive A.
La preuve découle directement de l'approximation quadratique de la vraisemblance (1.8).

Considérons maintenant les deux exemples décrits ci-dessus :
Si θ_0 est un point intérieur de Θ et Θ_0 et que nous voulons tester que $\theta_{01} = 0$, alors $C_{\theta_0} = \mathbb{R}^p$ et

$\tilde{C}_{\theta_0} = \{0\} \times \mathbb{R}^{p-1}$ et

$$\lim_{N\to\infty} \text{lrt}(y_{1:N}) = Z_1^T I(\theta_0) Z_1 \sim \chi_1^2$$

Maintenant, si nous faisons le même test qu'avant mais tel que θ_{01} est contraint à être positif, alors :

$$\begin{aligned}\text{lrt}(y_{1:N}) &\underset{N\to\infty}{\overset{d}{\to}} Z_1^2 I_{11} - \inf_{\theta_{01}\in[0,+\infty[} (Z_1 - \theta_{01})^2 I_{11} \\ &= Z_1^2 I_{11} \mathbb{1}(Z_1 > 0)\end{aligned}$$

Où $I_{11} = (I(\theta_0))_{11}$ est l'inverse de la variance de Z_1, donc $\sqrt{I_{11}} Z_1$ est une variable gaussienne standardisée.
Ainsi, la distribution asymptotique de $\text{lrt}(y_{1:N})$ est un mélange d'une distribution χ^2 avec un degré de liberté et d'une distribution Dirac en 0 avec des poids égaux à $\frac{1}{2}$.

Dans le cadre des modèles à effets mixtes, différents auteurs se sont intéressés aux tests de nullité des variances des effets aléatoires, en utilisant le rapport de vraisemblance ou celui du score. Nous pouvons citer Stram and Lee (1994); Zhang and Lin (2008); Molenberghs and Verbeke (2007); Qu et al. (2013) qui proposent des tests des composantes de la variance dans les modèles linéaires (ou linéaires généralisés) dans des cadres spécifiques. Plus récemment, Baey et al. (2019) se sont intéressés à ces tests, quelle que soit la structure de corrélation des effets aléatoires, et ont dérivé la distribution asymptotique de la statistique du rapport de vraisemblance comme étant une distribution du *chi-bar square*, dont les poids dépendent directement de la structure de covariance des effets aléatoires. Un package *R* (Baey and Kuhn, 2020) permet d'appliquer ce test asymptotique.

1.2.3 . Limites des tests asymptotiques

Cette section présente les limites des tests asymptotiques, qui peuvent parfois, rendre ceux–ci inapplicables. Les deux limites principales sont les suivantes :

- Si des variances non testées sont nulles, les tests asymptotiques présentés dans la section précédente sont inapplicables car le cône tangent à Θ_0 en θ_0 est inconnu, et négliger cet aspect conduit à l'application d'un test potentiellement incorrect. Les variances non testées égales à zéro seront appelées paramètres de nuisance.
- Dans les modèles non linéaires, des problèmes de singularité de l'information de Fisher modifient la distribution asymptotique de la statistique du rapport de vraisemblance, et rendent donc le test difficile, ou impossible à appliquer. Les statistiques telles que le score qui mettent en jeu l'inverse de l'information de Fisher sont donc également inutilisables.

Cette partie illustre en détail ces deux aspects.

Présence de paramètres de nuisance

Considérons ici le modèle linéaire à effets mixtes présenté dans l'exemple (1.3) :

$$\begin{cases} y_i = X_i\beta + Z_i b_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_J) \\ b_i \sim \mathcal{N}(0, \Gamma) \end{cases}$$

avec une matrice de covariance Γ diagonale :

$$\Gamma = \begin{pmatrix} \gamma_1 & 0 & \cdots & 0 \\ 0 & \gamma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_p \end{pmatrix}$$

Les paramètres du modèle sont $\beta \in \mathbb{R}^{d_\beta}$, $\gamma = (\gamma_1, ..., \gamma_p)^T \in \mathbb{R}_+^p$ et $\sigma^2 \in \mathbb{R}_+^*$, et donc l'espace du paramètre $\theta = (\beta, \gamma, \sigma^2)^T$ est :

$$\Theta = \mathbb{R}^{d_\beta} \times \mathbb{R}_+^p \times \mathbb{R}_+^*$$

Supposons que l'on souhaite tester la nullité des r dernières variances des effets aléatoires. L'espace restreint est donc :

$$\Theta_0 = \mathbb{R}^{d_\beta} \times \mathbb{R}_+^{p-r} \times \{0\}^r \times \mathbb{R}_+^*$$

Soit $\theta_0 \in \Theta_0$ le vrai paramètre. Supposons également qu'en plus des dernières r variances nulles, m variances non testées soient également nulles. On supposera sans perdre de généralité que les m variances non testées égales à zéro sont ordonnées :

$$\begin{aligned} \theta_0 &= (\beta_0, \gamma_0, \sigma_0^2)^T \\ \gamma_0 &= (\gamma_{0,1}, ...\gamma_{0,p-r-m}, \underbrace{0, ..., 0}_{m}, \underbrace{0, ..., 0}_{r})^T \end{aligned}$$

Le cône tangent à Θ en θ_0 est :

$$C_{\theta_0} = \mathbb{R}^{d_\beta} \times \mathbb{R}^{p-r-m} \times \mathbb{R}_+^m \times \mathbb{R}_+^r \times \mathbb{R} \tag{1.11}$$

En pratique, les m variances non testées égales à 0 sont à des localisations inconnues et en nombre inconnu. Ainsi, en présence de tels paramètres de nuisance, le cône tangent est inconnu et donc la distribution asymptotique (1.10) également. Ainsi en présence de variances non testées égales à 0, le test asymptotique correct de Baey et al. (2019) n'est pas réalisable.

Singularité de la matrice d'Information de Fisher

La vraisemblance (1.2) du modèle (2.1) n'est pas définie lorsque une variance est nulle. Soit

$i = 1, ..., N$, $j = 1, ..., J$, considérons le modèle simplifié suivant :

$$\begin{cases} y_{ij} & = g(x_j, b_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1) \\ b_i & \sim \mathcal{N}(0, \gamma^2) \end{cases}$$

où g est une fonction non linéaire de b_i, connue.

La vraisemblance d'un individu i, pour un paramètre $\gamma^2 > 0$ est définie par :

$$L(\gamma^2; y_i) = \int_{\mathbb{R}} \sqrt{2\pi}^{-J} \exp\left\{ \frac{-\sum_{j=1}^{J} (y_{ij} - g(x_j, b))^2}{2} \right\} \sqrt{2\pi\gamma^2}^{-1} \exp\left\{ \frac{-b^2}{2\gamma^2} \right\} db$$

La vraisemblance n'est pas définie pour $\gamma^2 = 0$ et converge vers 0 quand γ^2 tend vers 0. Le modèle ne convient donc pas à la problématique de test de nullité de variance étudiée.

Un modèle alternatif est le suivant :

$$\begin{cases} y_{ij} & = g(x_j, \lambda\xi_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1) \\ \xi_i & \sim \mathcal{N}(0,1) \end{cases} \tag{1.12}$$

où λ est un paramètre d'échelle positif.

En utilisant cette nouvelle paramétrisation, la vraisemblance est correctement définie en $\lambda = 0$. On retrouve cette paramétrisation dans Ekvall and Bottai (2021).

Cependant en $\lambda = 0$, le score est nul :

$$\begin{aligned} \frac{\partial \log L(\lambda, y_i)}{\partial \lambda}|_{\lambda=0} &= L(\lambda, y_i)^{-1} \int -\sum_{j=1}^{J} \xi \frac{\partial g}{\partial \lambda}(x_j, 0)(y_{ij} - g(x_j, 0)) \sqrt{2\pi}^{-1} \exp\left\{ \frac{-\xi^2}{2} \right\} d\xi \\ &= -L(\lambda, y_i)^{-1} \sum_{j=1}^{J} \frac{\partial g}{\partial \lambda}(x_j, 0)(y_{ij} - g(x_j, 0)) \int \xi \sqrt{2\pi}^{-1} \exp\left\{ \frac{-\xi^2}{2} \right\} d\xi \\ &= 0 \end{aligned}$$

La dernière égalité découle du fait que l'on retrouve l'espérance d'une variable aléatoire Gaussienne centrée. Le score étant systématiquement nul pour $\lambda = 0$, la matrice d'Information de Fisher est également nulle car définie comme la variance du score. Cette problématique empêche l'approximation quadratique permettant d'obtenir la distribution asymptotique de la statistique du rapport de vraisemblance, et empêche donc l'application de la théorie asymptotique jusqu'ici présentée (Self and Liang, 1987; Andrews, 1999; Silvapulle and Sen, 2005; Baey et al., 2019).

Parallèlement, des approches non asymptotiques ont été proposées, basées sur des méthodes par rééchantillonnage, qui présentent deux avantages :

- ces méthodes ont souvent de meilleurs performances lorsque le nombre de données disponibles est petit,
- ces méthodes permettent de ne pas avoir à déterminer la distribution asymptotique, potentiellement complexe, de la statistique d'intérêt.

Sinha (2009) a proposé une approche boostrap d'un test du score, pour des modèles linéaires ou linéaires généralisés, mais qui requiert la non singularité de la matrice d'information de Fisher, et ne prend pas en compte les paramètres de nuisance. Drikvandi et al. (2013a) propose une approche de tests par permutations qui présentent de très bons résultats mais qui se limite aux modèles linéaires. Enfin Crainiceanu and Ruppert (2004) ont déterminé la distribution exacte de la statistique du rapport de vraisemblance dans le modèle linéaire à un seul effet aléatoire.

Pour conclure, le développement d'un test d'hypothèses sur les composantes de la variance dans les modèles à effets mixtes, s'appliquant indifféremment aux modèles linéaires et non linéaires, et prenant en compte les paramètres de nuisance, semble d'un grand intérêt.

1.3 . Quelques éléments sur les tests par Bootstrap

Cette section introduit le Bootstrap, et plus spécifiquement, les tests Bootstrap.

Introduit pour la première fois par Efron (Efron, 1992), le Bootstrap a suscité beaucoup d'intérêt tant du point de vue théorique (Beran, 1997; Bickel and Freedman, 1981) que pratique (Davidson and MacKinnon, 2006).

Le principe du Bootstrap est le suivant. Soit $y_{1:N} = (y_1, ..., y_N)^T$ un échantillon issu d'une distribution de probabilité P inconnue. Soit T une statistique d'intérêt. Afin d'approximer la véritable distribution de $T(y_{1:N})$, le Bootstrap consiste à considérer une distribution $\hat{P}$ approchant P, et de simuler de nouvelles données $y^*_{1:N} \sim \hat{P}$, afin d'obtenir de nouvelles réalisations de la statistique Bootstrap $T(y^*_{1:N})$.

Le choix de $\hat{P}$ est très important car sa proximité avec P déterminera la qualité de l'approximation de la distribution de la statistique d'intérêt. Le choix le plus commun est, dans le contexte indépendant et identiquement distribué (i.i.d.), la distribution empirique de l'échantillon :

$$\hat{P} = \sum_{i=1}^{N} \frac{1}{N} \delta_{y_i}$$

où δ_x représente la mesure Dirac au point x. Ce choix de $\hat{P}$ correspond au Bootstrap non paramétrique. Échantillonner selon cette distribution revient à tirer aléatoirement avec remise dans le jeu de données initial.

Une autre approche possible est le Bootstrap paramétrique. Si la véritable distribution P est supposée appartenir à une famille paramétrique $\{P_\theta; \quad \theta \in \Theta\}$, $\hat{P}$ peut être choisie comme $P_{\hat{\theta}}$ avec $\hat{\theta}$ un estimateur du vrai paramètre θ_0 tel que $P = P_{\theta_0}$.

Le Bootstrap non paramétrique est souvent préféré car il ne requiert aucune hypothèse paramétrique, et est donc plus flexible. Cependant, il n'est pas toujours utilisable pour des tests d'hypothèses. En effet, pour réaliser un test Bootstrap, la distribution de $T(y^*_{1:N})$ doit vérifier l'hypothèse nulle (Hall and Wilson, 1991). Dans la plupart des modèles complexes cette contrainte ne peut pas être vérifiée. Le Bootstrap paramétrique ne rencontre pas la même contrainte, car il suffit de choisir $\hat{\theta}$ dans l'espace des paramètres contraints pour que cette hypothèse soit vérifiée.

La consistance du Bootstrap (paramétrique ou non) a été étudiée dans la littérature (Bickel and Freedman, 1981; Beran, 1997; Hall, 2013). La consistance mentionnée ici est celle de la convergence faible en loi de la statistique Bootstrap, vers la même distribution asymptotique que celle la statistique d'intérêt. De manière plus formelle, on cherche à choisir $\hat{\theta}$ de telle sorte que pour tout $t \in \mathbb{R}$:

$$P_{\hat{\theta}}\left(T(y^*_{1:N}) \leq t\right) - P\left(T(y_{1:N}) \leq t\right) \underset{N\to+\infty}{\xrightarrow{p}} 0 \tag{1.13}$$

où la convergence est en probabilité, car $\hat{\theta}$ dépend des observations $y_{1:N}$ et donc $P_{\hat{\theta}}\left(T(y^*_{1:N}) \leq t\right)$ est une variable aléatoire.

Le problème qui se pose dans notre contexte est que, comme discuté dans Beran (1997) et souligné dans Andrews (2000), il est connu que le Bootstrap est inconsistant lorsque la valeur réelle du paramètre est sur la frontière de l'espace des paramètres. Pour illustrer ce point, nous considérons les deux exemples suivants.

Estimation sous contrainte de la moyenne dans un échantillon Gaussien i.i.d : Nous présentons un premier exemple utilisé dans Andrews (2000) :

Soit $Y_1, ..., Y_N$ un échantillon i.i.d. de variables aléatoires réelles suivant une loi $\mathcal{N}(\mu, 1)$ où $\mu \in \mathbb{R}^+$. Considérons comme statistique d'intérêt la moyenne tronquée :

$$\hat{\mu}_N = \max\left(0, \bar{Y}_N\right)$$

où $\bar{Y}_N = \frac{1}{N}\sum_{i=1}^N Y_i$.

Supposons que la vraie valeur du paramètre μ_0 soit nulle. Par continuité de $x \mapsto \max(0, x)$

et en appliquant le théorème de la limite centrale :

$$\sqrt{N}(\hat{\mu}_N - \mu_0) \xrightarrow{d} \max(0, Z), \quad Z \sim \mathcal{N}(0,1)$$

Soit $Y^* = \{X_i^*; i = 1, ..., N\}$ un échantillon simulé de manière i.i.d. (conditionnellement à $\hat{\mu}_N$) selon la loi $\mathcal{N}(\hat{\mu}_N, 1)$.

Soit $\mu_N^* = \max\left(0, \bar{Y}_N^*\right)$ la statistique Bootstrap correspondant à l'échantillon Bootstrap Y^*. La consistance Bootstrap telle que définie en (1.13) est vérifiée si :

$$\sqrt{N}(\mu_N^* - \hat{\mu}_N) \xrightarrow{d} \max(0, Z), \quad Z \sim \mathcal{N}(0,1)$$

conditionnellement à $X_1, .., X_N$, avec probabilité 1.

$$\begin{aligned}\sqrt{N}(\mu_N^* - \hat{\mu}_N) &= \sqrt{N}\max\left(0, \bar{Y}_N^*\right) - \sqrt{N}\max\left(0, \bar{Y}_N\right) \\ &= \max\left(0, \sqrt{N}(\bar{Y}_N^* - \bar{Y}_N) + \sqrt{N}\bar{Y}_N\right) - \max\left(0, \sqrt{N}\bar{Y}_N\right)\end{aligned}$$

Soit $c > 0$, par la loi des logarithmes itérés, $P(\limsup_N \sqrt{N}\bar{Y}_N \geq c) = 1$.

Soit $A_c = \{\limsup_N \sqrt{N}\bar{Y}_N \geq c\}$.

Conditionnellement à A_c, on peut extraire une sous-suite $\{\phi(N); N \in \mathbb{N}\}$ telle que $\sqrt{\phi(N)}\bar{Y}_{\phi(N)} \geq c$, pour tout N. Ainsi :

$$\begin{aligned}\sqrt{\phi(N)}(\mu_{\phi(N)}^* - \hat{\mu}_{\phi(N)}) &\leq \max\left(0, \sqrt{\phi(N)}(\bar{Y}_{\phi(N)}^* - \bar{Y}_{\phi(N)}) + \sqrt{\phi(N)}\bar{Y}_{\phi(N)}\right) - \sqrt{\phi(N)}\bar{Y}_{\phi(N)} \\ &= \max\left(-\sqrt{\phi(N)}\bar{Y}_{\phi(N)}, \sqrt{\phi(N)}(\bar{Y}_{\phi(N)}^* - \bar{Y}_{\phi(N)})\right) \\ &\leq \max\left(-c, \sqrt{\phi(N)}(\bar{Y}_{\phi(N)}^* - \bar{Y}_{\phi(N)})\right) \\ &= \max(-c, Z) \quad \text{avec} \quad Z \sim \mathcal{N}(0,1)\end{aligned}$$

où la dernière ligne est vraie conditionnellement à $X_1, ..., X_N$ car :

$$\sqrt{\phi(N)}\bar{Y}_{\phi(N)}^* \sim \mathcal{N}\left(\max(0, \sqrt{\phi(N)}\bar{Y}_{\phi(N)}), 1\right)$$

et $\max\left\{0, \sqrt{\phi(N)}\bar{Y}_{\phi(N)}\right\} = \sqrt{\phi(N)}\bar{Y}_{\phi(N)}$ conditionnellement à A_c.

Finalement, $\max(-c, Z) < \max(0, Z)$ avec probabilité $P(Z \leq 0) > 0$.

Ainsi, conditionnellement à A_c, la suite $\sqrt{\phi(N)}(\mu_{\phi(N)}^* - \hat{\mu}_{\phi(N)})$ ne converge pas en distribution vers $\max(0, Z)$, donc $\sqrt{N}(\mu_N^* - \hat{\mu}_N)$ non plus, et ce résultat est vrai avec probabilité $P(A_c) = 1$, ce qui montre l'inconsistance du Bootstrap dans cet exemple.

Statistique du rapport de vraisemblance sous contraintes

Cette partie reprend les lignes de Cavaliere et al. (2020). Le cadre ici est celui de l'étude d'un maximum de vraisemblance contraint, où l'approximation quadratique (1.8) est vérifiée, permettant de déterminer la distribution asymptotique de la statistique du rapport de vraisemblance (1.10).

Soient $\hat{\theta}_N$ l'estimateur du maximum de vraisemblance non restreint (sous H_1), $y^*_{1:N}$ les observations Bootstrap de loi paramétrée par $\hat{\theta}_N$, et $\mathrm{lrt}(y^*_{1:N})$ la statistique Bootstrap correspondante. Afin de déterminer la distribution asymptotique de cette statistique, on procède à la même approximation quadratique de la vraisemblance autour du paramètre $\hat{\theta}_N$:

$$\begin{aligned} l(\theta; y^*_{1:N}) = l(\hat{\theta}_N; y^*_{1:N}) + (\theta - \hat{\theta}_N)^T s^*_N(\hat{\theta}_N) \\ - \frac{1}{2}(\theta - \hat{\theta}_N)^T I^*_N(\hat{\theta}_N)(\theta - \hat{\theta}_N) + R^*_N(\theta) \end{aligned} \tag{1.14}$$

où s^*_N, I^*_N et R^*_N correspondent aux versions Bootstrap des objets définis dans l'approximation quadratique (1.8), mais dépendants des observations Bootstrap $y^*_{1:N}$.

Après quelques manipulations, on obtient :

$$l(\theta; y^*_{1:N}) - l(\hat{\theta}_N; y^*_{1:N}) = \left(U^*_N(\hat{\theta}_N) - \sqrt{N}(\theta - \hat{\theta}_N)\right)^T I^*(\hat{\theta}_N) \left(U^*_N(\hat{\theta}_N) - \sqrt{N}(\theta - \hat{\theta}_N)\right) + \tilde{R}^*_N(\theta)$$

où $U^*_N(\hat{\theta}_N) = \frac{1}{\sqrt{N}} I^{*-1}(\hat{\theta}_N) s^*_N(\hat{\theta}_N)$ et $\tilde{R}^*_N(\theta)$ contient à la fois les termes négligeables de l'expansion et ceux ne dépendant pas de θ.

Sous hypothèses de régularités du modèle, $U^*_N(\hat{\theta}_N) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta_0))$. En appliquant les mêmes outils que pour obtenir la distribution asymptotique de l'estimateur du maximum de vraisemblance (1.9), l'estimateur du maximum de vraisemblance Bootstrap non restreint est asymptotiquement équivalent au maximum sur le cône $C_{\hat{\theta}_N}$ de :

$$\left(Z - \sqrt{N}(\theta - \hat{\theta}_N)\right)^T I^*(\hat{\theta}_N) \left(Z - \sqrt{N}(\theta - \hat{\theta}_N)\right) \quad \text{où} \quad Z \sim \mathcal{N}(0, I^{-1}(\theta_0))$$

En appliquant le même développement pour l'estimateur restreint (sous H_0), et en suivant le même raisonnement que pour obtenir l'expression (1.10), on obtient un équivalent asymptotique de la statistique du rapport de vraisemblance Bootstrap $\mathrm{lrt}(y^*_{1:N})$:

$$\begin{aligned} \mathrm{lrt}(y^*_{1:N}) = \inf_{\theta \in \hat{C}_{\hat{\theta}_N}} \|Z - (\theta - \hat{\theta}_N)\|_{I(\theta_0)} \\ - \inf_{\theta \in C_{\hat{\theta}_N}} \|Z - (\theta - \hat{\theta}_N)\|_{I(\theta_0)} + o_{p^*}(1) \end{aligned}$$

En remarquant que si $\theta \in C_{\hat{\theta}_N}$ alors $\theta - \hat{\theta}_N + \theta_0 \in C_{\theta_0}$ et en notant V la variable aléatoire telle que $\sqrt{N}(\hat{\theta}_n - \theta_0) \xrightarrow{d} V$ et V_0 sa projection sur Θ_0, on obtient finalement que conditionnellement

à $y_{1:N}$:

$$\begin{aligned} \text{lrt}(y^*_{1:N}) \xrightarrow[N\to+\infty]{d} & \inf_{\theta\in\tilde{C}_{\theta_0}} \|Z+V_0-(\theta-\theta_0)\|_{I(\theta_0)} \\ & - \inf_{\theta\in C_{\theta_0}} \|Z+V-(\theta-\theta_0)\|_{I(\theta_0)} \end{aligned} \tag{1.15}$$

Remarque 1.2. *Pour un ensemble $K \subset \mathbb{R}^d$, et un point $x \in \mathbb{R}^d$, on note $x + K$ l'ensemble $\{x + y; \quad y \in K\}$.*

Si θ_0 est un point intérieur de Θ alors $C_{\theta_0} - V = \mathbb{R}^{dim(\Theta)} = C_{\theta_0}$ (idem pour $\tilde{C}_{\theta_0} - V_0$) et finalement le Bootstrap est consistant. Cependant, dans le cadre des modèles à effets mixtes, si des variances sont égales à 0, cette égalité ne tient plus. En reprenant l'exemple de la section 1.2.3 on a :

$$C_{\theta_0} = \mathbb{R}^{d_\beta} \times \mathbb{R}^{p-r-m} \times \mathbb{R}^m_+ \times \mathbb{R}^r_+ \times \mathbb{R}$$

En décomposant $V = (V_\beta, V_1, V_{nuis}, V_{H_0}, V_{\sigma^2})$:

$$C_{\theta_0} - V = \mathbb{R}^{d_\beta} \times \mathbb{R}^{p-r-m} \times \{\mathbb{R}^m_+ - V_{nuis}\} \times \{\mathbb{R}^r_+ - V_{H_0}\} \times \mathbb{R}$$

Si les variables aléatoires V_{nuis} et V_{H_0} ne sont pas presque sûrement nulles, $C_{\Theta_0} - V \neq C_{\Theta_0}$ et le Bootstrap est inconsistant.

Dans Andrews (2000), deux approches sont proposées pour rétablir la consistance du Bootstrap.

La première est le "*m out of* N" Bootstrap, qui consiste à sous-échantillonner les données Bootstrap, et à considérer du rééchantillonnage de taille $m < N$. Avec un choix théorique de m tel que $m \to +\infty$ et $\frac{m}{N} \to 0$, on peut rétablir la consistance de la procédure.

La seconde approche, celle considérée dans le chapitre 2, consiste à seuiller le paramètre Bootstrap. Cette idée est également considérée dans Cavaliere et al. (2020), où cette approche est comparée au *"m out of N"* Bootstrap. Dans leur étude de simulation, le seuillage présentait de meilleures performances. Ici n'est présentée que la problématique de l'estimation contrainte, cependant la singularité de la matrice d'Information de Fisher présentée dans la section 1.2.3 est une autre source d'inconsistance du Bootstrap.

1.4 . Calcul de ratios de constantes de normalisation

L'objectif de cette section vise à introduire le chapitre (3.3) qui traite la question du calcul de ratios de constantes de normalisation.
Le test présenté dans la section 1.2.1 de ce manuscrit est un test du rapport de vraisemblance.

La définition de la statistique de test du rapport de vraisemblance est définie en 1.5. Comme introduit dans la section 1.1.2, dans les modèles non linéaires à effets mixtes, la vraisemblance (1.2) est définie comme une intégrale par rapport aux variables latentes. La statistique du rapport de vraisemblance peut donc s'écrire :

$$LRT(y_{1:N}) = -2\sum_{i=1}^{N}\log\left(\frac{\int f_i(y_i,b_i;\tilde{\theta}_N)db_i}{\int f_i(y_i,b_i;\hat{\theta}_N)db_i}\right) \tag{1.16}$$

Le calcul de cette statistique requiert le calcul des ratios $\frac{\int f_i(y_i,b_i;\tilde{\theta}_N)db_i}{\int f_i(y_i,b_i;\hat{\theta}_N)db_i}$. L'approche naturelle consiste à calculer séparément les vraisemblances marginales au numérateur et au dénominateur, en utilisant par exemple les outils à disposition dans les packages R. Dans *nlme* les approches proposées sont basées sur des méthodes de quadrature gaussienne ou d'approximation de Laplace. Le lecteur est renvoyé à Pinheiro and Bates (2006) chapitre 7 pour plus de détails. Parmi les limites de ces méthodes, la principale est l'absence de résultats théoriques. Le package *saemix* présenté dans Comets et al. (2017) propose également une approche par quadrature, mais aussi une approche stochastique d'importance sampling. L'importance sampling (IS) est une méthode de Monte-Carlo permettant d'estimer une espérance par rapport à une loi inconnue, en introduisant une loi de proposition connue. Pour un analyse approfondie le lecteur est renvoyé à Robert and Casella (1999). Soit π une densité de probabilité connue, strictement positive sur $\mathbb{R}^p$. L'estimateur par IS de la vraisemblance marginale individuelle $L(\theta;y_i)$ est basé sur l'identité suivante :

$$L(\theta;y_i) = \int f_i(y_i,b_i;\theta)db_i = \mathbb{E}_\pi\left[\frac{f_i(y_i,Z;\theta)}{\pi(Z)}\right] \tag{1.17}$$

où l'espérance est prise par rapport à la variable aléatoire Z suivant une loi à densité π. Dans la suite nous ferons l'amalgame entre la loi de probabilité et la densité de probabilité correspondante. Ainsi nous écrirons $Z \sim \pi$. L'estimateur naturel qui découle de cette identité est celui de la moyenne empirique à partir d'un échantillon i.i.d. de loi π. Soit $Z_1,...,Z_K \overset{i.i.d.}{\sim} \pi$ l'estimateur IS de $L(\theta;y_i)$ est défini par :

$$\hat{c}_K^{IS,\pi} = \frac{1}{K}\sum_{k=1}^{K}\frac{f_i(y_i,Z_k;\theta)}{\pi(Z_k)} \tag{1.18}$$

Cet estimateur est sans biais, fortement consistant (car $\mathbb{E}_\pi\left[\frac{f_i(y_i,Z;\theta)}{\pi(Z)}\right] < +\infty$), et asymptotiquement normal (sous l'hypothèse $\mathbb{E}_\pi\left[\left|\frac{f_i(y_i,Z;\theta)}{\pi(Z)}\right|^2\right] < +\infty$) :

$$\sqrt{K}\left(\hat{c}_K^{IS,\pi} - L(\theta;y_i)\right) \underset{K\to+\infty}{\xrightarrow{d}} \mathcal{N}\left(0,V_{IS}^{\pi}\right)$$

où

$$V_{IS}^{\pi} = \mathbb{E}_{\pi}\left[\left(\frac{f_i(y_i, Z; \theta)}{\pi(Z)}\right)^2\right] - L(\theta; y_i)^2$$

Cette expression permet de choisir la densité de proposition π comme étant celle minimisant la variance asymptotique de l'estimateur. En remarquant qu'une variance est toujours positive, et nulle si et seulement si la variable aléatoire est constante, la fonction π^{opt} minimisant V_{IS}^{π} est :

$$z \mapsto \pi^{opt}(z) \propto f_i(y_i, z; \theta)$$

La fonction π^{opt} étant une densité de probabilité, alors $\pi^{opt}(z) = \frac{f_i(y_i,z;\theta)}{L(\theta;y_i)}$ qui est la densité de la loi des effets aléatoires b_i conditionnellement aux observations, souvent appelée loi *a posteriori* des effets aléatoires. La vraisemblance marginale à estimer est la constante de normalisation de la loi *a posteriori* des effets aléatoires. Estimer la statistique du rapport de vraisemblance (1.16) revient donc à estimer des ratios de constantes de normalisation. Cette question se pose également en statistiques bayésiennes où la vraisemblance marginale est souvent non explicite, et est également la constante de normalisation de la densité *a posteriori* des paramètres.

La question de l'estimation de ratios de constantes de normalisation permet d'une part d'estimer des quantités d'intérêt comme la statistique du rapport de vraisemblance, et d'autre part de généraliser la méthode d'importance sampling d'estimation d'une unique constante de normalisation comme détaillé plus haut. La suite de cette section présente les principales méthodes d'estimation de ratios de constantes de normalisation. Pour cela, on considère, de manière plus générale, deux densités de probabilité sur un ensemble $\mathcal{Z} \subset \mathbb{R}^p$, connues à une constante près :

$$p_i(z) = \frac{f_i(z)}{c_i}, \quad i = 0, 1$$

où pour $i = 0, 1$, f_i est une fonction connue telle que $\int_{\mathcal{Z}} f_i(z)dz = c_i > 0$. L'objectif est celui de l'estimation du ratio des constantes de normalisation inconnues c_0 et c_1 :

$$r^* = \frac{c_0}{c_1} \tag{1.19}$$

Nous présentons ici des résultats et des méthodes de Meng and Wong (1996); Chen and Shao (1997b) qui considèrent directement l'estimation de ce ratio. Le lecteur est renvoyé à ces articles pour plus de détails.
La première approche est le bridge sampling, basée sur l'identité suivante :

$$r^* = \frac{\mathbb{E}_1\left[f_0(Z)\alpha(Z)\right]}{\mathbb{E}_0\left[f_1(Z)\alpha(Z)\right]}$$

où α est une fonction positive définie sur $\mathcal{Z}$ vérifiant $\int_{\mathcal{Z}} \alpha(z)p_1(z)p_0(z)dz < +\infty$, et $\mathbb{E}_i$ est l'espérance par rapport à la densité de probabilité p_i ($i = 0, 1$).

L'estimateur naturel de r^* basé sur cette identité est le suivant :

$$r_K^{bridge} = \frac{\sum_{k=1}^K f_0(Z_k^1)\alpha(Z_k^1)}{\sum_{k=1}^K f_1(Z_k^0)\alpha(Z_k^0)}$$

où $Z_1^0, .., Z_K^0 \overset{iid}{\sim} p_0$ et $Z_1^1, .., Z_K^1 \overset{iid}{\sim} p_1$. De la même manière que pour l'IS, il est possible de trouver la fonction α^{opt} optimale qui minimise la variance asymptotique de l'estimateur bridge sampling. Elle est définie comme :

$$z \mapsto \quad \alpha_{bridge}^{opt}(z) \propto \frac{1}{p_1(z) + p_0(z)} \propto \frac{1}{r^* f_1(z) + f_0(z)}$$

Cette fonction n'est pas directement utilisable, cependant le schéma optimal peut être obtenu (asymptotiquement) en considérant l'estimateur r_{bridge}^{opt} comme solution de l'équation :

$$\sum_{k=1}^K \frac{f_0(Z_k^1)}{r f_1(Z_k^1) + f_0(Z_k^1)} - \sum_{k=1}^K \frac{r f_1(Z_k^0)}{r f_1(Z_k^0) + f_0(Z_k^0)} = 0$$

Il est intéressant de noter que le bridge sampling peut également être utilisé pour estimer une seule constante de normalisation. En effet, supposons que l'on souhaite uniquement estimer c_0, en introduisant une densité de probabilité π connue ($c_1 = \int_{\mathcal{Z}} \pi(z)dz = 1$). Le ratio $r^* = \frac{c_0}{c_1} = c_0$ peut être estimé par bridge sampling. En choisissant la fonction $\alpha = \pi^{-1}$ l'identité bridge sampling devient :

$$\begin{aligned} r^* &= \frac{\mathbb{E}_\pi \left[f_0(Z)\alpha(Z) \right]}{\mathbb{E}_0 \left[\pi(Z)\alpha(Z) \right]} \\ &= \frac{\mathbb{E}_\pi \left[f_0(Z) \times \pi(Z)^{-1} \right]}{\mathbb{E}_0 \left[\pi(Z)\pi(Z)^{-1} \right]} \\ &= \mathbb{E}_\pi \left[\frac{f_0(Z)}{\pi(Z)} \right] \end{aligned}$$

qui correspond à l'identité IS. Ce résultat suggère, que toute procédure d'IS peut être améliorée en lui préférant sa version bridge sampling optimale.

Une seconde identité intéressante est celle du ratio importance sampling (RIS), qui consiste à utiliser l'identité IS au numérateur et au dénominateur. Encore une fois, on considère une densité de probabilité g sur $\mathcal{Z}$ qui vérifie pour tout $z \in \mathcal{Z}$: $\frac{f_i(z)}{\pi(z)} \neq 0$ ($i = 0, 1$). L'identité du RIS est la suivante :

$$r^* = \frac{\mathbb{E}_\pi \left[\frac{f_0(z)}{\pi(z)} \right]}{\mathbb{E}_\pi \left[\frac{f_1(z)}{\pi(z)} \right]}$$

qui définit un estimateur naturel de r^* à partir d'un échantillon $Z_1, ..., Z_K \underset{iid}{\sim} \pi$:

$$r_K^{ris} = \frac{\sum_{k=1}^K \frac{f_0(Z_k)}{\pi(Z_k)}}{\sum_{k=1}^K \frac{f_1(Z_k)}{\pi(Z_k)}}$$

Il est possible d'identifier la densité π_{ris}^{opt} qui minimise la variance asymptotique de l'estimateur RIS, définie pour tout $z \in \mathcal{Z}$ comme : $\pi_{ris}^{opt}(z) \propto |f_0(z) - r^* f_1(z)|$. A nouveau, cette distribution n'est pas utilisable, et cette fois-ci, la procédure optimale ne présente pas, comme pour le bridge sampling, de schéma alternatif. Cependant il est intéressant de noter que la variance asymptotique correspondante est inférieure à la variance asymptotique optimale du bridge sampling (Chen and Shao, 1997a).
De la même manière on peut retrouver l'estimateur IS à partir d'un estimateur RIS, ce qui montre encore une fois que ces méthodes d'estimation de ratios de constantes de normalisation sont d'un grand intérêt. En effet elles améliorent les méthodes d'IS et guident les choix de distributions de proposition, en ayant des schémas optimaux explicites, sans être pour autant systématiquement applicables. Le chapitre 3 présente une nouvelle méthodologie pour estimer des ratios de constantes de normalisation.

1.5 . Contributions de la thèse

Les contributions de ce travail de thèse sont regroupées en trois chapitres. Les deux premiers chapitres présentent de nouvelles approches méthodologiques, dont certaines propriétés théoriques sont étudiées. Le troisième présente un travail appliqué, à l'interface avec la biologie.

Le chapitre 2 propose une procédure de tests des composantes de la variance dans les modèles à effets mixtes. Ce travail explore deux problématiques qui empêchent l'utilisation des tests asymptotiques classiques présentés dans l'introduction. Dans un premier temps cette procédure intègre la présence de paramètres de nuisance qui modifient la distribution asymptotique de la statistique de test. Ensuite, ce travail traite la problématique de singularité de la matrice d'information de Fisher qui apparait dans le cadre non linéaire. Une procédure de test basée sur du Bootstrap paramétrique est proposée. Une étude théorique de la consistance du test, et une étude de simulation sont présentées.

Le chapitre 3 propose une nouvelle méthodologie pour estimer des ratios de constantes de normalisation dans un cadre général. La motivation de cette seconde contribution est l'estimation des statistiques de test du rapport de vraisemblance de la procédure présentée au chapitre précédent. Cette nouvelle méthodologie repose sur le principe de l'approximation stochastique. Dans ce chapitre il est montré que l'estimation du ratio peut être considéré comme la recherche du zéro d'une certaine identité. L'estimateur proposé est consistant et atteint la même variance asymptotique que celle de l'estimateur RIS optimal présenté dans la section 1.4 qui n'est pas calculable. Nous montrons également que notre procédure itérative peut être intégrée dans un processus joint d'inférence dans les modèles à variables latentes.

Le chapitre 4 est dédié à la problématique biologique présentée en section 1.1.1. La question de l'étude de la variabilité génotypique chez *Arabidopsis thaliana* est détaillée. Le modèle mécaniste ARNICA décrivant les échanges d'azote et de carbone entre la plante et son environnement au cours de son processus de croissance est décrit. Ce modèle déterministe permet de décrire différents traits phénotypiques de la plante au cours du temps, à l'échelle d'un unique génotype, à partir de différents paramètres ayant un sens biologique précis. Une approche statistique populationnelle, qui consiste à intégrer le modèle ARNICA dans un modèle à effets mixtes est proposée. Une procédure d'inférence des paramètres du modèle statistique est développée, et illustrée sur des données simulées, dans le cadre d'un modèle restreint. Cette approche est également appliquée au jeu de données réelles des 48 écotypes d'*Arabidopsis thaliana*.

Enfin, le chapitre 5 conclura ce manuscrit et présentera les perspectives de recherche relatives à chacun des travaux.

Chapitre 2

Une procédure Bootstrap pour tester la nullité des composantes de la variance dans les modèles à effets mixtes

Ce chapitre présente l'article publié :

Guédon, T., Baey, C., & Kuhn, E. (2024). Bootstrap test procedure for variance components in nonlinear mixed effects models in the presence of nuisance parameters and a singular Fisher information matrix. Biometrika, asae025. **Guédon et al. (2024a)**

Le problème considéré est celui du test des composantes de variance dans les modèles à effets mixtes. Nous considérons N individus, chacun mesuré J fois. Nous notons y_{ij} la j-ème observation du i-ème individu pour $i = 1, ..., N$ et $j = 1, ..., J$. Nous considérons le modèle à effets mixtes suivant :

$$y_i = g(x_i, \beta, \Lambda\xi_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_J), \quad \xi_i \sim \mathcal{N}(0, I_p)$$

où g est une fonction potentiellement non linéaire connue, x_i regroupe toutes les covariables de l'individu i, β est le vecteur inconnu des effets fixes, Λ est une matrice triangulaire supérieure servant de paramètre d'échelle pour l'effet aléatoire ξ_i, et σ^2 est la variance positive du bruit. Cette modélisation généralise le modèle simplifié présenté en 1.12, qui rend la vraisemblance définie, même si des variances d'effets aléatoires sont nulles.

Nous considérons la statistique du rapport de vraisemblance pour tester si certaines composantes de la matrice de covariance des effets individuels $(\Lambda\xi_i)_{i=1,...,N}$ sont nulles. Dans le contexte considéré, deux problèmes principaux surviennent.

D'une part, sous l'hypothèse nulle, le vrai paramètre n'est pas un point intérieur de l'espace des paramètres, ce qui empêche l'utilisation de la théorie asymptotique habituelle de l'estimation par maximum de vraisemblance. Ce problème a été considéré dans la littérature (voir Andrews (1999), Self and Liang (1987) par exemple).

D'autre part, dans le contexte spécifique des modèles à effets mixtes non linéaires, la matrice d'information de Fisher est singulière lorsque l'on considère des composantes de Λ égales à zéro et lorsque des variances inconnues non testées, également appelées paramètres de nuisance, sont nulles. Ces deux problèmes rendent difficile, voire impossible, la mise en oeuvre du test asymptotique du rapport de vraisemblance.

Nous proposons une procédure basée sur le Bootstrap paramétrique pour effectuer ce test. L'avantage du Bootstrap dans notre contexte est double : d'une part, il fonctionne bien avec de petites tailles d'échantillons ; d'autre part, il permet de prendre en compte la présence de paramètres de nuisance. Cependant, les deux problèmes mentionnés plus haut peuvent conduire à une inconsistance du Bootstrap. Le problème de frontière peut être résolu en utilisant un paramètre de Bootstrap modifié qui seuille à zéro les paramètres de nuisance, comme développé dans Cavaliere et al. (2020). Nous montrons que le problème de singularité peut être résolu de la même manière, à condition que le paramètre de seuillage converge suffisamment rapidement

vers zéro.

Les principaux résultats théoriques sont le Théorème 2.1 qui établit des hypothèses sur le choix du paramètre Bootstrap permettant d'assurer la consistance de la procédure de test sous l'hypothèse nulle, et le Théorème 2.2 qui étend ce résultat au cadre non identiquement distribué. La Proposition 2.5 explicite une manière de choisir le paramètre Bootstrap afin que celui-ci vérifie les hypothèses des théorèmes 2.1 et 2.2. Des conditions, vérifiables en pratique, sur la fonction g sont enfin données en section 2.3.4. Celles-ci permettent de vérifier les hypothèses de régularité requises à la validité des résultats théoriques. Cette propriété permet de montrer que différents modèles usuels, souvent utilisés dans la littérature vérifient les hypothèses requises.

La principale contribution de ce travail est une procédure facile à appliquer même aux modèles non linéaires et qui traite le problème des paramètres de nuisance, ce qui, à notre connaissance, n'a pas été considéré dans la littérature sur les modèles à effets mixtes. Nous illustrons la performance de notre procédure sur des données simulées et des données réelles.

Bootstrap test procedure for variance components in nonlinear mixed effects models in the presence of nuisance parameters and a singular Fisher information matrix

2.1 . Introduction

Mixed effects models are a powerful statistical tool to model longitudinal studies with repeated measurements or data with an underlying unknown latent structure as hierarchical data. There are many fields of applications, e.g. pharmacokinetic-pharmacodynamic (Bonate, 2011), medicine (Brown and Prescott, 2015), agriculture (Zhou et al., 2022), ecology (Bolker et al., 2009), psychology (Meteyard and Davies, 2020) or educational and social sciences (Gordon, 2019). These models allow to take into account two types of variabilities, between different individuals in a population and between several measurements made on the same individual, also called inter and intra variabilities. These are modeled by two types of effects : on the one hand, random effects that vary from one individual to another, and on the other hand, fixed effects, common to all individuals in the population (Pinheiro and Bates, 2006; Davidian and Giltinan, 2017).

From a modeling point of view, being able to distinguish among all effects those that can be modeled as fixed effects would allow one to reduce the number of model parameters. This would also help to better identify the processes that are the cause of the variability observed in the population. Two main approaches have been developed to tackle this task. On the one hand, some authors suggested methods based on variable selection, using a Bayesian procedure as in Chen and Dunson (2003) or a penalized likelihood approach as in Ibrahim et al. (2011) or Groll and Tutz (2014). Specific selection criteria for mixed-effects models were also developed by Vaida and Blanchard (2005); Gurka (2006) and Delattre et al. (2014). On the other hand, other authors focused on hypothesis testing for the nullity of some variance components of the random effects.

Such a test is equivalent to comparing two nested models, and standard tools to address this question include the likelihood ratio, the score and the Wald tests statistics (Van der Vaart, 2000). However two issues arise when testing the nullity of variance components, that prevent from using the usual asymptotic results of Wilks (1938). The first issue results from the true value of the variance parameter lying on the boundary of the parameter space, while the second is due to the singularity of the Fisher information matrix.

In the specific context of mixed effects models, Crainiceanu and Ruppert (2004) derived the exact distribution of the likelihood ratio test statistic in a linear model with one random effect. Stram and Lee (1994) derived an asymptotic likelihood ratio test for linear models in some specific cases. Lin (1997) proposed a score test in several specific cases in generalized linear mixed

effects models. Wood (2013) proposed a way to treat variance components testing in generalized models using the linear case machinery. Baey et al. (2019) derived the asymptotic distribution of the likelihood ratio test statistic for testing that any subset of the variances of the random effects is null. However, for applicability, these references assume that the untested parameters do not lie on the boundary of the parameter space.

In a more general framework, as far as the boundary issue is concerned, several authors studied the asymptotic of the likelihood ratio test statistic in this context. Chernoff (1954), Chant (1974), Self and Liang (1987) and Geyer (1994b) derived the asymptotic distribution of the likelihood ratio test statistic in specific cases. Andrews (1999), Silvapulle and Sen (2005) gave a more general way of dealing with hypothesis testing when the true parameter is not constrained to be an interior point of the ambient space. When this is the case, the asymptotic distribution of the likelihood ratio test statistic is intractable as it depends on the unknown location of these nuisance parameters on the boundary (Self and Liang, 1987). Therefore, procedures that do not involve the asymptotic distribution of the test statistic can be preferred. In particular, resampling methods, such as those based on the Bootstrap or permutations, are powerful tools to address this issue. In addition, these methods are usually more robust in small samples context. Dealing with variance components testing,Sinha (2009) proposed a Bootstrap-based score test in the context of generalized linear mixed models with one single random effect. Drikvandi et al. (2013b) proposed a permutation-based test for any subset of the covariance matrix of the random effects in linear mixed models. The latter method is very easy to use in practice but is restricted to the context of linear models. The former requires the computation of the Fisher information matrix, which can be heavy in practice, especially in the context of nonlinear models. Moreover, the presence of nuisance parameters on the boundary of the parameter space is not considered in the aforementioned works, even though it can be a source of inconsistency for the Bootstrap procedures. Indeed, as discussed in Beran (1997), and highlighted in Andrews (2000), the Bootstrap is known to be inconsistent when the true parameter value is a boundary point. When estimating the expectation of a Gaussian distribution, restricted to be nonnegative, Andrews (2000) proposed a parametric procedure that shrinks the parameter used to generate the Bootstrap data near the boundary. Following this idea, Cavaliere et al. (2020) proposed a more general parametric Bootstrap test procedure based on the Likelihood Ratio Test statistic with parameters lying on the boundary. Their method consists in shrinking the Bootstrap parameter in order to accelerate its rate of convergence toward the boundary.

The second issue is the singularity of the Fisher information matrix that arises specifically in the context of mixed effect models, as discussed in Ekvall and Bottai (2021).This phenomenon is studied in Rotnitzky et al. (2000) when the rank of the Fisher Information Matrix is full minus one. Following the development of Hiroyuki et al. (2012) to derive the new asymptotic distribution of the likelihood ratio test statistic, we show that this singularity issue is another source of

inconsistency of the Bootstrap procedure.

In this work we propose a shrinked parametric Bootstrap test procedure for variance components in nonlinear mixed effects models that addresses the two issues mentioned above. We show that given an appropriate choice of the Bootstrap parameter, the procedure is consistent as the number of individuals grows to infinity. Our contribution is twofold : first, our procedure can be applied to linear, generalized linear and nonlinear models, and second, it takes into account the presence of nuisance parameters at unknown locations. We also provide a verifiable criterion to check the required regularity conditions. Finally, we illustrate our results on simulated and real data, exhibiting the good finite sample properties of the procedure and its applicability in practice.

2.2 . Proposed methodology

2.2.1 . Mixed effects models

Consider N individuals each measured $J_i < J$ times, where N and J_i are nonnegative integers. We denote by y_{ij} ($i = 1, ..., N; j = 1, ..., J_i$) the jth observation of the ith individual and we define $y_i = (y_{i1}, \dots, y_{iJ_i})$ and $y_{1:N} = (y_1^T, \dots, y_N^T)$. In the sequel, $\mathcal{L}_p^+$ denotes the space of lower triangular matrices of size $p \times p$ with positive diagonal coefficients, $\mathbb{S}_+^p$ denotes the space of symmetric, positive semi-definite $p \times p$ matrices, I_p is the identity matrix of size $p \times p$, $[A]_{ij}$ is the element on the ith line and jth column of matrix A, and $\mathcal{N}(\mu, V)$ denotes the multivariate Gaussian distribution with expectation $\mu \in \mathbb{R}^p$ and covariance matrix V of size $p \times p$. We consider the following nonlinear mixed effects model

$$\begin{cases} y_{ij} & = g(x_{ij}, \beta, \Lambda\xi_i) + \varepsilon_{ij} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \xi_i & \sim \mathcal{N}(0, I_p) \end{cases}, \tag{2.1}$$

where $(\xi_i)_{i=1,\dots,N}$ and $(\varepsilon_{ij})_{i=1,..,N, j=1,\dots,J_i}$ are mutually independent random variables, g is a known nonlinear function, x_{ij} gathers all the covariates of the jth observation of the ith individual, $\beta \in \mathbb{R}^b$ is the vector of fixed effects, $\Lambda \in \mathcal{L}_p^+$ is a scaling parameter for random effect ξ_i, and σ^2 is the positive noise variance. The main advantage of this formulation for a nonlinear mixed effects model is that the random effect distribution is parameter-free, which will be particularly well adapted for the theoretical analysis of the proposed procedure.

Remark 2.1. *The covariance matrix of the scaled random effect $b_i = \Lambda\xi_i$ is equal to $\Gamma = \Lambda\Lambda^T$ which is positive semi-definite. Therefore, a natural choice for Λ is the lower triangular matrix in the Cholesky decomposition of the scaled random effects covariance matrix. This reparametrization is used for instance in Chen and Dunson (2003). When Γ is positive definite, the Cholesky decomposition and hence the matrix Λ, is uniquely defined. When Γ is positive semi-definite, there exists a permutation of*

rows and columns of Γ such that the permuted matrix has a unique Cholesky decomposition (Higham, 1990). This permutation orders the rows and columns of Γ so that Λ is lower triangular, with some diagonal blocks that can be equal to 0.

Remark 2.2. *The definition of model* (2.1) *is slightly more general than the usual terminology of mixed effects models (Pinheiro and Bates, 2006, p. 306) that defines $y_{ij} = g(v_{ij}, \phi_i) + \varepsilon_{ij}$, with $\phi_i = A_{ij}\beta + B_{ij}b_i$ the ith individual parameter, β the vector of fixed effects associated with random effect $b_i \sim \mathcal{N}(0, \Gamma)$, and where v_{ij}, A_{ij} and B_{ij} are known covariates. Model* (2.1) *covers this definition by taking $x_{ij} = (v_{ij}, B_{ij}, A_{ij})$ and $b_i = \Lambda\xi_i$. For example, when considering a linear mixed effects model one writes*

$$g(x_{ij}, \beta, \Lambda\xi_i) = x_{ij}^T(\beta + \Lambda\xi_i)$$

where x_{ij} are known covariates, β is a unknown vector of fixed effects parameters, Λ is an unknown scaling parameter and ξ_i is the random effect. One can also consider the logistic growth model (Pinheiro and Bates, 2006) given by :

$$g(x_{ij}, \beta, \Lambda\xi_i) = \frac{\beta_1 + \lambda_1 b_{i1}}{1 + \exp\left\{-\frac{x_{ij} - (\beta_2 + \lambda_2 b_{i2})}{\beta_3 + \lambda_3 b_{i3}}\right\}}$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$ is supposed diagonal and $(x_{ij})_{j=1,\dots,J_i}$ are the times of measurements of individual i. A more detailed development of the general differences between those two parameterizations is given in section (2.7.1) *of the supplementary material.*

Let us denote by $\theta = (\beta, \Lambda, \sigma^2)$ the unknown vector of model parameters taking values in Θ, by $f_i(\cdot; \theta)$ the density of the ith individual response y_i given a parameter $\theta \in \Theta$, by $f_i(y_i; \xi_i, \theta)$ the conditional density of y_i given the random effect ξ_i and a parameter θ, and by $\pi_p(\cdot)$ the density of the p-dimensional standard Gaussian density. With these notations we can define the log-likelihood of the model given the N-sample $y_{1:N}$ by

$$l(\theta; y_{1:N}) = \log\{L_\theta(y_{1:N})\} = \log\{\prod_{i=1}^{N} f_i(y_i; \theta)\} = \sum_{i=1}^{N} \log\{\int f_i(y_i; \xi_i, \theta)\pi_p(\xi_i)d\xi_i\} \qquad (2.2)$$

We consider the marginal likelihood defined as the complete likelihood integrated over the distribution of the random effects, since the random effects ξ_i are unobserved. We recall that contrary to the usual formulation of nonlinear mixed effects models, where the random effects are defined as the scaled version b_i, the one considered in (2.1) has the advantage to lead to a parameter-free distribution for the random effects ξ_i. Indeed, with the former definition, the distribution of the latent variables depends on Λ, and is not defined on the entire parameter space since we only constrain $\Gamma = \Lambda\Lambda^T$ to be positive semi-definite. When dealing with linear models, since the variance of the random effects adds up with the noise variance, the fact that some

diagonal components in Γ are null is not an issue. However, the change of variables $b_i = \Lambda\xi_i$ is a $\mathcal{C}^1$ diffeomorphism if and only if the diagonal coefficients of Λ are strictly nonnegative. Without this assumption the two parametrizations are no longer equivalent as illustrated in the supplementary material (see section 2.7.1). Our parametrization is similar to the so-called re-parametrization trick proposed by Kingma and Welling (2020) to train variational autoencoders with back-propagation.

2.2.2 . Variance components testing

Let $r \in \{1, \dots, p\}$ be the number of variances to be tested. Without loss of generality we assume that we test the nullity of the last r variances in $\Gamma = \Lambda\Lambda^T$. Therefore let us consider the following block matrix notation

$$\Lambda = \left(\begin{array}{c|c} \Lambda_1 & 0_{(p-r)\times r} \\ \hline \Lambda_{12} & \Lambda_2 \end{array} \right),$$

where $\Lambda_1 \in \mathcal{L}^+_{p-r}$, $\Lambda_2 \in \mathcal{L}^+_r$ and $\Lambda_{12} \in \mathcal{M}_{r\times(p-r)}(\mathbb{R})$. We write θ_0 the true parameter on which we consider the following test :

$$H_0 : \theta_0 \in \Theta_0 \quad \text{against} \quad H_1 : \theta_0 \in \Theta, \tag{2.3}$$

where

$$\Theta_0 = \{\theta \in \mathbb{R}^q \mid \beta \in \mathbb{R}^b, \Lambda_1 \in \mathcal{L}^+_{p-r}, \Lambda_2 = 0, \Lambda_{12} = 0, \sigma^2 \in \mathbb{R}^+_*\},$$
$$\Theta = \{\theta \in \mathbb{R}^q \mid \beta \in \mathbb{R}^b, \Lambda \in \mathcal{L}^+_p, \sigma^2 \in \mathbb{R}^+_*\}.$$

Remark 2.3. *We do not impose the diagonal of Λ_1 to be strictly non-negative, which enables the case where some untested variances of the scaled random effects are in fact equal to zero. This will be discussed in more details in section* (2.3.1) *with the definition of nuisance parameters.*

The likelihood ratio test statistic is defined as

$$\text{lrt}(y_{1:N}) = 2\left\{ \sup_{\theta\in\Theta} \; l(\theta; y_{1:N}) - \sup_{\theta\in\Theta_0} \; l(\theta; y_{1:N}) \right\}.$$

In order to test (2.3) with a nominal level $0 < \alpha < 1$, we define the rejection region as $R_\alpha = \{\text{lrt}(y_{1:N}) \geq q_\alpha\}$ with q_α being the $(1-\alpha)$th quantile of the distribution of $\text{lrt}(y_{1:N})$. Unfortunately, this distribution is often intractable.

In the following section, we detail the proposed shrinked parametric Bootstrap procedure to test (2.3).

2.2.3 . Testing procedure

Following the lines of Cavaliere et al. (2020), we propose a parametric Bootstrap procedure using a Bootstrap parameter θ^*_N and $B \in \mathbb{N}^*$ Bootstrap replications to test (2.3) with a type I

error $0 < \alpha < 1$. As introduced in remark 2.3 and then detailed in section (2.3.1), some untested variances, at unknown locations, can be null. Therefore using $\hat{\theta}_N = (\hat{\beta}_N, \hat{\Lambda}_N, \hat{\sigma}^2_N)$ the maximum likelihood estimator as a Bootstrap parameter over Θ would fail to asymptotically mimic the true distribution of the likelihood ratio test statistic. Indeed there are elements in Λ_0 which are supposed to be zero, but that are non null in $\hat{\Lambda}_N$.Since we require that $\theta^*_N \in \Theta_0$ we can choose θ^*_N to be the unrestricted maximum likelihood estimator projected on Θ_0 or the restricted one. Furthermore, we use a shrinking parameter c_N to fix to zero the untested components of Λ that are smaller than c_N. The proposed algorithm is described in algorithm 2.1, and the theoretical justification of this shrinking procedure is described in section 3.3.2.

Algorithm 2.1 Shrinked parametric Bootstrap for variance components testing

Input : $c_N > 0$, $B \in \mathbb{N}^*$, $0 < \alpha < 1$
Set $\beta^*_N = \hat{\beta}_N$, $\Lambda^*_N = \hat{\Lambda}_N$, and ${\sigma^*_N}^2 = \hat{\sigma}^2_N$
Set $\Lambda^*_{2,N} = \Lambda^*_{12,N} = 0$
Set $[\Lambda^*_{1,N}]_{mn} = [\hat{\Lambda}_{1,N}]_{mn} \mathbb{1}_{[\hat{\Lambda}_{1,N}]_{mn} > c_N}$
For $b = 1, ..., B$
  For $i = 1, ..., N$, draw independently $\varepsilon_i^{*,b} \sim \mathcal{N}(0, {\sigma^*_N}^2 I_{J_i})$ and $\xi_i^{*,b} \sim \mathcal{N}(0, I_p)$
  Build the ith value of the bth Bootstrap sample $y_i^{*,b} = g(x_i, \beta^*_N, \Lambda^*_N \xi_i^{*,b}) + \varepsilon_i^{*,b}$
  Compute the likelihood ratio statistic $\mathrm{lrt}(y^{*,b}_{1:N})$
Compute the Bootstrap p-value as $p_{boot} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\mathrm{lrt}(y^{*,b}_{1:N}) > \mathrm{lrt}(y_{1:N})}$
Reject H_0 if $p_{boot} < \alpha$

The next section is dedicated to the asymptotic validity of this testing procedure.

2.3 . Theoretical results

2.3.1 . Notations and theoretical setting

In this section we are interested in the theoretical consistency of the Bootstrap procedure presented in section 2.2.3. We consider the asymptotic as the number of individuals N grows to infinity, while the number of measurements per individual remains fixed and bounded by some value J. We denote by $\theta_0 \in \Theta_0$ the true value of the parameter, such that the density of the response $y_{1:N}$ is $L_{\theta_0}(y_{1:N})$. We denote by $E\{T(y_{1:N})\}$ the expectation of any measurable function T of $y_{1:N}$ if there is no confusion about the distribution of $y_{1:N}$. Otherwise we specify $E_\theta\{T(y_{1:N})\}$ to emphasize that the expectation is with respect to the density $L_\theta(y_{1:N})$, for any $\theta \in \Theta$. As commonly used in the Bootstrap literature, we denote by X^* the Bootstrap version of a random variable X. We write $E^*\{T(y^*_{1:N})\} = E_{\theta^*_N}\{T(y^*_{1:N}) \mid y_{1:N}\}$. Similarly, for any mea-

surable subset A we write $\text{pr}^*\{T(y^*_{1:N}) \in A\} = E_{\theta^*_N}\{\mathbb{1}_{T(y^*_{1:N})\in A} \mid y_{1:N}\}$.

We want to show that the proposed Bootstrap procedure is asymptotically valid which means that $\text{lrt}(y^*_{1:N})$ converges weakly in probability to the same limiting distribution as the one of $\text{lrt}(y_{1:N})$. More precisely, we want to show that, under some conditions, if there exists a random variable lrt_∞ such that $\text{lrt}(y_{1:N})$ converges weakly to lrt_∞ then for every $t \in \mathbb{R}$, as $N \to +\infty$, it holds in probability that

$$\text{pr}^*\{\text{lrt}(y^*_{1:N}) \leq t\} \longrightarrow \text{pr}(\text{lrt}_\infty \leq t). \tag{2.4}$$

We also use the notations $o_p(1)$ and $O_p(1)$ for random sequences that respectively converge toward zero and are bounded in probability. More generally, this notation is used to compare two random sequences, using the definition of Van der Vaart (2000, section 2.2). We also use their Bootstrap versions o_{p^*} and O_{p^*} defined as follows : for a random quantity X^*_N computed on the Bootstrap data, $X^*_N = o_{p^*}(1)$ means that for any $\varepsilon > 0$, $\text{pr}^*(X^*_N > \varepsilon) \to 0$ in probability as $N \to +\infty$. Similarly $X^*_N = O_{p^*}(1)$ means that for any $\varepsilon > 0$ there exists a real $M > 0$ and an integer N_0 such that for all $N > N_0$, the event $\{\text{pr}^*(\|X^*_N\| > M) < \varepsilon\}$ is arbitrary close to one in probability.

We now formalize what we call nuisance parameters. We suppose that, in addition to the last r tested variances, m untested variances are null. Without loss of generality we suppose that the last $m + r$ variances of the individual parameters are null therefore Λ_0 is of the form

$$\Lambda_0 = \left(\begin{array}{c|c|c} \Lambda_1^{nonuis} & 0_{(p-r-m)\times m} & 0_{(p-r-m)\times r} \\ \hline \Lambda_{12}^{nuis} & \Lambda_1^{nuis} & 0_{m\times r} \\ \hline \Lambda_{12,1} & \Lambda_{12,2} & \Lambda_2 \end{array}\right) = \left(\begin{array}{c|c|c} \Lambda_1^{nonuis} & 0_{(p-r-m)\times m} & 0_{(p-r-m)\times r} \\ \hline 0_{m\times(p-r-m)} & 0_{m\times m} & 0_{m\times r} \\ \hline 0_{r\times(p-r-m)} & 0_{r\times m} & 0_{r\times r} \end{array}\right).$$

It is important to notice that in real life applications the m rows inducing nuisance parameters are located at unknown positions in matrix Λ, and that the remaining $p - m - r$ variances are strictly non-negative which is equivalent to the diagonal coefficients of Λ_1^{nonuis} being strictly non-negative.

Following Self and Liang (1987) and Cavaliere et al. (2020), we now split the parameter as $\theta = (\psi, \delta, \lambda)$, where λ stands for all the coefficients of Λ_2 and $\Lambda_{12,2}$, δ represents the coefficients in Λ_1^{nuis} and ψ gathers all the remaining parameters. The dimension of λ is $d_\lambda = r(r+1)/2+r(p-r-m)$, the dimension of δ is $d_\delta = m(m+1)/2+r\times m$ and the dimension of ψ is $d_\psi = d_\theta - d_\lambda - d_\delta$. Moreover, $\theta_0 = (\psi_0, \delta_0, \lambda_0) = (\psi_0, 0_{d_\delta}, 0_{d_\lambda})$. Before introducing our results we first state a set of general conditions on the model that will be required in this work.

Assumption 2.1. *$(i)\Theta$ is compact, (ii) the model is identifiable, (iii) for all $i \in \mathbb{N}, y \in \mathbb{R}_i^J$, $\xi \in \mathbb{R}^p$ the conditional likelihood $\theta \mapsto f_i(y;\xi,\theta)$ is 4-times differentiable on the interior of Θ, and directional derivatives exist on the boundary, $iv)$ each partial derivative of $\theta \mapsto f_i(y;\xi,\theta)$ is bounded by a positive function which does not depend on θ and is integrable with respect to the distribution of the*

random effects.

Remark 2.4. *The compactness assumption is not verified for Θ. However in practice it only requires that $\sigma^2 \geq \rho$ for some non-negative number ρ and that each component of θ is upper and lower bounded, which is reasonable in real data applications.Assumption (ii) is usual in the context of estimation theory. Assumption (iii) is needed to perform a Taylor expansion of the log likelihood and (iv) is needed to differentiate under the integral sign in* (2.2)*. These assumptions are discussed in section* (2.3.4)*.*

The following proposition induces that if the Fisher Information Matrix exists, it will present blocks equal to zero, and will therefore be singular. This result extends the one of Rotnitzky et al. (2000) stated when the rank of the Fisher Information Matrix is full minus 1 to more general settings where the rank of the Fisher Information Matrix is full minus $d_\lambda + d_\delta$.

Proposition 2.1. *Under assumption* (2.1)*, for $k = 0, 1$, for all $i \in \mathbb{N}$ and for all $y \in \mathbb{R}^{J_i}$, $\nabla_\delta^{2k+1} \log\{f_i(y;\theta_0)\} = 0_{d_\delta^{2k+1}}$ and $\nabla_\lambda^{2k+1} \log\{f_i(y;\theta_0)\} = 0_{d_\lambda^{2k+1}}$. In particular,* var$\{\nabla_\delta l(\theta; y_{1:N})\} = 0_{d_\delta \times d_\delta}$ *and* var$\{\nabla_\lambda l(\theta; y_{1:N})\} = 0_{d_\lambda \times d_\lambda}$.

Remark 2.5. *If $\theta \mapsto l(\theta; y_{1:N})$ admits higher order derivatives, the first part of Proposition 2.1 is true for every odds order derivatives. This comes from the null odds moments of the standard normal distribution of the random effects.*

Remark 2.6. *As shown in the proof of proposition 2.1, in section 2.7.2 of the Appendix, by considering the kth column $[\Lambda]_{.k} = ([\Lambda]_{1k}, ..., [\Lambda]_{pk})^T$ of Λ, for all $j = 1, ..., p$, $\partial l(\theta; y_{1:N})/\partial[\Lambda]_{jk}|_{[\Lambda]_{.k}=0_p} = 0$. That explains why the coefficients of $\Lambda_{12,2}$ are part of the definition of λ.*

2.3.2 . Consistency of the Bootstrap procedure in the identically distributed setting

We first deal with the simpler identically distributed case. In model (2.1) it corresponds to the case where $(x_{ij})_{j=1,...,J_i}$ are common to every individual i. The next section is devoted to extending the results to the non identically distributed setting presented before.

Before studying the consistency of the test procedure, we first need to ensure the consistency of the restricted (respectively unrestricted) maximum likelihood estimator, i.e. computed over Θ_0 (respectively Θ). We first state the regularity conditions required for the asymptotic theory that follows.

Assumption 2.2. *For every $k, l, s, t = 1, \ldots, d_\theta$ and for every $i = 1, \ldots, N$:*

$$
\begin{aligned}
(i) \quad & \sup_{\theta' \in \Theta} \mathbb{E}_{\theta'} \{ \sup_{\theta \in \Theta} |\log f(y_i;\theta)|^2 \} < +\infty \\
(ii) \quad & \sup_{\theta' \in \Theta} \mathbb{E}_{\theta'} \{ \sup_{\theta \in \Theta} \left| \frac{\partial \log f(y_i;\theta)}{\partial \theta_k} \right|^3 \} < +\infty \\
(iii) \quad & \sup_{\theta' \in \Theta} \mathbb{E}_{\theta'} \{ \sup_{\theta \in \Theta} \left| \frac{\partial^2 \log f(y_i;\theta)}{\partial \theta_k \partial \theta_l} \right|^3 \} < +\infty \\
(iv) \quad & \sup_{\theta' \in \Theta} \mathbb{E}_{\theta'} \{ \sup_{\theta \in \Theta} \left| \frac{\partial^3 \log f(y_i;\theta)}{\partial \theta_k \partial \theta_l \partial \theta_s} \right|^2 \} < +\infty \\
(v) \quad & \sup_{\theta' \in \Theta} \mathbb{E}_{\theta'} \{ \sup_{\theta \in \Theta} \left| \frac{\partial^4 \log f(y_i;\theta)}{\partial \theta_k \partial \theta_l \partial \theta_s \partial \theta_t} \right|^2 \} < +\infty
\end{aligned}
$$

Assumption (2.2) (i) is needed to ensure the consistency of the maximum likelihood estimators. Indeed it enables to derive a uniform law of large numbers. Assumptions (ii) and (iii) are similar to assumption (N8') in Hoadley (1971). It is required to apply a central limit theorem to the score function, and the pseudo score function $\tilde{S}_N(\theta)$ that appears in the quadratic expansion (see equation (2.22) in the Appendix). Assumptions (iv) and (v) are needed to control the rest of the quadratic approximation. All the suprema are needed to control the consistency of the Bootstrap distributions.

We now derive the consistency of the maximum likelihood estimators, following the result of Moran (1971).

Proposition 2.2. *Under assumptions* (2.1)–(2.2) $i)$ *:*

$$
\begin{aligned}
\arg\max_{\theta \in \Theta} \quad & l(\theta; y_{1:N}) = \theta_0 + o_p(1) \\
\arg\max_{\theta \in \Theta_0} \quad & l(\theta; y_{1:N}) = \theta_0 + o_p(1)
\end{aligned}
$$

A natural choice for the Bootstrap parameter θ_N^* is the maximum likelihood estimator. However the Bootstrap fails in presence of the nuisance parameters summarized in vector δ. This is why care must be taken when choosing δ_N^*. To explain and solve this issue we first need to derive the speed of convergence of the maximum likelihood estimator.

Proposition 2.3. *Let* $\hat{\theta}_N = (\hat{\psi}_N, \hat{\delta}_N, \hat{\lambda}_N)$ *and* $\tilde{\theta}_N = (\tilde{\psi}_N, \tilde{\delta}_N, 0_{d_\lambda})$ *be respectively the unrestricted and restricted maximum likelihood estimators of* θ*. Under assumptions* (2.1) *and* (2.2) $(\sqrt{N}(\hat{\psi}_N - \psi_0), \sqrt{N}(\tilde{\psi}_N - \psi_0)) = O_p(1), (\hat{\delta}_N, \tilde{\delta}_N, \hat{\lambda}_N) = O_p(N^{-1/4})$.

We emphasize that this result achieves the same rate of convergence as in Rotnitzky et al. (2000). However their result is not applicable here as the number of vanishing score components is greater than one, therefore the exact asymptotic distribution of the maximum likelihood estimators is unknown. The usual way to derive the asymptotic distribution of the likelihood ratio

statistic is to consider a quadratic approximation of the log-likelihood around the true value of the parameter, based on a second-order Taylor expansion. However in our case, due to the vanishing score property stated in proposition 2.1, this quadratic expansion is degenerate with respect to parameters δ and λ. Using a reparametrization of parameter θ as in Hiroyuki et al. (2012), we obtain a new quadratic approximation based on a higher-order expansion of the log-likelihood. We then apply results from Andrews (1999) and Silvapulle and Sen (2005) to obtain an explicit formula for lrt_∞. This new quadratic approximation involves a new matrix $\tilde{I}(\theta_0)$ which plays the role of the Fisher information matrix, and which is defined explicitly in equation (2.22) of the supplementary material. This new matrix is no longer systematically degenerate, we can therefore state the usual assumption which must be verified case by case in real life applications.

Assumption 2.3. $\tilde{I}(\theta_0) \succ 0$

Matrix $\tilde{I}(\theta_0)$ is the asymptotic variance of a modified score function (see Remark 2.10 in the Appendix), which ensures that it is positive semi-definite. This matrix depends on derivatives up to the fourth order of the log-likelihood, and is no longer degenerate.

Before showing that the proposed test procedure is consistent, we first need to show that in the boostrap world, if the Bootstrap parameter is consistent, the Bootstrap maximum likelihood estimators are, conditionally to the data, consistent.

Proposition 2.4. *Under assumptions* (2.1)–(2.2) i*), if* θ_N^* *the parameter used to generate the data is consistent, then :*

$$\begin{aligned} \arg\max_{\theta\in\Theta} \quad & l(\theta; y_{1:N}^*) = \theta_0 + o_{p^*}(1) \\ \arg\max_{\theta\in\Theta_0} \quad & l(\theta; y_{1:N}^*) = \theta_0 + o_{p^*}(1). \end{aligned}$$

We can now state the main result that guarantees the consistency of the Bootstrap procedure.

Theorem 2.1. *Under assumptions (1)–(3), if* θ_N^* *is chosen such that* $\theta_N^* \in \Theta_0$, $\theta_N^* = \theta_0 + o_p(1)$ *and* $N^{1/4}\delta_N^* = o_p(1)$ *then, for every* $t \in \mathbb{R}$ *as* $N \to +\infty$*, it holds in probability that*

$$\mathsf{pr}^*\{\mathsf{lrt}(y_{1:N}^*) \leq t\} \longrightarrow \mathsf{pr}(\mathsf{lrt}_\infty \leq t). \tag{2.5}$$

where the expression of lrt_∞ *is given in the Appendix.*

A way of choosing θ_N^* that fulfills the hypotheses of theorem 2.1 is to follow the idea of Cavaliere et al. (2020) and shrinks the parameter toward 0. However here the rate of convergence of the shrinking parameter (c_N) is not the same due to the singularity issue. The following lemma gives a procedure to choose θ_N^* and justify the way it is chosen in algorithm 2.1.

Proposition 2.5. *Let $(c_N)_{N\in\mathbb{N}}$ be a sequence such that $\lim_{N\to+\infty} c_N = 0$ and $\lim_{N\to+\infty} N^{\frac{1}{4}} c_N = +\infty$. Let $\hat{\theta}_N = (\hat{\psi}_N, \hat{\delta}_N, \hat{\lambda}_N)$ be a maximum likelihood estimator (restricted or not) of $\theta_0 = (\psi_0, 0_{d_\delta}, 0_{\delta_\lambda})$. Under assumptions* (2.1)–(2.3)*, by choosing $\theta_N^* = (\psi_N^*, \delta_N^*, \lambda_N^*)$ such that : $\forall k = 1, .., d_\psi \ \psi_{N,k}^* = \hat{\psi}_{N,k} \mathbb{1}(\hat{\psi}_{N,k} > c_N)$, $\forall k = 1, .., d_\delta \ \delta_{N,k}^* = \hat{\delta}_{N,k} \mathbb{1}(\hat{\delta}_{N,k} > c_N)$ and $\lambda_N^* = 0_{d_\lambda}$ then, θ_N^* verifies the hypothesis of theorem 2.1.*

As we do not know which parameters are part of δ, it is important to deal with every potential nuisance parameters. This is why we also consider a shrinkage Bootstrap parameter for ψ. In the proof of this proposition we show that the shrinkage does not change the limit of the estimate, but only speeds up its convergence toward 0.

In addition to the boundary issue, the singularity is another source of inconsistency for the Bootstrap procedure. As highlighted in the proof of theorem (2.1), this inconsistency comes from the polluting random variables due to the asymptotic distribution of $N^{\frac{1}{4}}\delta_N^*$, that does not appear in the asymptotic distribution of LRT_∞. The shrinkage enables to enforce that $N^{\frac{1}{4}}\delta_N^* = o_p(1)$ and no longer $O_p(1)$.

2.3.3 . Extension to the non identically distributed setting

As in the previous section we first derive the consistency of the maximum likelihood estimator. To do so, we need the regularity required in assumption (2.2) to hold uniformly over the different distributions of the individuals.

Assumption 2.4. *We suppose that assumptions* (2.2) (i)–(v) *hold uniformly over the different individuals $i \in \mathbb{N}$.*

In addition to that, as discussed in Hoadley (1971) an additional assumption is required to ensure the unicity of the maximum of the asymptotic objective function.

Assumption 2.5. *For every $\theta \neq \theta_0$:*

$$\lim_{N\to+\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}\left[\log\left\{\frac{f_i(y_i;\theta)}{f_i(y_i;\theta_0)}\right\}\right] < 0$$

.

Proposition 2.6. *Under assumptions* (2.1), (2.4) *and* (2.5)*, propositions 2.2 and 2.4 still hold in the non identically distributed case.*

Following the same lines as in the last section the result of theorem 2.1 still holds.

Theorem 2.2. *Under assumptions* (2.1)–(2.5)*, if θ_N^* is chosen such that $\theta_N^* \in \Theta_0$, $\theta_N^* = \theta_0 + o_p(1)$, i.e. $\lambda_N^* = 0$ and $N^{\frac{1}{4}}\delta_N^* = o_p(1)$ then as $N \to +\infty$, it holds in probability that*

$$\mathsf{pr}^*\{\mathsf{lrt}(y_{1:N}^*) \leq t\} - \mathsf{pr}(\mathsf{lrt}(y_{1:N}) \leq t) = o_p(1). \tag{2.6}$$

2.3.4 . Sufficient verifiable conditions for regularity assumptions

Assumptions (2.2) state regularity conditions on the model. These assumptions can be straightforward to verify in some models (see details of the calculation in a linear mixed effects model in Appendix 2.7.13). Nevertheless, in most of the nonlinear cases these assumptions are very difficult to check, in particular due to the non explicit integrated form of the likelihood in (2.2). Nie (2006) proposed some verifiable conditions under which the maximum likelihood estimators in nonlinear mixed models is strongly consistent. However in his work he considered that the true parameter is an interior point of the parameter space, and that the Fisher information matrix is nonsingular. Furthermore in our context the conditions required are even more difficult to verify as we deal not only with maximum likelihood estimator consistency but also with likelihood ratio and Bootstrap statistic consistency.

We propose an analytical sufficient criterion for nonlinear mixed models that only depends on the regularity of the known function g in model (2.1) when g is nonlinear in the random effects ξ_i. We first state a regularity condition on the derivatives of g.

Assumption 2.6. *For every ξ, g is 4 times differentiable on Θ, and for $k_1 = 0, ..., 4$ and $k_2 \in \mathbb{N}$:*

$$\sup_{i\in\mathbb{N},j=1,...,J_i} \mathbb{E}\{ \sup_{\theta\in\Theta} \|\nabla_\theta^{k_1} g(x_{ij},\beta,\Lambda\xi)\|^{k_2}\} < +\infty, \quad \xi \sim \mathcal{N}(0, I_p). \tag{2.7}$$

Remark 2.7. *This assumption seems very strong but in practice it only requires that the derivatives of g are not exponential in $\|\xi\|^2$ which is verified by almost every commonly used models.*

We now state the regularity condition on the function g, which is the proposed criterion to be verified case by case in real life applications.

Assumption 2.7. *For every $\varepsilon > 0$, there exists a compact set $K \subset \mathbb{R}^p$ such that*

$$\forall \xi \in \mathbb{R}^p \backslash K \quad \sup_{i\in\mathbb{N},,j=1,...,J_i} \sup_{\theta\in\Theta} \frac{\|g(x_{ij},\beta,\Lambda\xi)\|}{\|\xi\|} \leq \varepsilon. \tag{2.8}$$

Proposition 2.7. *Suppose that assumption* (2.1) *holds, and that the function g verifies assumption* (2.6)–(2.7)*, then assumption* (2.4) *is verified.*

Remark 2.8. *Models with a bounded function g verify this property for any compact set K (see for example a common pharmacokinetic model presented in (Davidian and Giltinan, 2017) and detailed in Appendix 2.7.15). Regarding models with an unbounded function g, for example as the logistic growth model (Pinheiro and Bates, 2006) used in the experiments section 2.4.1, one can verify case by case that this criterion is satisfied (details of calculations are given in Appendix 2.7.14).*

2.4 . Experiments

2.4.1 . Simulation study

We denote by $\theta_0 = (\beta_0, \Lambda_0, \sigma_0^2)^T$ the true parameter used to generate the data. We use the notation β_k for the kth component of the fixed effects vector β, and we write diag$(x_1, ..., x_p)$ for a diagonal $p \times p$ matrix, with a diagonal being equal to $(x_1, ..., x_p)^T$. When it is not explicitly written we consider diagonal matrices $\Lambda = \text{diag}(\lambda_1, ...\lambda_p)^T$. The same way we write $\xi_i = (\xi_{i1}, ..., \xi_{ip})^T$ for the vector of random effects. We consider a linear and a nonlinear mixed effects models, with a varying number of random effects to account for the presence of nuisance parameters. Results were obtained using the `lme4` and `saemix` packages in R. Codes are available upon request from the first author.

We first consider the linear case. We denote by m_1 the linear model with two independent random effects, i.e. with $g(x_{ij}, \beta, \Lambda\xi_i) = \beta_1 + \lambda_1\xi_{i1} + (\beta_2 + \lambda_2\xi_{i2})x_{ij}$. We set $\beta_0 = (0, 7)^T$, $\lambda_{01} = 1.3, \lambda_{02} = 0$. In this model, we consider the test $H_0 : \lambda_2 = 0$ against $H_1 : \lambda_2 \geq 0$. We then denote by m_2 the linear model with three independent random effects, i.e. with $g(x_{ij}, \beta, \Lambda\xi_i) = \beta_1 + \lambda_1\xi_{i1} + (\beta_2 + \lambda_2\xi_{i2})x_{ij} + (\beta_3 + \lambda_3\xi_{i3})x_{ij}^2$. We set $\beta_0 = (0, 7, 3)^T$, $\lambda_{01} = 1.3, \lambda_{02} = 0, \lambda_{03} = 0$. In this model, we consider the test $H_0 : \lambda_3 = 0$ against $H_1 : \lambda_3 \geq 0$, so that in this simulation λ_2 is a nuisance parameter. For the choice of the shrinkage Bootstrap parameter, we set $c_N = aN^{-\nu}$ with $a = 0.5$ and $\nu = 0.2$, similarly to Cavaliere et al. (2020). This choice is motivated by the theoretical convergence assumptions on c_N in Proposition 2.5. This parameter shrinks to zero the variances of the individual parameters $\beta_2 + \lambda_2\xi_{i2}$ with a relative standard deviation lower than 4%. In both settings, we set $x_{ij} = j$, $J = 5$ and $\sigma_0^2 = 1.5$. Finally, we denote by m_3 the linear model with $p = 8$ random effects and a varying number s of nuisance parameters, i.e. with $g(x_{ij}, \beta, \Lambda\xi_i) = \sum_{k=1}^{p} x_{ijk}\lambda_k\xi_{ik}$. Here we set $N = 40$, $J_i = 9$, $\sigma^2 = 1$, and every untested variance to 1. Finally we draw independently the covariates from a normal distribution with mean 2 and standard deviation 0.5. We want here to illustrate the effect of an increasing number of nuisance parameters on the performance of the test. We use three different values for the shrinkage parameter $c_N \in \{0; 0.24; 0.9\}$. We chose those values to consider three cases : first $c_N = 0$ is equivalent to the parametric Bootstrap procedure without shrinkage, then $c_N = 0.5 \times 40^{-0.2} \approx 0.24$ shrinks most of the nuisance parameters toward 0, and finally $c_N = 0.9$ shrinks systematically the nuisance parameters (as if we were using the true model), but can also shrink some non-zero variances of the model. We consider the test $H_0 : \lambda_1 = 0$ against $H_1 : \lambda_1 \geq 0$.

Next, we consider the nonlinear logistic model with three random effects denoted by m_4, where

$$g(x_{ij}, \beta, \Lambda\xi_i) = \frac{\beta_1 + \lambda_1\xi_{i1}}{1 + \exp\left\{-\frac{x_{ij} - (\beta_2 + \lambda_2\xi_{i2})}{\beta_3 + \lambda_3\xi_{i3}}\right\}}. \tag{2.9}$$

Table 2.1 – Empirical levels (expressed as percentages) of the test that one variance is null in a linear model with two independent random effects, for $K = 5000$ simulated datasets and $B = 500$ Bootstrap replicates. The last column gives the maximal standard deviation value obtained in each row

Level α	$N = 10$		$N = 20$		$N = 30$		$N = 40$		$N = 100$		max sd
	boot	asym	boot	asym	boot	asym	boot	asym	boot	asym	
1%	1.14	0.68	0.98	0.68	1.20	0.94	0.74	0.70	0.86	0.72	0.15
5%	5.20	3.64	5.22	3.82	5.74	4.30	4.86	3.94	5.26	4.50	0.33
10%	10.72	7.16	10.80	7.98	10.30	8.40	10.80	8.44	10.34	8.86	0.44

boot., parametric Bootstrap procedure; asym., asymptotic procedure; sd, standard deviation .

We set $\beta_0 = (200, 500, 150)^T$, $\lambda_{01} = \lambda_{02} = 10$, $\lambda_{03} = 0$ and $\sigma_0^2 = 5^2$. We set $(x_{i1}, ..., x_{iJ}) = (50, 287.5, 525, 762, 1000, 1100, 1200, 1300, 1400, 1500)$ for all i. In this model, we consider the test $H_0 : \lambda_3 = 0$ against $H_1 : \lambda_3 \geq 0$.

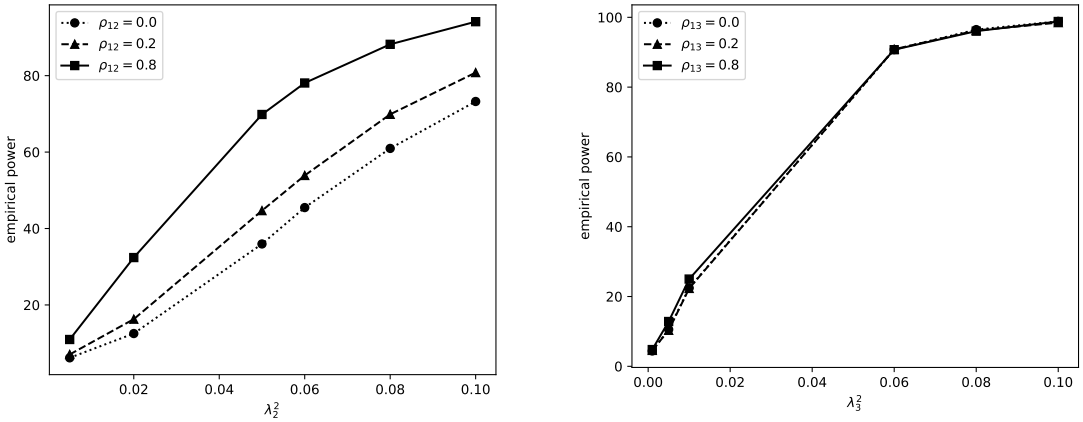
First, we study the finite sample size properties of our procedure using models m_1, m_2 and m_4. We compute the empirical levels by generating K datasets under the null hypothesis as described in the previous paragraph, and by computing the proportion of these datasets for which we reject the null hypothesis, for a nominal level α in $\{0.01, 0.05, 0.10\}$ and a sample size N in $\{10, 20, 30, 40, 100\}$ for m_1, N in $\{20, 30, 40\}$ for m_2 and $N = 40$ for m_4. Results are given in tables 2.1, 2.2 and 2.3. We compare the empirical level of the test associated with our Bootstrap procedure with those obtained using the asymptotic distribution which is a $0.5 - 0.5$ mixture between a Dirac distribution at zero and a chi–squared distribution with one degree of freedom (Baey et al., 2019). We observe that the empirical levels obtained with our Bootstrap procedure are closer to the nominal ones than those obtained with the asymptotic procedure, and that good results are already obtained for small values of N in the linear case. As expected, our procedure exhibits better small sample size properties than the asymptotic procedure, both in the linear and the nonlinear cases. It is noteworthy to mention that the existing non-asymptotic test procedures such as the one proposed by Drikvandi et al. (2013b) can not be used in the latter case since they rely on explicit expressions for the parameter estimates, hence requiring the linearity assumption. We also observe that the presence of nuisance parameters also deteriorate the asymptotic results. It is not a surprise as it modifies the true asymptotic distribution. However we observe that the standard parametric Bootstrap procedure is robust to the presence of a single nuisance parameter, and so the choice of c_N does not have a significant effect on this example. This must be due to the low number of nuisance parameters and the few number of parameters of the model.

We then study the empirical power of our procedure using models m_1 and m_2 for $N = 30$. To this end, we consider a non diagonal matrix Λ, introducing a correlation between the components of the scaled random effects b_i. We denote by ρ_{kl} the correlation coefficient between

Table 2.2 – Empirical levels (expressed as percentages) of the test that one variance is null in a linear model with three random effects and one nuisance parameter, for $K = 5000$ simulated datasets and $B = 500$ Bootstrap replicates. The last column gives the maximal standard deviation value obtained in each row.

Level α	$N = 20$		$N = 30$		$N = 40$		max sd
	boot	asym	boot	asym	boot	asym	
1%	0.82	0.66	0.72	0.58	0.90	0.62	0.13
5%	4.46	3.54	3.96	3.28	4.14	3.08	0.29
10%	8.88	6.78	7.52	6.34	8.40	6.98	0.40

boot., parametric Bootstrap procedure ; asym., asymptotic procedure ; sd, standard deviation.



(a) Model m_1 (b) Model m_2

Figure 2.1 – Empirical power of the test that one variance is null in a linear model with (a) two random effects and (b) three random effects, for varying value of the tested variance and of the correlation coefficient, for $K = 2500$ simulated datasets and $B = 500$ boostrap replicates.

the scaled random effects b_{ik} and b_{il}. We then consider increasing values of λ_2 and ρ_{12} in m_1, and increasing values of λ_3 and ρ_{13} in m_2. Results are given in figure 2.1. As expected, we observe that, for fixed values of the correlation coefficient, the empirical power increases when the true value of the tested variance increases, and that, for fixed values of the variance, the power increases when the correlation coefficient increases. In m_1, since $\beta_2 = 7$, we obtain an empirical power of at least 70% for a relative standard deviation of 4.5% (i.e. when $\lambda_2^2 = 0.1$). In m_2, since $\beta_3 = 3$, the empirical power is greater than 12.5% for a relative standard deviation of 4.7% (i.e. when $\lambda_3^2 = 0.02$), and above 90% for a relative standard deviation of 10% (i.e. when $\lambda_3^2 = 0.1$).

We then study the effect of shrinkage on the type I error using model m_3. Results are presented in Figure 2.2 for a theoretical level of 5%. We see that the procedure is sensitive to extreme

Table 2.3 – Comparison of the bootsrap procedure and the asymptotic procedure in the test in m_4, using $K = 1000$ datasets of size $N = 40$ and $B = 300$ Bootstrap replicates.

Level α	boot	asym	max sd
1%	0.80	0.80	0.28
5%	5.10	3.60	0.70
10%	10.30	7.00	0.96

boot., parametric Bootstrap procedure; asym., asymptotic procedure; sd, standard deviation.

values of c_N. The performances of the shrinked Bootstrap procedure are stable as the number of nuisance parameters increases, provided that the shrinkage parameter c_N is carefully chosen, whereas the performances of the regular Bootstrap procedure with no shrinkage are downgraded in this context. Indeed choosing a value of $c_N \approx 0.24$, i.e. that shrinks most of the nuisance parameters, provides good results while choosing $c_N = 0.9$, i.e. of the same order of magnitude as the non-zero variances of the model (here, $\lambda = 1$) deteriorates the results. On the other hand, neglecting the nuisance parameters, which corresponds to the case $c_N = 0$, also as an influence on the results and leads to an empirical level which is smaller than the theoretical one.

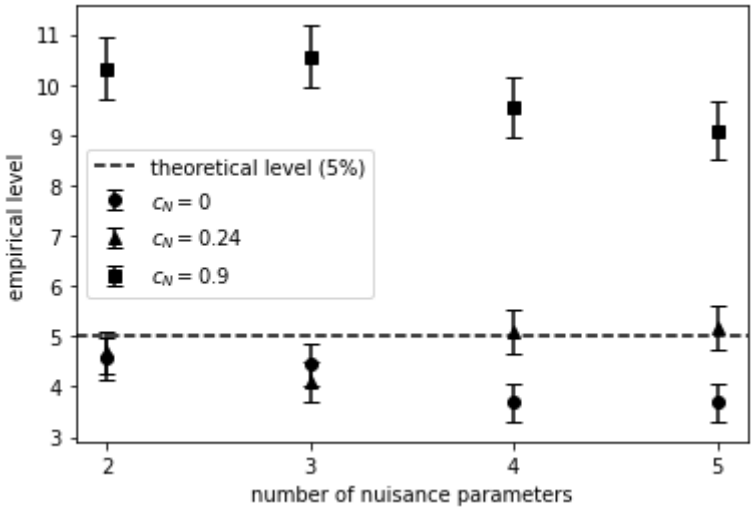


Figure 2.2 – Comparison of the parametric Bootstrap procedure and the shrinked parametric Bootstrap procedure in m_3, using $K = 2500$ datasets of size $N = 30$ and $B = 300$ Bootstrap replicates for varying numbers of nuisance parameters, and different shrinkage parameters c_N.

The previous experiment suggests that shrinking nonzero variances deteriorates the performance of the procedure. However it is difficult to distinguish between nonzero estimations of nuisance parameters and nonzero estimations of small variances. Therefore we carry out a simulation study to investigate the robustness of the procedure. In particular, we are inter-

Table 2.4 – Empirical levels (expressed as percentages) of the test that one variance is null in a model with three independent random effects (m_2), with one growing nonzero variance parameter shrinked, for $K = 2500$ simulated datasets and $B = 500$ Bootstrap replicates. case(1) corresponds to the case where the shrinkage parameter is equal to the true value of the growing variance, case (2) corresponds to a systematic shrinkage of the growing variance. On the left, the growing variance is λ_2^2, on the right it is λ_1^2.

λ_1^2	case (1)	case (2)	max sd	λ_2^2	case (1)	case (2)	max sd
0.001	4.96	5.08	0.44	0.001	4.16	4.84	0.43
0.01	8.92	9.04	0.57	0.01	5.4	5.48	0.46
0.1	19.44	20.12	0.80	0.1	5.24	5.76	0.47

sd, standard deviation

ested in evaluating through simulation the behavior of the type one error when considering a shrinking parameter of the same order than the true value of some of the variances. The idea is to mimic a practical shrinking procedure were the threshold would be chosen equal to the estimates of the variance parameter. To this end, we consider two experiments using model m_2 with $\beta_0 = (0, 7, 3)^T$, $\lambda_{03} = 0$ and $\sigma_0^2 = 1.5$. In both cases we test $H_0 : \lambda_3 = 0$ against $H_1 : \lambda_3 \geq 0$ at the level 5%. In the first experiment we set $\lambda_{01} = 1.3$ and consider for λ_{02} the different values $\{0.001, 0.01, 0.01\}$, while in the second experiment we set $\lambda_{02} = 1.3$ and consider for λ_{01} the different values $\{0.001, 0.01, 0.01\}$. In the first (respectively second) experiment, λ_2 (respectively λ_1) is the supposed nuisance parameter that will be mistakenly shrinked. We study the effect of two different shrinkage acting only on the supposed nuisance parameter. In the first case, the shrinkage parameter is chosen of the same order of magnitude as the small untested variance, i.e. we set $c_N = \lambda_{02}$ (respectively $c_N = \lambda_{01}$) in the first (respectively second) experiment. This leads to a procedure that will often shrink the small variance parameter, even though its true value is strictly nonnegative. In the second case, we set the Bootstrap parameter $\lambda_{2N}^* = 0$ (respectively $\lambda_{1N}^* = 0$) so that the nuisance parameter is always shrinked. Results are presented in Table 2.4. These experiments suggest that shrinking nonzero variances can highly deteriorate the level of the test, especially when the corresponding variance contributes heavily to the total variance of the model. Indeed, the impact of shrinking a small variance in model m_2 is higher when considering λ_2 rather than λ_1. This might be due to the fact that the variance associated to one observation is given by $\text{var}(y_{ij}) = \lambda_1^2 + j^2\lambda_2^2 + \sigma^2$ (with $1 \leq j \leq 5$), so that the part of the total variance that can be attributed to λ_2 is higher than the part associated to λ_1. However, it also shows that the level can still be preserved when the values of the shrinked variances remain small compared to the total variance given by the model. Overall, this suggests that an expert-based point of view can be adopted for the choice of c_N, guided by the applications. For example, it could be defined as a proportion of the total variance above which the variability of the effect should be taken into account.

2.4.2 . Real data application

We then apply our procedure to a study of white-browed coucal growth rates, available as a Dryad package (Goymann et al., 2016). We use the logistic growth model defined in (2.9) to describe the evolution of the body mass of the nestlings as a function of their age. More precisely, if we denote by y_{ij} the body mass of nestling i at age t_j for $i = 1, ..., 292$, $j = 1, ..., N_i$, we have that $\beta_1 + \lambda_1\xi_{i1}$ is the asymptotic nestling body mass, $\beta_2 + \lambda_2\xi_{i2}$ is the age (in days) at which nestling i reaches half its asymptotic body mass and $\beta_3 + \lambda_3\xi_{i3}$ is the growth rate of nestling i. When fitting the complete model the estimated scaling matrix is $\hat{\Lambda} = \text{diag}(\sqrt{212.34}, \sqrt{0.89}, \sqrt{0.02})$, which motivates the test that λ_2 and λ_3 are null.

In order to test for the presence of randomness in the inflexion point and in the growth rate, we proceed sequentially. First, we test if the variances of both the inflexion point and the growth rate are null, i.e. we consider the test $T_1 : H_0 : \lambda_2 = 0, \lambda_3 = 0$ against $H_1 : \lambda_2 \geq 0, \lambda_3 \geq 0$. If the null hypothesis in T_1 is rejected, we perform two univariate tests, one for each variance tested in T_1. More precisely, we consider the tests $T_2 : H_0 : \lambda_3 = 0$ against $H_1 : \lambda_3 \geq 0$ and $T_3 : H_0 : \lambda_2 = 0$ against $H_1 : \lambda_2 \geq 0$. If the null hypothesis is rejected in T_1, it means that at least one of the two tested variances is nonzero, thus we can then consider the univariate tests T_2 and T_3, where λ_2 and λ_3 might be nuisance parameters. For each test, we consider two procedures, one where the estimate of the potential nuisance parameter (λ_2 in T_2 and λ_3 in T_3) is shrinked toward zero, and another one without shrinkage. This choice enables to consider the two possible cases : $\lambda_2 > c_N$ and $\lambda_2 < c_N$ for T_2 and $\lambda_3 > c_N$ and $\lambda_3 < c_N$ for T_3. In practice the practitioner can choose the threshold according to his own level of significance of variability desired. For instance by saying that under X% of relative variability of a parameter, we consider it as a fixed effect. Table 2.5 compiles the results of the three tests.

Table 2.5 – Comparison of the p–value (in %) among the three tests $T1, T2$ and $T3$, using $B = 1000$ Bootstrap replicates.

Test	T1		T2		T3
lrt	302.5		1.6		278.2
procedure	no shrink	shrink	no shrink	shrink	no shrink
p–value	0	8.9	7.9	0	0

shrink., shrinked parametric Bootstrap procedure; no shrink., regular parametric Bootstrap procedure without shrinkage

We can see that the procedures lead to different p-values, and can thus, in practice, lead to different conclusions with respect to the null hypothesis depending on the type I error considered.

2.5 . Discussion

This work can lead to several future developments, both from a theoretical and a practical point of view. In particular, the choice of the shrinking parameter c_N is of interest for practitioners that would want to apply our procedure. We showed that in some cases it can have a large impact on the estimated level, especially when it leads to the shrinkage of variances that amount for a large part of the total variance. This suggests that the choice of this tuning parameter could be tackled from an expert-based point of view, based on the expected variability of each parameter. For example, c_N could be defined as a threshold under which the variability of a random effect is not relevant for the task of interest, as long as it does not account for a significant part of the total variance. However, this compromise is model-dependent and should be considered according to each specific application. From a methodological point of view, it would be interesting to propose an automated procedure, following for example the idea of Bickel and Sakov (2008).

In a model building approach, our procedure presents the advantage of dealing with nuisance parameters, therefore sequential tests can be performed as we did in the real data applications. This idea is promising to select the exact number of random effects to consider, however such sequential tests present other issues which require further development to be addressed carefully such as multiple testing and post selection concerns that are beyond the scope of this paper.

From a computational point of view, our algorithm can be used indifferently for linear and nonlinear mixed effects models, however the computation time can be prohibitive in the latter case, especially for complex models. It would be interesting to find a criterion to optimize the choice of the Bootstrap sample size. Another issue when dealing with complex models is the computation of the likelihood ratio test statistic. Indeed, it involves a ratio of likelihoods which are usually estimated separately using Monte Carlo approaches, leading to a biased estimate. This point is crucial since this quantity is calculated at each iteration of the Bootstrap procedure.

2.6 . Acknowledgment

This work was funded by the Stat4Plant project ANR-20-CE45-0012.

2.7 . Supplementary material

We recall that we consider the following nonlinear mixed effects model, for any $i = 1, ..., N$ and any $j = 1, .., J_i$:

$$\begin{cases} y_{ij} & = g(x_{ij}, \beta, \Lambda\xi_i) + \varepsilon_{ij} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \xi_i & \sim \mathcal{N}(0, I_p) \end{cases} \tag{2.10}$$

We also recall that the log–likelihood of the model given the data $y_{1:N}$ writes :

$$l(\theta; y_{1:N}) = \log\{L_\theta(y_{1:N})\} = \log\{\prod_{i=1}^{N} f_i(y_i; \theta)\} = \sum_{i=1}^{N} \log\{\int f_i(y_i; \xi_i, \theta)\pi(\xi_i)d\xi_i\}. \tag{2.11}$$

Finally we write

$$\hat{\theta}_N = \arg\sup_{\theta \in \Theta} \quad l(\theta; y_{1:N}). \tag{2.12}$$

2.7.1 . Different parametrizations for mixed effects models

The commonly used parametrization for nonlinear mixed effects models (Pinheiro and Bates (2006) page 306) is for $i = 1, ..., N$, $j = 1, ..., J_i$:

$$\begin{cases} y_{ij} & = g(v_{ij}, \phi_i) + \varepsilon_{ij} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \phi_i & = A_{ij}\beta + B_{ij}b_i \quad b_i \sim \mathcal{N}(0, \Gamma) \end{cases} \tag{2.13}$$

where v_{ij}, A_{ij} and B_{ij} are known covariates. With this definition, the log–likelihood is defined as follows :

$$l(\theta; y_{1:N}) = \log\{L_\theta(y_{1:N})\} = \log\{\prod_{i=1}^{N} f_i(y_i; \theta)\} = \sum_{i=1}^{N} \log\{\int f_i(y_i; b_i, \theta)\pi(b_i; 0, \Gamma)db_i\}, \tag{2.14}$$

where $f_i(y_i; b_i, \theta)$ is the density of the conditional distribution of y_i given b_i, and $\pi(b_i; 0, \Gamma)$ is the density of the p-dimensional centered Gaussian distribution with covariance Γ.

The change of variable $b_i = \Lambda\xi_i$ is a $\mathcal{C}^1$ diffeomorphism if and only if the diagonal coefficients of Λ are strictly nonnegative. Therefore in our setting where this condition is not verified these two parametrizations are not equivalent.

In particular taking $\Lambda = 0$ in (2.2) is equivalent to considering a fixed-effects nonlinear model, while in (2.14), taking $\Gamma \to 0$ makes $l(\theta; y_{1:N}) \to -\infty$.

2.7.2 . Proof of proposition 2.1

To prove the proposition we only have to prove that if the mth column of Λ is 0, then the odds orders of the partial derivatives of the log-likelihood with respect to the elements of this column are null regardless of the values of the $y_{1:N}$. We have, for all $n = 1, .., p$:

$$\frac{\partial f_i(y_i; \theta)}{\partial[\Lambda]_{nm}}|_{\theta=\theta_0} = \int_{\xi_1} \cdots \int_{\xi_p} \frac{\partial f_i(y_i; \xi, \theta_0)}{\partial[\Lambda]_{nm}} \pi(\xi_1)d\xi_1 ... \pi(\xi_p)d\xi_p$$

Given the definition of model 2.1 :

$$f_i(y_i;\xi,\theta) = (2\pi\sigma^2)^{-\frac{J_i}{2}} exp\left[-\frac{\sum_{j=1}^{J_i}\{y_{ij}-g(x_{ij},\beta,\Lambda\xi)\}^2}{2\sigma^2}\right]$$

therefore,

$$\begin{aligned}\frac{\partial f_i(y_i;\xi,\theta)}{\partial[\Lambda]_{nm}} &\propto \frac{\partial\Lambda\xi}{\partial[\Lambda]_{mn}}\nabla_{\Lambda\xi}\exp\left[-\frac{\sum_{j=1}^{J_i}\{y_{ij}-g(x_{ij},\beta,\Lambda\xi)\}^2}{2\sigma^2}\right]\\ &\propto \xi_m\nabla_{\Lambda\xi}\exp\left[-\frac{\sum_{j=1}^{J_i}\{y_{ij}-g(x_{ij},\beta,\Lambda\xi)\}^2}{2\sigma^2}\right]\end{aligned}$$

Evaluated at $\theta=\theta_0$ the last term (the gradient) no longer depends on ξ_m as the mth column of Λ is null.

$$\begin{aligned}\frac{\partial f_i(y_i;\theta)}{\partial[\Lambda]_{nm}}|_{\theta=\theta_0} \propto &\int_{\xi_m}\xi_m\pi(\xi_m)d\xi_m\\ &\times\int\nabla_{\Lambda\xi}\exp\left[-\frac{\sum_{j=1}^{J_i}\{y_{ij}-g(x_{ij},\beta,\Lambda\xi)\}^2}{\sigma^2}\right]|_{\theta=\theta_0}\prod_{l\neq m}\pi(\xi_l)d\xi_l\end{aligned}$$

which is equal to 0 as the first term of the right hand side equation is expectation of a standard gaussian distribution. We can apply the same reasoning to every odds order derivatives.

2.7.3 . Proof of proposition 2.2

Due to assumption (2.1) we have that :

$$\sup_{\theta\in\Theta\setminus\{\theta_0\}}\mathbb{E}\{l(\theta;y_1)\} < \mathbb{E}\{l(\theta_0;y_1)\} \tag{2.15}$$

which comes from the identifiability of the model and the positivity of the Kullback-Leibler divergence. Assumption (2.2)$i)$ enables to apply the uniform law of large number to the log–likelihood. Then the result follows from arguments as in Andrews (1993) lemma A.1.

2.7.4 . Proof of proposition 2.4

To prove the consistency of the Bootstrap maximum likelihood estimator, we will use the same reasoning, as in the proof of proposition 2.2. The sketch of the proof is similar to the one of Cavaliere et al. (2020).

We first want to show that :

$$\sup_{\theta\in\Theta}|\frac{1}{N}l(\theta;y^*_{1:N}) - \mathbb{E}\{l(\theta;y_1)\}| = o_{p^*}(1)$$

First of all we have that :

$$\sup_{\theta\in\Theta}|\frac{1}{N}l(\theta;y^*_{1:N}) - \mathbb{E}\{l(\theta;y_1)\}| \leq \sup_{\theta\in\Theta}A^*_N(\theta) + \sup_{\theta\in\Theta}A_N(\theta)$$

where :

$$A_N^*(\theta) = |\frac{1}{N} l(\theta; y_{1:N}^*) - \mathbb{E}^* \{l(\theta; y_1^*)\}|$$
$$A_N(\theta) = |\mathbb{E}^* \{l(\theta; y_1^*)\} - \mathbb{E} \{l(\theta; y_1)\}|$$

We now want to apply the uniform law of large numbers to $A_N(\theta)$, and it's Bootstrap version to $A_N^*(\theta)$. Therefore we shall show that both term converges toward 0 and that they are lipshitz. We first consider $A_N(\theta)$ for a given $\theta \in \Theta$.

$$\begin{aligned}\mathbb{E}^* \{l(\theta; y_1^*)\} &= \mathbb{E}\{l(\theta; y_1^*)|y_{1:N}\} \\ &= \int l(\theta; y_1) f(y_1; \theta_N^*) dy_1\end{aligned}$$

Using assumption (2.1), we can state that there exist θ^+ between θ_0 and θ_N^* such that,

$$\begin{aligned}|f(y_1; \theta_N^*) - f(y_1; \theta_0)| &\leq \|\theta_0 - \theta_N^*\| \|\nabla_\theta f(y_1; \theta^+)\| \\ &\leq \|\theta_0 - \theta_N^*\| \|\nabla_\theta \log\{f(y_1; \theta^+)\}\| f(y_1; \theta^+)\end{aligned}$$

therefore,

$$\begin{aligned}A_N(\theta) &\leq \int |l(\theta; y_1)| |f(y_1; \theta_N^*) - f(y_1; \theta_0)| dy_1 \\ &\leq \int |l(\theta; y_1)| \|\theta_0 - \theta_N^*\| \|\nabla_\theta \log\{f(y_1; \theta^+)\}\| f(y_1; \theta^+) dy_1 \\ &\leq \|\theta_0 - \theta_N^*\| \int |l(\theta; y_1)| \|\nabla_\theta l(\theta^+; y_1)\}\| f(y_1; \theta^+) dy_1\end{aligned}$$

Due to assumption (2.2)(i)–(ii), $|l(\theta; y_1)| \|\nabla_\theta l(\theta^+; y_1)\}\|$ is integrable with respect to the density $f(y_1; \theta^+)$. And finally using the elementary inequality ,

$$\frac{1}{2}(a^2 + b^2) \geq |ab|, \quad \forall a, b \in \mathbb{R} \tag{2.16}$$

we can state that

$$A_N(\theta) \leq \frac{1}{2}\|\theta_0 - \theta_N^*\| \sup_{\theta^+ \in \Theta} \int \sup_{\Theta \in \Theta} |l(\Theta; y_1)|^2 + \sup_{\theta_2 \in \Theta} \|\nabla_\theta l(\theta_2; y_1)\}\|^2 f(y_1; \theta^+) dy_1$$

Finally thanks to assumption (2.2), as $N \to +\infty$, it holds in probability that :

$$A_N(\theta) \to 0$$

We now consider :

$$A_N^*(\theta) = |\frac{1}{N} \sum_{i=1}^{N} l(\theta; y_i^*) - \mathbb{E}^* \{l(\theta; y_1^*)\}|$$

This quantity is a sum of conditionally independent and centered random variables. We can't directly apply a law of large number as the parameter θ_N^* and the index of the sum depends both on N.

For every real nonnegative number t, it holds almost surely that :

$$\text{pr}^*(A_N^* > t) \leq \text{pr}^*\{\frac{1}{N}\sum_i |l(y_i^*;\theta) - \mathbb{E}^*\{l(\theta;y_1^*)\}| > t\} \leq \frac{\sup_{\theta'\in\Theta}\mathbb{E}_{\theta'}\{\sup_{\theta\in\Theta}|l(\theta;y_1)|^2\}}{Nt^2}$$

by applying first triangular inequality and then Chebychev inequality, using assumption (2.2)i). And finally $A_N^*(\theta) \to 0$ in probability, as $N \to +\infty$, which concludes the pointwise convergence. And we note that this result holds uniformly over Θ so :

$$\sup_{\theta\in\Theta}|\frac{1}{N}l(\theta;y_{1:N}^*) - \mathbb{E}\{l(\theta;y_1)\}| = o_{p^*}(1)$$

Let now use this result to show that the Bootstrap maximum likelihood estimator is consistent.

Let $\varepsilon > 0$, using equation (2.15), there exists $\delta > 0$ such that :

$$\inf_{\|\theta-\theta_0\|>\varepsilon}\mathbb{E}\{l(\theta_0;y_1)) - \mathbb{E}(l(\theta;y_1)\} \geq \delta$$

Let us introduce $\theta_{mle}^B = \arg\max_{\theta\in\Theta} \quad l(\theta;y_{1:N}^*)$.

By writing $V_\varepsilon = \{\theta \in \Theta : \|\theta - \theta_0\| > \varepsilon\}$, we have that :

$$\begin{aligned}
\text{pr}^*(\theta_{mle}^B \in V_\varepsilon) &\leq \text{pr}^*\left[\mathbb{E}\{l(\theta_0;y_1)\} - \mathbb{E}\{l(\theta_{mle}^B;y_1^*)\} \geq \delta\right] \\
&= \text{pr}^*\left[\mathbb{E}\{l(\theta_0;y_1)\} - \frac{1}{N}l(\theta_{mle}^B;y_{1:N}^*) + \frac{1}{N}l(\theta_{mle}^B;y_{1:N}^*) - \mathbb{E}\left\{l(\theta_{mle}^B;y_1^*)\right\} \geq \delta\right] \\
&\leq \text{pr}^*\left[\mathbb{E}\{l(\theta_0;y_1)\} - \frac{1}{N}l(\theta_0;y_{1:N}^*) + \frac{1}{N}l(\theta_{mle}^B;y_{1:N}^*) - \mathbb{E}\{l(\theta_{mle}^B;y_1^*)\} \geq \delta\right] \\
&\leq \text{pr}^*\left\{2\sup_{\theta\in\Theta}|\frac{1}{N}l(\theta;y_{1:N}^*) - \mathbb{E}\{l(\theta;y_1)\}| \geq \delta\right\} \\
&\leq o_p(1)
\end{aligned}$$

Which concludes the proof that $\theta_{mle}^B = \theta_0 + o_{p^*}(1)$. The exact same proof still holds for the restricted Bootstrap maximum likelihood estimator by replacing Θ by Θ_0.

2.7.5 . Quadratic approximation of the log–likelihood

To derive the asymptotic distribution of LRT_N, we expand the log-likelihood around θ_0 (see Andrews (1999) theorem 6). Under assumption (2.1), following the lines of Hiroyuki et al. (2012), we can write :

$$
\begin{aligned}
l(\theta;y_{1:N}) - l(\theta_0;y_{1:N}) &= (\psi-\psi_0)^T\nabla_\psi l(\theta_0;y_{1:N}) + \frac{1}{2}(\psi-\psi_0)^T\nabla^2_\psi l(\theta_0;y_{1:N})(\psi-\psi_0) \\
&+ (1/2)\sum_{i,j=1,\dots,d_\delta}\delta_i\delta_j\frac{\partial^2 l(\theta_0;y_{1:N})}{\partial\delta_i\partial\delta_j} + (3/3!)(\psi-\psi_0)^T\sum_{i,j=1,\dots,d_\delta}\delta_i\delta_j\frac{\partial^3 l(\theta_0;y_{1:N})}{\partial\delta_i\partial\delta_j\partial\psi} \\
&+ (1/2)\sum_{i,j=1,\dots,d_\lambda}\lambda_i\lambda_j\frac{\partial^2 l(\theta_0;y_{1:N})}{\partial\lambda_i\partial\lambda_j} + (3/3!)(\psi-\psi_0)^T\sum_{i,j=1,\dots,d_\lambda}\lambda_i\lambda_j\frac{\partial^3 l(\theta_0;y_{1:N})}{\partial\lambda_i\partial\lambda_j\partial\psi} \\
&+ (6/4!)\sum_{i,j=1,\dots,d_\delta}\sum_{k,l=1,\dots,d_\lambda}\delta_i\delta_j\lambda_k\lambda_l\frac{\partial^4 l(\theta_0;y_{1:N})}{\partial\delta_i\partial\delta_j\partial\lambda_k\partial\lambda_l} \\
&+ (1/4!)\sum_{i,j,k,l=1,\dots,d_\delta}\delta_i\delta_j\delta_k\delta_l\frac{\partial^4 l(\theta_0;y_{1:N})}{\partial\delta_i\partial\delta_j\partial\delta_k\partial\delta_l} \\
&+ (1/4!)\sum_{i,j,k,l=1,\dots,d_\lambda}\lambda_i\lambda_j\lambda_k\lambda_l\frac{\partial^4 l(\theta_0;y_{1:N})}{\partial\lambda_i\partial\lambda_j\partial\lambda_k\partial\lambda_l} + R_N(\theta)
\end{aligned}
$$

With $R_N(\theta)$ being the rest in the Taylor expansion . We define for all integers i,j $c_{ij} = \frac{1}{2}$ if $i=j$ and 1 otherwise. We also define $\mathcal{I}_\lambda = \{11,22,..,d_\lambda-1,d_\lambda\}$ and $\mathcal{I}_\delta = \{11,22,..,d_\delta-1,d_\delta\}$ that respectively index $v(\lambda) = (\lambda_i\lambda_j)_{ij\in\mathcal{I}_\lambda}$ and $v(\delta) = (\delta_i\delta_j)_{ij\in\mathcal{I}_\delta}$. For $ij\in\mathcal{I}_\lambda$ (respectively $\mathcal{I}_\delta$), we write $\frac{\partial^2 l(\theta;y_{1:N})}{\partial v(\lambda)_{ij}} = \frac{\partial^2 l(\theta;y_{1:N})}{\partial\lambda_i\lambda_j}$ (the same with δ). With these notations we have that :

$$
\begin{aligned}
l(\theta;y_{1:N}) - l(\theta_0;y_{1:N}) =& (\psi-\psi_0)^T\nabla_\psi l(\theta_0;y_{1:N}) + \frac{1}{2}(\psi-\psi_0)^T\nabla^2_\psi l(\theta_0;y_{1:N})(\psi-\psi_0) \\
&+ \sum_{i\in\mathcal{I}_\lambda} v(\lambda)_i c_i\frac{\partial^2 l(\theta;y_{1:N})}{\partial v(\lambda)_i} + (\psi-\psi_0)^T\sum_{i\in\mathcal{I}_\lambda} v(\lambda)_i c_i\frac{\partial^3 l(\theta_0;y_{1:N})}{\partial v(\lambda)_i\partial\psi} \\
&+ \sum_{i\in\mathcal{I}_\delta} v(\delta)_i c_i\frac{\partial^2 l(\theta;y_{1:N})}{\partial v(\delta)_i} + (\psi-\psi_0)^T\sum_{i\in\mathcal{I}_\delta} v(\delta)_i c_i\frac{\partial^3 l(\theta_0;y_{1:N})}{\partial v(\delta)_i\partial\psi} \\
&+ 4\times(6/4!)\sum_{i\in\mathcal{I}_\lambda,j\in\mathcal{I}_\delta} c_i c_j v(\lambda)_i v(\delta)_j\frac{\partial^4 l(\theta_0;y_{1:N})}{\partial v(\lambda)_i\partial v(\delta)_j} \\
&+ (4/4!)\sum_{i\in\mathcal{I}_\delta,j\in\mathcal{I}_\delta} c_i c_j v(\delta)_i v(\delta)_j\frac{\partial^4 l(\theta_0;y_{1:N})}{\partial v(\delta)_i\partial v(\delta)_j} \\
&+ (4/4!)\sum_{i\in\mathcal{I}_\lambda,j\in\mathcal{I}_\lambda} c_i c_j v(\lambda)_i v(\lambda)_j\frac{\partial^4 l(\theta_0;y_{1:N})}{\partial v(\lambda)_i\partial v(\lambda)_j} + R_N(\theta)
\end{aligned}
$$

We define the reparametrization of θ : $\phi(\theta) = (\psi, v(\delta), v(\lambda))$.

By writing $\tilde{\nabla}_{v(\lambda)} l(\theta;y_{1:N}) = \left(c_i\frac{\partial^2 l(\theta;y_{1:N})}{\partial v(\lambda)_i}\right)_{i\in\mathcal{I}_\lambda}$ (similarly for $\tilde{\nabla}_{v(\delta)} l(\theta;y_{1:N})$) and $\tilde{\nabla}^2_{v(\lambda)} l(\theta;y_{1:N}) = (c_i c_j\frac{\partial^4 l(\theta;y_{1:N})}{\partial v(\lambda)_i\partial v(\lambda)_j})_{i,j=1,\dots,d_\lambda}$ (similarly for $\tilde{\nabla}^2_{v(\delta)} l(\theta;y_{1:N})$), we also define :

$$
\tilde{S}_N(\theta_0) = \sqrt{N}^{-1}\left(\ \nabla_\psi l(\theta_0;y_{1:N})^T, \tilde{\nabla}_{v(\delta)} l(\theta_0;y_{1:N})^T, \tilde{\nabla}_{v(\lambda)} l(\theta_0;y_{1:N})^T\ \right)^T
$$

$$\tilde{I}_N(\theta_0) = \begin{pmatrix} I_{N,\psi}(\theta_0) & I_{N,\psi,v(\delta)}(\theta_0) & I_{N,\psi,v(\lambda)}(\theta_0) \\ I_{N,\psi,v(\delta)}(\theta_0)^T & I_{N,v(\delta)}(\theta_0) & I_{N,v(\delta),v(\lambda)}(\theta_0) \\ I_{N,\psi,v(\lambda)}(\theta_0)^T & I_{N,v(\delta),v(\lambda)}(\theta_0)^T & I_{N,v(\lambda)}(\theta_0) \end{pmatrix} \tag{2.17}$$

Where $I_{N,\psi}(\theta_0) = -\frac{1}{N}\nabla^2_\psi l(\theta_0; y_{1:N})$, $I_{N,\psi,v(\lambda)}(\theta_0) = \left(-\frac{c_i}{N}\frac{\partial^3 l(\theta_0;y_{1:N})}{\partial\psi\partial v(\lambda)_i}\right)_{.,i\in\mathcal{I}_\lambda}$, $I_{N,\psi,v(\delta)}(\theta_0) = \left(-\frac{c_i}{N}\frac{\partial^3 l(\theta_0;y_{1:N})}{\partial\psi\partial v(\delta)_i}\right)_{.,i\in\mathcal{I}_\delta}$, $I_{N,v(\lambda)}(\theta_0) = -\frac{1}{3N}\tilde{\nabla}^2_{v(\lambda)} l(\theta_0; y_{1:N})$, $I_{N,v(\delta)}(\theta_0) = -\frac{1}{3N}\tilde{\nabla}^2_{v(\delta)} l(\theta_0; y_{1:N})$, $I_{N,v(\lambda),v(\delta)}(\theta_0) = (-\frac{1}{N}c_i c_j \frac{\partial^4 l(\theta_0;y_{1:N})}{\partial v(\lambda)_i \partial v(\delta)_j})_{i\in\mathcal{I}_\lambda, j\in\mathcal{I}_\delta}$.

Where the notation $(a_i)_{.,i\in\mathcal{I}}$ stands for a $dim(a_i) \times card(\mathcal{I})$ matrix whose columns are a_i for $i \in \mathcal{I}$.

With these new notations we obtain the following quadratic approximation of the log–likelihood :

$$\begin{aligned} l(\theta; y_{1:N}) - l(\theta_0; y_{1:N}) = &\sqrt{N}(\phi(\theta) - \phi(\theta_0))^T \tilde{S}_N(\theta_0) \\ &- \frac{1}{2}\sqrt{N}(\phi(\theta) - \phi(\theta_0))^T \tilde{I}_N(\theta_0)\sqrt{N}(\phi(\theta) - \phi(\theta_0)) + R_N(\theta) \end{aligned} \tag{2.18}$$

2.7.6 . Asymptotic distribution of the likelihood ratio test statistic

We start from the expansion of the log–likelihood (2.18) derived in the last section, that we rewrite :

$$\begin{aligned} l(\theta; y_{1:N}) - l(\theta_0; y_{1:N}) = &\frac{1}{2} Z_N(\theta_0)^T \tilde{I}_N(\theta_0) Z_N(\theta_0) \\ &- \frac{1}{2}\left(t_N(\theta) - Z_N(\theta_0)\right)^T \tilde{I}_N(\theta_0)\left(t_N(\theta) - Z_N(\theta_0)\right) + R_N(\theta) \end{aligned}$$

where $t_N(\theta) = \sqrt{N}(\phi(\theta) - \phi(\theta_0))$ and $Z_N(\theta_0) = \tilde{I}_N(\theta_0)^{-1}\tilde{S}_N(\theta_0)$

Remark 2.9. *We consider the quantity $\tilde{I}_N(\theta_0)^{-1}$ which implies that $\tilde{I}_N(\theta_0)$ is non-singular which may not be always true. However under assumption* (2.3)*, the probability that $\tilde{I}_N(\theta_0)$ is non-singular tends to 1 as $N \to +\infty$.*

The set a feasible values for $t_N(\theta)$ is :

$$t_N(\Theta) = \{\sqrt{N}\left(\Theta_\psi - \psi_0\right)\} \times \{\sqrt{N}\left(v(\Theta_\lambda) - v(\lambda_0)\right)\} \times \{(v(\Theta_\delta) - v(\delta_0))\} \tag{2.19}$$

$t_N(\Theta)$ does not depend on N (it is a cartesian product whose terms are whether $\mathbb{R}$ or $[0, +\infty[$), and it is locally approximated by a cone (in fact it is a cone), we write it $\mathcal{C}(\Theta)$. Therefore if we prove that $R_N(\theta)$ is $o_p(1)$ when evaluated at the maximum likelihood estimator, we can apply the result from Andrews (1999). For more details see Andrews (1999, Theorem 3) and paragraph 4.3.

In order to apply the theory of Andrews, one shall prove that $\tilde{I}_N(\theta_0)$ converges in probability toward a nonnegative matrix $\tilde{I}(\theta_0)$, and that $\tilde{S}_N(\theta_0)$ converges weakly to a random variable $U(\theta_0)$.

1) Let $d_{\tilde{I}} \in \mathbb{N}$ such that $\tilde{I}_N(\theta_0) \in \mathbb{R}^{d_{\tilde{I}} \times d_{\tilde{I}}}$, let $1 \leq m, n \leq d_{\tilde{I}}$. We can write :

$$\left[\tilde{I}_N(\theta_0)\right]_{m,n} = \frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)}(\theta_0)$$

where $h_{m,n}^{(i)}(\theta_0)$ is of the form :

$$h_{m,n}^{(i)}(\theta_0) = c_{m,n}\frac{\partial^k \log f(y_i;\theta)}{\partial\theta_{i_1}...\partial\theta_{i_k}}|_{\theta=\theta_0}$$

with $c_{m,n} \in \mathbb{R}$, $k \in \{2, 4\}$, $1 \leq i_1, ..., i_k \leq d_\lambda + d_\psi + d_\delta$.

With assumption (2.2) we can apply the law of large numbers to this empirical mean and therefore it holds in probability that :

$$\tilde{I}_N(\theta_0) \underset{N\to+\infty}{\longrightarrow} \left[\mathbb{E}\{h_{m,n}^{(1)}(\theta_0)\}\right]_{1\leq m,n\leq q_I} = \tilde{I}(\theta_0)$$

2) We now consider $\tilde{S}_N(\theta_0)$.

First the score is centered,

$$\mathbb{E}\left[\nabla_\psi \log f(x_i;\theta_0)\right] = 0$$

then, due to proposition (2.1), $\forall m, n = 1, ..., q_\lambda$,

$$\mathbb{E}\left[\frac{\partial^2 \log f(y_i;\theta_0)}{\partial\lambda_m\partial\lambda_n}\right] = -\mathbb{E}\left[\frac{\partial \log f(y_i;\theta_0)}{\partial\lambda_m}\frac{\partial \log f(y_i;\theta_0)}{\partial\lambda_n}\right] = 0$$

We apply the central limit theorem to $\tilde{S}_N(\theta_0)$ which is a sum of independent and identically distributed centered random variables with finite variances, and therefore $\tilde{S}_N(\theta_0)$ converges weakly to a random variable $U(\theta_0)$

By doing the exact same quadratic approximation and development but considering the parameter space Θ_0, combining (2.19)–(2.21), we use Andrews (1999) to obtain :

$$LRT_\infty = \inf_{t\in\mathcal{C}(\Theta_0)} \|t - \tilde{I}(\theta_0)^{-1}U(\theta_0)\|_{\tilde{I}(\theta_0)} - \inf_{t\in\mathcal{C}(\Theta)} \|t - \tilde{I}(\theta_0)^{-1}U(\theta_0))\|_{\tilde{I}(\theta_0)} \tag{2.20}$$

Which leads to the expression of LRT_∞.

Remark 2.10. *On the matrix $\tilde{I}(\theta_0)$, it is obviously symmetric as the limit of* (2.17)*.It is also positive semi-definite because one can show that $\tilde{I}(\theta_0)$ is the asymptotic variance of $\tilde{S}_N(\theta_0)$. The proof can be found in a simpler case in the proof of proposition 2 (b) of Hiroyuki et al. (2012). The proof is based on the Fisher's identity and the fact that the odd derivatives with respect to λ and δ are zero.*

2.7.7 . Proof of proposition 2.3

To obtain an explicit form for $R_N(\theta)$ we use the multivariate version of Taylor-Lagrange formula, which is for instance defined in Andrews (1999) Theorem 6.

This way we have that $R_N(\theta)$ is a sum of higher order derivatives with respect to ψ and the fourth crossed derivatives with respect to λ. Given assumption (2.2) all the derivatives of the log-likelihood are $\mathcal{O}_p(N)$, using Cauchy-Schwartz inequality and the fact that $\|t_N(\theta)\|^2 = N\left(\|\psi-\psi_0\|^2+\|v(\delta)\|^2+\|v(\lambda)\|^2\right) = N\left(\|\psi-\psi_0\|^2+\|\lambda\|^4+\|\delta\|^4\right)\mathcal{O}(1)$ we have that :

$$\begin{aligned}|R_N(\theta)| \leq \mathcal{O}_p(N)(&\|\psi-\psi_0\|^3+\|\psi-\psi_0\|^4+\|\psi-\psi_0\|^2\|\lambda\|^2+\|\psi-\psi_0\|^2\|\delta\|^2)\\ &+\|\delta\|^4\left|\sum_{m,n,o,p=1,\dots,d_\delta}\frac{\partial^4 l(\theta^+;y_{1:N})}{\partial\delta_m\partial\delta_n\partial\delta_o\partial\delta_p}-\frac{\partial^4 l(\theta_0;y_{1:N})}{\partial\delta_m\partial\delta_n\partial\delta_o\partial\delta_p}\right|\\ &+\|\lambda\|^4\left|\sum_{m,n,o,p=1,\dots,d_\lambda}\frac{\partial^4 l(\theta^+;y_{1:N})}{\partial\lambda_m\partial\lambda_n\partial\lambda_o\partial\lambda_p}-\frac{\partial^4 l(\theta_0;y_{1:N})}{\partial\lambda_m\partial\lambda_n\partial\lambda_o\partial\lambda_p}\right|\end{aligned}$$

and then :

$$\begin{aligned}|R_N(\theta)| \leq \mathcal{O}_p(1)\|t_N(\theta)\|^2(o_p(1)&+\frac{1}{N}\left|\sum_{m,n,o,p=1,\dots,d_\lambda}\frac{\partial^4 l(\theta^+;y_{1:N})}{\partial\lambda_m\partial\lambda_n\partial\lambda_o\partial\lambda_p}-\frac{\partial^4 l(\theta_0;y_{1:N})}{\partial\lambda_m\partial\lambda_n\partial\lambda_o\partial\lambda_p}\right|\\ &+\frac{1}{N}\left|\sum_{m,n,o,p=1,\dots,d_\delta}\frac{\partial^4 l(\theta^+;y_{1:N})}{\partial\delta_m\partial\delta_n\partial\delta_o\partial\delta_p}-\frac{\partial^4 l(\theta_0;y_{1:N})}{\partial\delta_m\partial\delta_n\partial\delta_o\partial\delta_p}\right|)\end{aligned}$$

where $\theta^+ = \theta_0 + t(\theta-\theta_0)$ for some $0<t<1$.

To show that the last two terms tend to zero we proceed as follows :

$$\begin{aligned}\frac{1}{N}\Big|\sum_{m,n,o,p=1,\dots,d_\lambda}&\frac{\partial^4 l(\theta^+;y_{1:N})}{\partial\lambda_m\partial\lambda_n\partial\lambda_o\partial\lambda_p}-\frac{\partial^4 l(\theta_0;y_{1:N})}{\partial\lambda_m\partial\lambda_n\partial\lambda_o\partial\lambda_p}\Big| =\\ &\frac{1}{N}\Big|\sum_{m,n,o,p=1,\dots,d_\lambda}\frac{\partial^4 l(\theta^+;y_{1:N})}{\partial\lambda_m\partial\lambda_n\partial\lambda_o\partial\lambda_p}-\mathbb{E}\left[\sum_{m,n,o,p=1,\dots,d_\lambda}\frac{\partial^4 l(\theta^+;y_1)}{\partial\lambda_m\partial\lambda_n\partial\lambda_o\partial\lambda_p}\right]\\ &+\mathbb{E}\left[\sum_{m,n,o,p=1,\dots,d_\lambda}\frac{\partial^4 l(\theta^+;y_{1:N})}{\partial\lambda_m\partial\lambda_1\partial\lambda_o\partial\lambda_p}\right]-\mathbb{E}\left[\frac{\partial^4 l(\theta_0;y_1)}{\partial\lambda_m\partial\lambda_n\partial\lambda_o\partial\lambda_p}\right]\\ &+\mathbb{E}\left[\frac{\partial^4 l(\theta_0;y_1)}{\partial\lambda_m\partial\lambda_n\partial\lambda_o\partial\lambda_p}\right]-\frac{\partial^4 l(\theta_0;y_{1:N})}{\partial\lambda_m\partial\lambda_n\partial\lambda_o\partial\lambda_p}\Big|\end{aligned}$$

We then apply triangular inequality to separate the 3 terms. The first and third terms are empirical means of centered random variables with bounded variances (assumption (2.2)v)). Therefore we can use each time Chebychev's inequality to obtain a weak law of large number and obtain the consistency toward 0. For the second term,

$$|\sum_{m,n,o,p=1,\dots,d_\lambda} \frac{\partial^4 l(\theta^+;y_1)}{\partial\lambda_m\partial\lambda_1\partial\lambda_o\partial\lambda_p} - \sum_{m,n,o,p=1,\dots,d_\lambda} \frac{\partial^4 l(\theta;y_1)}{\partial\lambda_m\partial\lambda_1\partial\lambda_o\partial\lambda_p}| \leq 2C\sup_{\theta\in\Theta}\|\nabla^4_\theta l(\theta;y_1)\|$$

where C is the nonnegative constant that appears in the equivalence between the L1 and L2 norm. When we evaluate at $\theta = \hat{\theta}_N$, $\theta^+ = \theta_0 + t(\hat{\theta}_N - \theta_0) \to \theta_0$ in probability, as $N \to +\infty$. And by continuity of $\nabla^4_\theta l(\cdot;y_1)$, and dominated convergence, the second term also converges toward 0. Finally we obtain :

$$|R_N(\hat{\theta}_N)| \leq o_p(1)\|t_N(\hat{\theta}_N)\|^2$$

And then : $0 \leq l(\hat{\theta}_N;y_{1:N}) - l(\theta_0;y_{1:N})$
$\leq \|\tilde{S}_N(\theta_0)\|\|t_N(\hat{\theta}_N)\| - \frac{1}{2}\|t_N(\hat{\theta}_N)\|^2_{\tilde{I}_N(\theta_0)} + o_p(\|t_N(\hat{\theta}_N)\|^2)$
$\leq \|\tilde{S}_N(\theta_0)\|\|t_N(\hat{\theta}_N)\| - \frac{1}{2}(o_p(1)+a)\|t_N(\hat{\theta}_N)\|^2$
where

$$a = \inf_{N>n_0}\inf_{x\neq 0}\frac{\|x\|^2_{\tilde{I}_N(\theta_0)}}{\|x\|^2}$$

where $\|x\|_A$ stands for x^TAx, with A being a positive definite symmetric matrix.

By taking n_0 large enough so that for every $N > n_0$, $\tilde{I}_N(\theta_0) \succ 0$ (assumption (2.3)) we have that $0 < a < +\infty$. The last inequality, shows that for N large enough, this polynomial of degree 2 in $\|t_N(\hat{\theta}_N)\|$ is upper bounded (dominant coefficient negative) and lower bounded by 0. Which shows that

$$t_N(\hat{\theta}_N) = \mathcal{O}_p(1)$$

which concludes the proof, and :

$$R_N(\hat{\theta}_N) = o_p(1)\mathcal{O}_p(1) = o_p(1) \tag{2.21}$$

which is fundamental for the proof of theorem (2.1)

2.7.8 . Proof of theorem 2.1

Now that we derived the expression of LRT_∞, it remains to show that the Bootstrap statistic also converges weakly in probability to this random variable. To do so we first derive a Bootstrap quadratic approximation as in (2.22), where the expansion is done around θ^*_N, as it is the true parameter of the Bootstrap data. We obtain that :

$$\begin{aligned} l(\theta;y^*_{1:N}) - l(\theta^*_N;y^*_{1:N}) = &\frac{1}{2}Z^*_N(\theta^*_N)^T\tilde{I}^*_N(\theta^*_N)Z^*_N(\theta^*_N) \\ &- \frac{1}{2}\left(t^*_N(\theta) - Z^*_N(\theta^*_N)\right)^T\tilde{I}^*_N(\theta^*_N)\left(t^*_N(\theta) - Z^*_N(\theta^*_N)\right) + R^*_N(\theta) \end{aligned} \tag{2.22}$$

where the exponent * stands for "evaluated on the Bootstrap data". The following proof will follow three main steps.

First we show that the Bootstrap version of $\tilde{I}_N(\theta_0)$ and of $\tilde{S}_N(\theta_0)$ have the correct conditional limiting distribution which means that :

$$\begin{cases} \tilde{I}_N^*(\theta_N^*) - \tilde{I}_N(\theta_0) = o_{p^*}(1) \\ \forall t \in \mathbb{R}^{d_{\tilde{I}}} \quad \text{pr}^*\{\tilde{S}_N^*(\theta_N^*) < t\} - \text{pr}\{\tilde{S}_N(\theta_0) < t\} = o_p(1) \end{cases} \tag{2.23}$$

Second, we show that the new terms in the quadratic approximation that are supposed to be null converge toward 0 thanks to the shrinkage parameter.

Finally we show that the rest in the Bootstrap quadratic approximation is a $o_{p^*}(1)$.

We first consider the case with no nuisance parameters i.e. $\theta = (\psi, \lambda)$ to lighten the notations and work in two steps.

We first consider $\tilde{I}_N(\theta_N^*)$. We write :

$$\left[\tilde{I}_N(\theta_N^*)\right]_{m,n} = \frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)*}(\theta_N^*)$$

where $h_{m,n}^{(i)}(\theta_N^*)$ is of the form :

$$h_{m,n}^{(i)*}(\theta_N^*) = c_{m,n}\frac{\partial^k \log f(y_i;\theta_N^*)}{\partial\theta_{i_1}...\partial\theta_{i_k}}$$

with $c_{m,n} \in \mathbb{R}$, $k \in \{2,4\}$, $1 \leq i_1, ..., i_k \leq q_\lambda + q_\psi$.

We want to show that :

$$\frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)*}(\theta_N^*) - \mathbb{E}\{h_{m,n}^{(1)}(\theta_0)\} = o_{p^*}(1)$$

To show that we decompose the difference :

$$\frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)*}(\theta_N^*) - \mathbb{E}\left[h_{m,n}^{(1)}(\theta_0)\right]$$

as :

$$\overbrace{\frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)*}(\theta_N^*) - \mathbb{E}^*\left[h_{m,n}^{(1)*}(\theta_N^*)\right]}^{(T1)} + \overbrace{\mathbb{E}^*[h_{m,n}^{(1)*}(\theta_N^*)] - \mathbb{E}^*[h_{m,n}^{(1)*}(\theta_0)]}^{(T2)} + \overbrace{\mathbb{E}^*[h_{m,n}^{(1)*}(\theta_0)] - \mathbb{E}\left[h_{m,n}^{(1)}(\theta_0)\right]}^{(T3)}$$

The term $(T1)$ is a sum of centered random variables, we can apply Chebychev's inequality to have the convergence towards 0 (using assumption (2.2) to have that the variance is $\mathcal{O}_p(1)$.

The term $(T2)$ is controlled as follows :

$$\begin{aligned}|\mathbb{E}^*[(h^{(1)*}_{m,n}(\theta^*_N) - h^{(1)*}_{m,n}(\theta_0))]| &\leq \mathbb{E}^*\left[|(h^{(1)*}_{m,n}(\theta^*_N) - h^{(1)*}_{m,n}(\theta_0))|\right] \\ &\leq \sup_{\theta\in\Theta}\mathbb{E}_\theta\left[|(h^{(1)*}_{m,n}(\theta^*_N) - h^{(1)*}_{m,n}(\theta_0))|\right]\end{aligned}$$

as $|(h^{(1)*}_{m,n}(\theta^*_N) - h^{(1)*}_{m,n}(\theta_0))| \leq 2\sup_{\theta\in\Theta}|h^{(1)}_{m,n}(\theta)|$ almost surely, using assumption 2.2 and thanks to the consistency of θ^*_N we can use the dominated convergence theorem to prove the convergence in probability toward 0.

Finally we deal with $(T3)$ as follows :

$$\begin{aligned}|\mathbb{E}^*[h^{(1)*}_{m,n}(\theta_0)] - \mathbb{E}\left[h^{(1)}_{m,n}(\theta_0)\right]| &= |\int h^{(1)*}_{m,n}(\theta_0)\{f(y;\theta^*_N) - f(y;\theta_0)\}dy| \\ &\leq \int |h^{(1)*}_{m,n}(\theta_0)||f(y;\theta^*_N) - f(y;\theta_0)|dy\end{aligned}$$

To show that this term tends toward 0 we use another time the equation (2.16), first we use a taylor expansion, there exist θ^+ between θ_0 and θ^*_N such that :

$$\begin{aligned}f(y;\theta^*_N) - f(y;\theta_0) &= (\theta^*_N - \theta_0)^T\nabla_\theta f(y;\theta^+) \\ &= (\theta^*_N - \theta_0)^T\nabla_\theta \log f(y;\theta^+)f(y;\theta^+)\end{aligned}$$

therefore,

$$|f(y;\theta^*_N) - f(y;\theta_0)| \leq \|\theta^*_N - \theta_0\|\|\nabla_\theta \log f(y;\theta^+)\|f(y;\theta^+)$$

and,

$$\begin{aligned}|\mathbb{E}^*[h^{(1)*}_{m,n}(\theta_0)] - \mathbb{E}\left[h^{(1)}_{m,n}(\theta_0)\right]| &\leq \|\theta^*_N - \theta_0\|\int |h^{(1)*}_{m,n}(\theta_0)|\|\nabla_\theta \log f(y;\theta^+)\|f(y;\theta^+)dy \\ &\leq \|\theta^*_N - \theta_0\|\int \frac{1}{2}\left\{|h^{(1)*}_{m,n}(\theta_0)|^2 + \|\nabla_\theta \log f(y;\theta^+)\|^2\right\}f(y;\theta^+)dy\end{aligned}$$

Thanks to assumption (2.2), as $\theta^*_N - \theta_0 = o_p(1)$ this last term is $o_p(1)$ as $N \to +\infty$.

We then consider $\tilde{S}^*_N(\theta^*_N)$, which is $\sqrt{N}$ times a sum of (conditionally) independent, centered, with finite variance random variables. Therefore, by proving that :

$$\mathbb{E}^*[\tilde{S}^*_N(\theta^*_N)\tilde{S}^*_N(\theta^*_N)^T] - \mathbb{E}\left[\tilde{S}_1(\theta_0)^T\tilde{S}_1(\theta_0)^T\right] = o_p(1) \tag{2.24}$$

and applying a multivariate version of the conditional central limit theorem of Bulinski (2017), it will conclude the second part of (13). Our third moment condition, and the consistency of the variance matrix of the score is much stronger then the Lindberg Feller conditions.

For each m, n, $\left[\tilde{S}_N^*(\theta_N^*)\tilde{S}_N^*(\theta_N^*)^T\right]_{m,n}$ can also be written as $\frac{1}{N}\sum_{i=1}^N h_{m,n}^{(i)*}(\theta_N^*)$, where $h_{m,n}^{(i)*}(\theta_N^*) = [\tilde{S}_N^*(\theta_N^*)]_m \times [\tilde{S}_N^*(\theta_N^*)]_n$. Due to assumption (2.2)(ii)–(iii), and the elementary equation (2.16), $\mathbb{E}\left[h_{m,n}^{(i)*}(\theta_N^*)^{3/2}\right] < +\infty$.

In order to prove (2.24), we once again split the sum :

$$\mathbb{E}^*[h_{m,n}^{(1)*}(\theta_N^*)] - \mathbb{E}\left[h_{m,n}^{(1)}(\theta_0)\right] = \overbrace{\mathbb{E}^*[h_{m,n}^{(1)*}(\theta_N^*)] - \mathbb{E}^*\left[h_{m,n}^{(1)*}(\theta_0)\right]}^{(S1)} + \overbrace{\mathbb{E}^*[h_{m,n}^{(1)*}(\theta_0)] - \mathbb{E}\left[h_{m,n}^{(1)}(\theta_0)\right]}^{(S2)}$$

We first deal with $(S1)$:

$$\begin{aligned}
|\mathbb{E}^*[h_{m,n}^{(1)*}(\theta_N^*)] - \mathbb{E}^*\left[h_{m,n}^{(1)*}(\theta_0)\right]| &\leq \sup_{\theta\in\Theta} \quad |\mathbb{E}_\theta[h_{m,n}^{(1)*}(\theta_N^*) - h_{m,n}^{(1)*}(\theta_0)]| \\
&\leq \sup_{\theta\in\Theta} \quad \mathbb{E}_\theta[|h_{m,n}^{(1)*}(\theta_N^*) - h_{m,n}^{(1)*}(\theta_0)|] \\
&\leq 2\sup_{\theta\in\Theta} \quad \mathbb{E}_\theta[\sup_{\theta'\in\Theta}|h_{m,n}^{(1)*}(\theta')|] \\
&< +\infty
\end{aligned}$$

Which enables, to apply dominated convergence to the first term, as $\hat{\theta}_N$ is consistent.

For $(S2)$ we proceed as follows :

$$\begin{aligned}
|\mathbb{E}^*[h_{m,n}^{(1)*}(\theta_0)] - \mathbb{E}\left[h_{m,n}^{(1)}(\theta_0)\right]| &= |\int h_{m,n}^{(1)*}(\theta_0)\{f(y;\theta_N^*) - f(y;\theta_0)\}dy| \\
&\leq \int |h_{m,n}^{(1)*}(\theta_0)||f(y;\theta_N^*) - f(y;\theta_0)|dy
\end{aligned}$$

To show that this term tends toward 0 we use the same reasonning as before, first we use a Taylor expansion : there exist θ^+ between θ_0 and θ_N^* such that :

$$\begin{aligned}
f(y;\theta_N^*) - f(y;\theta_0) &= (\theta_N^* - \theta_0)^T\nabla_\theta f(y;\theta^+) \\
&= (\theta_N^* - \theta_0)^T\nabla_\theta \log f(y;\theta^+)f(y;\theta^+)
\end{aligned}$$

therefore,

$$|f(y;\theta_N^*) - f(y;\theta_0)| \leq \|\theta_N^* - \theta_0\|\|\nabla_\theta \log f(y;\theta^+)\|f(y;\theta^+)$$

and,

$$|\mathbb{E}^*[h_{m,n}^{(1)*}(\theta_0)] - \mathbb{E}\left[h_{m,n}^{(1)}(\theta_0)\right]| \leq \|\theta_N^* - \theta_0\| \int |h_{m,n}^{(1)*}(\theta_0)|\|\nabla_\theta \log f(y;\theta^+)\|f(y;\theta^+)dy$$

We can't directly use equation (2.16) here because $h_{m,n}^{(1)}(\theta)$ doesn't admit second order moments ($\tilde{S}_N(\theta)$ admits third order moments). Thanks to Holder's inequality using $p = \frac{3}{2}$ and $q = 3$ we have that :

$$
\begin{aligned}
&\|\theta_N^* - \theta_0\| \int |h_{m,n}^{(1)*}(\theta_0)| \|\nabla_\theta \log f(y;\theta^+)\| f(y;\theta^+) dy \\
&\leq \|\theta_N^* - \theta_0\| \left(\int \|\nabla_\theta \log f(y;\theta^+)\|^3 f(y;\theta^+) dy \right)^{\frac{1}{3}} \left(\int |h_{m,n}^{(1)*}(\theta_0)|^{\frac{3}{2}} f(y;\theta^+) dy \right)^{\frac{2}{3}} \\
&= o_p(1)
\end{aligned}
$$

the last inequality holds thanks to assumption (2.2)(ii)–(iii) that enables to state that the two integrals are finite. That concludes the proof that (2.24) holds.

We now deal with the nuisance parameters. The proof starts the same way as before. The issue is that the odd derivatives with respect to the parameter δ are not 0 when evaluated at θ_N^* because $\delta_N^* \neq 0$. We recall that $\theta = (\psi, \delta, \lambda)$ and $\theta_N^* = (\psi_N^*, \delta_N^*, 0)$

To lighten the calculations we write $dx^* = x - x_N^*$, for any quantity x.

After expanding the likelihood as (2.22), new terms appear in $R_N^*(\theta)$: the odds order derivatives with respect to δ which would be 0 if $\delta_N^* = 0$.

We want to show that : $d\delta^{*T} \nabla_\delta l(\theta_N^*; y_{1:N}^*) = o_{p^*}(1)$ and $\sum_{ijk} d\delta_i^* d\delta_j^* d\delta_k^* \frac{\partial^3 l(\theta_N^*; y_{1:N}^*)}{\partial \delta_i \partial \delta_j \partial \delta_k} = o_{p^*}(1)$.

As said before the issue comes from the fact that the boostrap parameter of δ is not 0 . Indeed the "good" Bootstrap parameter would be $u_N^* = (\psi_N^*, 0, 0) = \theta_N^* - (0, \hat{\delta}_N, 0)$ (the Bootstrap parameter that we would use if we knew where were located the nuisance parameters) .

We are now going to expand the terms around u_N^*, so that the odds order derivatives evaluated at u_N^* will be zero. And we will use the fact that δ_N^* converges very fast to zero.

We write $\theta^+ = t\theta_N^* + (1-t)u_N^*$:

$$
d\delta^{*T} \nabla_\delta l(\theta_N^*; y_{1:N}^*) = d\delta^{*T} \left(0 + \nabla_\delta^2 l_N^*(u_N^*) \delta_N^* + \sum_{ij} \delta_{Ni}^* \delta_{Nj}^* \times 0 + \sum_{ijk} \delta_{Ni}^* \delta_{Nj}^* \delta_{Nk}^* \frac{\partial^4 l(\theta^+; y_{1:N}^*)}{\partial \delta_i \partial \delta_j \partial \delta_k \partial \delta} \right)
$$

the first non zero term is a centered random variable with finite variance (assumption (2.2)(iii)), therefore by the central limit theorem it is $\mathcal{O}_{p^*}(\sqrt{N})$, and the last term is $\mathcal{O}_{p^*}(1)$ by the law of large number. Therefore :

$$
\begin{aligned}
|d\delta^{*T} \nabla_\delta l(\theta_N^*; y_{1:N}^*)| &\leq \|d\delta^*\| \left(\mathcal{O}_p(\sqrt{N}) o_p(N^{-\frac{1}{4}}) + \mathcal{O}_p(N) o_p(N^{-\frac{3}{4}}) \right) \\
&\leq \|d\delta^*\| o_p(N^{\frac{1}{4}})
\end{aligned}
$$

The same way we have :

$$
\begin{aligned}
|\sum_{ijk} d\delta_i^* d\delta_j^* d\delta_k^* \frac{\partial^3 l(\theta_N^*; y_{1:N}^*)}{\partial \delta_i \partial \delta_j \partial \delta_k}| &= 0 + |{\delta_N^*}^T \sum_{ijk} d\delta_i^* d\delta_j^* d\delta_k^* \frac{\partial^4 l(\theta^+; y_{1:N}^*)}{\partial \delta \partial \delta_i \partial \delta_j \partial \delta_k}| \\
&\leq o_{p^*}(N^{\frac{-1}{4}}) \|d\delta^*\|^3 \times \mathcal{O}_{p^*}(N) \\
&\leq \|d\delta^*\|^3 o_{p^*}(N^{\frac{3}{4}})
\end{aligned}
$$

By using the same reasoning as in section (2.7.6) for the derivation of the likelihood ratio statistic, evaluating the expansion of the Bootstrap likelihood at the Bootstrap maximum likelihood estimator (restricted or not) $\hat{\theta}^*$, we have that almost surely

$$\begin{aligned}
0 &\leq l(\hat{\theta}^*; y^*_{1:N}) - l(\theta^*_N; y^*_{1:N}) \\
&\leq \|\tilde{S}^*_N(\theta^*_N)\| \|t_N(\hat{\theta}^*)\| - \frac{1}{2}\|t_N(\hat{\theta}^*)\|^2_{\tilde{I}^*_N(\theta^*_N)} + R^*_N(\hat{\theta}^*) \\
&\leq \|\tilde{S}^*_N(\theta^*_N)\| \|t_N(\hat{\theta}^*)\| - \frac{1}{2}(o_{p^*}(1) + a^*)\|t_N(\hat{\theta}^*)\|^2 + o_{p^*}(1)(\|t_N(\hat{\theta}^*)\|^{\frac{1}{2}} + \|t_N(\hat{\theta}^*)\|^{\frac{3}{2}})
\end{aligned}$$

Even if this quantity is no longer a polynomial, the dominant term remain the same, therefore this quantity is lower bounded by 0 and upper bounded in probability as a upper bounded function of $\|t_N(\hat{\theta}^*)\|$. Which implies that $\|t_N(\hat{\theta}^*)\| = \mathcal{O}_{p^*}(1)$ (in this proof we don't show that it is not a $o_{p^*}(1)$ but it is not important here as we already showed that the Bootstrap score and the Bootstrap FIM converge toward the correct limit). But the important is that we showed that $R^*_N(\hat{\theta}^*) = o_{p^*}(1)$ which conclude the proof.

2.7.9 . Proof of proposition 2.5

We recall that $\hat{\theta}_N = (\hat{\psi}_N, \hat{\delta}_N, \hat{\lambda}_N) = \arg\max_{\theta \in \Theta} \quad l(\theta; y_{1:N})$, and (c_N) is a sequence defined as in proposition (2.5) .

Consider $\theta^*_N = (\psi^*_N, \delta^*_N, \lambda^*_N)$ such that $\forall k = 1, .., d_\psi \ \psi^*_{N,k} = \hat{\psi}_{N,k} \ \mathbb{1}(\hat{\psi}_{N,k} > c_N)$, $\forall k = 1, .., d_\delta$ $\delta^*_{N,k} = \hat{\delta}_{N,k} \ \mathbb{1}(\hat{\delta}_{N,k} > c_N)$ and $\lambda^*_N = 0_{d_\lambda}$.

The proof of this proposition follows exactly the lines of Cavaliere et al. (2020) lemma 1. First the fact that $N^{1/4}c_N \to +\infty$ as $N \to +\infty$ implies also that $\sqrt{N}c_N \to +\infty$.

Let us establish a technical result that will then be applied to our proposition. Let (x_N) a real valued random sequence and $x_0 \in \mathbb{R}$ such that $r_N(x_N - x_0) = O_p(1)$, for r_N being whether $\sqrt{N}$ of $N^{1/4}$.

If $x_0 = 0$,

$$\text{pr}(x_N > c_N) = \text{pr}(r_N x_N > r_N c_N) = \text{pr}\{O_p(1) > r_N c_N\} = o(1)$$

The last equality holds because $r_N c_N \to +\infty$. Therefore $\mathbb{1}(x_N > c_N) = o_p(1)$ and finally $r_N x_N \mathbb{1}(x_N > c_N) = O_p(1)o_p(1) = o_p(1)$.

If $x_0 \neq 0$,

$$\text{pr}(|x_N| > c_N) = \text{pr}(|x_N - x_0 + x_0| > c_N) \geq \text{pr}\{||x_N - x_0| - |x_0|| > c_N\}$$

$$\text{pr}\{||x_N - x_0| - |x_0|| > c_N\} = \text{pr}\{|x_0| - |x_N - x_0| > c_N\} + \text{pr}\{|x_N - x_0| - |x_0| > c_N\}$$

First, $|x_N - x_0| + c_N = o_p(1)$ and $|x_0| > 0$ therefore $\text{pr}\{|x_0| - |x_N - x_0| > c_N\} \to 1$ as $N \to +\infty$. And $r_N(|x_0| + c_N) \to +\infty$ so $\text{pr}\{|x_N - x_0| - |x_0| > c_N\} \to 0$ as $N \to +\infty$. Therefore :

$$\mathbb{1}(x_N > c_N) - 1 = o_p(1) \tag{2.25}$$

Finally,

$$\begin{aligned} r_N\{x_N \mathbb{1}(x_N > c_N) - x_0\} &= r_N(x_N - x_0)\mathbb{1}(x_N > c_N) - x_0 \mathbb{1}(x_N \leq c_N) \\ &= r_N(x_N - x_0)\mathbb{1}(x_N > c_N) - x_0\{1 - \mathbb{1}(x_N > c_N)\} \\ &= O_p(1) + o_p(1) \end{aligned}$$

using equation (2.25), it concludes with Slutsky's theorem that $r_N\{x_N \mathbb{1}(x_N > c_N) - x_0\}$ and $r_N(x_n - x_0)$ have the same limiting distribution.

Applying this result to $x_N = \hat{\delta}_N$ and $r_N = N^{1/4}$ we have that $\delta_N^* = o_p(1)$. the same holds for $\hat{\psi}_N$: $\sqrt{N}(\hat{\psi}_N - \psi_0) = O_p(1)$ (if $\psi_0 = 0$ then $\sqrt{N}\hat{\psi}_N = o_p(1)$).

Finally it is obvious that $\theta_N^* \in \Theta_0$ as $\lambda = 0_{d_\lambda}$. Which concludes the proof.

2.7.10 . Proof of proposition 2.6

We verify that our hypothesis imply the conditions required in Hoadley (1971).
We show easily that the assumptions C(1), C(2), C(3'),C(4'), C(5) are verified. Assumptions C(1)–(2) are verified with assumption (2.1). Assumption C(3') is weaker than assumption (2.4). C(4') is equivalent to assumption (2.5). C(5) is verified as we the continuity of the likelihood with respect to θ, for every y and the measurability with respect to y for every θ. The result is then discussed for instance in Giné and Nickl (2021) exercise 7.2.3.

2.7.11 . Proof of theorem 2.2

This proof is very similar to the one of theorem (2.1).

First we have to derive the asymptotic distribution of the likelihood ratio test statistic. We start from the quadratic expansion (2.22). We first show that $\tilde{I}_N(\theta_0)$ converges in probability toward a non random matrix $\tilde{I}(\theta_0)$.

As in the proof of theorem (2.1) we write

$$\left[\tilde{I}_N(\theta_0)\right]_{m,n} = \frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)}(\theta_0)$$

where $h_{m,n}^{(i)}(\theta_0)$ is of the form :

$$h_{m,n}^{(i)}(\theta_0) = c_{m,n}\frac{\partial^k \log f_i(y_i;\theta_0)}{\partial\theta_{i_1}...\partial\theta_{i_k}}$$

with $c_{m,n} \in \mathbb{R}$, $k \in \{2,4\}$, $1 \leq i_1, ..., i_k \leq d_\psi + d_\lambda + d_\delta$.

As a consequence of assumption (2.4), using Chebychev's inequality :

$$\frac{1}{N}\sum_{i=1}^{N} |h_{m,n}^{(i)}(\theta_0) - \mathbb{E}[h_{m,n}^{(i)}(\theta_0)]| = o_p(1)$$

which enables to define $\tilde{I}(\theta_0) = \left[\lim_{N\to+\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[h_{m,n}^{(i)}(\theta_0)]\right]_{m,n}$ which is a nonrandom matrix that is supposed to be positive definite (assumption (2.3)). Furthermore,

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[|h_{m,n}^{(i)}(\theta_0)|\right] &\leq \frac{1}{N}\sum_{i=1}^{N} \sup_{i\in\mathbb{N}} \quad \mathbb{E}\left[|h_{m,n}^{(i)}(\theta_0)|\right] \\
&\leq \sup_{i\in\mathbb{N}} \quad \mathbb{E}\left[|h_{m,n}^{(i)}(\theta_0)|\right] \\
&< +\infty
\end{aligned}$$

which holds for every $N \geq 0$. This last inequality enables to invert the sum and the integral :

$$\begin{aligned}
\left[\tilde{I}(\theta_0)\right]_{m,n} &= \lim_{N\to+\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}[h_{m,n}^{(i)}(\theta_0)] \\
&= \mathbb{E}\left[\lim_{N\to+\infty} \frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)}(\theta_0)\right] \\
&= \lim_{N\to+\infty} \frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)}(\theta_0)
\end{aligned}$$

where the last equality holds as we consider a non random quantity.

We consider now $\tilde{S}_N(\theta_0)$ which is a sum of centered random variables with finite variances, we want to apply theorem 6.5 of Hansen (2022). Assumption (2.2)(ii)–(iii) and assumption (2.4) enables to state that :

$$\lim_{N\to+\infty} \mathbb{E}\left[\tilde{S}_N(\theta_0)\tilde{S}_N(\theta_0)^T\right] < +\infty$$

which is a direct consequence of theorem A.5 of Hoadley (1971).Furthermore, still thanks to assumption (2.2)(ii)–(iii) and assumption (2.4) equation (6.3) in Hansen (2022) theorem 6.5 is verified for $\delta = 1$, and therefore $\tilde{S}_N(\theta_0)$ is $O_p(1)$ and converges in distribution toward a random variable that we call $U(\theta_0)$.

The next step of the proof is to prove (2.23).

We first deal with $\tilde{I}_N^*(\theta_N^*)$, we still write :

$$\left[\tilde{I}_N(\theta_N^*)\right]_{m,n} = \frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)*}(\theta_N^*)$$

We proceed as in the proof of theorem (2.1), and we split :

$$\frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)*}(\theta_N^*) - \left[\tilde{I}(\theta_0)\right]_{m,n} = \frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)*}(\theta_N^*) - \lim_{N\to+\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[h_{m,n}^{(i)}(\theta_0)]$$
$$= \frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)*}(\theta_N^*) - \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[h_{m,n}^{(i)}(\theta_0)] + o(1)$$

as

$$\overbrace{\frac{1}{N}\sum_{i=1}^{N} h_{m,n}^{(i)*}(\theta_N^*) - \mathbb{E}^*\left[h_{m,n}^{(i)*}(\theta_N^*)\right]}^{(U1)} + \overbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}^*[h_{m,n}^{(i)*}(\theta_N^*)] - \mathbb{E}^*[h_{m,n}^{(i)*}(\theta_0)]}^{(U2)}$$
$$+ \overbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}^*[h_{m,n}^{(i)*}(\theta_0)] - \mathbb{E}\left[h_{m,n}^{(i)}(\theta_0)\right]}^{(U3)} + o(1)$$

The term $(U1)$ is a sum of centered random variables with finite variance uniformly bounded over $i \in$ and therefore is $o_{p^*}(1)$.

We deal with $(U2)$ as before :

$$\left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}^*[h_{m,n}^{(i)*}(\theta_N^*)] - \mathbb{E}^*[h_{m,n}^{(i)*}(\theta_0)]\right| \leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}^*[|h_{m,n}^{(i)*}(\theta_N^*) - h_{m,n}^{(i)*}(\theta_0)|]$$
$$\leq \sup_{i\in\mathbb{N}}\quad\sup_{\theta\in\Theta}\quad \mathbb{E}_\theta\left[|h_{m,n}^{(i)*}(\theta_N^*) - h_{m,n}^{(i)*}(\theta_0)|\right]$$

and for every i, $|h_{m,n}^{(i)*}(\theta_N^*) - h_{m,n}^{(i)*}(\theta_0)| \leq 2\sup_{\theta'\in\Theta}|h_{m,n}^{(i)*}(\theta')|$, thanks to assumption (2.4), we can apply dominated convergence.

For the term $(U3)$, we apply the exact same reasoning as in the proof of theorem (2.1) to show that

$$\left|\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}^*[h_{m,n}^{(i)*}(\theta_0)] - \mathbb{E}\left[h_{m,n}^{(i)}(\theta_0)\right]\right|$$

is almost surely smaller than

$$\|\theta_N^* - \theta_0\|\quad\sup_{i\in\mathbb{N}}\quad\int\frac{1}{2}\left\{|h_{m,n}^{(i)*}(\theta_0)|^2 + \|\nabla_\theta\log f_i(y;\theta^+)\|^2\right\}f_i(y;\theta^+)dy$$

which is almost surely smaller than

$$\frac{\|\theta_N^* - \theta_0\|}{2}\left\{\sup_{\theta^+\in\Theta}\mathbb{E}_{\theta^+}\left[\sup_{\theta\in\Theta}\quad|h_{m,n}^{(i)}(\theta)|^2\right] + \sup_{\theta^+\in\Theta}\mathbb{E}_{\theta^+}\left[\sup_{\theta\in\Theta}\quad\|\nabla_\theta\log f_i(y;\theta)\|^2\right]\right\}$$

which is $o_p(1)$ du to the consistency of θ_N^* and assumption (2.2). Which concludes the proof of theorem (2.2).

2.7.12 . Proof of proposition 2.7

Recall that we want to show that :

$$\sup_{\theta' \in \Theta} \mathbb{E}_{\theta'} \{ \sup_{\theta \in \Theta} \| \nabla_\theta^k \log f_i(y_i; \theta) \|^\gamma \} < +\infty$$

with $\gamma = 2$ for $k = 0, 3, 4$ and $\gamma = 3$ for $k = 1, 2$, for every $i \in \mathbb{N}$.

For a sake of clarity, we consider the following simplified notations :

$g(x_{ij}, \beta, \Lambda \xi_i) = g_\theta^{ij}(\xi_i)$ and $g_\theta^i(\xi_i) = (g_\theta^{ij}(\xi_i))_{j=1,\dots,J_i}$

$f(y_i; \theta) = \mathbb{E}\{f(y_i; \xi, \theta)\} \propto \mathbb{E}[\exp\{-V(\theta, y_i, \xi_i)\}]$ with

$$V(\theta, y_i, \xi_i) = \frac{\sum_j (y_{ij} - g_\theta^j(\xi_i))^2}{2\sigma^2} = \frac{\|y_i - g_\theta^i(\xi_i)\|^2}{2\sigma^2}$$

where the expectation is taken with respect to the random variable ξ.

As these quantities are individual, we get rid of the subscript i to lighten the notations.

As we consider the parameter space Θ compact, the residual variance σ is restricted to lie in a segment $[\sigma_{min}; \sigma_{max}]$ with $0 < \sigma_{min} < \sigma_{max} < +\infty$. That is why we consider the gaussian density up to a constant that won't change the reasonning. From now on we won't write $\propto$ and we make the shortcut $f(y; \xi, \theta) = e^{-V(\theta, y, \xi)}$.

We also write $\mathbb{E}^Z\{\cdot\}$ when the expectation is taken with respect to the random variable Z.

We suppose now that $\|g_\theta(\xi)\| \to +\infty$ as $\|\xi\| \to +\infty$ which is the most complicated case.

We first deal with the simplest case $k = 0$, we want to show that

$$\mathbb{E}_{\theta'} \left\{ \sup_{\theta \in \Theta} |\log f(y; \theta)|^2 \right\} < +\infty$$

Let $\theta, \theta' \in \Theta$, let $M > 0$,

$$\begin{aligned}
f(y; \theta) &= \mathbb{E}^\xi \{f(y; \xi, \theta)\} \\
&\geq \mathbb{E}^\xi \{f(y; \xi, \theta) \mathbb{1}(\xi \leq M)\} \\
&\geq \mathbb{E}^\xi \{e^{-V(\theta, y, \xi)} \mathbb{1}(\xi \leq M)\} \\
&= \mathbb{E}^\xi \{e^{-\frac{\|y - g_\theta(\xi)\|^2}{2\sigma}} \mathbb{1}(\xi \leq M)\} \\
&= \mathbb{E}^\xi \{e^{-\frac{\|y\|^2 + \|g_\theta(\xi)\|^2 - 2y^T g_\theta(\xi)}{2\sigma}} \mathbb{1}(\xi \leq M)\} \\
&\geq \mathbb{E}^\xi \{e^{-\frac{\|y\|^2 + \|g_\theta(\xi)\|^2 + 2\|y\| \|g_\theta(\xi)\|}{2\sigma}} \mathbb{1}(\xi \leq M)\}
\end{aligned}$$

where the last inequality is a direct application of Cauchy Schwartz's inequality.

The quantity in the exponential is a polynomial in $\|g_\theta(\xi)\|$ that goes to $-\infty$ when $\|\xi\| \to +\infty$. Its minimal value is achieved at $\alpha_M(\theta) = \sup_{\xi : \|\xi\| \leq M} \|g_\theta(\xi)\| = \|g_\theta(\tilde{\xi})\|$

Therefore we obtain that for every $M > 0$,

$$f(y;\theta) \geq \text{pr}(\xi \leq M) e^{-\frac{(\|y\|+\alpha_M(\theta))^2}{2\sigma^2}} \tag{2.26}$$

yet, by writing $\kappa_M = \text{pr}(\xi \leq M)$,

$$\begin{aligned}
& 1 > f(y;\theta) > \kappa_M e^{-\frac{(\|y\|+\alpha_M(\theta))^2}{2\sigma^2}} \\
\Leftrightarrow \quad & 0 > \log f(y;\theta) > \log(\kappa_M) - \frac{(\|y\|+\alpha_M(\theta))^2}{2\sigma^2} \\
\Leftrightarrow \quad & 0 < |\log f(y;\theta)|^2 < \log(\kappa_M)^2 + \frac{(\|y\|+\alpha_M(\theta))^4}{4\sigma^2} - \frac{\log(\kappa_M)(\|y\|+\alpha_M(\theta))^2}{\sigma^2} \\
\Leftrightarrow \quad & 0 < |\log f(y;\theta)|^2 < \log(\kappa_M)^2 + \frac{\{\|y\|+\alpha_M(\theta)\}^4}{4\sigma^4} \\
\Leftrightarrow \quad & 0 < |\log f(y;\theta)|^2 < \log(\kappa_M)^2 + \frac{\{\|y\|+\alpha_M(\bar{\theta})\}^4}{4\sigma^4_{min}}
\end{aligned}$$

where $\bar{\theta} = \underset{\theta\in\Theta}{\text{argsup}}\ \alpha(\theta) \in \Theta$ by continuity of $\alpha(\cdot)$ (continuity of $\theta \to g_\theta(\tilde{\xi})$) and compactness of Θ

We define the quantity $P_M(y) = \log(\kappa_M)^2 + \frac{\{\|y\|+\alpha_M(\bar{\theta})\}^4}{4\sigma^4_{min}}$ that does not depend on θ.

Therefore we have that :

$$\begin{aligned}
\mathbb{E}_{\theta'}\left\{\sup_{\theta\in\Theta} \quad |\log f(y;\theta)|^2\right\} &\leq \mathbb{E}_{\theta'}\left\{P_M(y)\right\} \\
&= \int_y P_M(y) f(y;\theta')dy \\
&= \int_y P_M(y) \int_\xi f(y;\xi,\theta')\pi_p(\xi)d\xi dy \\
&= \int_\xi \int_y P_M(y) f(y;\xi,\theta')\pi_p(\xi)dyd\xi \\
&= \int_\xi \int_u 2\sigma'^2 P_M(2\sigma'^2 u + g_{\theta'}(\xi))\pi_J(u)\pi_p(\xi)dud\xi \\
&\leq \int_\xi \int_u 2\sigma^2_{max} P_M(2\sigma^2_{max} u + g_{\theta'}(\xi))\pi_J(u)\pi_p(\xi)dud\xi
\end{aligned}$$

using first Fubini-Tonelli's theorem and then a change of variable : $u = \frac{y-g_{\theta'}(\xi)}{2\sigma'^2}$.

$P_M(2\sigma'^2 u + g_{\theta'}(\xi))$ is a polynomial of degree 4 in $\|u\|$ and $\|g_{\theta'}(\xi)\|$. Thanks to the assumption (2.6) of proposition (2.7) with $k_1 = 0$ and $k_2 = 4$, we have that :

$$\sup_{\theta'\in\Theta} \mathbb{E}_{\theta'}\left\{\sup_{\theta\in\Theta} \quad |\log f(y;\theta)|^2\right\} \leq \int_\xi \int_u 2\sigma^2_{max} \sup_{\theta'\in\Theta} P_M(2\sigma^2_{max} u + g_{\theta'}(\xi))\Psi_J(u)\Psi_p(\xi)dud\xi$$

which concludes the first part of this proof.

We now consider the case $k = 1, 2, 3, 4$, and $\gamma = 2, 3$ (the fact of considering $\gamma = 2$ or 3 is the same, therefore we will consider $\gamma = 3$ as it is stronger).

$$\|\nabla_\theta^k \log f(y;\theta)\|^3 \leq \sup_{\substack{I\in\{1,...,d_\theta\}^k \\ I=(i_1,...,i_k)}} d_\theta^k \| \frac{\partial^k \log f(y;\theta)}{\prod_{j=1}^k \partial\theta_{i_j}} \|^3$$

Let $I_0 = (i_1, ..., i_k)$ the subset of indexes where the sup in the right hand side is achieved.

We consider the quantity

$$\frac{\partial^k \log f(y;\theta)}{\prod_{j=1}^k \partial\theta_{i_j}}$$

as the derivatives of the comoposition

$$\theta \mapsto f(y;\theta) \mapsto \log f(y;\theta)$$

and we use Faa di Bruno's formula to develop this expression :

$$\frac{\partial^k \log f(y;\theta)}{\prod_{j=1}^k \partial\theta_{i_j}} = \sum_{\Psi\in\mathcal{P}(\{i_1,...,i_k\})} \alpha_\Psi f(y;\theta)^{-|\Psi|} \prod_{B\in\Psi} \frac{\partial^{|B|} f(y;\theta)}{\prod_{b\in B}\partial\theta_b}$$

where $\mathcal{P}(K)$ stands for all partitions of a set K, and $(\alpha_\Psi)_\Psi$ are constants. A sufficient condition for this quantity to be $\mathbb{L}^3$ is that each term of the sum is $\mathbb{L}^3$.

Let $\Psi \in \mathcal{P}(\{i_1, ..., i_k\})$, we write m the cardinal of Ψ.
We recall that :

$$f(y;\theta) = \mathbb{E}^\xi \left\{ e^{-V(\theta,y,\xi)} \right\}$$

therefore, for every $B \in \Psi$,

$$\begin{aligned} \frac{\partial^{|B|} f(y;\theta)}{\prod_{b\in B}\partial\theta_b} &= \mathbb{E}^\xi \left\{ \frac{\partial^{|B|} f(y;\xi,\theta)}{\prod_{b\in B}\partial\theta_b} \right\} \\ &= \mathbb{E}^\xi \left[P^B\{V(\theta,y,\xi)\} e^{-V(\theta,y,\xi)} \right] \end{aligned}$$

where $P^B\{V(\theta,y,\xi)\}$ is a polynomial of degree m in y and the partial derivatives of $g_\theta(\xi)$.

$$\begin{aligned} |f(y;\theta)^{-m} \prod_{B\in\Psi} \frac{\partial^{|B|} f(y;\theta)}{\prod_{b\in B}\partial\theta_b}| &= f(y;\theta)^{-m} \prod_{B\in\Psi} |\frac{\partial^{|B|} f(y;\theta)}{\prod_{b\in B}\partial\theta_b}| \\ (Jensen) &\leq f(y;\theta)^{-m} \prod_{B\in\Psi} \mathbb{E}^\xi \left[|P^B\{V(\theta,y,\xi)\}| e^{-V(\theta,y,\xi)} \right] \end{aligned}$$

the random variables ξ can be renamed ξ_B for each term of the product so that :

$$|f(y;\theta)^{-m} \prod_{B\in\Psi} \frac{\partial^{|B|} f(y;\theta)}{\prod_{b\in B}\partial\theta_b}| \leq f(y;\theta)^{-m} \prod_{B\in\Psi} \mathbb{E}^{\xi_B} \left[|P^B\{V(\theta,y,\xi_B)\}| e^{-V(\theta,y,\xi_B)} \right]$$

We introduce $\Xi = (\xi_{B_1}^T, ..., \xi_{B_m}^T)^T \sim \mathcal{N}(0, I_{m\times p})$, so that we can write :

$$\begin{aligned} f(y;\theta)^{-m} \prod_{B\in\Psi} \mathbb{E}^{\xi_B}\left[|P^B\{V(\theta,y,\xi_B)\}|e^{-V(\theta,y,\xi_B)}\right] \\ = f(y;\theta)^{-m}\mathbb{E}^{\Xi}\left[\prod_{B\in\Psi}|P^B\{V(\theta,y,\xi_B)\}|e^{-V(\theta,y,\xi_B)}\right] \\ = f(y;\theta)^{-m}\mathbb{E}^{\Xi}\left(\left[\prod_{B\in\Psi}|P^B\{V(\theta,y,\xi_B)\}|\right]e^{-\sum_{B\in\Psi}V(\theta,y,\xi_B)}\right) \end{aligned}$$

every terms in the product inside the expectation is nonnegative, therefore when rising this quantity to the power of 3 we can use Jensen by convexity on $\mathbb{R}^+$ of $x \mapsto x^3$. And we find that :

$$\begin{aligned} |f(y;\theta)^{-m}\prod_{B\in\Psi}\frac{\partial^{|B|}f(y;\theta)}{\prod_{b\in B}\partial\theta_b}|^3 &\leq f(y;\theta)^{-3m}\mathbb{E}^{\Xi}\left(\left[\prod_{B\in\Psi}|P^B\{V(\theta,y,\xi_B)\}|\right]e^{-\sum_{B\in\Psi}V(\theta,y,\xi_B)}\right)^3 \\ &\leq f(y;\theta)^{-3m}\mathbb{E}^{\Xi}\left(\left[\prod_{B\in\Psi}|P^B\{V(\theta,y,\xi_B)\}|^3\right]e^{-\sum_{B\in\Psi}3V(\theta,y,\xi_B)}\right) \end{aligned}$$

using Jensen's inequality.

By using equation (2.26), we know that :

$$f(y;\theta) \geq \kappa_M e^{-\frac{(\|y\|+\alpha_M(\bar{\theta}))^2}{2\sigma^2}}$$

κ_m is a constant that has no impact on the reasoning therefore we neglect it to lighten the notations. We write $V(y,\sigma) = \frac{(\|y\|+\alpha_M(\bar{\theta}))^2}{2\sigma^2}$ so that we have :

$$|f(y;\theta)^{-m}\prod_{B\in\Psi}\frac{\partial^{|B|}f(y;\theta)}{\prod_{b\in B}\partial\theta_b}|^3 \leq e^{3mV(y,\sigma)}\mathbb{E}^{\Xi}\left(\left[\prod_{B\in\Psi}|P^B\{V(\theta,y,\xi_B)\}|^3\right]e^{-\sum_{B\in\Psi}3V(\theta,y,\xi_B)}\right)$$

We recall that the cardinal of Ψ is equal to m so :

$$|f(y;\theta)^{-m}\prod_{B\in\Psi}\frac{\partial^{|B|}f(y;\theta)}{\prod_{b\in B}\partial\theta_b}|^3 \leq \mathbb{E}^{\Xi}\left(\left[\prod_{B\in\Psi}|P^B\{V(\theta,y,\xi_B)\}|^3\right]e^{-3\sum_{B\in\Psi}V(\theta,y,\xi_B)-V(y,\sigma)}\right) \quad (2.27)$$

Once more to lighten the notations we define $M_\theta^\Psi(y,\Xi) = \prod_{B\in\Psi}|P^B\{V(\theta,y,\xi_B)\}|^3$

Let $\theta' \in \Theta$,

$$
\begin{aligned}
&\mathbb{E}^y_{\theta'}\left\{|f(y;\theta)^{-m}\prod_{B\in\Psi}\frac{\partial^{|B|}f(y;\theta)}{\prod_{b\in B}\partial\theta_b}|^3\right\}\\
&\leq \mathbb{E}^y_{\theta'}\left\{\mathbb{E}^{\Xi}\left(M^{\Psi}_{\theta}(y,\Xi)e^{-3\sum_{B\in\Psi}V(\theta,y,\xi_B)-V(y,\sigma)}\right)\right\}\\
&= \mathbb{E}^{\Xi}\left\{\mathbb{E}^y_{\theta'}\left(M^{\Psi}_{\theta}(y,\Xi)e^{-3\sum_{B\in\Psi}V(\theta,y,\xi_B)-V(y,\sigma)}\right)\right\}\\
&= \mathbb{E}^{\Xi}\left\{\int_y M^{\Psi}_{\theta}(y,\Xi)e^{-3\sum_{B\in\Psi}V(\theta,y,\xi_B)-V(y,\sigma)}f(y;\theta')dy\right\}\\
&= \mathbb{E}^{\Xi}\left[\int_y M^{\Psi}_{\theta}(y,\Xi)e^{-3\sum_{B\in\Psi}V(\theta,y,\xi_B)-V(y,\sigma)}\mathbb{E}^{\xi}\{e^{-V(\theta',y,\xi)}\}dy\right]\\
&= \mathbb{E}^{\Xi,\xi}\left\{\int_y M^{\Psi}_{\theta}(y,\Xi)e^{-3\sum_{B\in\Psi}V(\theta,y,\xi_B)-V(y,\sigma)}e^{-V(\theta',y,\xi)}dy\right\} \qquad (2.28)
\end{aligned}
$$

using once more Fubini Tonelli's theorem.

We focus on the exponential :

$$
\begin{aligned}
-3\sum_{B\in\Psi}\{V(\theta,y,\xi_B)-V(y,\sigma)\}-V(\theta',y,\xi) &= -3\sum_{B\in\Psi}\{V(\theta,y,\xi_B)-V(y,\sigma)\}-V(\theta',y,\xi)\\
&= -3\sum_{B\in\Psi}\left\{\frac{\|y-g_\theta(\xi_B)\|^2}{2\sigma^2}-\frac{\{\|y\|+\alpha(\bar{\theta})\}^2}{2\sigma^2}\right\}-\frac{\|y-g_{\theta'}(\xi)\|^2}{2\sigma'^2}\\
&= -3\sum_{B\in\Psi}\frac{1}{2\sigma^2}\{\|g_\theta(\xi_B)\|^2-2y^Tg_\theta(\xi_B)+2\|y\|\alpha(\bar{\theta})-\alpha(\bar{\theta})^2\}\\
&\quad -\frac{1}{2\sigma'^2}\{\|y\|^2+\|g_{\theta'}(\xi)\|^2-2y^Tg_{\theta'}(\xi)\}
\end{aligned}
$$

We use Cauchy-Schwartz's inequality to get rid of the scalar product and consider scalar quantities :

$$
\begin{aligned}
-3\sum_{B\in\Psi}\{V(\theta,y,\xi_B)-V(y,\sigma)\}-V(\theta',y,\xi) &= -3\sum_{B\in\Psi}\{V(\theta,y,\xi_B)-V(y,\sigma)\}-V(\theta',y,\xi)\\
&\leq -3\sum_{B\in\Psi}\frac{1}{2\sigma^2}\{\|g_\theta(\xi_B)\|^2-2\|y\|\|g_\theta(\xi_B)\|+2\|y\|\alpha(\bar{\theta})-\alpha(\bar{\theta})^2\}\\
&\quad -\frac{1}{2\sigma'^2}\{\|y\|^2+\|g_{\theta'}(\xi)\|^2-2\|y\|\|g_{\theta'}(\xi)\|\}
\end{aligned}
$$

Therefore by taking the integrated form of (2.28) we have that :

$$
\begin{aligned}
&\mathbb{E}^y_{\theta'}\left\{|f(y;\theta)^{-m}\prod_{B\in\Psi}\frac{\partial^{|B|}f(y;\theta)}{\prod_{b\in B}\partial\theta_b}|^3\right\}\\
&\leq\int_{\Xi,\xi}\int_y M^\Psi_\theta(y,\Xi)e^{-3\sum_{B\in\Psi}V(\theta,y,\xi_B)-V(y,\sigma)}e^{-V(\theta',y,\xi)}dy e^{-\frac{\|\Xi\|^2}{2}-\frac{\|\xi\|^2}{2}}d\Xi d\xi\\
&\leq\int_{\Xi,\xi,y} M^\Psi_\theta(y,\Xi)e^{-3\sum_{B\in\Psi}V(\theta,y,\xi_B)-V(y,\sigma)-V(\theta',y,\xi)-\frac{\|\Xi\|^2}{2}-\frac{\|\xi\|^2}{2}}dyd\Xi d\xi\\
&\leq\int_{\Xi,\xi,y} M^\Psi_\theta(y,\Xi)e^{H(\theta,\theta',y,\Xi,\xi)}dyd\Xi d\xi
\end{aligned}
$$

where

$$
\begin{aligned}
H(\theta,\theta',y,\Xi,\xi)=-3\sum_{B\in\Psi}\frac{1}{2\sigma^2}\{\|g_\theta(\xi_B)\|^2-2\|y\|\|g_\theta(\xi_B)\|+2\|y\|\alpha(\bar\theta)-\alpha(\bar\theta)^2\}\\
-\frac{1}{2\sigma'^2}\{\|y\|^2+\|g_{\theta'}(\xi)\|^2-2\|y\|\|g_{\theta'}(\xi)\|-\frac{\|\Xi\|^2}{2}-\frac{\|\xi\|^2}{2}
\end{aligned}
$$

by splitting $\|\Xi\|^2$ and rearranging the terms we get :

$$
\begin{aligned}
H(\theta,\theta',y,\Xi,\xi)=\|y\|\left\{\frac{\|g_{\theta'}(\xi)\|}{\sigma'^2}-\frac{3m}{\sigma^2}\alpha(\bar\theta)\right\}+\frac{3m}{2\sigma^2}\alpha(\bar\theta)^2+\sum_{B\in\Psi}\frac{3}{\sigma^2}\|y\|\|g_\theta(\xi_B)\|\\
-\frac{1}{2\sigma'^2}\|y\|^2-\frac{1}{2}\left\{\|\xi\|^2+\frac{1}{\sigma'^2}\|g_{\theta'}(\xi)\|^2\right\}-\frac{1}{2}\sum_{B\in\Psi}\left\{\frac{3}{\sigma^2}\|g_\theta(\xi_B)\|^2+\|\xi_B\|^2\right\}
\end{aligned}
$$

As it is constant we can omit $\frac{3m}{2\sigma^2}\alpha(\bar\theta)^2$ in the development as it can be taken out from the integrals. Furhtermore $-\alpha(\bar\theta\|y\|/\sigma^2<0$ therefore it can be upper bounded by zero.

Then we define for $\theta\in\Theta$, $\xi\in\mathbb{R}^p$ $r_\theta(\xi)$ the ratio $\frac{\|g_\theta(\xi)\|}{\|\xi\|}$ such that,

$$
\begin{aligned}
H(\theta,\theta',y,\Xi,\xi)\leq&\frac{r_{\theta'}(\xi)}{\sigma'^2}\|\xi\|\|y\|+\sum_{B\in\Psi}\frac{3r_\theta(\xi_B)}{\sigma^2}\|\xi_B\|\|y\|\\
&-\frac{1}{2\sigma'^2}\|y\|^2-\frac{1}{2}\left(1+\frac{r_{\theta'}(\xi)^2}{\sigma'^2}\right)\|\xi\|^2-\frac{1}{2}\sum_{B\in\Psi}\left\{1+\frac{3r_\theta(\xi_B)^2}{\sigma^2}\right\}\|\xi_B\|^2\\
&\leq-\frac{1}{2}(\frac{y^T}{\sigma_{min}},\tilde\xi^T,\tilde\Xi^T)^T\begin{pmatrix}
1 & -\frac{\sigma_{min}r_{\theta'}(\xi)}{\sigma^2_{max}} & \dots & -\frac{3\sigma_{min}r_\theta(\xi_B)}{\sigma^2_{max}} & \dots\\
-\frac{\sigma_{min}r_{\theta'}(\xi)}{\sigma^2_{max}} & 1 & 0 & \cdots & 0\\
\vdots & 0 & \ddots & 0 & \vdots\\
-\frac{3\sigma_{min}r_\theta(\xi_B)}{\sigma^2_{max}} & \vdots & 0 & \ddots & 0\\
\vdots & 0 & \cdots & 0 & 1
\end{pmatrix}\begin{pmatrix}\frac{y}{\sigma_{min}}\\ \tilde\xi\\ \tilde\Xi\end{pmatrix}
\end{aligned}
$$

Where $\tilde\xi=\sqrt{\left(1+\frac{r_{\theta'}(\xi)^2}{\sigma'^2}\right)}\times\xi$ and $\tilde\Xi=\left(\sqrt{\left\{1+\frac{3r_\theta(\xi_B)^2}{\sigma^2}\right\}}\times\xi_B\right)_{B\in\Psi}$ Let $\Omega(\theta,\theta',y,\xi,\Xi)$ the symmetric matrix involved in the previous equation. In order to bound this last quantity we study

the spectrum of this matrix. To do so we determine its characteristic. Let $x \in \mathbb{R}$. We call $n = n$ By using the cofactor expansion formula for the first line we find :

$$\begin{aligned} det\left\{xI_n - \Omega(\theta,\theta',y,\xi,\Xi)\right\} &= (x-1)^n - \frac{\sigma_{min}^2 r_{\theta'}^2(\xi)}{\sigma_{max}^4}(x-1)^{n-2} - \sum_{B\in\Psi}\frac{9\sigma_{min}^2 r_{\theta}^2(\xi_B)}{\sigma_{max}^4}(x-1)^{n-2} \\ &= (x-1)^{n-2}\left[(x-1)^2 - \left\{\frac{\sigma_{min}^2 r_{\theta'}^2(\xi)}{\sigma_{max}^4} + \sum_{B\in\Psi}\frac{9\sigma_{min}^2 r_{\theta}^2(\xi_B)}{\sigma_{max}^4}\right\}\right] \end{aligned}$$

therefore the spectrum of $\Omega(\theta,\theta',y,\xi,\Xi)$ is

$$\left\{1; 1-\sqrt{\frac{\sigma_{min}^2 r_{\theta'}^2(\xi)}{\sigma_{max}^4} + \sum_{B\in\Psi}\frac{9\sigma_{min}^2 r_{\theta}^2(\xi_B)}{\sigma_{max}^4}}; 1+\sqrt{\frac{\sigma_{min}^2 r_{\theta'}^2(\xi)}{\sigma_{max}^4} + \sum_{B\in\Psi}\frac{9\sigma_{min}^2 r_{\theta}^2(\xi_B)}{\sigma_{max}^4}}\right\}$$

and we get :

$$\begin{aligned} H(\theta,\theta',y,\Xi,\xi) &\leq -\frac{1}{2}\left\{1-\sqrt{\frac{\sigma_{min}^2 r_{\theta'}^2(\xi)}{\sigma_{max}^4} + \sum_{B\in\Psi}\frac{9\sigma_{min}^2 r_{\theta}^2(\xi_B)}{\sigma_{max}^4}}\right\}\left(\frac{\|y\|^2}{\sigma_{min}^2} + \|\tilde{\xi}\|^2 + \|\tilde{\Xi}\|^2\right) \\ &\leq -\frac{1}{2}\left\{1-\sqrt{\frac{\sigma_{min}^2 r_{\theta'}^2(\xi)}{\sigma_{max}^4} + \sum_{B\in\Psi}\frac{9\sigma_{min}^2 r_{\theta}^2(\xi_B)}{\sigma_{max}^4}}\right\}\left(\frac{\|y\|^2}{\sigma_{min}^2} + \|\xi\|^2 + \|\Xi\|^2\right) \end{aligned}$$

where the last equality holds because $\|\tilde{\xi}\| > \|\xi\|$ and $\|\tilde{\Xi}\| > \|\Xi\|$.

Let $\varepsilon > 0$ small enough such that :

$$0 < 1-\sqrt{\frac{\sigma_{min}^2\varepsilon^2}{\sigma_{max}^4} + \sum_{B\in\Psi}\frac{9\sigma_{min}^2\varepsilon^2}{\sigma_{max}^4}} < 1$$

Let $\omega = \sqrt{\frac{\sigma_{min}^2\varepsilon^2}{\sigma_{max}^4} + \sum_{B\in\Psi}\frac{9\sigma_{min}^2\varepsilon^2}{\sigma_{max}^4}}$

Let K a compact set such that assumption (2.7) is verified, and $K_M = \{y \in \mathbb{R}^J \leq M\}$ for some $M > 0$

we finally get to :

$$\begin{aligned} &\mathbb{E}_{\theta'}^y\left\{|f(y;\theta)^{-m}\prod_{B\in\Psi}\frac{\partial^{|B|}f(y;\theta)}{\prod_{b\in B}\partial\theta_b}|^3\right\} \\ &\quad\leq \int_{(\Xi,\xi,y)\in K^{m+1}\times K_M} M_\theta^\Psi(y,\Xi)e^{H(\theta,\theta',y,\Xi,\xi)}dyd\Xi d\xi \\ &\quad+ \int_{(\Xi,\xi,y)\in\mathbb{R}^{(m+1)p+J}\setminus K^{m+1}\times K_M} M_\theta^\Psi(y,\Xi)e^{-\frac{1}{2}(1-\omega)\left(\frac{\|y\|^2}{\sigma_{min}^2}+\|\xi\|^2+\|\Xi\|^2\right)}dyd\Xi d\xi \end{aligned}$$

Now, to be precise, we restart from equation (2.27) and the following calculation leading to equation (2.28), to show that the sup can be considered inside the integral :

$$\begin{aligned}
&\mathbb{E}^y_{\theta'}\left\{\sup_{\theta\in\Theta} \quad |f(y;\theta)^{-m}\prod_{B\in\Psi}\frac{\partial^{|B|}f(y;\theta)}{\prod_{b\in B}\partial\theta_b}|^3\right\}\\
&\leq \mathbb{E}^y_{\theta'}\left\{\sup_{\theta\in\Theta} \quad \mathbb{E}^{\Xi}\left(M^{\Psi}_{\theta}(y,\Xi)e^{-3\sum_{B\in\Psi}V(\theta,y,\xi_B)-V(y,\sigma)}\right)\right\}\\
&= \int_y \sup_{\theta\in\Theta}\left[\mathbb{E}^{\Xi}\left\{M^{\Psi}_{\theta}(y,\Xi)e^{-3\sum_{B\in\Psi}V(\theta,y,\xi_B)-V(y,\sigma)}\right\}\right]f_{\theta'}(y)dy\\
(f'_\theta(y)>0) &= \int_y \sup_{\theta\in\Theta}\left[\mathbb{E}^{\Xi}\left\{M^{\Psi}_{\theta}(y,\Xi)e^{-3\sum_{B\in\Psi}V(\theta,y,\xi_B)-V(y,\sigma)}\right\}f_{\theta'}(y)\right]dy\\
(Jensen) &\leq \int_y \mathbb{E}^{\Xi}\left[\sup_{\theta\in\Theta}\left\{M^{\Psi}_{\theta}(y,\Xi)e^{-3\sum_{B\in\Psi}V(\theta,y,\xi_B)-V(y,\sigma)}\right\}\right]f_{\theta'}(y)dy
\end{aligned}$$

where the second equality holds because $f_{\theta'}(y)$ is nonnegative and does not depend on θ. And finally :

$$\begin{aligned}
&\mathbb{E}^y_{\theta'}\left\{\sup_{\theta\in\Theta}|f(y;\theta)^{-m}\prod_{B\in\Psi}\frac{\partial^{|B|}f(y;\theta)}{\prod_{b\in B}\partial\theta_b}|^3\right\}\\
&\leq \int_{(\Xi,\xi,y)\in K^{m+1}\times K_M}\sup_{\theta\in\Theta}M^{\Psi}_{\theta}(y,\Xi)e^{H(\theta,\theta',y,\Xi,\xi)}dyd\Xi d\xi\\
&+\int_{(\Xi,\xi,y)\in\mathbb{R}^{(m+1)p+J}\setminus K^{m+1}\times K_M}\sup_{\theta\in\Theta}M^{\Psi}_{\theta}(y,\Xi)e^{-\frac{1}{2}(1-\omega)\left(\frac{\|y\|^2}{\sigma^2_{min}}+\|\xi\|^2+\|\Xi\|^2\right)}dyd\Xi d\xi \qquad (2.29)\\
&\leq \sup_{\substack{\theta\in\Theta\\(\Xi,\xi,y)\in K^{m+1}\times K_M}} Vol(K^{m+1}\times K_M)M^{\Psi}_{\theta}(y,\Xi)e^{H(\theta,\theta',y,\Xi,\xi)}\\
&+\int_{(\Xi,\xi,y)\in\mathbb{R}^{(m+1)p+J}\setminus K^{m+1}\times K_M}\sup_{\theta\in\Theta}M^{\Psi}_{\theta}(y,\Xi)e^{-\frac{1}{2}(1-\omega)\left(\frac{\|y\|^2}{\sigma^2_{min}}+\|\xi\|^2+\|\Xi\|^2\right)}dyd\Xi d\xi
\end{aligned}$$

The first term of the last quantity is a suprema of a continuous function over a compact set and is therefore finite.

$M^{\Psi}_{\theta}(y,\Xi)$ is a polynomial with respect to $\|y\|$ and the partial derivatives of $\theta\mapsto g_\theta(\xi_B)$ for each $B\in\Psi$, therefore, thanks to assumption (2.6), considering ε small enough such that $\varepsilon<\delta$ where δ is defined in proposition (2.7), we finally get to the conclusion :

$$\mathbb{E}^y_{\theta'}\left\{\sup_{\theta\in\Theta}|f(y;\theta)^{-m}\prod_{B\in\Psi}\frac{\partial^{|B|}f(y;\theta)}{\prod_{b\in B}\partial\theta_b}|^3\right\}<+\infty$$

In equation (2.29) is is important to notice that the last term no longer depends on θ', and that the first term in the sup is a continuous function of θ', therefore by compacity of Θ the suprema can be taken with respect to θ' and

$$\sup_{\theta' \in \Theta} \quad \mathbb{E}^y_{\theta'} \left\{ \sup_{\theta \in \Theta} |f(y;\theta)^{-m} \prod_{B \in \Psi} \frac{\partial^{|B|} f(y;\theta)}{\prod_{b \in B} \partial \theta_b}|^3 \right\} < +\infty$$

which concludes the proof.

2.7.13 . Linear model specific case

when considering a linear model, for every $i = 1, ..., N$ the model writes :

$$y_{ij} = x_i^T \beta + Z_i \Lambda \xi_i + \varepsilon_i$$

where $\xi_i \sim \mathcal{N}(0, I)$ and independent from $\varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{J_i})$. Therefore we have :

$$y_{ij} \sim \mathcal{N}\left(x_i^T \beta, Z_i \Lambda \Lambda^T Z_i^T + \sigma^2 I_{J_i}\right)$$

And the log-likelihood is explicit :

$$\log f(y_i;\theta) = \frac{-J_i}{2} \log det\left(Z_i \Lambda \Lambda^T Z_i^T + \sigma^2 I_{J_i}\right) - (y_i - X_i\beta)^T \left\{Z_i \Lambda \Lambda^T Z_i^T + \sigma^2 I_{J_i}\right\}^{-1} (y_i - X_i \beta)$$

y_i is Gaussian, therefore have all its moments finite. Given that σ^2 is strictly nonnegative, $\log f(y_i;\theta)$ is infinitely differentiable on Θ, and each of its derivatives are quadratic forms of y_i, and therefore have finite moments. Which verifies assumption (2.2).

2.7.14 . Logistic growth model example

We consider here the logistic growth model that is commonly used in many fields. We show how to use the criteria given in proposition (2.7).

We recall the definition of the logistic growth function :

$$g(x_{ij}, \beta, \Lambda\xi_i) = \frac{\beta_1 + \lambda_1 \xi_{i1}}{1 + \exp\left(-\frac{x_{ij} - (\beta_2 + \lambda_2 \xi_{i2})}{\beta_3 + \lambda_3 \xi_{i3}}\right)}. \tag{2.30}$$

To verify that assumption (2.6) is verified, we need to calculate the derivatives of g with respect to θ. To avoid heavy calculations we consider the parameter λ_3 :

$$\begin{aligned}\frac{\partial g(x_{ij}, \beta, \Lambda\xi_i)}{\partial \lambda_3} =& -\xi_{i3} \left(\frac{j - (\beta_2 + \lambda_2 \xi_{i2})}{(\beta_3 + \lambda_3 \xi_{i3})^2}\right) \exp\left(-\frac{j - (\beta_2 + \lambda_2 \xi_{i2})}{\beta_3 + \lambda_3 \xi_{i3}}\right) \\ &\times \frac{\beta_1 + \lambda_1 \xi_{i1}}{\left(1 + \exp\left(-\frac{j - (\beta_2 + \lambda_2 \xi_{i2})}{\beta_3 + \lambda_3 \xi_{i3}}\right)\right)^2}\end{aligned}$$

The only issue of this quantity occurs at $\xi_{i3} = -\frac{\beta_3}{\lambda_3}$ that tends toward 0 as $\xi_{i3} \to -\frac{\beta_3}{\lambda_3}$. Finally this quantity is integrable with respect to the Gaussian distribution. By iterating the derivatives, we find expressions similar to this one and assumption (2.6) is verified.

Let $\varepsilon > 0, M > 0$, and we define the set :

$$K_M = \{\xi \in \mathbb{R}^3 : |\xi_1| \leq M, |\xi_2| \leq M|\xi_1|, |\xi_3| \leq M|\xi_1|\}$$

Therefore for every $\xi \in \mathbb{R}^3 \backslash K_M$:

$$\begin{aligned}\frac{\|g(x_{ij}, \beta, \Lambda\xi)\|^2}{\|\xi\|^2} &\leq \frac{\lambda_1^2\xi_1^2}{\xi_1^2 + \xi_2^2 + \xi_3^2} \\ &\leq \frac{\lambda_1^2\xi_1^2}{M^2\xi_1^2} \\ &\leq \frac{\lambda_1^2}{M^2}\end{aligned}$$

By taking $M > \frac{\lambda_1}{\varepsilon}$, it verifies assumption (2.7).

2.7.15 . Pharmacokinetic model

The pharmokinetic study of theophylline concentration along time is a well known example in the literature of mixed effects models , see for example Davidian and Giltinan (2017). The simplified model is defined as follows, the theophylline concentration of patient i ($i = 1, ..., N$) at time t_j ($j = 1, ..., J$) is modeled as :

$$\begin{cases} y_{ij} = \frac{Dk_{ai}}{V_i k_{ai} - Cl_i} \left\{ \exp\left(-k_{ai} t_j\right) - \exp\left(-\frac{Cl_i}{V_i} t_j\right) \right\} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ (k_{ai}, Cl_i, V_i) = \left(\exp\left\{\beta_1 + \lambda_1 \xi_{i1}\right\}, \exp\left\{\beta_2 + \lambda_2 \xi_{i2}\right\}, \exp\left\{\beta_3 + \lambda_3 \xi_{i3}\right\}\right) \\ (\xi_{i1}, \xi_{i2}, \xi_{i3})^T \sim \mathcal{N}\left(0, I_3\right) \end{cases}$$

where D is a specific constant of the experiment, and (k_{ai}, Cl_i, V_i) are individual parameters describing the biological process (rate of drug absorption, rate of drug elimination, complete volume of blood with drugs). The exponentials enforce positivity of these parameters. In this setting,

$$g(x_{ij}, \beta, \Lambda\xi_i) = \frac{Dk_{ai}}{V_i k_{ai} - Cl_i} \left\{ \exp\left(-k_{ai} t_j\right) - \exp\left(-\frac{Cl_i}{V_i} t_j\right) \right\}$$

(where $x_{ij} = t_j$ and Λ is diagonal with diagonal $(\lambda_1, \lambda_2, \lambda_3)^T$) is a bounded function of $(\xi_{i1}, \xi_{i2}, \xi_{i3})$, and therefore the criterion is straightforward to verify.

Chapitre 3

SARIS : une nouvelle procédure de calcul de ratios de constantes de normalisation

Ce chapitre présente le travail pré-publié suivant :

Guédon, T., Baey, C., & Kuhn, E. (2024). Estimation of ratios of normalizing constants using stochastic approximation : the SARIS algorithm. arXiv preprint arXiv :2408.13022. (Guédon et al., 2024b)

Le calcul de ratios de constantes de normalisation de densités de probabilité est souvent nécessaire lorsque des études statistiques sont menées. Deux exemples communs sont le calcul de la statistique du rapport de vraisemblance dans les modèles à variables latentes, et celui du facteur de Bayes en statistiques bayésiennes. Ces ratios dépendent d'intégrales n'ayant souvent pas de forme analytique, ce qui rend leur approximation complexe. Usuellement, afin de répondre à cette problématique des méthodes de Monte-Carlo sont utilisées. Ce chapitre présente une nouvelle méthode basée sur le principe d'approximation stochastique afin de calculer des ratios de constantes de normalisation.

Les méthodes d'approximation stochastique (Robbins and Monro, 1951; Duflo, 1996; Benveniste et al., 2012) consistent à approcher de manière itérative un zéro d'une fonction définie par une espérance. L'algorithme Robbins-Monro présenté dans Robbins and Monro (1951) est à la base des méthodes d'approximation stochastique. Il vise à résoudre, en θ, une équation de la forme :

$$\mathbb{E}\left[H(X, \theta)\right] = 0$$

où $X \sim P$ est une variable aléatoire et H une fonction connue.

L'algorithme est défini de manière itérative et fournit, à l'itération k, l'estimateur θ_k de la manière récursive suivante :

$$\theta_{k+1} = \theta_k + \gamma_{k+1} H(X_{k+1}, \theta_k), \quad X_{k+1} \sim P$$

où $(\gamma_k)_k$ est une suite de pas positifs.

Sous certaines conditions, la séquence $(\theta_k)_{k\geq 0}$ converge presque sûrement vers θ^* défini de telle sorte que :

$$\mathbb{E}\left(H(X, \theta^*)\right) = 0$$

Les méthodes de gradient stochastique font partie de cette famille d'algorithmes.

L'algorithme proposé est appelé SARIS pour *Stochastic Approximation of Ratio Importance Sampling*. Il est basé sur l'identité suivante :

$$\mathbb{E}_\pi \left[\frac{f_0(Z) - r^* f_1(Z)}{\pi(Z)} \right] = 0$$

où $r^* = \frac{\int_{\mathcal{Z}} f_0(z)\mu(dz)}{\int_{\mathcal{Z}} f_1(z)\mu(dz)}$ avec f_i ($i = 0, 1$) sont des densités non normalisées connues et π est une densité d'échantillonnage préférentiel. Dans le contexte des tests du rapport de vraisemblance dans un modèle à variables latentes, les fonctions f_i ($i = 0, 1$) sont les vraisemblances complètes évaluées aux maxima de vraisemblance restreint et non restreint. Les variables z correspondent aux variables latentes. Dans un cadre bayésien, les variables z correspondent aux paramètres, et les densités non normalisées sont les distributions *a posteriori* non normalisées des paramètres sous différents modèles.

Ce chapitre présente l'algorithme SARIS permettant d'estimer le ratio r^* par approximation stochastique. L'estimateur obtenu est fortement consistant et asymptotiquement normal. La densité de proposition π minimisant la variance asymptotique est déterminée. La densité de proposition optimale n'étant connue qu'à une constante près, un algorithme étendu SARIS-EXT est également proposé, permettant d'utiliser des densités de proposition non normalisées, tant qu'il est possible d'en obtenir des échantillons. Il est démontré que l'estimateur obtenu présente les mêmes propriétés asymptotiques que l'estimateur SARIS.

La méthode proposée est comparée à deux méthodes de la littérature : le bridge sampling de Meng and Wong (1996) et le ratio importance sampling de Chen and Shao (1997a). Ce dernier possède les mêmes propriétés théoriques que SARIS étant basé sur la même identité, son schéma optimal n'est cependant pas applicable facilement. Un schéma non optimal inspiré du bridge sampling, utilisant des échantillons simulés selon les distributions non normalisées f_i ($i = 0, 1$) est présenté dans Chen and Shao (1997a). Une approche similaire est également permise par l'algorithme SARIS, appelée SARIS-MIXT. Une comparaison théorique de ces trois méthodes est proposée en section 3.4.

Par ailleurs, l'aspect itératif de la procédure proposée permet de l'intégrer dans une procédure d'estimation des paramètres dans des modèles à variables latentes de type gradient stochastique ou EM stochastique, et d'ainsi profiter de l'effort computationnel réalisé pour l'inférence. Cette procédure jointe est présentée en section 3.4.2.

Des simulations comparent les différentes méthodes présentées, et illustrent la robustesse de l'algorithme SARIS à un faible recouvrement des supports entre les densités non normalisées f_0 et f_1. Les performances de la procédure jointe sont également illustrées dans un modèle de régression à variables manquantes.

Estimation of ratios of normalizing constants using stochastic approximation : the SARIS algorithm

3.1 . Introduction

In statistical modeling, comparing models often hinges on estimating ratios of integrals, which frequently serve as normalizing constants of posterior distributions. For example in latent variables models, such ratios emerge when choosing between two nested models via the likelihood ratio test. Each marginal likelihood represents the normalizing constant of the posterior distribution density of the latent variables given the data. However, computing marginal likelihoods defined as integrals becomes infeasible when the relationship between observed data and latent variables is complex. Similarly, in Bayesian statistics, model selection can be performed by comparing evidences (or marginal likelihoods) between competing models through the Bayes factor. In this setting, marginal likelihoods serve as the normalizing constants of the posterior distribution of the parameters given the data. Therefore, being able to efficiently compute ratios of normalizing constants is of significant practical interest.

This topic has motivated many fields of applications such as phylogenetics (Lartillot and Philippe, 2006), astrophysics (Russel et al., 2018), psychology (Annis et al., 2019) or chemical physics (Shirts and Chodera, 2008). To tackle this task one can either separately estimate the two likelihoods, or directly compute the ratio. As far as marginal likelihood estimation is concerned, several classical methods exist such as importance sampling (Robert and Casella, 1999), the harmonic mean estimator (Newton and Raftery, 1994) or the generalized harmonic mean estimator (Gelfand and Dey, 1994), using samples from the posterior distributions. It has been shown that those methods are often particular cases of estimators used to directly compute ratios of normalizing constants such as Bridge sampling (Meng and Wong, 1996) and ratio importance sampling (Chen and Shao, 1997b), both first introduced in the physics literature by respectively Bennett (1976) and Torrie and Valleau (1977). Gronau et al. (2017) highlights that the Bridge sampling estimator is superior to both the importance sampling and the general harmonic mean estimators, as it is more robust to the choice of the proposal distribution. However, its performance deteriorates as the overlap between the two densities associated with the normalizing constants decreases. To circumvent this issue, refinements were proposed in Meng and Schilling (2002) and Wang et al. (2022). They rely on a modification of the samples and their associated densities in order to increase the overlap between the densities. However, these refinements require some knowledge of the distributions under consideration, making the approach less attractive for practitioners. The optimal ratio importance sampling estimator presents the smallest asymptotic variance among the estimators mentioned here. However, the optimal scheme is not tractable in practice, which might explain why it is not often considered

in the literature. Several other refined methods have been developed in the last decades, based on intermediate distributions that create a path between the two distributions. We can mention for example annealed importance sampling (Neal, 2001), sequential Monte Carlo (Del Moral et al., 2006), path sampling (Gelman and Meng, 1998) (or thermodynamic integration), stepping stone sampling (Xie et al., 2011) and generalized stepping stone sampling (Fan et al., 2011). These methods received a lot of interest and are presented and reviewed for instance in Friel and Wyse (2012) and more exhaustively in Llorente et al. (2023). However, most of these schemes can be seen as refinements of elementary methods to compute ratios of normalizing constants presented in Chen and Shao (1997b). For example, path sampling is an extension of Bridge sampling (Gelman and Meng, 1998), and stepping stone sampling is an extension of the pre-umbrella identity using intermediate power posteriors distributions.

We propose here a new approach to calculate ratios of normalizing constants based on a stochastic approximation algorithm. Our procedure is also related to ratio importance sampling, and is therefore called Stochastic Approximation of Ratio Importance Sampling (SARIS). It benefits from the different refinements available in these two fields. After showing that estimating a ratio of normalizing constants is equivalent to finding the root of a function defined as an expectation, we develop an iterative stochastic approximation scheme to compute this root. We show that the sequence generated by the proposed algorithm converges almost surely towards the targeted ratio. The main advantage of our approach is that, thanks to its iterative construction, there is no need to fix the computational effort ahead since the procedure can be stopped once a convergence criterion has been reached. We also express the optimal proposal distribution in terms of asymptotic variance using convergence results from stochastic approximation theory. Moreover our method allows to reach the same asymptotic variance as the theoretical one of the optimal ratio importance sampling estimator.

This paper is organized as follows : the next section describes the context and the objective and gives a quick review of existing methods. Section 3.3 presents the proposed SARIS procedures and studies their theoretical properties. Section 3.4 is dedicated to algorithms using solely samples from both distributions involved in the targeted ratio. An extension for the simultaneous estimation of model parameters and likelihood ratio test statistic in latent variables models is also proposed. Section 3.5 is dedicated to numerical experiments and practical guidelines. We conclude and discuss the perspectives in Section 3.6. The proofs are postponed to the appendix.

3.2 . Ratios of normalizing constants

In this section we introduce ratios of normalizing constants, the notations and their uses in statistics. We illustrate and motivate our purpose through two concrete examples : the likeli-

hood ratio test statistic in latent variables models and the Bayes factor in Bayesian statistics.

3.2.1 . Statistical setting and objective

Let d be a positive integer and μ a σ-finite positive Borel measure on a subspace $\mathcal{Z}$ of $\mathbb{R}^d$. Assume that f_0 and f_1 are two positive integrable Borel functions on $\mathcal{Z}$ such that $c_0 = \int_{\mathcal{Z}} f_0(z)\mu(dz) > 0$ and $c_1 = \int_{\mathcal{Z}} f_1(z)\mu(dz) > 0$. We assume that these normalizing constants c_0 and c_1 are unknown and introduce the two probability densities with respect to μ denoted by p_0 and p_1 defined for $i \in \{0, 1\}$ and for all z in $\mathcal{Z}$ by

$$p_i(z) = \frac{f_i(z)}{c_i}.$$

The objective is to estimate the ratio r^* of normalizing constants defined as :

$$r^* = \frac{c_0}{c_1} = \frac{\int_{\mathcal{Z}} f_0(z)\mu(dz)}{\int_{\mathcal{Z}} f_1(z)\mu(dz)}. \tag{3.1}$$

We first motivate our contribution by two practical examples which require the computation of such ratios.

Example 3.1 (**computation of likelihood ratio test statistic in latent variables model**). *Let us consider a general latent variables model where the observed variable is given by the random variable Y, taking values in $\mathcal{Y}$ and the latent variable by Z, taking values in $\mathcal{Z}$. We denote by y the observation of Y. We assume that the random vector (Y, Z) follows a parametric distribution with density f_θ parameterized by $\theta \in \Theta$. The objective is to test the hypotheses :*

$$H_0 : \quad \theta \in \Theta_0 \quad against \quad H_1 : \quad \theta \in \Theta_1,$$

where $\Theta_0 \subset \Theta_1 \subset \Theta$. A natural popular test is the likelihood ratio test (Van der Vaart, 2000) which statistic is defined by :

$$LR = -2 \log \left(\frac{L(\hat{\theta}_0; y)}{L(\hat{\theta}_1; y)} \right),$$

and where the marginal likelihood $L(\theta; y)$ and the maximum likelihood estimates $\hat{\theta}_i$, for $i \in \{0, 1\}$, are defined respectively by

$$L(\theta; y) = \int_{\mathcal{Z}} f_\theta(y; z)\mu(dz)$$

and

$$\hat{\theta}_i = \arg\max_{\theta \in \Theta_i} \quad L(\theta; y).$$

Accurately estimating LR is crucial as its value determines whether H_0 is rejected or not.

Example 3.2 (**computation of Bayes factor for Bayesian model choice**). *In Bayesian statistics, the parameter $\theta \in \Theta$ is considered as a random variable with a known prior distribution $p(\theta)$. The*

posterior distribution of θ given a dataset D is defined as the product of the prior density and the likelihood of the model. To compare two models M_1 and M_2 and choose the one that better fits the data, the Bayes factor B_{12} (Gelfand and Dey, 1994) is a powerful tool. It is defined as :

$$B_{12} = \frac{p(D \mid M_1)}{p(D \mid M_2)}$$

where $p(D \mid M_i) = \int_\Theta p(D \mid \theta, M_1)p(\theta)d\theta$ is the marginal likelihood of model M_i, for $i \in \{1, 2\}$.

3.2.2 . State of the art

When the ratio (3.1) has no explicit expression, its computation can be performed by evaluating separately the numerator and the denominator. This can be done, for instance, using the harmonic mean estimator of Newton and Raftery (1994). In a Bayesian context, this estimator uses draws from the posterior distribution to compute the inverse of the marginal likelihood. However, it is known to overestimate the marginal likelihood and can have infinite variance. Another solution is to use importance sampling (Robert and Casella, 1999), which requires the introduction of a proposal distribution close to the integrand. However, it may be challenging to build such a proposal in complex settings. Furthermore, importance sampling is very sensitive to a misspecification of the proposal with respect to the density of interest (Gronau et al., 2017). These methods are specific cases of more general ones that aim at estimating ratios of normalizing constants.

In this section we focus on two existing methods that will serve as comparison for the proposed methodology : i) the Bridge sampling which is particularly popular (Meng and Schilling, 2002; Frühwirth-Schnatter, 2004; Gronau et al., 2017, 2020), and ii) the ratio importance sampling (Chen and Shao, 1997b) which is strongly linked to our approach. For a more precise and exhaustive review we refer the reader to Llorente et al. (2023, sections 4.1 and 4.2), and to Chen and Shao (1997b). In the sequel, we denote by Z any random variable defined on a probability space $(\Omega, \mathcal{A}, P)$ taking values in $\mathcal{Z}$, and by $\mathbb{E}_i$ the expectation with respect to density p_i for $i \in \{0, 1\}$.

Bridge sampling First introduced in Bennett (1976) and later reintroduced in Meng and Wong (1996), the Bridge sampling is based on the following identity :

$$r^* = \frac{\mathbb{E}_1 \left[f_0(Z)\alpha(Z)\right]}{\mathbb{E}_0 \left[f_1(Z)\alpha(Z)\right]},$$

where α is a non-negative function defined on $\mathcal{Z}$ verifying $0 < \int_{\mathcal{Z}} \alpha(z)p_0(z)p_1(z)\mu(dz) < +\infty$. Let K be a fixed positive integer. Then the Bridge sampling estimator of r^* is obtained using

two K-samples $(Z_k^0)_{1\leq k\leq K}$ and $(Z_k^1)_{1\leq k\leq K}$ from p_0 and p_1 respectively as follows :

$$\hat{r}_K^{BS} = \frac{\sum_{k=1}^K f_0(Z_k^1)\alpha(Z_k^1)}{\sum_{k=1}^K f_1(Z_k^0)\alpha(Z_k^0)}. \tag{3.2}$$

This approach is particularly popular since in most statistical contexts it is straightforward to apply once a first inference step has been performed. This is the case for instance in the two examples introduced in the previous section. In the first example dealing with hypotheses testing in latent variables models, p_i is the posterior distribution of the latent variables given the data under hypothesis H_i. In the second example of Bayesian model choice, p_i is the posterior distribution of the parameter given the data under model M_i. In both contexts, sampling from these distributions is part of the entire estimation process, therefore no additional work is required regardless of the complexity of the distributions.

Meng and Wong (1996) showed that the optimal choice of α, that minimizes the mean square error of the estimator and its asymptotic variance, is given by :

$$\alpha_{bridge}^{opt}(z) \propto \frac{1}{p_1(z)+p_0(z)} \propto \frac{1}{r^* f_1(z)+f_0(z)}. \tag{3.3}$$

It reaches the following optimal normalized asymptotic variance :

$$V_{bridge}^{opt} = 4r^{*2}\left[\left(\int_{\mathcal{Z}} \frac{2p_0(z)p_1(z)}{p_0(z)+p_1(z)}\mu(dz)\right)^{-1} - 1\right]. \tag{3.4}$$

As α_{bridge}^{opt} depends on the unknown ratio r^*, it is not possible to use it directly in practice. Therefore the authors propose an iterative scheme to reach the optimal asymptotic variance. Starting from an initial guess $\hat{r}_K^{(0)}$, and using the two K-samples $(Z_k^0)_{1\leq k\leq K}$ and $(Z_k^1)_{1\leq k\leq K}$ from p_0 and p_1 defined above, we get :

$$\hat{r}_K^{(t+1)} = \left(\sum_{k=1}^K \frac{f_0(Z_k^1)}{\hat{r}_K^{(t)} f_1(Z_k^1)+f_0(Z_k^1)}\right) \Bigg/ \left(\sum_{k=1}^K \frac{f_1(Z_k^0)}{\hat{r}_K^{(t)} f_1(Z_k^0)+f_0(Z_k^0)}\right). \tag{3.5}$$

As t grows to infinity this estimator converges towards $\tilde{r}_K^{BS}$ defined as :

$$\tilde{r}_K^{BS} = \left(\sum_{k=1}^K \frac{f_0(Z_k^1)}{\tilde{r}_K^{BS} f_1(Z_k^1)+f_0(Z_k^1)}\right) \Bigg/ \left(\sum_{k=1}^K \frac{f_1(Z_k^0)}{\tilde{r}_K^{BS} f_1(Z_k^0)+f_0(Z_k^0)}\right)$$

which can be rewritten as :

$$\sum_{k=1}^K \frac{f_0(Z_k^1)}{\tilde{r}_K^{BS} f_1(Z_k^1)+f_0(Z_k^1)} - \sum_{k=1}^K \frac{\tilde{r}_K^{BS} f_1(Z_k^0)}{\tilde{r}_K^{BS} f_1(Z_k^0)+f_0(Z_k^0)} = 0.$$

This last equation shows that the optimal Bridge sampling estimator can also be defined as the root of a function. Note that the solution to this equation nullifies the score function of Geyer's likelihood described in Geyer (1994a), leading to the reverse logistic regression estimator.

Even if Bridge sampling is more robust than other methods as mentioned above, it still suffers from a too small overlap between distributions p_0 and p_1. Indeed, when this overlap vanishes, the optimal variance grows to infinity. In such cases, more refined methods have been developed (see Meng and Schilling (2002); Wang et al. (2022)) that modify the two distributions considered without changing the normalizing constant. However, these methods are more involved since they require some additional effort to work properly.

Ratio importance sampling The ratio importance sampling (RIS) estimator was first introduced in the physics literature in Torrie and Valleau (1977) as umbrella sampling and then rediscovered in Chen and Shao (1997b). It generalizes the importance sampling estimator to compute a ratio of normalizing constants. Considering a positive density function π on $\mathcal{Z}$, dominated by p_0 and p_1, equation (3.1) can be written as follows :

$$r^* = \frac{\int_{\mathcal{Z}} f_0(z)\mu(dz)}{\int_{\mathcal{Z}} f_1(z)\mu(dz)} = \frac{\mathbb{E}_\pi\left[\frac{f_0(Z)}{\pi(Z)}\right]}{\mathbb{E}_\pi\left[\frac{f_1(Z)}{\pi(Z)}\right]}. \tag{3.6}$$

Equation (3.6) is called the *ratio importance sampling identity*, which will also be the basis of the methodology proposed in this paper. The ratio importance sampling estimator can be obtained using a $2K-$sample $(Z_k)_{1\leq k\leq 2K}$ from π as follows :

$$\hat{r}_K^{RIS} = \frac{\sum_{k=1}^{2K} f_0(Z_k)/\pi(Z_k)}{\sum_{k=1}^{2K} f_1(Z_k)/\pi(Z_k)}. \tag{3.7}$$

Remark 3.1. *Contrary to the Bridge sampling estimator that uses two samples, the Ratio Importance sampling estimator only requires one. That is why it is presented here using a sample of size $2K$.*

Note that identity (3.6) is very interesting and very general. For example, by taking $\pi = p_1$ it leads to $r^* = \mathbb{E}_1\left[\frac{f_0(Z)}{f_1(Z)}\right]$ which gives an unbiased estimator of r^*. However, this estimator does not reach the optimal asymptotic variance. Indeed, Chen and Shao (1997b) showed that the optimal proposal density that minimizes the asymptotic variance of the estimator is given by :

$$\pi_{ris}^{opt}(z) \propto |p_1(z) - p_0(z)| \propto |f_1(z)r^* - f_0(z)|. \tag{3.8}$$

It reaches the following optimal asymptotic variance :

$$V_{ris}^{opt} = r^{*2}\left(\int_{\mathcal{Z}} |p_1(z) - p_0(z)|\mu(dz)\right)^2. \tag{3.9}$$

Chen and Shao (1997b, Theorem 3.3) showed that the variance of the optimal ratio importance sampling estimator is smaller than the variance of the optimal Bridge sampling estimator. Furthermore, when the overlap between p_0 and p_1 goes to 0, the optimal variance V_{ris}^{opt} converges to $4r^*$ which is bounded. This is a clear advantage compared to the previous methodology. However, contrary to the Bridge sampling setting, there is no straightforward procedure to approximate the optimal scheme and reach the optimal asymptotic variance. Chen and Shao (1997b) suggested a two-stage approach consisting in building a first consistent estimator $\hat{r}$ of r^* based on a chosen method, and then use it as a plug-in estimator in (3.8). Since the first step can be difficult to achieve, the optimal scheme can be difficult to implement in practice. This might partly explain why RIS is not as popular as Bridge sampling.

3.3 . Stochastic approximation procedures to compute ratio of normalizing constants

Given the limitations raised by the two approaches presented in the previous section, namely the sensitivity to a small overlap of the two distributions p_0 and p_1 for Bridge sampling, and the little practical applicability for RIS, there is a need for a new method that could address these issues.

In this section we propose an approach based on stochastic approximation principles. Our procedure convert the ratio importance sampling identity (3.6) into a root finding program, which brings several advantages. First, and contrary to usual Monte Carlo computation, there is no need to fix the sample size ahead of the procedure. Second, the stochastic approximation framework enables the use of sampling distributions that depend on the current estimate of the unknown ratio r^*, circumventing the main obstacle of RIS, but still enjoying its good theoretical properties. Indeed, the obtained sequence of estimates is almost surely convergent and asymptotically Gaussian. It reaches the same asymptotic variance as the one of the optimal RIS estimator, with an applicable scheme. As a consequence, it is much more robust to little overlap between p_0 and p_1, avoiding the major drawback of Bridge sampling.

3.3.1 . Description of the SARIS algorithm

Let π be a positive density function on $\mathcal{Z}$. Starting from equation (3.1), rewritten as :

$$\int_{\mathcal{Z}} (f_0(z) - r^* f_1(z))\mu(dz) = 0, \tag{3.10}$$

we can write :

$$\mathbb{E}_\pi \left[\frac{f_0(Z) - r^* f_1(Z)}{\pi(Z)} \right] = 0, \tag{3.11}$$

where $\mathbb{E}_\pi$ stands for the expectation with respect to the density π.

Calculating r^* is now equivalent to finding the root of a function defined as an expectation, and can therefore be solved using stochastic approximation algorithms. Assuming that r_0 is given in $\mathbb{R}$ and that one can sample independent draws from π, we thus consider the sequence $(r_k)_{k\geq 0}$ of estimators of r^*, defined by the following recursion for every positive integer k :

$$r_{k+1} = r_k + \gamma_{k+1} \frac{f_0(Z_{k+1}) - r_k f_1(Z_{k+1})}{\pi(Z_{k+1})}, \quad \text{with } Z_{k+1} \sim \pi, \tag{3.12}$$

and where $(\gamma_k)_{k\geq 0}$ is a sequence of positive decreasing step sizes.

The main task in the construction of the sequence is the choice of the proposal distribution π. Most of the relevant choices for π might depend on the true ratio r^*. For example the optimal choice for the Bridge sampling involves the quantity $p_0 + p_1$, depending on r^* which can not be evaluated. This is also the main drawback to the use of optimal ratio importance sampling. Thanks to its iterative nature, our methodology allows to consider proposal distributions which might depend on r^*. More precisely, let us consider a positive density function π_r on $\mathcal{Z}$ which depends on r. Equation (3.11) can be written as :

$$\mathbb{E}_{\pi_{r^*}}\left[\frac{f_0(Z) - r^* f_1(Z)}{\pi_{r^*}(Z)}\right] = 0. \tag{3.13}$$

Such equations can be solved using the Robbins-Monro algorithm (Robbins and Monro, 1951) that is based on the following stochastic recursion defined for every positive integer k :

$$r_{k+1} = r_k + \gamma_{k+1} \frac{f_0(Z_{k+1}) - r_k f_1(Z_{k+1})}{\pi_{r_k}(Z_{k+1})}, \quad \text{with } Z_{k+1} \sim \pi_{r_k} \tag{3.14}$$

The general algorithm is called SARIS (Stochastic Approximation Ratio Importance Sampling) and is summarized in Algorithm 3.1.

Algorithm 3.1 SARIS algorithm

Input : $(\gamma_k)_{k\geq 0}$, r_0, stopping criterion
Until stopping criterion :
  Draw Z_{k+1} from π_{r_k}
  Update $r_{k+1} = r_k + \gamma_{k_1} \frac{f_0(Z_{k+1}) - r_k f_1(Z_{k+1})}{\pi_{r_k}(Z_{k+1})}$
  $k = k + 1$
Return r_k

Remark 3.2. *The iterative structure of our procedure enables to introduce a stopping criterion. This is not the case in many other methods, in particular for ratio importance sampling and Bridge sampling.*

Indeed in those two cases it is not possible to compute r_{k+1} given r_k, therefore one should fix the sampling size at the beginning of the procedure.

Remark 3.3. *In most real-life applications, it is difficult to independently and exactly simulate from complex distributions. Therefore, the simulation step in* (3.14) *might be intractable. It is however possible to use the transition kernel of an ergodic Markov Chain having π_r as invariant distribution. One common practical choice for such Markov Chain Monte Carlo (MCMC) sampling scheme is the Metropolis-Hastings or the Metropolis-within-Gibbs algorithm (Robert and Casella, 1999). The recursive scheme* (3.14) *can therefore be generalized as follows for every positive integer k :*

$$r_{k+1} = r_k + \gamma_{k+1} \frac{f_0(Z_{k+1}) - r_k f_1(Z_{k+1})}{\pi_{r_k}(Z_{k+1})} \qquad Z_{k+1} \sim \Pi_{r_k}(.; Z_k) \tag{3.15}$$

where $\Pi_r(.,.)$ is a transition kernel of an ergodic Markov chain with invariant distribution π_r.

Remark 3.4. *We emphasize that the SARIS algorithm can also be used to compute a single marginal likelihood. If we know a density p only up to a normalizing constant c, $p = f/c$, then by introducing a known normalized density g, the $SARIS$ algorithm can be used to compute the ratio $r^* = c$.*

Remark 3.5. *Finally, the proposed procedure enables to directly estimate any strictly monotonous and invertible transformation g of the ratio. Suppose that the objective is to compute $g(r^*)$ (for example $g = -2\log$ to obtain a likelihood ratio statistic). The recursive scheme* (3.14) *can be easily modified to estimate $g(r^*)$ with the sequence $(g_k)_{k\geq 0}$ defined as follows :*

$$g_{k+1} = g_k + \gamma_{k+1} \frac{f_0(Z_{k+1}) - g^{-1}(g_k) f_1(Z_{k+1})}{\pi_{g^{-1}(g_k)}(Z_{k+1})} \qquad Z_{k+1} \sim \pi_{g^{-1}(g_k)} \tag{3.16}$$

Algorithm 3.1 can be easily adapted using this recursion, only changing the updating rule.

3.3.2 . Theoretical property of the SARIS algorithm

In this section we study the theoretical convergence property of the sequence $(r_k)_{k\geq 0}$ generated by the SARIS procedure described in Algorithm 3.1. We emphasize that such a setting is less general than the recursion defined in (3.15), however its theoretical study corresponds to the one of Robbins and Monro (1951). Moreover it allows a more fair comparison with the Monte Carlo methods presented in section 3.2.2, for which the theory was established for independent and identically distributed sampling.

We first state some regularity assumptions on the functions f_0, f_1 and on the densities $\{\pi_r, r \in \mathbb{R}\}$:

Assumption 3.1. *The functions f_0 and f_1 are positive integrable and for every $z \in \mathcal{Z}$, $r \mapsto \pi_r(z)$ is continuous.*

Assumption 3.2.

$$\mathbb{E}_0\left(\sup_{r\in\mathbb{R}}\left|\frac{f_0(Z)-rf_1(Z)}{\pi_r(Z)}\right|\right)+\mathbb{E}_1\left(\sup_{r\in\mathbb{R}}\left|\frac{f_0(Z)-rf_1(Z)}{\pi_r(Z)}\right|\right)<+\infty.$$

This assumption ensures the integrability of the main quantities involved in the algorithm. We also state a common assumption on the sequence of step sizes $(\gamma_k)_{k\geq 0}$.

Assumption 3.3. *The sequence $(\gamma_k)_{k\geq 0}$ is positive, decreasing and verifies $\sum_{k=0}^{+\infty}\gamma_k=+\infty$ and $\sum_{k=0}^{+\infty}\gamma_k^2<+\infty$.*

We can now state the almost sure (a.s.) convergence of the sequence (r_k).

Proposition 3.1. *Considering the sequence $(r_k)_{k\geq 0}$ generated by Algorithm 3.1, under Assumptions 3.1, 3.2 and 3.3, we get :*

$$\lim_{k\to+\infty} \quad r_k=r^* \quad \textit{a.s.}$$

The proof is postponed to the appendix. We now require a stronger regularity assumption on the functions f_0, f_1, π_r to derive the asymptotic distribution of the sequence $(r_k)_{k\geq 0}$.

Assumption 3.4. *The functions f_0 and f_1 are not proportional to each other.*

Assumption 3.5. *There exists $\delta>0$ such that*

$$\sup_{k\geq 0} \quad \mathbb{E}_0\left(\left|\frac{f_0(Z)-r_kf_1(Z)}{\pi_{r_k}(Z)}\right|^{1+\delta}\right)+\sup_{k\geq 0} \quad \mathbb{E}_1\left(\left|\frac{f_0(Z)-r_kf_1(Z)}{\pi_{r_k}(Z)}\right|^{1+\delta}\right)<+\infty.$$

Assumption 3.6. *There exists $\frac{1}{2}<\epsilon<1$, $a>0$, $b>0$ such that the sequence of step sizes (γ_k) is of the form $\gamma_k=\frac{a}{b+k^\epsilon}$.*

These integrability and step sizes assumptions are classical ones to obtain asymptotic normality results for martingales. The next result states the asymptotic normality of the averaged sequence defined as $(r_k^{AV})_{k\geq 0}=\left(\frac{1}{k}\sum_{j=0}^{k}r_j\right)_{k\geq 0}$.

Proposition 3.2. *Considering the sequence $(r_k)_{k\geq 0}$ generated by Algorithm 3.1 and its averaged version $(r_k^{AV})_{k\geq 0}$, under Assumptions 3.1, 3.4, 3.5 and 3.6, we get :*

$$\sqrt{k}\left(r_k^{AV}-r^*\right)\xrightarrow[k\to+\infty]{d}\mathcal{N}\left(0,V_{saris}(\pi_{r^*})\right)$$

with

$$V_{saris}(\pi_{r^*})=\frac{1}{{c_1}^2}\mathbb{E}_{r^*}\left[\left(\frac{f_0(Z)-r^*f_1(Z)}{\pi_{r^*}(Z)}\right)^2\right]$$

where the expectation $\mathbb{E}_{r^*}$ is taken with respect to the density π_{r^*}. Furthermore, the optimal proposal π_{saris}^{opt} defined as the one which minimizes the asymptotic variance is given as

$$\pi_{r^*}^{opt}(z) \propto |p_1(z) - p_0(z)|,$$

corresponding to the optimal variance :

$$V_{saris}^{opt} = r^{*2} \left(\int_{\mathcal{Z}} |p_1(z) - p_0(z)| \mu(dz) \right)^2 .$$

The proof is postponed to the appendix and relies on similar arguments as the derivation of optimal importance function in importance sampling (see Robert and Casella (1999) for more details). We achieve the same optimal variance and retrieve the optimal proposition density of ratio importance sampling. Moreover in Chen and Shao (1997b) the authors show that this variance is smaller than the variance of the optimal Bridge sampling estimator.

Remark 3.6. *If sampling is done through the use of a Markov transition kernel, similar theoretical results can be obtained assuming additional regularity conditions on the Markov kernel. For further details we refer to Allassonniere and Kuhn (2015) and Fort (2015).*

It is important to notice that the algorithm presented in this section is not always applicable. In particular, the analytical expression of the optimal proposal density $\pi_{r^*}^{opt}$ is unknown, and therefore the update rule of the sequence r_k defined in Algorithm 3.1 is not computable. The next section solves this issue.

3.3.3 . A practical extension of the SARIS algorithm

Suppose that for all z, $\pi_r(z) = \tilde{\pi}_r(z)/c(r)$ where the analytical expression of $\tilde{\pi}_r(z)$ is known, then equation (3.11) is equivalent to :

$$\mathbb{E}_{\pi_{r^*}} \left[\frac{f_0(Z) - r^* f_1(Z)}{\tilde{\pi}_{r^*}(Z)} \right] = 0. \tag{3.17}$$

Based on this identity, one can consider the following extension of Algorithm 3.1 called SARIS-EXT, that uses the known unnormalised density $\tilde{\pi}_r$:

Algorithm 3.2 SARIS-EXT algorithm

Input : $(\gamma_k)_{k\geq 0}$, r_0, stopping criterion
Until stopping criterion :
  Draw Z_{k+1} from $\pi_{r_k} \propto \tilde{\pi}_{r_k}$
  Update $r_{k+1} = r_k + \gamma_{k+1} \frac{f_0(Z_{k+1}) - r_k f_1(Z_{k+1})}{\tilde{\pi}_{r_k}(Z_{k+1})}$
  $k = k + 1$
Return r_k

We state mild additional conditions on $\tilde{\pi}_r$ similar to Assumptions 3.2 and 3.5, to prove that the sequence $(r_k)_{k\geq 0}$ generated by Algorithm 3.2 verifies the same properties as the one generated by Algorithm 3.1.

Assumption 3.7.

$$\mathbb{E}_0\left(\sup_{r\in\mathbb{R}}\left|\frac{f_0(Z) - r f_1(Z)}{\tilde{\pi}_r(Z)}\right|\right) + \mathbb{E}_1\left(\sup_{r\in\mathbb{R}}\left|\frac{f_0(Z) - r f_1(Z)}{\tilde{\pi}_r(Z)}\right|\right) < +\infty.$$

Proposition 3.3. *Considering the sequence $(r_k)_{k\geq 0}$ generated by Algorithm 3.2, under Assumptions 3.1, 3.3 and 3.7, we get :*

$$\lim_{k\to+\infty} \quad r_k = r^* \quad \textit{a.s.}$$

Assumption 3.8. *There exists $\delta > 0$ such that*

$$\sup_{k\geq 0} \quad \mathbb{E}_0\left(\left|\frac{f_0(Z) - r_k f_1(Z)}{\tilde{\pi}_{r_k}(Z)}\right|^{1+\delta}\right) + \sup_{k\geq 0} \quad \mathbb{E}_1\left(\left|\frac{f_0(Z) - r_k f_1(Z)}{\tilde{\pi}_{r_k}(Z)}\right|^{1+\delta}\right) < +\infty.$$

Assumption 3.9. *There exists a neighborhood U of r^* such that $r \mapsto h(r) = \frac{c_0 - r c_1}{c(r)}$ is continuously differentiable and $h'(r^*) < 0$.*

Assumption 3.10. *For all $r \in \mathbb{R}$ the set $\mathcal{A}_r = \{z \in \mathcal{Z}, \quad f_0(z) = r f_1(z)\}$ satisfies $\mu(\mathcal{A}_r) = 0$.*

Remark 3.7. *$i)$ Assumption 3.9 is required to apply a central limit theorem on the sequence $(r_k)_{k\geq 0}$. For the SARIS algorithm, $h(r) = c_0 - r c_1$ which does not require regularity conditions, as it is affine. $ii)$ Assumption 3.10 is required to ensure that $c^{opt}(r)$ defined as the normalizing constant of $z \mapsto \tilde{\pi}_r^{opt}(z) = |f_0(z) - r f_1(z)| \propto \pi_r^{opt}(z)$ verifies Assumption 3.9. The following proposition extends the theoretical results of the SARIS algorithm to the SARIS-EXT algorithm.*

Proposition 3.4. *Considering the sequence $(r_k)_{k\geq 0}$ generated by Algorithm 3.2 and its averaged version $(r_k^{AV})_{k\geq 0}$, under Assumptions 3.1, 3.4, 3.6, 3.8 and 3.9, we get :*

$$\sqrt{k}\left(r_k^{AV} - r^*\right) \xrightarrow[k\to+\infty]{d} \mathcal{N}\left(0, V_{ext}(\tilde{\pi}_{r^*})\right)$$

Moreover we have :

$$V_{ext}(\tilde{\pi}_{r^*}) = V_{saris}(\pi_{r^*}).$$

Furthermore, under Assumption 3.10, an optimal unnormalised proposal $\tilde{\pi}_r^{opt}$ defined as one of which minimizes the asymptotic variance is given as

$$\tilde{\pi}_{r^*}^{opt}(z) = |f_0(z) - r^* f_1(z)| \propto \pi_{r^*}^{opt}(z)$$

corresponding to the optimal variance :

$$V_{saris}^{opt} = r^{*2} \left(\int_{\mathcal{Z}} |p_1(z) - p_0(z)| \mu(dz) \right)^2 .$$

This proposition shows that theoretically, the extended algorithm has the same asymptotic performances as the initial SARIS algorithm.

3.4 . Non optimal methods using draws from p_0 and p_1

3.4.1 . Estimating ratios of normalizing constants using only distributions p_0 and p_1

It is interesting to compare methods that use draws solely from p_0 and p_1. As explained in Section 3.2.1 with the two examples, in latent variable models, these represent the posterior distributions of latent variables given data under two different hypotheses. In Bayesian inference, they denote the posterior distributions of parameters under two distinct models. In both contexts, draws from these distributions are essential for inference, making it convenient to consider proposal distributions based on them. Numerically, this allows for the reuse of already simulated samples. Practically, it simplifies the method by eliminating the need to build new samplers.

A natural choice is a mixture of p_0 and p_1, similar to the approach proposed by Chen and Shao (1997b) in Section 5, and closely related to bridge sampling. We define the proposal density based on this mixture as follows :

$$z \mapsto \pi_{r^*}^{mixt}(z) = \frac{1}{2} \{p_0(z) + p_1(z)\} \propto f_0(z) + r^* f_1(z) \tag{3.18}$$

As long as we know how to draw samples from p_1 and p_0, it is easy to sample from this mixture, by sampling uniformly randomly from one or the other distribution. Of course the analytical expression of π_r^{mixt} is unknown, but the extended algorithm is applicable considering the unnormalised density $\tilde{\pi}_r^{mixt}(z) = f_0(z) + r f_1(z)$. The simulating step in Algorithm 3.2

requires to sample from $\pi_{r_k}^{mixt}$, which verifies for every $z \in \mathcal{Z}$:

$$\begin{aligned}\pi_{r_k}^{mixt}(z) &\propto f_0(z) + r_k f_1(z) \\ &\propto c_0 p_0(z) + r_k c_1 p_1(z) \\ &\propto r^* p_0(z) + r_k p_1(z) \\ \iff \pi_{r_k}^{mixt} &= \left(1 - \frac{r_k}{r_k + r^*}\right) p_0(z) + \frac{r_k}{r_k + r^*} p_1(z) \end{aligned} \tag{3.19}$$

where the weights of the mixture depend on r^*. Therefore it is not possible to sample from $\pi_{r_k}^{mixt}$ using simply draws from p_0 and p_1.

Remark 3.8. *We emphasize that even if it is not possible to sample from $\pi_{r_k}^{mixt}$ using draws from p_0 and p_1, it is still possible to use a MCMC procedure to sample from $\pi_{r_k}^{mixt} \propto f_0 + r_k f_1$, but this loses the practical benefits of using separately the two distributions. This scheme is still illustrated in the simulation study.*

However, as simulating from the mixture $\pi_{r^*}^{mixt}$ is possible, one can consider the following alternative recursive scheme :

$$r_{k+1} = r_k + \gamma_{k+1} \frac{f_0(Z_{k+1}) - r_k f_1(Z_{k+1})}{\tilde{\pi}_{r_k}^{mixt}(Z_{k+1})}, \qquad Z_{k+1} \sim \pi_{r^*}^{mixt} \tag{3.20}$$

which defines a new estimation procedure that is summarized in Algorithm 3.3, and called SARIS-MIXT.

Algorithm 3.3 SARIS-MIXT algorithm

Input : $(\gamma_k)_{k \geq 0}$, r_0, stopping criterion
Until stopping criterion :
  Draw Z_{k+1} from $\pi_{r^*}^{mixt} = \frac{1}{2}(p_0 + p_1)$
  Update $r_{k+1} = r_k + \gamma_{k+1} \frac{f_0(Z_{k+1}) - r_k f_1(Z_{k+1})}{f_0(Z_{k+1}) + r_k f_1(Z_{k+1})}$
  $k = k + 1$
Return r_k

Remark 3.9. *Contrary to the SARIS-EXT algorithm, in this procedure the distribution used for the sampling step does not depend on the current ratio r_k.*

It can be shown that under mild conditions, the sequence generated by Algorithm 3.3 converges almost surely towards r^* and is asymptotically Gaussian. The proof follows the same lines as for the two other SARIS algorithms, the main steps are given in the proof of the next result.

Unfortunately, the estimator obtained from the algorithm SARIS-MIXT presents a higher asymptotic variance than the one obtained using SARIS-EXT with $\tilde{\pi}_r^{mixt}$ as proposal, $V_{ext}(\tilde{\pi}_{r^*}^{mixt})$. The following proposition formalizes this statement.

Proposition 3.5. *Let V_{saris}^{mixt} be the asymptotic variance of the averaged sequence generated by the SARIS-MIXT Algorithm 3.3. Let $\tilde{\pi}_r^{mixt}(z) = f_0(z) + rf_1(z)$ be the unnormalised mixture between p_0 and p_1. Let Ψ be the quantity defined as :*

$$\Psi = \int_{\mathcal{Z}} \frac{p_1(z)p_0(z)}{\frac{1}{2}\{p_1(z) + p_0(z)\}} dz$$

Under Assumptions 3.1, 3.4, 3.6, we get :

$$V_{ext}(\tilde{\pi}_{r^*}^{mixt}) = 4r^{*2}(1 - \Psi) = \Psi^2 V_{saris}^{mixt}.$$

The quantity Ψ can be seen as an overlap index between p_0 and p_1. It is easy to show that $0 \leq \Psi \leq 1$. If the two distributions have disjoint supports, then $\Psi = 0$. If $p_0 = p_1$ then $\Psi = 1$.

This index Ψ is very convenient to compare the asymptotic variances derived in Proposition 3.5 with those of the optimal Bridge sampling estimator and ratio importance sampling estimator using $\pi_{r^*}^{mixt}$ as proposal. For a detailed description of the later called Bridge-like ratio importance sampling method (RIS-MIXT), we refer to Chen and Shao (1997b), section 5. Let V_{ris}^{mixt} denote the asymptotic variance of the Bridge-like ratio importance sampling estimator. The following relationship exists between the various variances discussed in this paragraph :

$$V_{ext}(\tilde{\pi}_{r^*}^{mixt}) = \Psi \times V_{bridge}^{opt} = \Psi^2 \times V_{ris}^{mixt} = \Psi^2 \times V_{saris}^{mixt} \tag{3.21}$$

Note that the SARIS-MIXT estimator reaches the same asymptotic variance as the Bridge-like ratio importance sampling estimator, which is not surprising as they are both based on the same identity (3.6).

It is noticable that $V_{ext}(\tilde{\pi}_{r^*}^{mixt})$ is bounded, unlike the other variances that diverge as Ψ approaches 0. In fact, when $\Psi = 0$, π_r^{mixt} aligns with π_r^{opt}. Therefore, finding a way to sample from the distribution defined by (3.19) using only draws from p_0 and p_1 would be very beneficial. This would allow for variance reduction, as described in equation (3.21), while still maintaining the simplicity of simulating only from p_0 and p_1. This result indicates that Bridge sampling is still the best method when using only draws from p_0 and p_1. However, when Ψ is closed to one (which is the case with a lot of overlap between p_0 and p_1), these methods remain comparable.

The next section explores the extension of the SARIS-MIXT algorithm to a joint procedure with parameter inference in latent variables models.

3.4.2 . A joint procedure for model parameter estimation and LRT statistic compu-

tation in latent variables models

In this section, we extend the use of the SARIS estimator based on mixtures between p_0 and p_1 to the context of likelihood ratio test (LRT) in latent variables model and introduce a joint procedure for model parameter inference and LRT statistic computation.

Consider two random variables Y on $\mathcal{Y}$ and Z on $\mathcal{Z}$. Assume that the joint density of (Y, Z) belongs to a parametric family $\{f_\theta, \theta \in \Theta\}$, with $\Theta \subset \mathbb{R}^p$ with p a positive integer. We only observe a realization y of Y, the random variable Z being unobserved. The maximum likelihood estimator $\hat{\theta}$ is defined as :

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \quad L(\theta; y)$$

where the marginal likelihood $L(\theta; y)$ is equal to the complete likelihood integrated over the latent variable :

$$L(\theta; y) = \int_{\mathcal{Z}} f_\theta(y, z)\mu(dz)$$

The above integral is often untractable, which makes the optimization process difficult. To solve this issue, stochastic methods can be used. Two popular ones are the stochastic approximation expectation maximization (SAEM) algorithm (Delyon et al., 1999; Kuhn and Lavielle, 2004) and the stochastic gradient descent (SGD) algorithm (Baey et al., 2023). Both methods require draws from the posterior distribution of the latent variables given the data whose density is denoted by p_θ in the sequel.

Both SAEM and SGD are iterative algorithms that can be summarized as follows, at each step $k > 0$:

1. Draw Z_k from p_{θ_k}.
2. Update θ_k with a gradient step when using SGD or a maximization step when using SAEM.

Consider now the context of Example 3.1 where the objective is to test the hypotheses :

$$H_0 : \quad \theta \in \Theta_0 \quad against \quad H_1 : \quad \theta \in \Theta_1$$

where $\Theta_0 \subset \Theta_1 \subset \Theta$. The LRT statistic equals :

$$LR = -2 \log \left(\frac{L(\hat{\theta}_0; y)}{L(\hat{\theta}_1; y)} \right).$$

We propose to combine the estimation procedures for $\hat{\theta}_0$ and $\hat{\theta}_1$ with the computation of the marginal likelihood ratio $r^* = \exp(-LR/2)$, taking advantage of the computational effort of the inference task. The procedure is detailed in Algorithm 3.4.

Algorithm 3.4 Joint parameter and LRT statistic estimation in latent variables models

Input : $z_{0,0}, z_{1,0}, \theta_{0,0}, \theta_{1,0}, r_0, (\gamma_k)_{k\geq 0}$
$k = 0$
Until convergence criterion :
  Draw $z_{0,k+1}$ from $p_{\theta_{0,k}}$ and $z_{1,k+1}$ from $p_{\theta_{1,k}}$
  Update $\theta_{0,k+1}$ and $\theta_{1,k+1}$ using a SGD or SAEM step
  Draw $\tilde{z}_{k+1}$ from a uniform distribution on $\{z_{0,k+1}, z_{1,k+1}\}$
  Update r_{k+1} as :

$$r_{k+1} = r_k + \gamma_k \frac{f_{\theta_{0,k+1}}(y, \tilde{z}_{k+1}) - r_k f_{\theta_{1,k+1}}(y, \tilde{z}_{k+1})}{f_{\theta_{0,k+1}}(y, \tilde{z}_{k+1}) + r_k f_{\theta_{1,k+1}}(y, \tilde{z}_{k+1})}$$

  $k = k + 1$
Return $r_k, \theta_{0,k}, \theta_{1,k}$

Remark 3.10. *When the two estimation processes can not be applied jointly, this procedure can be carried out post-estimation, provided the sequences $(\theta_{i,k}, z_{i,k})_{i=0,1,k\geq 0}$ are kept in memory.*

Remark 3.11. *If only one marginal likelihood is to be estimated, the procedure applies by introducing a proposal density q (or a sequence of proposal densities $(q_k)_{k\geq 0}$) from which it is possible to sample from and to proceed as follows :*

1. *Draw z_{k+1} from p_{θ_k}*
2. *Update θ_{k+1} with SGD or SAEM step*
3. *With probability 0.5 define $\tilde{z}_{k+1} = z_{k+1}$, otherwise draw $\tilde{z}_{k+1}$ from $q_{k+1}(.)$*
4. *Update r_{k+1} as :*

$$r_{k+1} = r_k + \gamma_k \frac{f_{\theta_{k+1}}(y, \tilde{z}_{k+1}) - r_k q_{k+1}(\tilde{z}_{k+1})}{f_{\theta_{k+1}}(y, \tilde{z}_{k+1}) + r_k q_{k+1}(\tilde{z}_{k+1})}$$

The R package `bridgesampling` (Gronau et al., 2020) proposes to use a Gaussian approximation as a second distribution when considering the computation of a single marginal likelihood. However, in our procedure prior knowledge on the distribution of interest (mean and variance for example) is not available as inference has not been performed yet. A possible approach to overcome this issue could be to use a Gaussian proposal with adaptive mean m_k and variance σ_k^2, where m_k and σ_k^2 are defined at each step k as follows :

$$\begin{aligned} m_{k+1} &= m_k + \gamma_k(z_{k+1} - m_k) \\ v_{k+1} &= v_k + \gamma_k(z_{k+1}^2 - v_k) \\ \sigma_{k+1}^2 &= v_{k+1} - m_{k+1}^2 \end{aligned}$$

3.5 . Numerical experiments and practical considerations

This section is devoted to numerical experiments. We first illustrate the performances of the three SARIS estimators compared to the RIS estimator of Chen and Shao (1997a) and the optimal Bridge sampling estimator of Meng and Wong (1996). We then provide an example of the joint procedure introduced in Section 3.4.2.

3.5.1 . Simulation study in a one dimensional Gaussian setting

We first illustrate our method with a one dimensional Gaussian setting. We consider two Gaussian distributions, $f_0 = \phi$ and $f_1 = \phi(\cdot - \mu)$ where ϕ is the standardized Gaussian density and $\mu \in \mathbb{R}$. Since these densities are already normalized, we have $c_0 = c_1 = 1$ and therefore $r^* = 1$.

In the simulation study, we compare the performances of three SARIS estimators presented in this paper. These methods are compared with the optimal bridge sampling estimator (BRIDGE OPT) and the RIS estimator based on $\pi_{r^*}^{mixt}$ (RIS-MIXT). For the entire simulation study, the sampling steps are performed using one step of an adaptive Metropolis Hastings (MH) algorithm (see for example Roberts and Rosenthal (2009, section 3)) in order to stick to most real life applications, implemented manually.

The three SARIS estimators presented in Figure 3.1 are the following :

1) Optimal SARIS extended using $\tilde{\pi}_r^{opt}$ (SARIS-EXT opt) This is the estimator generated by Algorithm 3.2 using $\tilde{\pi}_r^{opt}(z) = |f_0(z) - r f_1(z)|$. In this procedure, the increment $\frac{f_0(Z_{k+1}) - r_k f_1(Z_{k+1})}{|f_0(Z_{k+1}) - r_k f_1(Z_{k+1})|} \in \{0; 1\}$ only takes two values. The drawback is that the increment has no intensity, i.e. it gives no indication on the order of magnitude of the descent step. However, it can still solve computational issues, in particular when the evaluation of the likelihood can be complicated. In order to use this method, one only need to know how to evaluate the unnormalised densities up to a non decreasing transformation, as only comparison of them is required, and not their evaluation.

2) SARIS using mixture between p_0 and p_1 (SARIS-MIXT) This distribution has already been discussed in Section 3.4. As long as we know how to draw samples from p_1 and p_0, it is easy to sample from the mixture by randomly sampling from one or the other distribution with probability 1/2 :1/2. This estimator corresponds to the one generated by Algorithm 3.3.

3) SARIS extended using $\tilde{\pi}_r^{mixt} = f_0(z) + r f_1(z)$ (SARIS-EXT-mixt) This proposal is not of practical use, as it neither uses draws from p_0 and p_1 nor is an approximation of the optimal scheme. However, it is interesting to distinguish it from the previous one, as they are very similar. This distribution is also a mixture between p_0 and p_1 as explained in Section 3.4. Even if it is

supposed to approximate $\pi_{r^*}^{mixt}$, it benefits from a significant gain in performance, as Proposition 3.5 theoretically justifies it, and Figure 3.1 illustrates it.

Note that the three different proposal considered above lead to a bounded increment. This property is crucial both from a practical point of view, as it enhances the numerical stability of the procedure, and a theoretical one, as it guarantees Assumptions 3.2 and 3.5 to be verified.

Remark 3.12. *The optimality criterion considered in this article is given by the asymptotic variance of the exact sampling scheme, which is most of the time intractable in practice. Indeed, in most real life applications sampling is performed through the use of transition kernels in MCMC algorithms. In this setting, even if central limit theorems may apply under regularity conditions, the asymptotic variance is in general not explicit and there is no guarantee that the proposal which minimizes this variance is the same as in the exact sampling case. Therefore it might still be of practical interest to consider other proposals. Some proposal distributions are worth mentioning, e.g. the geometric mean between p_0 and p_1 discussed in Meng and Wong (1996) and Chen and Shao (1997a) for example, or an element of the q-path between p_0 and p_1 that generalizes the harmonic mean and geometric mean between the two distributions (see Brekelmans et al. (2020) for details).*

For each of the SARIS procedure the following step size is considered :

$$\begin{cases} \gamma_k = 0.1 & \text{for } k < K_{heat} \\ \gamma_k = \frac{0.1}{1+k^{2/3}} & \text{otherwise,} \end{cases} \tag{3.22}$$

which is standard in the stochastic gradient descent literature. It verifies Assumption 3.3 and 3.6 while presenting a heating phase that enables a wider exploration of the parameter space at the beginning of the algorithm.

As discussed in Section 3.2.2, the optimal Bridge sampling estimator is not available in a closed form, so we relied on the recursive algorithm described in equation (3.5), which was run until convergence. For the RIS estimator, we used equation (3.7) using $\pi_{r^*}^{mixt}$ as proposal distribution.

We adjusted the sample sizes of each method to make sure that they are all comparable in terms of number of calls to functions f_0 and f_1. Four MH samplers were implemented. The first two are MH samplers whose invariant distributions are p_0 and p_1. These are used for BRIDGE-OPT, RIS-MIXT and SARIS-MIXT. Then two samplers generating non homogenous Markov chains with, at each step, invariant distributions being $\pi_{r_k}^{opt}$ and $\pi_{r_k}^{mixt}$, were built to compute respectively the SARIS-EXT-opt and the SARIS-EXT-mixt estimators.

For each of the estimators, a budget of $2K$+$2K_{heat}$ draws were allocated, with $2K$ samples used to compute the estimators and $2K_{heat}$ for heating the MH samplers. For the experiments, we used $K = 5000$ and $K_{heat} = 300$.

We also considered the estimation of $\log(r^*)$ using recursion (3.16). From a theoretical point of view, the use of this transformation is equivalent to using a delta method. Therefore, results obtained for the estimation of $\log r^*$ or r^* are comparable in terms of performances. However, since the numerical stability of the procedure is greater in the former case, in the sequel we only present results associated with the estimation of $\log r^*$.

We first consider two cases : $1)$ a strong overlap between p_0 and p_1 with $\mu = 1$ and $2)$ a small overlap between p_0 and p_1 with $\mu = 5$. Results are presented in Figure 3.1.

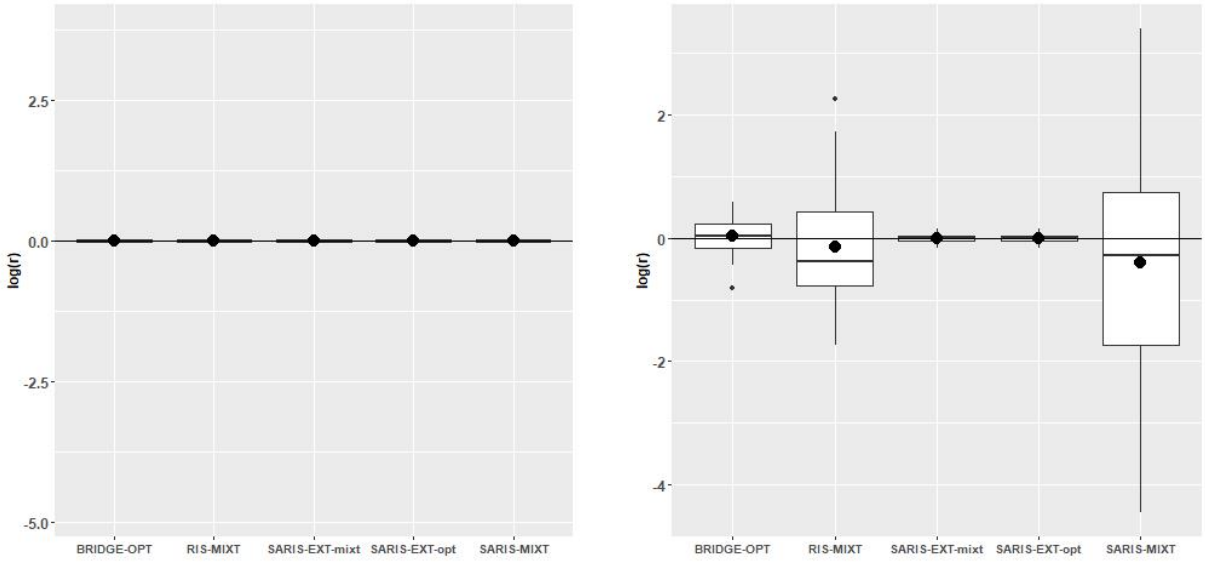


Figure 3.1 – Estimation of $\log r^*$ using the optimal bridge sampling estimator (BRIDGE-OPT), the bridge like ratio importance sampling (RIS-MIXT), SARIS-EXT-mixt, SARIS-EXT-opt and SARIS-MIXT. The black line represents the true value $log(r^*) = 0$. The boxplots were computed on 50 repetitions of the experiments. The two graphs illustrate strong overlap between p_0 and p_1 (on the left, in the case $\mu = 1$) and little overlap (on the right, in the case $\mu = 5$)

As the theory suggests, the optimal bridge sampling estimator outperforms the two other methods based on samples from p_0 and p_1. However, when considering the case $\mu = 1$, the five methods present performances of the same order. The fact that the bridge like ratio importance sampling performs better than the SARIS-MIXT estimator can be interpreted by the fact that the ratio importance sampling estimator imposes at each step the estimator to solve the empirical version of the SARIS identity (3.11), which might enhance the stability of the procedure. This difference should be diminished by considering adaptive step sizes, to mimic the differences between $\hat{r}_K^{ris}$ and $\hat{r}_{K+1}^{ris}$.

Figure 3.1 illustrates the fact that the SARIS-EXT estimators presented are much more robust to little overlap than ratio importance sampling and Bridge sampling. However, it also illustrates

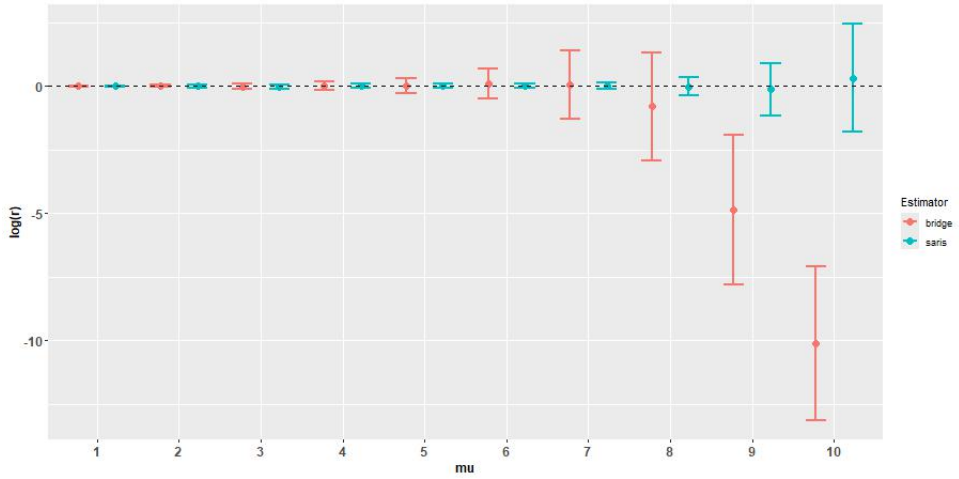


Figure 3.2 – Estimation of the log ratio of normalizing constants of the densities of a $\mathcal{N}(0,1)$ and the one of a $\mathcal{N}(\mu,1)$ for varying values of μ, using the optimal Bridge sampling estimator BRIDGE-OPT (red) and SARIS-EXT-opt to compute the ratio The dots represent the empirical means and the error bars the empirical standard deviations computed over 50 repetitions.

the fact that in more simple cases such as the one illustrated in Figure 3.1, methods that only use draws from p_0 and p_1, which are easier to apply might be sufficient.

To show intermediate results between the cases $\mu = 1$ and $\mu = 5$, and worse ones, we compare the optimal Bridge sampling estimator with the SARIS-EXT-opt estimator for varying values of μ between 1 and 10. For the simulations we use $K = 5000$ samples for each expectation in the bridge sampling estimation procedure, and therefore use $2K$ samples for the SARIS procedure. Results are displayed in Figure 3.2. For each value of μ, 50 repetitions were computed. The dots represent the empirical means and the errorbars the empirical standard deviations computed over the 50 repetitions. Figure 3.2 illustrates the robustness of the proposed procedure in comparison to the bridge sampling estimator that highly deteriorates when the overlap reduces. This makes the proposed procedure appealing when little information is available on the distributions.

3.5.2 . Joint estimation in latent variables models

To illustrate the joint procedure presented above, we consider the following model of linear regression with missing values. Let $i = 1, ..., n$, we observe the response $y_i \in \mathbb{R}$ modeled as :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \tag{3.23}$$

where $(\varepsilon_i)_{i=1,...,n}$ is a sequence of independent and identically distributed Gaussian noise with known variance σ^2, $\beta = (\beta_0, \beta_1, \beta_2)^T$ is an unknown vector of regression coefficients, and $(x_i)_{i=1,...,n} = (x_{i1}, x_{i2})_{i=1,...,n}^T$ is an independent and identically distributed sample of covariates

from a $\mathcal{N}\left((\mu_1,\mu_2)^T, \begin{pmatrix} \gamma_1^2 & 0 \\ 0 & \gamma_2^2 \end{pmatrix}\right)$. We suppose that for $i = 1, ..., r$ we only observe (y_i, x_{i1}) and for $i = r+1, ..., n$ we observe (y_i, x_{i1}, x_{i2}). This example is borrowed from the lecture notes of Julie Josse "Handling Missing values" available on this website [1]. Here the parameter to estimate is $\theta = (\beta, \gamma_1^2, \gamma_2^2, \mu_1, \mu_2)$. In order to make the notations as simple as possible, we will confound the notations of the random variables and their observed realizations. Furthermore, we are going to write $f_\theta(z)$ the density of the random variable Z evaluated at z. For example $f_\theta(y_i|x_{i1})$ is the conditional density of y_i given x_{i1}.

With these notations, the complete likelihood $L_n(\theta)$ is given by :

$$\begin{aligned} L_n(\theta) &= \prod_{i=1}^{n} f_\theta(y_i, x_i) \\ &= \prod_{i=1}^{n} f_\theta(y_i|x_i) f_\theta(x_i) \end{aligned}$$

However we do not observe the first r $(x_{i2})_i$ that are handled as latent variables, therefore the observed likelihood is marginalized over their distribution :

$$\begin{aligned} L_n(\theta) &= \left[\prod_{i=1}^{r} f_\theta(y_i, x_{i1})\right] \times \prod_{i=r+1}^{n} f_\theta(y_i, x_i) \\ &= \left[\prod_{i=1}^{r} f_\theta(y_i|x_{i1}) f_\theta(x_{i1})\right] \times \prod_{i=r+1}^{n} f_\theta(y_i|x_i) f_\theta(x_i) \\ &= \left[\prod_{i=1}^{r} \int_{x_{i2}} f_\theta(y_i|x_{i1}, x_{i2}) f_\theta(x_{i1}) f_\theta(x_{i2}) dx_{i2}\right] \times \prod_{i=r+1}^{n} f_\theta(y_i|x_i) f_\theta(x_i) \end{aligned}$$

In fact we can compute exactly the marginal likelihood as the conditional distribution of y_i given x_{1i} is a $\mathcal{N}\left(\beta_1 x_{i1}, \beta_2^2\gamma_2^2 + \sigma^2\right)$.
Given a realization $\mathbf{x}_2 = (x_{i2})_{i=1,..,r}$ of the unobserved variables, we introduce the complete log likelihood defined as :

$$l_n(x_{12}, ..., x_{r2}; \theta) = \sum_{i=1}^{n} \log\left(f_\theta(y_i, x_i)\right)$$

We consider here the following test :

$$H_0 : \quad \beta_0 = 0 \quad against \quad H_1 : \quad \beta_0 \neq 0$$

To apply the joint procedure described in Section 3.4.2 we consider the unconstrained parameter space Θ_1 corresponding to the alternative hypothesis, and the constrained parameter Θ_0 that corresponds to the case $\beta_0 = 0$. For the two estimation procedures, we used stochastic

1. https ://juliejosse.com/wp-content/uploads/2018/07/LectureMissing_Weij_modifAude.html

gradient descent. At each step k the procedure computes an estimator g_k of the log likelihood ratio, jointly to the estimators $\hat{\theta}_{0,k}$ and $\hat{\theta}_{1,k}$ of respectively the restricted and unrestricted maximum likelihood estimators as follows :

1. Draw $\mathbf{x}_{02}^{(k+1)}$ from the posterior distribution of $(x_{i2})_{i=1,\dots,r}$ given $((y_i, x_i)_{i=1,\dots,r}, \hat{\theta}_{0k})$ and $\mathbf{x}_{12}^{(k+1)}$ from the posterior distribution of $(x_{i2})_{i=1,\dots,r}$ given $((y_i, x_i)_{i=1,\dots,r}, \hat{\theta}_{1k})$
2. Update $\hat{\theta}_{0,k+1}$ and $\hat{\theta}_{1,k+1}$ each with a gradient step :

$$\hat{\theta}_{0,k+1} = \hat{\theta}_{0,k} - \gamma_k \nabla_\theta l_n(\mathbf{x}_{02}^{(k+1)}; \hat{\theta}_{0,k})$$
$$\hat{\theta}_{1,k+1} = \hat{\theta}_{1,k} - \gamma_k \nabla_\theta l_n(\mathbf{x}_{12}^{(k+1)}; \hat{\theta}_{1,k})$$

3. Draw $\tilde{\mathbf{x}}_2^{(k+1)}$ from a uniform distribution on $\{\mathbf{x}_{02}^{(k+1)}, \mathbf{x}_{12}^{(k+1)}\}$
4. Update g_{k+1} as :

$$g_{k+1} = g_k + \gamma_k \frac{\exp\left\{l_n(\tilde{\mathbf{x}}_2^{(k+1)}; \hat{\theta}_{0,k})\right\} - \exp\left\{l_n(\tilde{\mathbf{x}}_2^{(k+1)}; \hat{\theta}_{1,k}) + g_k\right\}}{\exp\left\{l_n(\tilde{\mathbf{x}}_2^{(k+1)}; \hat{\theta}_{0,k})\right\} + \exp\left\{l_n(\tilde{\mathbf{x}}_2^{(k+1)}; \hat{\theta}_{1,k}) + g_k\right\}}$$

For the following experiments, we used a constant step size of $\gamma_k = 0.1$ for the SARIS procedure and the estimation processes to keep the speed of convergence of the different iterations as similar as possible. The parameters used to generate the data are $\mu = (1, 1)^T$, $(\gamma_1, \gamma_2) = (1, 1)$, $\sigma^2 = 2$ and $\beta = (0.1, 1, 1)$. Finally, we considered $n = 200$ individuals and two different numbers of missing values. We considered $r = 20$ which corresponds to 10% of missing values for the second covariate, and then $r = 50$ which corresponds to 25%. We used $K = 250$ iterations for the estimation process. We reproduced the experiment 20.

Figure 3.3 displays the evolution of the exact likelihood ratio $-2 \log \frac{L_n(\hat{\theta}_{0,k})}{L_n(\hat{\theta}_{1,k})}$ over the iterations, compared to the one of the SARIS estimator obtained using the joint procedure described in Section 3.4.2. The dots represent the mean, and the errorbars the standard deviations computed over the 20 repetitions of the experiment. Figure 3.4 plots the mean squared error between the exact log likelihood ratio along the iteration process and its approximation using the SARIS joint procedure. The dots are the mean over the 20 repetitions, and the errorbars reach the 5% and 95% corresponding empirical quantiles.

In the case where $r = 10\%n$ we observe that the joint procedure accurately tracks the exact value of the likelihood ratio along the estimation process which is very encouraging. We observe that in the more complex case where $r = 25\%n$ the joint procedure correctly tracks the exact likelihood ratio statistic but with a small bias. There might be several sources for this bias such as the difference in convergence speed of both estimation processes or the higher variance due to the higher number of missing data. However, the approximation still tracks the exact

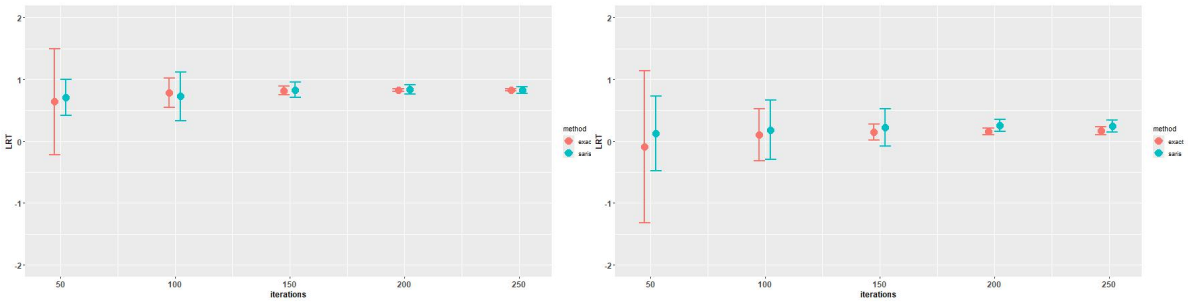


Figure 3.3 – Estimation of the log likelihood in the latent variable model of Section 3.5.2 in the case $r = 10\%n$ (on the left) and $r = 25\%n$ (on the right). Comparison over the iterations of the exact likelihood ratio $\left(-2\log\frac{L_n(\hat{\theta}_{0,k})}{L_n(\hat{\theta}_{1,k})}\right)_{k\geq 0}$ (*exact* in red) and its approximation $(-2g_k)_{k\geq 0}$ (*approx* in blue) using the joint procedure described in Section 3.4.2. The dots represents the means taken over 20 repetitions, the errorbars correspond to the empirical standard deviation.

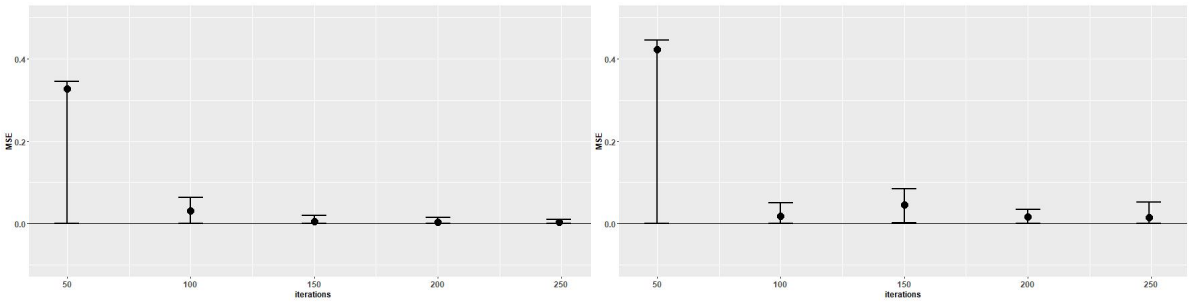


Figure 3.4 – Evolution of the empirical mean square error between the exact log likelihood ratio and its approximation using the joint procedure described in Section 3.4.2 in latent variable model of Section 3.5.2 in the case $r = 10\%n$ (on the left) and $r = 25\%n$ (on the right). The dots represents the means taken over 20 repetitions, the errorbars reach the 5% and 95% empirical quantiles.

value along iterations, which makes the proposed joined procedure appealing as a first step of a marginal likelihood computation. Indeed it gives a first estimator at the end of the estimation process that might be used as a starting point for a SARIS procedure, or an estimator $\hat{r}$ to use the approximated optimal scheme of Chen and Shao (1997b) to use as proposal distribution $\pi \propto |f_0 - \hat{r} f_1|$ in a ratio importance sampling procedure.

All the experiments were computed on R version 4.3.3 (2024-02-29 ucrt). For reproducibility the scripts are available at this link [2].

3.6 . Conclusion

We proposed a new methodology to compute ratios of normalizing constants that relies on the principle of stochastic approximation. Our procedure presents good theoretical properties which makes it competitive with the best methods from the literature. More precisely, our estimator is consistent and asymptotically Gaussian as the number of iterations goes to infinity. Moreover, the practical implementation of the algorithm reaches an asymptotic variance which is smaller than the optimal variance of the Bridge sampling estimator. Another important advantage is that our estimator does not require to fix in advance the computational effort thanks to its iterative nature. Indeed our procedure can be stopped in practice once a given convergence criterion is reached. Furthermore, our estimator seems more robust to little overlap between the two unnormalised distributions considered and outperforms the Bridge sampling estimator in some of the numerical examples considered. The proposed methodology also allows for the computation of single marginal likelihoods. Moreover, in the context of likelihood ratio test statistics in latent variables models, our procedure can be integrated in the parameter estimation process to reduce the computational effort.

Besides these positive points, there are several interesting perspectives to investigate. Thanks to the rich literature on stochastic approximation and more specifically on stochastic gradient descent, many refinements can be explored, such as acceleration, variance reduction or adaptive step sizes. Similarly, refinements used in classical Monte Carlo such as the Warp Bridge sampling (Meng and Schilling, 2002; Wang et al., 2022) could be applied to reduce the asymptotic variance of the estimator. Furthermore, as a method to compute ratios of normalizing constants, the SARIS procedure can benefit from the use of intermediate distributions that enable to decompose the problem into several simpler sub-problems, in the principle of stepping stone sampling.

2. https ://github.com/tguedon/saris

3.7 . Declarations

3.7.1 . Funding

This work was funded by the Stat4Plant project ANR-20-CE45-0012.

3.7.2 . Conflict of interest

The authors declare that they have no conflict of interest.

3.8 . Proofs

3.8.1 . Proof of Proposition 3.1

We first rewrite the iterative scheme (3.14) for every positive integer k :

$$r_{k+1} = r_k + \gamma_{k+1} H_{\pi_{r_k}}(Z_{k+1}), \qquad Z_{k+1} \sim \pi_{r_k}$$

introducing the notation $H_{\pi_r}(z, r) = \frac{f_0(z) - r f_1(z)}{\pi_r(z)}$.

We apply the almost sure convergence theorem of Robbins-Monro algorithms (Robbins and Monro, 1951) stated in section 5.1 of Benveniste et al. (2012) to prove Proposition 3.1.

Under Assumption 3.1 we can define the function h for every $r \in \mathbb{R}$ by $h(r) = \mathbb{E}_r \left[H_{\pi_r}(Z, r) \right]$ where $\mathbb{E}_r$ stands for $\mathbb{E}_{\pi_r}$ for sake of simplicity. It follows directly that $h(r) = c_0 - r c_1$ for every $r \in \mathbb{R}$ and $h(r^*) = 0$. We define the filtration $(\mathcal{F}_k)_{\geq 0}$ corresponding to the increasing family of σ-algebra generated by $(r_0, Z_1, ..., Z_k)$. We now verify the following assumptions of the theorem stated in section 5.1 of Benveniste et al. (2012) :

1. for any positive measurable function g defined on $\mathcal{Z} \times \mathbb{R}$ we have $\mathbb{E}\left[g(Z_{k+1}, r_k) | \mathcal{F}_k \right] = \mathbb{E}_{r_k}\left[g(Z, r_k) \right]$.

2. there exists $C > 0$, such that for every $r \in \mathbb{R}\ \mathbb{E}_r \left[H_{\pi_r}(Z, r)^2 \right] \leq C(1 + r^2)$

3. there exists $r^* > 0$ such that for every $r \in \mathbb{R} \setminus \{r^*\}$:

$$(r - r^*) h(r) < 0$$

4. $\sum \gamma_k = +\infty$ and $\sum \gamma_k^2 < +\infty$

The first point is straightforward considering the sampling scheme of equation (3.14).

We now consider point 2) :

$$\begin{aligned}
\mathbb{E}_r\left[H_{\pi_r}(Z,r)^2\right] &= \int_{\mathcal{Z}} \left(\frac{f_0(z) - r f_1(z)}{\pi_r(z)}\right)^2 \pi_r(z)\mu(dz) \\
&= \int_{\mathcal{Z}} \frac{(f_0(z) - r f_1(z))^2}{\pi_r(z)}\mu(dz) \\
&= \int_{\mathcal{Z}} |H_{\pi_r}(z,r)||f_0(z) - r f_1(z)|\mu(dz) \\
&\leq \int_{\mathcal{Z}} |H_{\pi_r}(z,r)| f_0(z)\mu(dz) + \int_{\mathcal{Z}} |H_{\pi_r}(z,r)| f_1(z)|r|\mu(dz) \\
&\leq c_0 \mathbb{E}_0\left[|H_{\pi_r}(z,r)|\right] + |r| c_1 \mathbb{E}_1\left[|H_{\pi_r}(z,r)|\right] \\
&\leq \tilde{C}(1+r^2)
\end{aligned}$$

thanks to Assumption 3.1.

Point 3) is straightforward as $h(r) = c_1(r^* - r)$. Finally point 4) is implied by Assumption 3.3. □

3.8.2 . Proof of Proposition 3.2

We apply the central theorem for Robbins-Monro algorithms (Duflo (1996) chapter 4) to prove Proposition 3.2. The additional assumptions to verify are :

1. There exists a neighborhood U of r^* such that the function h is continuously differentiable on U and for all $r \in U$, $h'(r) < 0$.
2. There exists $\Gamma > 0$, such that almost surely :

$$\lim_{k \to +\infty} \mathbb{E}\left[\left(H_{\pi_{r_k}}(Z_{k+1}, r_k) - h(r_k)\right)^2 |\mathcal{F}_k\right] = \Gamma$$

3. There exists $\delta > 0$ such that

$$\sup_k \quad \mathbb{E}\left[\left(H_{\pi_{r_k}}(Z_{k+1}, r_k) - h(r_k)\right)^{2+\delta} |\mathcal{F}_k\right] < +\infty$$

4. There exist $\frac{1}{2} < \epsilon < 1$, $a > 0$, $b > 0$ such that the sequence of step sizes (γ_k) is of the form $\gamma_k = \frac{a}{b + k^\epsilon}$.

The first point is straightforward as h is linear in r and h' is constant equal to $-c_1 < 0$. To prove the second point, let us introduce $\xi_{k+1} = H_{\pi_{r_k}}(Z_{k+1}, r_k) - h(r_k)$ for all integer k. After some calculation, we get :

$$\mathbb{E}(\xi_{k+1}^2 | \mathcal{F}_k) = c_0 \mathbb{E}_0\left[H_{\pi_{r_k}}(Z, r_k)\right] - r_k c_1 \mathbb{E}_1\left[H_{\pi_{r_k}}(Z, r_k)\right] - h(r_k)^2$$

Under Assumptions 3.1 and 3.6, the sequence $(r_k)_k$ converges almost surely to r^*, $h(r^*) = 0$,

and, applying the dominated convergence theorem, the following convergence holds almost surely :

$$\lim_{k\to+\infty} \quad \mathbb{E}(\xi_{k+1}^2|\mathcal{F}_k) = \mathbb{E}_{r^*}\left[H_{\pi_r^*}(Z,r^*)^2\right]$$

The right hand side term is positive under Assumption 3.4. Finally we show that Assumption 3.5 implies condition 3) using similar calculation and arguments. Assumption 3.6 corresponds to condition 4).

Applying the central theorem for Robbins-Monro algorithms (Duflo (1996) chapter 4) we get that :

$$V_{saris}(\pi_{r^*}) = \mathbb{E}_{r^*}(H_{\pi_{r^*}}(Z,r^*)^2)/c_1^2$$

We then apply Jensen inequality to get the minorization :

$$\begin{aligned} V_{saris}(\pi_{r^*}) &\geq \left(\mathbb{E}_{r^*}\left[|H_{\pi_{r^*}}(Z,r^*)|\right]\right)^2/c_1^2 \\ &= \left(\int_{\mathcal{Z}} |f_0(z) - r^* f_1(z)|\mu(dz)\right)^2/c_1^2 \\ &= r^{*2}\left(\int_{\mathcal{Z}} |p_0(z) - p_1(z)|\mu(dz)\right)^2 \end{aligned}$$

Moreover the equality case holds in Jensen inequality for $\pi_{saris}^{opt}(z) \propto |f_0(z) - r^* f_1(z)|$ which leads to the result. □

3.8.3 . Proof of Proposition 3.3

This proof follows the lines of the proof of Proposition 3.1. The only difference is that the objective function h to nullify is :

$$h : r \mapsto \frac{c_0 - rc_1}{c(r)}$$

that verifies for every $r \in \mathbb{R}$, $(r-r^*)h(r) = \frac{-c_1(r-r^*)^2}{c(r)} < 0$. The remaining hypothesis are verified thanks to Assumption 3.7.

Remark 3.13. *The fact that $c(r)$ is strictly positive for all r is implied by Assumption 3.7. $c(r) = 0$ would mean that $\tilde{\pi}_r(Z) = 0$ $\mu-$almost surely.*

□

3.8.4 . Proof of Proposition 3.4

In order to prove this proposition we proceed in two steps. The first step establishes the central limit theorem, the second step shows that the use of the unnormalized density $\tilde{\pi}_r^{opt}(z) = |f_0(z) - rf_1(z)|$ enables to reach the optimal asymptotic variance V_{saris}^{opt}.
The first step follows the same lines as the proof of Proposition 3.2, replacing Assumptions 3.2

and 3.5 by Assumptions 3.7 and 3.8. Assumption 3.9 ensures the differentiability of h in a neighborhood U of r^*, and the fact that $h'(r^*) \neq 0$. Furthermore, $h'(r^*) = \frac{-c_0 c(r^*) - c'(r^*)(c_0 - r^* c_1)}{c(r^*)^2} = \frac{-c_1}{c(r^*)}$. The continuity of c and $r \mapsto \tilde{\pi}_r(z)$ for every z are guaranteed by Assumption 3.1 and the differentiability of h. The central limit theorem of Duflo (1996) chapter 4 concludes this step, and :

$$\begin{aligned} V_{ext}(\tilde{\pi}_{r^*}) &= \frac{c(r^*)^2}{c_1^2} \mathbb{E}_{r^*}\left[\left(\frac{f_0(Z) - r^* f_1(Z)}{\tilde{\pi}_{r^*}(Z)}\right)^2\right] \\ &= \frac{1}{c_1^2} \mathbb{E}_{r^*}\left[\left(\frac{f_0(Z) - r^* f_1(Z)}{\pi_{r^*}(Z)}\right)^2\right] \\ &= V_{saris}(\pi_{r^*}) \end{aligned}$$

Let $r \in \mathbb{R}$, $\tilde{\pi}_r^{opt} : z \mapsto |f_0(z) - r f_1(z)|$ and $c^{opt} : r \mapsto \int_{\mathcal{Z}} \tilde{\pi}_r^{opt}(z) dz$. The second step proves that $\tilde{\pi}^{opt}$ verifies Assumptions 3.7, 3.8 and 3.9.

First for every $r \in \mathbb{R}$ and $z \in \mathcal{Z}$, the quantity $\frac{f_0(z) - r f_1(z)}{\tilde{\pi}_r^{opt}(z)}$ is bounded since its absolute value equals 1, and therefore Assumptions 3.7 and 3.8 are verified. The main point to verify is Assumption 3.9. This is not straightforward because of the absolute value in the definition of the c^{opt} function. Let $r \in \mathbb{R}$, $\mathcal{A}_r^+ = \{z \in \mathcal{Z} : f_0(z) > r f_1(z)\}$ and $\mathcal{A}_r^- = \{z \in \mathcal{Z} : f_0(z) < r f_1(z)\}$. We get

$$\begin{aligned} c(r) &= \int_{\mathcal{Z}} |f_0(z) - r f_1(z)| \mu(dz) \\ &= -\int_{\mathcal{A}_r^-} (f_0(z) - r f_1(z)) \, \mu(dz) + \int_{\mathcal{A}_r^+} (f_0(z) - r f_1(z)) \, \mu(dz) \\ &= -c_0 + r c_1 + 2 \int_{\mathcal{A}_r^+} (f_0(z) - r f_1(z)) \, \mu(dz) \end{aligned}$$

using Assumption 3.10, and that for $i = 0, 1$, $c_i = \int_{\mathcal{A}_r^+} f_i(z) \mu(dz) + \int_{\mathcal{A}_r^-} f_i(z) \mu(dz)$.

To study the differentiability of $m : r \mapsto \int_{\mathcal{A}_r^+} f_0(z) - r f_1(z) \mu(dz)$, we consider $r_0 \in \mathbb{R}$ and $\epsilon > 0$:

$$m(r_0 + \epsilon) - m(r_0) = \int_{\mathcal{A}_r^+ \setminus \mathcal{A}_{r+\epsilon}^+} (r f_1(z) - f_0(z)) \, \mu(dz) - \epsilon \int_{\mathcal{A}_{r+\epsilon}^+} f_1(z) \mu(dz)$$

Let study the first term.

$$\begin{aligned} \mathcal{A}_r^+ \setminus \mathcal{A}_{r+\epsilon}^+ &= \{z \in \mathcal{Z} : r f_1(z) < f_0(z) \leq (r + \epsilon) f_1(z)\} \\ &= \{z \in \mathcal{Z} : 0 < f_0(z) - r f_1(z) \leq \epsilon f_1(z)\} \end{aligned}$$

f_1 is $\mu-$almost surely upper-bounded as it is a (unnormalized) density, let $\bar{f}_1$ its upperbound.

$$\begin{aligned} \left| \int_{\mathcal{A}_r^+ \setminus \mathcal{A}_{r+\epsilon}^+} r f_1(z) - f_0(z) \mu(dz) \right| &\leq \int_{\mathcal{A}_r^+ \setminus \mathcal{A}_{r+\epsilon}^+} |r f_1(z) - f_0(z)| \mu(dz) \\ &\leq \epsilon \bar{f}_1 \mu(\mathcal{A}_r^+ \setminus \mathcal{A}_{r+\epsilon}^+) \end{aligned}$$

and $\mu(\mathcal{A}_r^+ \setminus \mathcal{A}_{r+\epsilon}^+) \underset{\epsilon \to 0}{\to} 0$ with Assumption 3.10. The same calculations hold with $\epsilon < 0$ and we

get that m is differentiable on $\mathbb{R}$ and for every $r \in \mathbb{R}$ $m'(r) = -\int_{\mathcal{A}_r^+} f_1(z)\mu(dz)$ and finally c is differentiable on $\mathbb{R}$ and specifically :

$$c'(r^*) = c_1 - 2\int_{\mathcal{A}_{r^*}^+} f_1(z)\mu(dz)$$

therefore h is differentiable on $\mathbb{R}$ which concludes the proof, as the first part of the proposition shows that $V_{ext}(\tilde{\pi}_{r^*}) = V_{saris}(\pi_{r^*})$.

3.8.5 . Proof of Proposition 3.5

We first derive the expression of $V_{ext}(\tilde{\pi}_{r^*}^{mixt})$.

We recall the notations : $\pi_{r^*}^{mixt}(z) = \frac{1}{2}\left(p_0(z) + p_1(z)\right) \propto f_0(z) + r^* f_1(z) = \tilde{\pi}_{r^*}^{mixt}(z)$.

We start from the result of Proposition 3.2 :

$$\begin{aligned}
V_{ext}(\tilde{\pi}_{r^*}^{mixt}) &= \int_{\mathcal{Z}} \frac{\left(f_0(z) - r^* f_1(z)\right)^2}{\pi_{r^*}^{mixt}(z)}\mu(dz)/c_1^2 \\
&= 2r^{*2}\int_{\mathcal{Z}} \frac{\left(p_0(z) - p_1(z)\right)^2}{p_0(z) + p_1(z)}\mu(dz) \\
&= 2r^{*2}\left(\int_{\mathcal{Z}} \frac{\left(p_0(z) + p_1(z)\right)^2 - 4p_0(z)p_1(z)}{p_0(z) + p_1(z)}\mu(dz)\right) \\
&= 2r^{*2}\left(2 - \int_{\mathcal{Z}} \frac{4p_0(z)p_1(z)}{p_0(z) + p_1(z)}\mu(dz)\right) \\
&= 4r^{*2}\left(1 - \Psi\right)
\end{aligned}$$

using $(a-b)^2 = (a+b)^2 - 4ab$ for all a and b reals. The result of Proposition 3.4 also holds as $h(r) = \frac{c_0 - rc1}{c_0 + rc_1}$ verifies Assumption 3.9.

We then present the main steps to prove the almost sure convergence and the asymptotic normality of the sequence obtained using Algorithm 3.3. The analysis of this procedure follows the same lines as the one of the SARIS and SARIS-ext procedure. We introduce the notation

$$\tilde{H}(z, r) = \frac{f_0(z) - r f_1(z)}{f_0(z) + r f_1(z)}$$

and we define under Assumption 3.1, with $\pi = \pi^{mixt}$, the function

$$\tilde{h}(r) = \mathbb{E}_{\pi_{r^*}^{mixt}}\left(\frac{f_0(Z) - r f_1(Z)}{f_0(Z) + r f_1(Z)}\right).$$

After some calculation, we get

$$\tilde{h}'(r^*) = -\frac{\Psi}{2r^*}$$

and

$$\tilde{\Gamma} = \mathbb{E}_{\pi_{r^*}^{mixt}}\left[\left(\frac{f_0(Z) - r^* f_1(Z)}{f_0(Z) + r^* f_1(Z)}\right)^2\right] = (1 - \Psi),$$

which is positive under Assumption 3.4. Following the same lines as in the proofs of Propositions 3.1 and 3.2, it is straightforward to check that all the assumptions required for the consistency and asymptotic normality hold. Applying these results to the sequence generated by Algorithm 3.3 leads to the following asymptotic variance V_{saris}^{mixt} :

$$V_{saris}^{mixt} = 4r^{*2}(1 - \Psi)/\Psi^2$$

The optimal bridge sampling asymptotic variance is given in equation (3.4) by $V_{bridge}^{opt} = 4r^{*2}(\Psi^{-1} - 1)$ and $V_{ris}^{mixt} = 4r^{*2}(1 - \Psi)/\Psi^2$ (see for example Chen and Shao (1997a, Theorem 5.2.)), which concludes the proof.

□

Chapitre 4

Étude de la variabilité génotypique chez *Arabidopsis thaliana*

Ce chapitre est issu d'un travail en cours, réalisé en collaboration avec Céline Richard-Molard, chercheure en écophysiologie végétale (INRAE, UMR Ecosys, Palaiseau) pour la partie biologique, et Jean-Benoist Leger, maître de conférences en statistiques (Université de Technologie de Compiègne, UMR Heudiasyc) pour l'implémentation informatique.

4.1 . Introduction

Sélectionner les variétés présentant les meilleures caractéristiques phénotypiques est un enjeu biologique important, notamment pour faire face aux nouveaux contextes environnementaux imposés par le changement climatique et la réduction des intrants. L'objectif est de s'inscrire dans une démarche agroécologique visant à produire, tout en préservant les équilibres naturels et la biodiversité, en utilisant de manière durable les ressources naturelles comme le sol, l'eau et l'énergie. Il existe, au sein d'une même espèce une grande variabilité dans la relation phénotype-génotype.
Le phénotype désigne l'ensemble des caractéristiques observables ou mesurables d'une plante, résultant de l'interaction entre son génotype (son patrimoine génétique) et l'environnement dans lequel il se développe. Ces caractéristiques peuvent inclure des aspects physiques comme la taille, la forme des feuilles, mais aussi des aspects fonctionnels comme la résistance aux maladies ou la production de fruits. En résumé, le phénotype est l'expression visible ou mesurable des gènes d'un individu dans un environnement. La figure 4.1 présente différentes plantes d'*Arabidopsis thaliana*, correspondant à des génotypes différents cultivés dans un même environnement et présentant des phénotypes contrastés. Elle illustre la grande variabilité existant dans la taille et le nombre de feuilles entre les différents génotypes.

Afin d'étudier cette relation génotype-phénotype, la modélisation mécaniste est une approche permettant de relier certains traits phénotypiques de la plante à un ensemble de paramètres ayant un sens biologique. Ce type de modèle mécaniste permet ainsi de représenter mathématiquement un génotype fixé, au travers d'un jeu de paramètres biologiques qui doit être estimé à partir de données. Si le modèle décrit correctement les traits d'intérêt considérés (variables de sortie), la variabilité phénotypique observée au sein de la population devrait être retranscrite au sein des paramètres du modèle correspondant à chaque génotype. Dans le cadre des modèles mécanistes, ces paramètres représentent des caractéristiques biologiques

Figure 4.1 – Illustration de la variabilité génotypique observé chez *Arabidopsis thaliana*, dans un environnement contrôlé (Weigel and Mott, 2009)

impliquées dans les processus complexes modélisés. Identifier, parmi les paramètres du modèle, ceux supportant la variabilité observée permet alors de déterminer des leviers de sélection pertinents. Cela permettrait également de réduire les coûts humains et expérimentaux relatifs aux expériences et analyses nécessaires à l'étude de cette variabilité, en réduisant le nombre de paramètres d'intérêt à caractériser.

Ce chapitre porte sur l'étude de l'adaptation d'*Arabidopsis thaliana* à un faible environnement azoté. *Arabidopsis thaliana* est une plante fréquemment considérée en recherche, car c'est un modèle idéal pour les études en biologie végétale en raison de son génome simple, de son cycle de vie court, de sa facilité de culture et de transformation, et de l'énorme quantité de ressources disponibles à son sujet.

Étudier son adaptation à un environnement faiblement azoté permettrait de réduire l'apport d'intrants azotés chimiques ayant des impacts potentiellement négatifs sur la qualité des sols, tout en restant efficace dans des sols à faible teneur en azote, et en optimisant les rendements. En effet, les carences azotées impactent différents caractères phénotypiques importants : croissance des racines, des feuilles, assimilation du carbone par photosynthèse pour ne citer que ceux-ci (Loudet et al., 2002).

Afin de comprendre les mécanismes impliqués dans la croissance de cette plante, il est essentiel d'examiner la production, l'allocation et le stockage du carbone et de l'azote nécessaires à cette croissance. Le modèle mécaniste ARNICA (Richard-Molard et al., 2007) permet de décrire et d'intégrer ces processus en détail. La section 4.2 présente la plante et les processus impliqués dans sa croissance, puis décrit le modèle mécaniste ARNICA, et présente son intégration dans un modèle statistique non linéaire à effets mixtes. La section 4.3 décrit le cadre statistique considéré et la procédure d'inférence associée. La section 4.4 présente les premiers résultats numériques obtenus à partir de données simulées et de données réelles. Enfin un exemple de calcul de la statistique du rapport de vraisemblance pendant l'inférence des paramètres, à partir de l'algorithme SARIS (chapitre 3 :Guédon et al. (2024b)) est présentée.

4.2 . Modélisation mécaniste et statistique

4.2.1 . Description d'*Arabidopsis thaliana* et de ses processus de croissance

Les plantes présentent une croissance complexe et coordonnée grâce à leurs deux principaux systèmes : le système aérien et le système racinaire. Le système aérien, constitué des feuilles et des tiges, joue un rôle crucial dans la photosynthèse, le processus par lequel les plantes captent le dioxyde de carbone (CO_2) de l'air et, en utilisant l'énergie solaire, le convertissent en sucres nécessaires à leur croissance. Cette production de carbone est essentielle pour la formation de nouvelles cellules et tissus dans les parties aériennes et racinaires de la plante. En parallèle, le système racinaire, composé des racines, est responsable de l'absorption de l'azote présent dans le sol (notamment). L'azote est un élément fondamental pour la synthèse des protéines et des acides nucléiques, contribuant ainsi à la croissance et au développement de la plante. Les racines captent l'azote sous différentes formes, telles que le nitrate et l'ammonium, l'utilisent pour la croissance racinaire et le transportent vers les parties aériennes où il est utilisé pour construire les structures cellulaires nécessaires.

Le modèle ARNICA est un modèle de plante entière, décrivant les processus expliqués dans le paragraphe précédent. Les grandeurs physiques modélisées sont au nombre de 15 et sont les suivantes :

- **Compartiment aérien** :

- LA (*leaf area*) : surface totale des feuilles
- PLA (*projected leaf area*) : surface des feuilles projetée, qui représente la surface active pour la photosynthèse
- SDW (*shoot dry weight*) : masse sèche des parties aériennes de la plante (feuilles, tiges)
- dLA_N (*the daily leaf area expansion allowed by the nitrogen quantity*) : variation journalière de la surface foliaire permise par la quantité d'azote disponible
- dLA (*the daily leaf area expansion*) : variation journalière de la surface foliaire
- QC_{prod} (*production of carbon*) : quantité de carbone produite
- R_QC_{avail} (*quantity of carbon available after leaf expansion*) : quantité de carbone disponible après l'expansion foliaire

- **Compartiment racinaire** :
  - RDW (*root dry weight*) : masse sèche des racines
  - $dRDW$ (*daily root growth*) : croissance journalière des racines
  - QN_{stor} (*quantity of nitrogen in the storage compartment*) : quantité d'azote stockée dans la plante
  - QN_{uptk} (*Nitrogen uptake (produced by the roots)*) : quantité d'azote absorbée par les racines
  - dQN_{stor} (*unused nitrogen after leaf expansion*) : quantité d'azote non utilisée pour la production de biomasse réallouée au compartiment de stockage
  - QN_{avail} (*quantity of nitrogen available*) : quantité d'azote disponible
  - TNQ (*total quantity of nitrogen accumulated*) : quantité totale d'azote accumulée depuis le début de la croissance
  - TDW (*total dry weight*) : masse sèche totale de la plante

Ces grandeurs seront par la suite appelées sorties du modèle, ou variables réponses. Afin de modéliser les échanges d'azote et de carbone au sein des compartiments, une hypothèse d'utilisation prioritaire du carbone par la partie aérienne et de l'azote par la partie racinaire est posée. La section suivante présente les équations décrivant la dynamique de ces grandeurs.

4.2.2 . ARNICA : un modèle mécaniste de croissance de plante

Initialement décrit dans Richard-Molard et al. (2007) comme un système d'équations différentielles, une version discrétisée basée sur un pas de temps journalier est présentée. Les

sorties du modèle à la date $t \in \mathbb{N}$, sont représentées par un vecteur $X_t \in \mathbb{R}^{15}$, et vérifient un système dynamique de la forme $X_{t+1} = G(X_t, \varphi)$ où G est une fonction non linéaire définissant le modèle ARNICA, paramétrisée par un vecteur φ de paramètres à estimer. Afin de faciliter la lecture des équations définissant le modèle, les paramètres du modèle composant le vecteur φ seront notés en gras, et les sorties décrites dans la section précédente présenteront une dépendance temporelle explicite.

Dans un premier temps, le modèle met à jour la masse sèche des racines (RDW) et la surface foliaire (LA), en leur ajoutant la quantité quotidienne de masse sèche produite ($dRDW$) et de surface foliaire produite (dLA) :

$$RDW(t) = RDW(t-1) + dRDW(t-1)$$

$$LA(t) = LA(t-1) + dLA(t-1)$$

Les quantités $dRDW$ et dLA sont produites grâce aux flux de carbone et d'azote. La quantité de carbone produite ($QCprod$) par photosynthèse est supposée proportionnelle à la surface foliaire projetée (PLA), qui est elle, reliée à LA suivant la loi de Beer-Lambert :

$$QCprod(t) = \boldsymbol{sca} \cdot PLA(t)$$

$$PLA(t) = \boldsymbol{g}(1 - e^{-\boldsymbol{k}LA(t)})$$

où $\boldsymbol{sca}$ est le taux d'assimilation spécifique du carbone, $\boldsymbol{g}$ est l'efficacité maximale d'interception de la lumière et $\boldsymbol{k}$ est le coefficient d'extinction de la loi de Beer-Lambert.

L'azote qui n'aurait pas été utilisé pour la croissance de la veille, est placé dans un compartiment de stockage ($QNStor$). La quantité d'azote présente dans le compartiment de stockage est définie quotidiennement comme la quantité stockée qui n'a pas encore été remobilisée, ajoutée à la quantité d'azote rentrée dans le compartiment de stockage le jour précédent ($dQNStor$).

$$QNStor(t) = (1 - \boldsymbol{remob\_NStor}) \cdot QNStor(t-1) + dQNStor(t-1)$$

La quantité d'azote absorbée par les racines ($QNuptk$), est dérivée de la masse sèche des

racines via le paramètre spécifique d'absorption d'azote $\boldsymbol{snu}$, qui donne l'efficacité de l'absorption d'azote par jour et par unité de biomasse racinaire.

$$QNuptk(t) = \boldsymbol{snu} \cdot RDW(t)$$

La quantité totale d'azote disponible ($QNavail$) est alors donnée comme la somme de l'azote absorbé et de l'azote remobilisé à partir du compartiment de stockage, à un taux $\boldsymbol{remob\_NStor}$. Nous définissons également la quantité totale d'azote accumulée par la plante au jour t, comme la somme de chaque absorption quotidienne d'azote (TNQ).

$$QNavail(t) = QNuptk(t) + \boldsymbol{remob\_NStor} \cdot QNStor(t)$$

$$TNQ(t) = TNQ(t-1) + QNuptk(t-1)$$

La masse sèche totale est mise à jour à partir de la production de carbone du jour précédent et de la teneur en carbone de la biomasse ($\boldsymbol{Ccont}$). La masse sèche des parties aériennes résulte ensuite de la différence entre la masse sèche totale et la biomasse racinaire.

$$TDW(t) = TDW(t-1) + \frac{QCprod(t-1)}{\boldsymbol{Ccont}}$$

$$SDW(t) = TDW(t) - RDW(t)$$

La quantité de carbone disponible après l'expansion foliaire ($R\_QCavail$) est définie comme la différence entre la quantité de carbone produite et celle consommée par la croissance des feuilles. Cette dernière est contrôlée par un paramètre de coût en carbone de l'expansion des feuilles $\boldsymbol{LA\_Ccost}$.

$$R\_QCavail(t) = \max(0, QCprod(t) - \boldsymbol{LA\_Ccost} \cdot dLA(t-1))$$

La croissance quotidienne des racines ($dRDW$) résulte de la confrontation entre la croissance permise par la quantité d'azote disponible ($\frac{QNavail}{\boldsymbol{R\_NCont\_min}}$), la croissance permise par la quantité de carbone disponible après l'expansion des feuilles ($\frac{R\_QCavail}{\boldsymbol{Ccont}}$) et un taux de croissance relatif maximal correspondant au paramètre $\boldsymbol{R\_RGRmax}$. Le paramètre $\boldsymbol{R\_NCont\_min}$ correspond à la teneur minimale en azote dans les racines.

$$dRDW(t) = \min\left(\frac{QNavail(t)}{\boldsymbol{R\_NCont\_min}}, \frac{R\_QCavail(t)}{\boldsymbol{Ccont}}, \boldsymbol{R\_RGRmax} \cdot RDW(t)\right)$$

L'expansion quotidienne de la surface foliaire (dLA_N) résulte de la confrontation entre la croissance permise par la quantité de carbone disponible provenant de la production quotidienne de carbone ($\frac{R\_QCavail}{\boldsymbol{Ccont}}$), la croissance permise par la quantité d'azote disponible $\frac{QNavail}{\boldsymbol{R\_NCont\_min}}$, en tenant compte de la teneur minimale en azote des racines via le paramètre $\boldsymbol{R\_NCont\_min}$ et d'un taux de croissance relatif maximal représenté par le paramètre ($\boldsymbol{LA\_RERmax}$). Le paramètre $\boldsymbol{\alpha}$ représente la surface foliaire produite avec une unité d'azote dans les parties aériennes.

$$dLAN(t) = \boldsymbol{\alpha} \cdot \boldsymbol{R\_NCont\_min} \cdot \max\left(0, \frac{QNavail(t)}{\boldsymbol{R\_NCont\_min}} - \frac{R\_QCavail(t)}{\boldsymbol{Ccont}}\right)$$

$$dLA(t) = \min\left(\frac{QCprod(t)}{\boldsymbol{LA\_Ccost}}, dLAN(t), \boldsymbol{LA\_RERmax} \cdot LA(t)\right)$$

Finalement, tout l'azote non utilisé pour la croissance foliaire est alloué au compartiment de stockage.

$$dQNStor(t) = \max\left(0, \frac{dLAN(t) - dLA(t)}{\boldsymbol{\alpha}}\right)$$

Initialisation du modèle : Quatre paramètres sont nécessaires à l'initialisation du modèle :

- la quantité de matière sèche totale : $\boldsymbol{DW\_init}$
- la surface foliaire $\boldsymbol{LA\_init}$
- la quantité totale d'azote $\boldsymbol{QN\_init}$
- le rapport entre la masse sèche des racines et la masse sèche totale $\boldsymbol{RT\_ratio\_init}$

À $t = 0$, la masse sèche totale est initialisée par $TDW(0) = \boldsymbol{DW\_init}$. Le rapport entre la masse sèche des racines et la masse sèche totale $\boldsymbol{RT\_ratio\_init}$ permet de calculer $RDW(0)$ et $SDW(0)$. La surface foliaire initiale $\boldsymbol{LA\_init}$ définit $LA(0)$. La quantité totale d'azote initiale $\boldsymbol{QN\_init}$ définit $TNQ(0)$. À partir de ces valeurs initiales, les autres sorties du modèle sont obtenues en utilisant les relations décrites plus haut. Le temps $t = 0$ correspond au jour 10 après le semis. Une version schématique du modèle ARNICA est présentée dans la figure (4.2), qui illustre les liens entre les différentes variables, et les échanges d'azote et de carbone entre les compartiments aériens et racinaires.

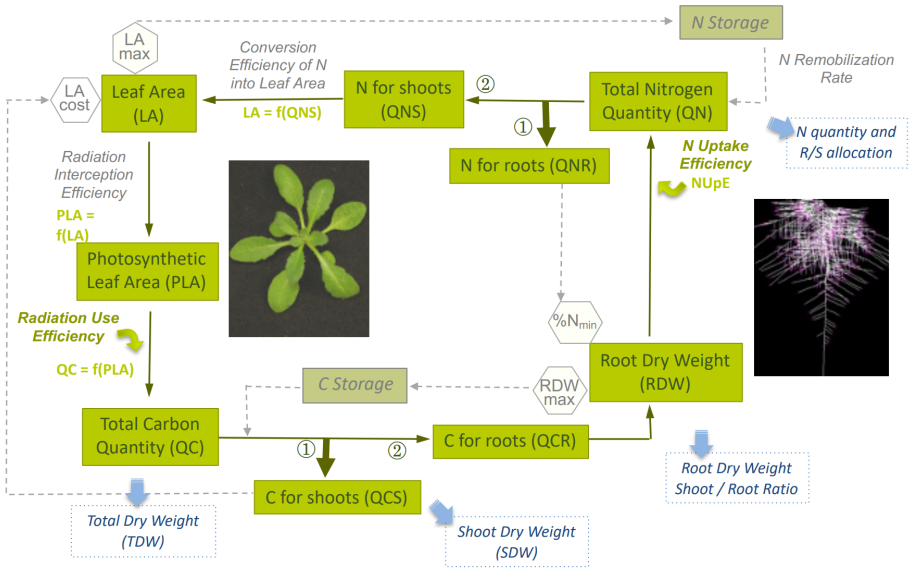


Figure 4.2 – Illustration schématique du modèle ARNICA, présentant le compartiment aérien (à gauche) et le compartiment racinaire (à droite).

Ainsi, le modèle ARNICA décrit 15 traits mis à jour quotidiennement par un système d'équations non linéaires. Ces 15 sorties sont représentées pour chaque temps $t \geq 0$ par un vecteur $X_t \in \mathbb{R}^{15}$. Les équations mettent en jeu 15 paramètres inconnus qui sont résumés dans un vecteur $\varphi \in \Phi = \Phi_1 \times ... \times \Phi_{15}$, où pour tout $k = 1, ..., 15$, $\Phi_k \subset \mathbb{R}$ est l'espace des valeurs admissibles pour le paramètre φ_k. En effet, les φ_k ayant un sens biologique précis, ils sont contraints dans une certaine gamme de valeurs, définie à partir des connaissances biologiques.

Le vecteur de sortie X_t est donc mis à jour suivant une relation de la forme $X_{t+1} = G(X_t, \varphi)$ à partir d'un paramètre $\varphi \in \Phi$ donné. De plus, comme expliqué plus haut, X_0 est entièrement déterminé par la donnée d'un paramètre φ. Plus précisément, on a $X_0 = h(\varphi)$ où h est définie par les relations présentées dans le paragraphe **Initialisation du modèle**. Ainsi pour tout temps t, nous avons la relation :

$$X_t = G^{(t)}(h(\varphi), \varphi)$$

où $f^{(k)}$ correspond à la fonction composée $\underbrace{f \circ ... \circ f}_{k \ fois}$ pour une fonction f quelconque.

Nous définissons ainsi pour un paramètre φ et un temps t donnés la fonction suivante :

$$(t, \varphi) \mapsto \tilde{g}(t, \varphi) := G^{(t)}(h(\varphi), \varphi)$$

Ainsi le modèle mécaniste ARNICA permettant de décrire un ensemble de 15 traits, notés X_t, au cours du temps, à partir d'un paramètre φ inconnu peut se résumer sous la forme :

$$X_t = \tilde{g}(t, \varphi)$$

Cependant cette modélisation déterministe décrit un unique génotype d'*Arabidopsis thaliana*, pour un paramètre φ spécifique à ce génotype. Deux génotypes différents présentant des caractéristiques phénotypiques différentes, seront décrits par des paramètres φ différents. Estimer un paramètre par génotype peut ne pas être la meilleur solution. La principale limite de cette approche est la quantité de données. Les données expérimentales peuvent être difficiles à obtenir ce qui les rend souvent rares et chères. Dans l'étude d'un processus de croissance, si chaque génotype est mesuré J fois, l'approche individuelle bénéficie d'un échantillon de taille J lorsqu'une approche populationnelle se base sur un échantillon de taille $N \times J$ où N est le nombre de génotypes considérés. La section suivante présente une approche populationnelle considérant l'ensemble des génotypes disponibles en même temps, en intégrant le modèle mécaniste ARNICA dans un modèle à effets mixtes. De cette manière la variabilité génotypique des paramètres du modèle est modélisée par des effets aléatoires.

4.2.3 . Une approche populationnelle : intégration du modèle mécaniste dans un modèle à effets mixtes

Nous considérons N génotypes, chacun mesuré J fois à des dates $0 < t_1 < t_2 < ... < t_J$, où les dates t_j, $(j = 1, ..., J)$ sont connues. Supposons que sont observées $q \in \{1, ..., 15\}$ sorties du modèle décrit dans la section 4.2.2.

Notons $y_{ij} \in \mathbb{R}^q$ l'observation correspondant au génotype i au temps t_j, $(j = 1, ..., J_i)$. Nous supposons que pour $i = 1, ..., N$ et $j = 1, ..., J_i$, les observations y_{ij} suivent le modèle log normal suivant :

$$\log y_{ij} = \log \tilde{g}(t_j, \varphi_i) + \varepsilon_{ij}$$

où $\varphi_i \in \Phi$ est un vecteur de paramètre spécifique au génotype i, et les ε_{ij}, $(j = 1, ..., J_i, i = 1, ..., N)$ sont des résidus gaussiens qu'on supposera indépendants et identiquement distribués selon une loi gaussienne multidimensionnelle de dimension q. Pour un même individu statistique i, cette étape modélise la variabilité intra-individuelle, c'est à dire la croissance des observations d'un seul génotype au cours du temps.

Remarque 4.1. *Un modèle log normal est considéré pour représenter au mieux les données. Cette transformation est courante dans les modèles temporels croissants, afin d'avoir une variabilité plus faible pour les petites valeurs et plus forte pour les plus grandes valeurs. Cette transformation est équivalente à considérer une erreur gaussienne multiplicative sur la réponse initiale y.*

Afin de modéliser la variabilité génotypique, qui correspond à la variabilité statistique inter-individuelle, nous supposons que le paramètre individuel φ_i, est la réalisation d'une variable latente non observée, modélisée comme suit :

$$\varphi_i = v(\beta + \Lambda\xi_i), \quad \xi_i \sim \mathcal{N}(0, I_{15})$$

où

- β est un vecteur de $\mathbb{R}^{15}$ commun à tous les génotypes
- Λ est une matrice diagonale, semi-définie positive
- v est une fonction bijective de $\mathbb{R}^{15}$ dans $\Phi_1 \times ... \times \Phi_{15}$

Pour résumer, on considère le modèle non linéaire à effets mixtes multidimensionnel suivant :

$$\begin{cases} \log y_{ij} = \log g(t_j, \beta, \Lambda\xi_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \Sigma) \\ \xi_i \sim \mathcal{N}(0, I_{15}) \end{cases} \tag{4.1}$$

où les résidus $(\varepsilon_{ij})_{ij}$ et les variables latentes $(\xi_i)_i$ sont des variables aléatoires mutuellement indépendantes, Σ est une matrice symétrique définie positive et diagonale, et la fonction g est définie comme $g(t_j, \beta, \Lambda\xi_i) = \tilde{g}(t, v(\beta + \Lambda\xi_i))$. Le paramètre statistique inconnu, est donc :

$$\theta = (\beta, \Lambda, \Sigma)$$

La section suivante présente les données disponibles. La procédure d'inférence et le cadre de simulation seront construits pour correspondre aux données disponibles.

4.2.4 . Présentation des données et de l'objectif biologique

Le jeu de données sur *Arabidopsis thaliana* comporte 48 écotypes, représentant 100% de la diversité allélique de l'espèce. Un écotype est une sous-catégorie de l'espèce, spécifique à un habitat donné, présentant des caractéristiques homogènes et donc une diversité génétique réduite. Pour des raisons de clarté, nous ferons l'amalgame entre génotype et écotype comme individu statistique d'intérêt. Ainsi par la suite le terme "individu" (au sens statistique) correspondra à un génotype.

On considère $N = 48$ individus, chacun mesuré à trois dates : 21, 28 et 32 jours après le semis. On observe $q = 6$ variables réponses : *TDW, RDW, TNQ, QNStor, LA* et *PLA*. Les données sont représentées graphiquement dans la figure 4.3. Dans chacun des graphiques, chaque couleur représente un génotype. Un paramètre φ^{ref} servira de paramètre de référence. La ligne continue présente dans la figure 4.4 correspond aux sorties du modèle ARNICA, obtenues à partir de φ^{ref}. Ce paramètre a été estimé sur des données d'*Arabidopsis thaliana* différentes, ne comprenant pas une aussi grande diversité génétique. L'estimation a été réalisée, avant le début de ce travail de thèse, en ajustant courbe par courbe les valeurs de paramètres, afin de représenter au mieux les traits phénotypiques modélisés. Les données utilisées pour l'inférence de ce paramètre étaient plus propices à une telle approche. En effet ces dernières étaient constituées de peu de génotypes, mais d'un nombre plus raisonnable de mesures par génotypes. L'ajustement individuel était donc plus envisageable. Un tel ajustement sur des courbes ne comprenant que 3 valeurs serait beaucoup plus difficile.

Ces figures illustrent la variabilité génotypique observée dans les données de croissance de *Arabidopsis thaliana*.

Comme introduit précédemment l'objectif est d'identifier les paramètres porteurs de cette variabilité. L'expertise biologique établit des hypothèses quant aux paramètres les plus variables. Dans cette étude, les paramètres supposés les plus variables sont le paramètre d'absorption spécifique d'azote **snu**, et le taux d'assimilation spécifique du carbone **sca** (Richard-Molard et al., 2007).

Le paramètre **snu** fait référence à la capacité d'une plante à absorber l'azote du sol par unité de masse racinaire. C'est une mesure de l'efficacité avec laquelle la plante peut prélever l'azote de son environnement pour l'utiliser dans sa croissance et son métabolisme. Le paramètre **sca**

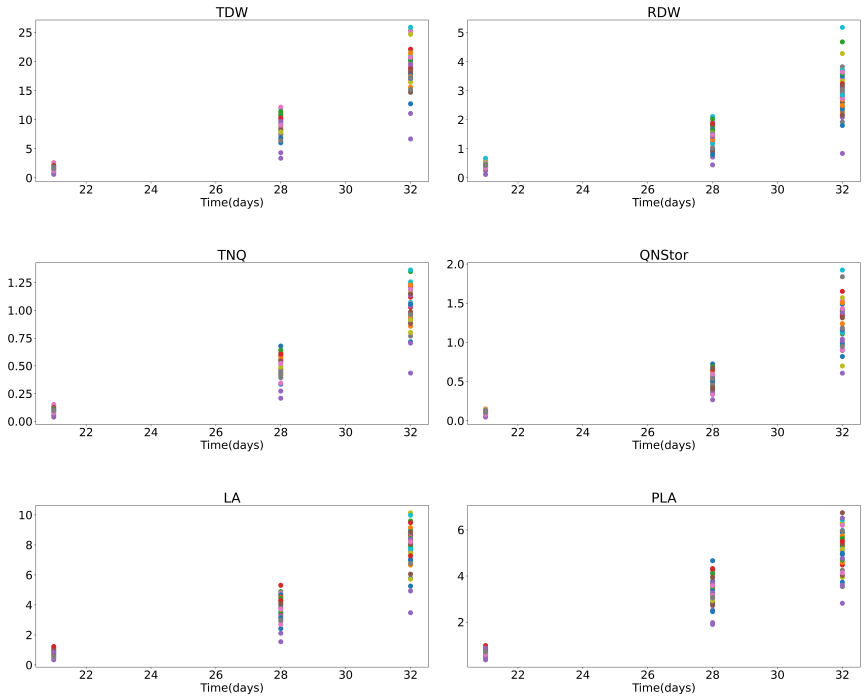


Figure 4.3 – Données réelles d'*Arabidopsis thaliana*. Chaque couleur représente la réponse d'un génotype. Les trois dates observées sont les jours 21, 28, 32 après l'ensemencement

est lié à la production d'énergie et de matière organique par photosynthèse. Ces deux paramètres sont donc essentiels car ils reflètent l'efficacité spécifique du génotype à utiliser des ressources de l'environnement.

4.3 . Modèle statistique et procédure d'inférence

4.3.1 . Modèle statistique restreint

L'espace des paramètres du modèle 4.1 décrit précédemment est :

$$\Theta^{tot} = \left\{ \theta = (\beta, \Lambda, \Sigma); \quad \beta \in \mathbb{R}^{15}, \Lambda \in D_{+}^{15}, \Sigma \in D_{++}^{15} \right\}$$

où D_{++}^{k} (respectivement D_{+}^{k}) représente l'ensemble des matrices diagonales définies positives (respectivement semi-définies positives). Cependant, à cause de la complexité du modèle, nous avons choisi de considérer dans un premier temps un nombre restreint de paramètres, afin d'expérimenter la procédure d'inférence proposée. On considère ici l'estimation de $nindiv$

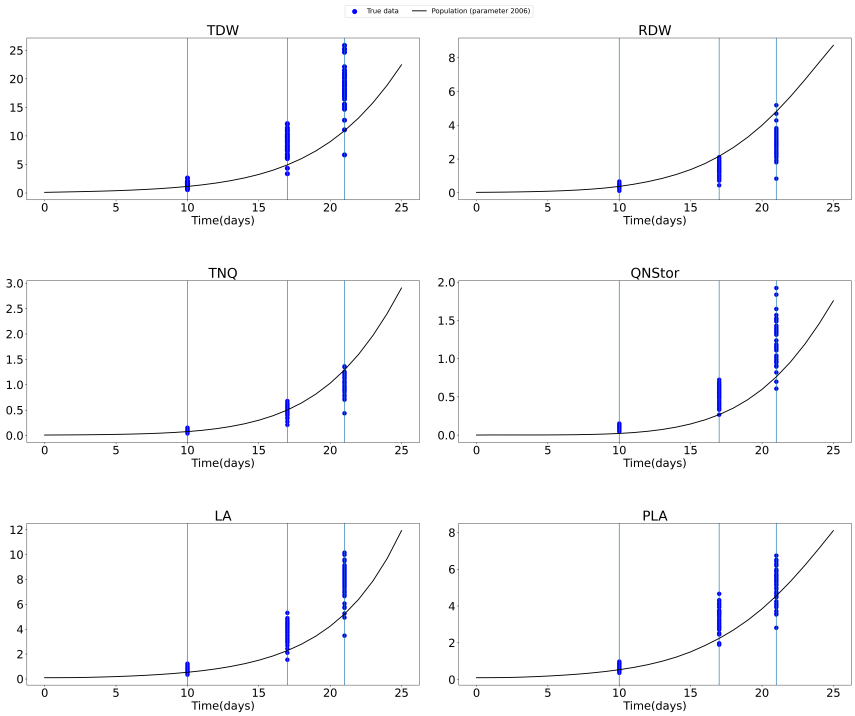


Figure 4.4 – Comparaison des données réelles d'*Arabidopsis thaliana* (points bleus) et des réponses simulées avec le modèle ARNICA, à partir des paramètres φ^{ref} (ligne noire).

paramètres où $0 < nindiv \leq 15$, supposés à priori variables génétiquement, et $npop$ paramètres supposés non variables, où $0 < npop \leq 15$. On note Ω_{indiv} et Ω_{pop} les ensembles contenant les noms des paramètres supposés variables, et ceux des paramètres supposés fixes. Si tous les paramètres ne sont pas estimés : $nindiv + npop \leq 15$. Les paramètres non estimés sont fixés à leurs valeurs de référence (celles du paramètre φ^{ref}, et donc celles utilisées dans la figure 4.4). De plus, tel que dans les données disponibles, les 15 variables réponses ne sont pas forcément observées. Ainsi, on considère dans le modèle la possibilité de ne prendre en compte que q variables réponses ($(0 < q \leq 15)$), dont les noms sont contenus dans Ω_{output}. Trois types de paramètres statistiques sont donc considérés :

- les paramètres de population $\beta_{pop} \in \mathbb{R}^{npop}$ communs à tous les individus, ne présentant par hypothèse aucune variabilité génotypique
- les paramètres $(\beta_{indiv}, \Lambda) \in \mathbb{R}^{nindiv} \times D_+^{nindiv}$ définissant les paramètres individuels
- les variances résiduelles $\Sigma \in D_{++}^q$

Tel que décrit ici, l'espace des paramètres du modèle réduit est :

$$\Theta = \left\{ \theta = (\beta_{pop}, \beta_{indiv}, \Lambda, \Sigma); \quad \beta_{pop} \in \mathbb{R}^{npop}, \beta_{indiv} \in \mathbb{R}^{nindiv}, \Lambda \in D_+^{nindiv}, \Sigma \in D_{++}^{q} \right\} \tag{4.2}$$

Remarque 4.2. *Le modèle restreint et le modèle complet sont en réalité imbriqués. En effet, les paramètres de population peuvent être vus comme des paramètres individuels avec des facteurs d'échelle correspondants (les éléments diagonaux de Λ) égaux à 0. De plus, les paramètres non estimés peuvent être interprétés comme des paramètres de population contraint à être dans le singleton correspondant à leur valeur de référence. Ainsi, le modèle considéré est toujours celui décrit en 4.1 en considérant le niveau intra individuel comme* $\log y_{ij} = \log g(\beta_{indiv}, \beta_{pop}, \Lambda\xi_i) + \varepsilon_{ij}$.

La section suivante présente la procédure d'inférence considérée.

4.3.2 . Inférence des paramètres du modèle à effets mixtes

Le modèle 4.1 permet de définir la vraisemblance complète du modèle, relative aux variables $(y_i, \xi_i)_{i=1,...,N}$, pour un paramètre $\theta \in \Theta$ donné. On note $f(y_i, \xi_i, \theta)$ la vraisemblance complète relative à un individu $i = 1, ..., N$. Cette vraisemblance est définie par la structure hiérarchique du modèle 4.1 de la manière suivante :

$$f(y_i, \xi_i; \theta) = \left(\prod_{j=1}^{J_i} \phi\left(\log y_{ij}; \log g(\beta_{indiv}, \beta_{pop}, \Lambda\xi_i), \Sigma\right) \right) \times \phi\left(\xi_i; 0, I_{nindiv}\right)$$

où la notation $\phi\left(x; \mu, V\right)$ représente la densité gaussienne multivariée, d'espérance μ, de matrice de variance V évaluée au point x.

Cependant le modèle 4.1 est un modèle à variables latentes. Les effets aléatoires $(\xi_i)_{i=1,...,N}$ ne sont pas observés et la vraisemblance complète relative aux variables $(y_i, \xi_i)_{i=1,...,N}$ est inconnue. On considère donc la vraisemblance relative uniquement aux observations $y_{1:N} = (y_i)_{i=1,...,N}$, qui correspond à une vraisemblance marginalisée sur les variables latentes. La log vraisemblance marginale d'un paramètre $\theta \in \Theta$ est définie de la manière suivante :

$$l(\theta; y_{1:N}) = \sum_{i=1}^{N} \log \left(\int_{\mathbb{R}^{nindiv}} f_i(y_i, \xi_i; \theta) d\xi_i \right) \tag{4.3}$$

L'inférence par maximum de vraisemblance consiste à trouver $\hat{\theta}_N$ défini par :

$$\hat{\theta}_N = \arg\max_{\theta \in \Theta} \quad l(\theta; y_{1:N})$$

Une approche populaire pour l'inférence dans les modèles à variables latentes est l'algorithme EM (Dempster et al., 1977) ou sa version stochastique SAEM (Delyon et al., 1999; Kuhn and Lavielle, 2004, 2005). Cependant, hors de la famille exponentielle, cet algorithme est difficile à mettre en place et ne présente plus de garanties théoriques (Debavelaere and Allassonnière, 2021). Une autre approche possible est la descente de gradient stochastique. Nous considérons ici un algorithme de descente de gradient stochastique préconditionné, directement inspiré de Baey et al. (2023). Le préconditionnement utilisé est basé sur une estimation de la matrice d'information de Fisher. La procédure utilisée est la suivante :

1. Entrées : $\xi_{1:n}^{(0)}$, $\Delta_{1:n}^{(0)}$, $\theta^{(0)}$, $y_{1:N}$, $\bar{\theta}^{(0)} = \theta^{(0)}$ critère d'arrêt, $k = 0$,
2. Jusqu'à la vérification du critère d'arrêt :
   - Pour chaque $i = 1, ..., N$, simuler $\xi_i^{(k+1)}$ à l'aide d'un noyau de transition $\Pi_{\theta^{(k)}}(.; \xi_i^{(k)})$
   - Mettre à jour les jacobiennes individuelles, pour chaque $i = 1, ..., N$:

$$\Delta_i^{(k+1)} = (1 - \gamma_{1,k})\Delta_i^{(k)} + \gamma_{1,k}\nabla_\theta \log f(y_i, \xi_i^{(k+1)}; \theta^{(k)})$$

     où $(\gamma_{1,k})_{k\geq 0}$ est une suite de pas positifs
   - Calculer le gradient actuel : $v_{k+1} = \frac{1}{N}\sum_{i=1}^{N} \Delta_i^{(k+1)}$
   - Calculer la matrice d'information de Fisher actuelle :

$$F_{k+1} = (1 - \gamma_{2,k})\frac{1}{n}\sum_{i=1}^{n} \Delta_i^{(k+1)} \left(\Delta_i^{(k+1)}\right)^T + \gamma_{2,k} Id_{d_\theta}$$

     où $(\gamma_{2,k})_{k\geq 0}$ est une suite de pas positifs
   - Mettre à jour : $\theta_{k+1} = \theta_k + \gamma_{3k} F_{k+1}^{-1} v_{k+1}$ où $(\gamma_{3,k})_{k\geq 0}$ est une suite de pas positifs
   - Mettre à jour $\bar{\theta}^{(k+1)} = \frac{k}{k+1}\bar{\theta}^{(k)} + \frac{1}{k+1}\theta_{k+1}$
   - k=k+1

où Id_{d_θ} est la matrice identité de taille $d_\theta = \dim(\Theta)$.

Remarque 4.3. • *L'étape de calcul de l'estimation de la matrice d'Information de Fisher est basée sur l'estimateur de Delattre and Kuhn (2019). Le terme $\gamma_{2,k} Id_{d_\theta}$ n'est considéré que durant une première phase de chauffe dans Baey et al. (2023) pour s'assurer de la non singularité de la matrice de préconditionnement. Ici ce terme est conservé pour l'intégralité de la phase*

d'inférence à des fins de stabilité, et afin de s'assurer de la non singularité de la matrice de conditionnement. En effet comme présenté dans le chapitre 2, si des composantes de Λ sont nulles, des problèmes de singularité se poseront.

- *Le gradient $\nabla_\theta \log f(y_i, \xi_i^{(k+1)}, \theta^{(k)})$, est calculé par différentiation automatique. Afin de considérer un modèle non contraint dans $\mathbb{R}^{d_\theta}$, l'inférence est réalisée sur une transformation bijective de Θ dans $\mathbb{R}^{d_\theta}$. Cette reparamétrisation est faite à l'aide du parametrization cookbook introduit dans Leger (2023)*
- *Lors de l'étape de simulation, la distribution exacte souhaitée est la distribution a posteriori de la variable latente sachant les observations. Cette distribution a une densité proportionnelle, à chaque étape k, à la vraisemblance complète $z \mapsto f(y_i, z, \theta_k)$. Simuler de manière exacte selon cette distribution n'est pas possible dans notre cas, une méthode de Monte-Carlo par chaîne de Markov est utilisée. Un choix populaire est l'utilisation d'une étape d'algorithme de Metropolis-Hastings (voir Robert and Casella (1999); Roberts and Rosenthal (2009) pour plus de détails).*

4.4 . Expériences numériques

L'implémentation du modèle et de la procédure d'inférence ainsi que l'ensemble des simulations ont été codés avec Python 3.11.9. L'ensemble des scripts sont disponibles sur Git-Hub [1].

4.4.1 . Cadre d'estimation et simulation de données

Afin d'effectuer une étude de simulation on considère le cadre de modélisation suivant :

- on considère $N = 48$ individus
- les temps d'observation sont $t_j \in \{11, 18, 22\}$ pour chacune des q réponses observées
- on considère $q = 6$ sorties du modèle ARNICA, les réponses observées sont regroupées dans $\Omega_{output} = \{$*TDW, RDW, TNQ, QNStor, LA , PLA*$\}$
- les paramètres individuels considérés sont regroupés dans $\Omega_{indiv} = \{$**snu**, **sca**, **R_RGRmax**, **LA_RERmax**$\}$
- aucun paramètre de population n'est considéré

1. https ://github.com/tguedon/arnica

En plus des paramètres **snu** et **sca** déjà présentés dans le paragraphe précédent, nous considérons également ici les paramètres **R_RGRmax** et **LA_RERmax** qui sont les taux de croissances relatifs maximaux des racines et des feuilles. Le contexte présenté ici tend à mimer celui des données d'*Arabidopsis thaliana* disponibles.

Valeurs des paramètres utilisées pour les simulations :

Pour tout $\nu \in \Omega_{indiv}$,

$$\beta_\nu = v_\nu^{-1}(\varphi_\nu^{ref})$$

où v_ν est la transformation bijective de $\mathbb{R}$ dans la gamme de valeurs possibles pour le paramètre $\nu \in \Omega_{indiv}$ et φ_ν^{ref} la valeur du paramètre de référence du paramètre ν.

Le paramètre $\Lambda = \text{diag}(\lambda_{snu}, \lambda_{sca}, \lambda_{R\_RGRmax}, \lambda_{LA\_RERmax})$ a été choisi de manière arbitraire, en considérant une variabilité nulle pour les paramètres **R_RGRmax** et **LA_RERmax**, afin de se conformer à l'hypothèse biologique d'intérêt. Nous appellerons par la suite hypothèse nulle, le cadre d'estimation dans lequel les paramètres **R_RGRmax** et **LA_RERmax** sont considérés comme paramètres de population i.e. sous hypothèse d'absence de variabilité génotypique.

Le paramètre Σ a été choisie à partir de l'ordre de grandeur des variances des valeurs des réponses considérées.

La figure 4.5 présente des données simulées (lignes rouges sur la figure) selon le modèle à effets mixtes 4.1. Cette simulation semble indiquer que le modèle proposé représente correctement les données réelles (points bleus sur la figure). La ligne noire correspond aux réponses simulées à partir du paramètre de référence.

4.4.2 . Étude de simulation

On considère les données simulées présentées en figure 4.5. Les variances utilisées pour simuler les données sont $\Lambda = \text{diag}(0.4, 0.2, 0.0, 0.0)$ et $\Sigma = \text{diag}(0.1, 0.05, 0.008, 0.015, 0.05, 0.03)$. Les pas $(\gamma_{s,k})_{k\geq 0}$ ($s = 1, 2, 3$) ont été choisis tels que $(\gamma_{1,k})_k = (\gamma_{3,k})_k$, et :

$$\begin{cases} \gamma_{1,k} = 0.9 & \text{si } k < K_{heat} \\ \gamma_{1,k} = \frac{0.9}{(k - K_{heat})^{-2/3}} & \text{sinon} \end{cases}$$

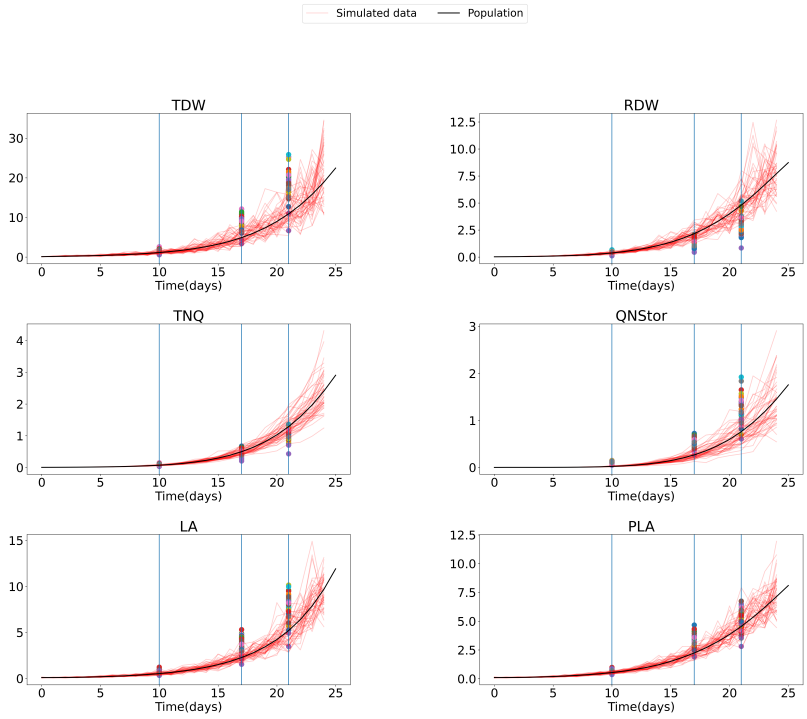


Figure 4.5 – Comparaison des données réelles d'*Arabidopsis thaliana* (points bleus), et des données individuelles simulées à partir du modèle ARNICA à effets mixte (courbes rouges). La ligne bleue représente les réponses du modèle ARNICA à partir des valeurs de paramètre de référence estimées manuellement.

et $\gamma_{2,k} = 0.09$. Une valeur de $K_{heat} = 300$ a été choisie. La valeur initiale θ_0 a été simulée selon une loi $\mathcal{N}(0, I_{d_\theta})$. Pour tout $\nu \in \Omega_{indiv}$, l'espace de valeurs admissibles a été choisi $\Phi_\nu = \mathbb{R}_*^+$. Ce choix est le moins restrictif car basé sur aucune hypothèse biologique *a priori*. Il est possible ici, car nous considérons un modèle restreint. Dans le cas d'un modèle statistique plus complexe intégrant un plus grand nombre de paramètres du modèle ARNICA, des choix de la forme $\Phi_\nu =]a_\nu; b_\nu[$ permettraient de simplifier l'inférence et de limiter les divergences et les compensations entre valeurs de paramètres, avec en contrepartie un besoin plus important de connaissances *a priori* sur les valeurs des paramètres.

L'inférence a été réalisée sur 100 jeux de données différents. Le critère d'arrêt d'inférence a été fixé à un nombre d'itérations maximal de $N_{max} = 750$. Afin d'illustrer les résultats, nous présentons les valeurs des paramètres au cours des itérations dans les figures 4.6, 4.7 et 4.8. Dans ces figures sont représentées les boxplots des valeurs des paramètres toutes les 50 itérations.

La courbe rouge en pointillés correspond à la valeur utilisée pour simuler les données.

Concernant les composantes fixes communes à tous les individus β_ν, pour $\nu \in \Omega_{indiv}$ dont les résultats sont présentés dans la figure 4.6, l'estimation est précise et rapide. Les variances résiduelles sont également correctement estimées. On observe cependant que pour quelques jeux de données, les paramètres ne convergent pas vers la vraie valeur. Ce phénomène peut être dû à la présence de maxima locaux de la vraisemblance, ou à des zones de très faible courbure de la vraisemblance. Ce dernier aspect est d'autant plus marqué dans la figure 4.7 présentant les résultats d'inférence des paramètres de variance des effets alétoires λ_ν pour $\nu \in \Omega_{indiv}$. Outre la convergence moins précise et moins rapide des paramètres, le nombre de procédures ne convergeant pas du tout est plus important. Par exemple, dans le graphique correspondant à l'inférence du paramètre λ_{R_RGRmax}, d'importantes sur-estimations sont observées, correspondant à des trajectoires d'estimation évoluant très peu. L'hypothèse expliquant la cause de ces mauvaises trajectoires est celle de zones de la vraisemblance très peu sensibles à ces paramètres. En annexes 4.6, les figures 4.12, 4.13 et 4.14 montrent les trajectoires d'estimation complètes. En annexes sont également présentées les mêmes figures, pour l'inférence sous l'hypothèse nulle : $\lambda_{R\_RGRmax}, \lambda_{LA\_RERmax} = 0$ (figures 4.15 à 4.20). Enfin, afin d'illustrer la consistance de l'algorithme d'inférence utilisé, les figure 4.21 à 4.26 montrent les figures d'estimation des paramètres sur un même jeu de données, avec des initialisations aléatoires.

4.4.3 . Inférence à partir des données réelles

Il est ensuite intéressant d'appliquer la procédure d'inférence présentée au jeu de données réelles, afin $i)$ d'estimer la variabilité inter-individuelle et $ii)$ de comparer les valeurs des paramètres estimés β_ν^{est}, pour $\nu \in \Omega_{indiv}$, avec les valeurs de référence.

Le même cadre d'estimation et la même procédure d'inférence sont considérés. Le tableau 4.1 compare les valeurs du paramètre de référence avec les valeurs estimées.

Remarque 4.4. *On rappelle que les paramètres β_ν correspondent aux paramètres non contraints, estimés dans $\mathbb{R}$. Afin de récupérer les valeurs des paramètres biologiques estimés, il faut appliquer la transformation inverse v_ν^{-1}.*

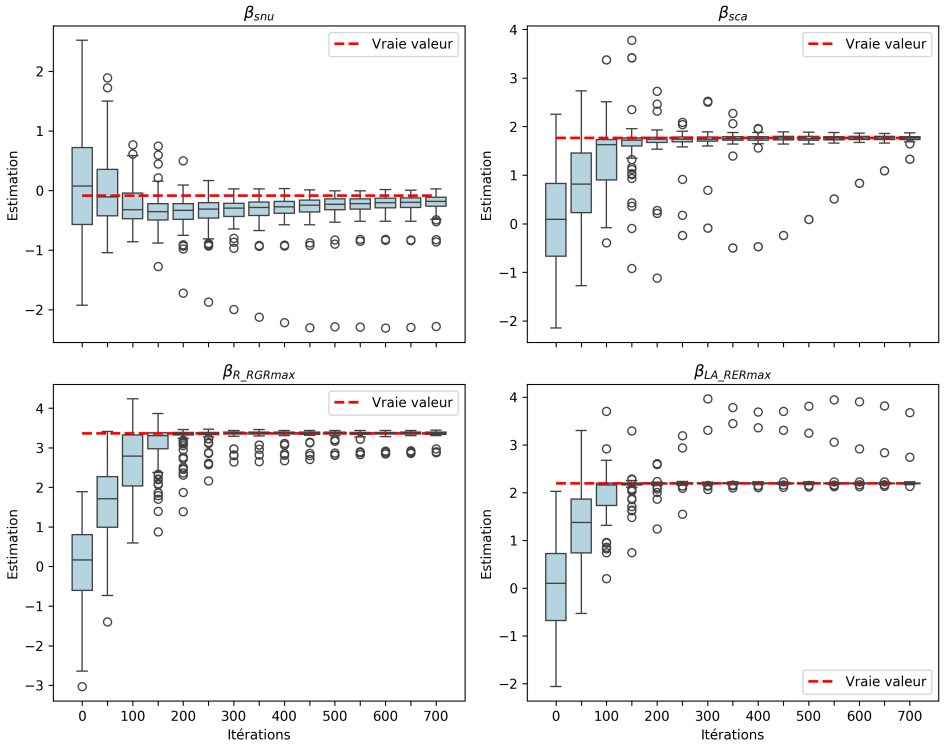


Figure 4.6 – Evolution de la valeur des paramètres de localisation $\beta_\nu, \nu \in \Omega_{indiv}$ pendant la procédure d'inférence, au cours des itérations. Les boxplots sont affichés toutes les 50 itérations, la ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 jeux de données simulés.

Paramètre	Valeur estimée	Valeur de référence
$v_{sca}^{-1}(\beta_{sca})$	0.146	0.192
$v_{snu}^{-1}(\beta_{snu})$	0.124	0.065
$v_{R\_RGRmax}^{-1}(\beta_{R\_RGRmax})$	1.423	0.34
$v_{LA\_RERmax}^{-1}(\beta_{LA\_RERmax})$	0.234	0.23

Table 4.1 – Comparaison des valeurs estimées sur les données réelles des paramètres du modèle ARNICA, avec les valeurs de référence estimées manuellement.

Il est intéressant de comparer les valeurs des paramètres estimées à celles de référence. L'estimation de β_{LA_RERmax} est égale à sa valeur de référence, alors que celles de β_{R_RGRmax} et β_{snu} sont très supérieures à leurs valeurs de référence. Une des limites de ces résultats réside dans l'absence de mesure d'incertitude. Une approche possible serait de considérer des intervalles de confiance asymptotiques, basés sur l'estimation de la matrice d'information de Fisher. Cependant, avec seulement $N = 48$ individus, l'approche asymptotique est questionnable. De

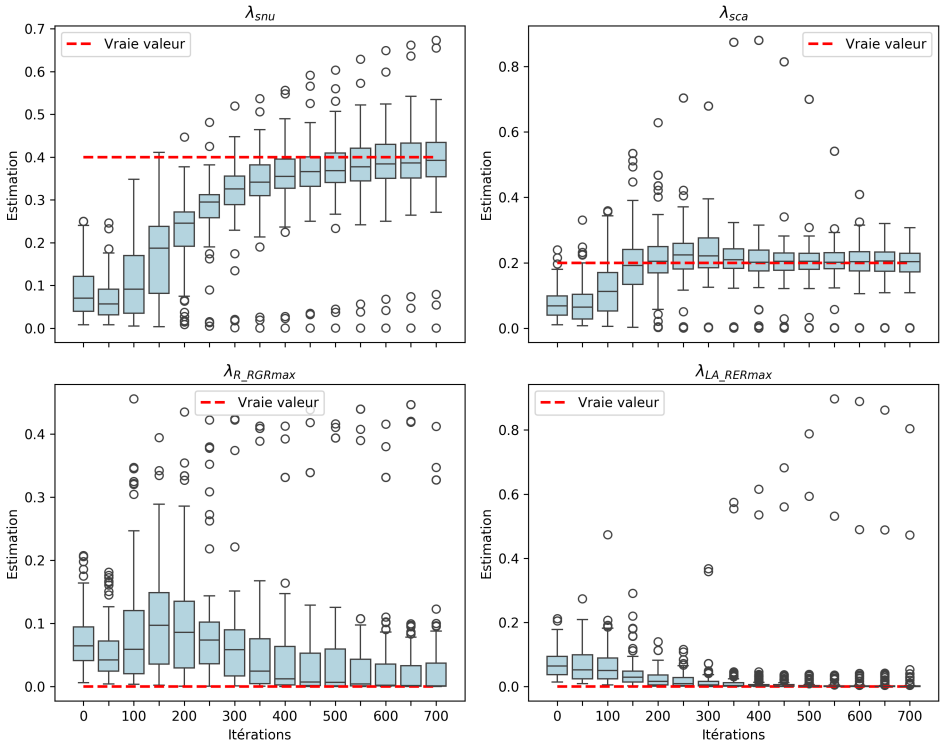


Figure 4.7 – Evolution de la valeur des paramètres d'échelle $\lambda_\nu, \nu \in \Omega_{indiv}$ pendant la procédure d'inférence, au cours des itérations. Les boxplots sont affichés toutes les 50 itérations, la ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 jeux de données simulés.

plus, l'estimation de cette matrice requiert une approche par Monte-Carlo, à cause de la forme intégrée de la vraisemblance. L'algorithme proposée par Baey et al. (2023) retourne un estimateur de cette matrice d'information de Fisher, cependant l'approche utilisée, qui considère une version modifiée de cet algorithme, retourne un estimateur biaisée de cette estimation.

La figure 4.9 présente les données simulées à partir du paramètre estimé. Les courbes rouges correspondent aux sorties du modèle ARNICA générées à partir du paramètre φ^{est}. Ce paramètre est défini de la manière suivante :

$$\begin{cases} \varphi_\nu^{est} & = v_\nu(\beta_\nu^{est}), \quad \nu \in \Omega_{indiv} \\ \varphi_\nu^{est} & = \varphi_\nu^{ref}, \quad \text{sinon} \end{cases}$$

Les courbes noires correspondent aux réponses simulées avec les valeurs du paramètre de référence φ^{ref}. Les points bleus sont les données réelles. Il est très encourageant de constater

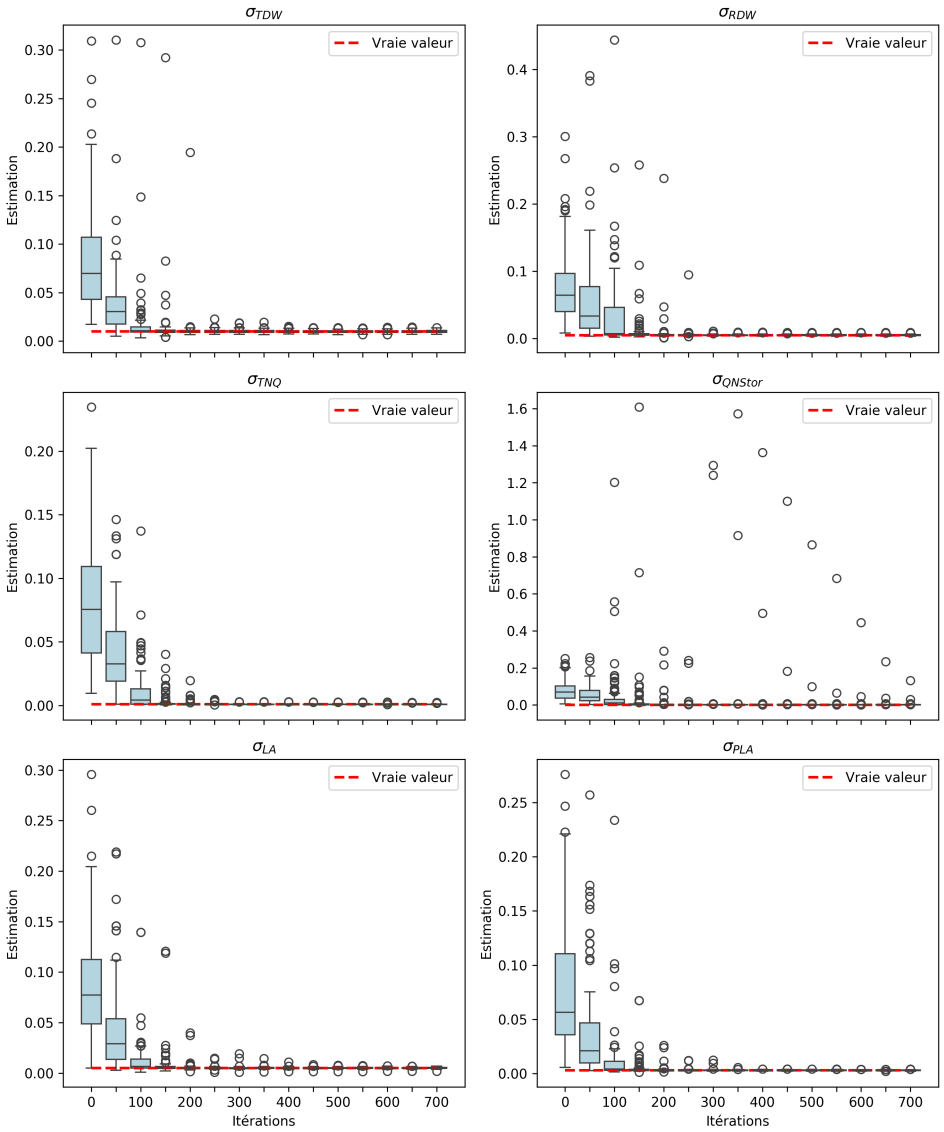


Figure 4.8 – Evolution de la valeur des paramètres de variance des résidus $\sigma_\nu, \nu \in \Omega_{output}$ pendant la procédure d'inférence, au cours des itérations. Les boxplots sont affichés toutes les 50 itérations, la ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 jeux de données simulés.

que les dynamiques de croissance de la masse sèche racinaire (RDW) et de la quantité d'azote stockée dans le compartiment de stockage (QNStor) semblent mieux représentées par les paramètres estimés par le modèle statistique proposé, que par le paramètre φ^{ref}. Il est cependant attendu que l'ajustement ne soit pas aussi bon pour l'ensemble des sorties, étant donné que beaucoup de paramètres du modèle ARNICA sont fixés et non estimés.

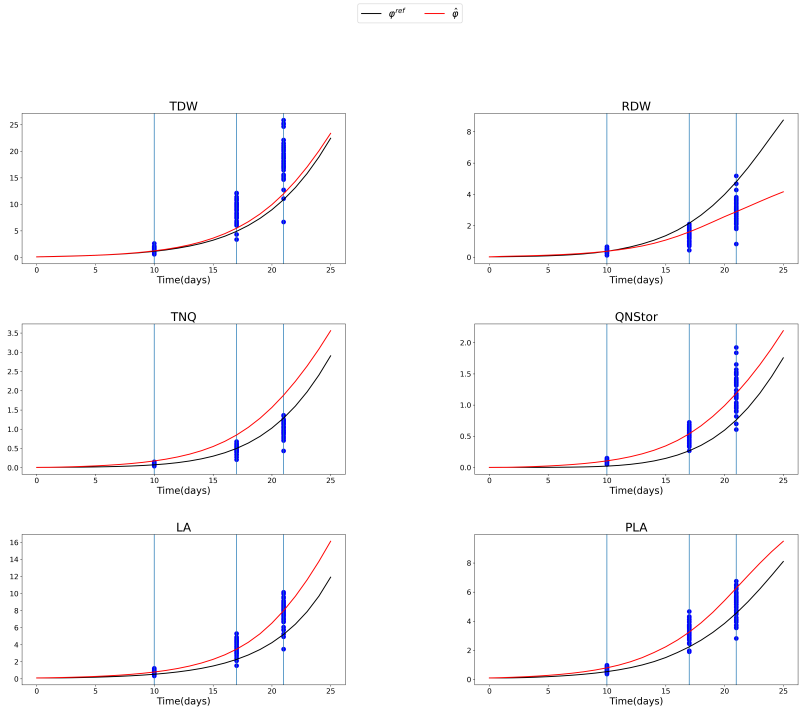


Figure 4.9 – Comparaison des données réelles d'*Arabidopsis thaliana* (points bleus) avec les sorties du modèle ARNICA obtenues avec le paramètre de référence φ^{ref} (courbe noire) et le paramètre estimé sur les données réelles φ^{est} (courbe rouge)

De manière assez naturelle, plus le nombre de paramètres à estimer augmente, plus l'inférence est complexe. Lorsque l'on considère trop de paramètres, $i)$ des phénomènes de compensation entre paramètres apparaissent, ce qui peut être dû à des problèmes d'identifiabilité du modèle, $ii)$ des paramètres d'échelle surestimés biaisent l'estimation des paramètres de localisation, $iii)$ des compensations entre variances résiduelles et paramètres d'échelle peuvent avoir lieu. Des raffinements de la procédure d'inférence seront donc nécessaires pour enrichir le modèle, et tendre vers un modèle statistique non restreint, modélisant la totalité des paramètres du modèle ARNICA.

Enfin, même si l'approche proposée ne considère pas explicitement de paramètres individuels (mais une distribution de paramètres individuels), il est tout de même possible d'estimer ces paramètres individuels, afin d'obtenir des ajustements individuels des données.

Soit $i = 1, ..., N$ un individu, le paramètre φ_i^{est} individuel est défini de la manière suivante :

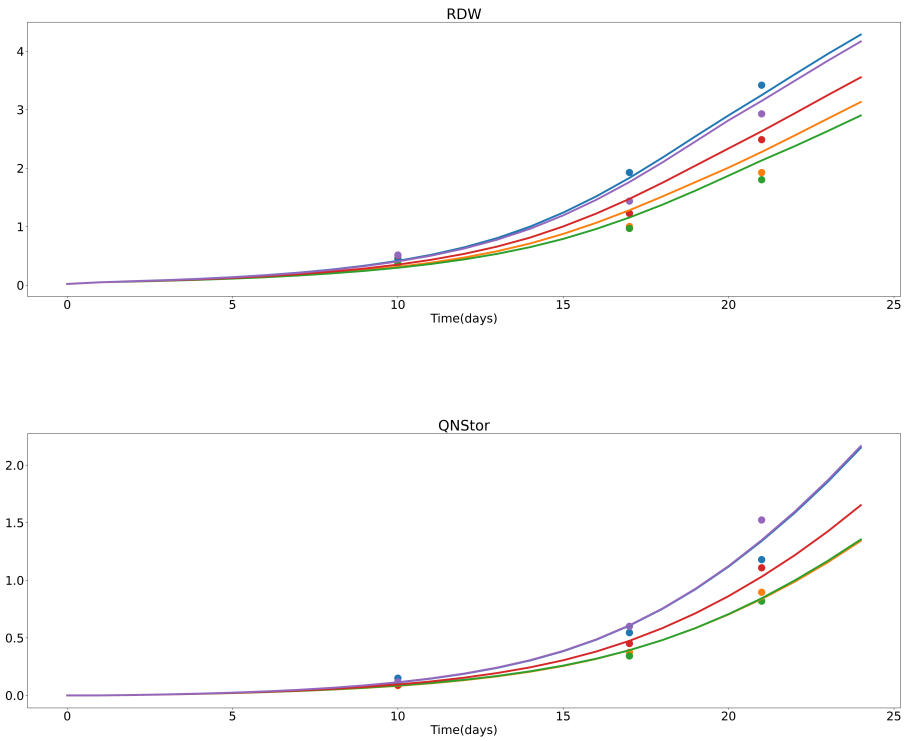


Figure 4.10 – Différentes courbes d'ajustement individuel pour 5 individus, et les données réelles correspondantes (les points) associées par couleur. Les courbes sont obtenues comme des sorties du modèle ARNICA à partir d'un paramètre individuel estimé.

pour tout $\nu \in \Omega_{indiv}$, $\varphi_{i,\nu}^{est} = v_\nu(\beta_\nu^{est} + \lambda_\nu^{est}\hat{\xi}_{\nu,i})$ où $\hat{\xi}_{\nu,i}$ correspond à l'effet aléatoire individuel estimé. Cet effet aléatoire est calculé comme la moyenne empirique d'échantillons de variables latentes simulées selon leur distribution *a posteriori*, conditionnelle aux données.

La figure 4.9 montre qu'il n'est pour le moment pas possible d'obtenir des ajustements individuels satisfaisants sur toutes les variables de sortie. Cependant sur celles comme RDW et QNStor, dont la tendance de la population a été correctement captée par le paramètre φ^{est}, des ajustements individuels précis sont possibles. La figure 4.10 présente ces ajustements sur quelques individus. Dans cette figure les courbes de couleur correspondent aux sorties du modèle mécaniste ARNICA obtenues à partir des paramètres individuels estimés $(\varphi_i^{est})_{i=1,..,N}$. Les points correspondent aux données réelles, les couleurs sont associées par individu.

Il est très encourageant de constater que l'approche populationnelle permet également d'estimer les paramètres individuels. Cet aspect est important afin de modéliser spécifiquement un génotype. Avoir un outil de modélisation et de prédiction fiable permettrait de limiter

les expériences réelles, non systématiquement réalisables. De telles outils sont ensuite généralisables à d'autres plantes.

4.4.4 . Estimation du ratio de vraisemblance

Dans cette section, la procédure jointe d'inférence des paramètres du modèle statistique sous les hypothèses nulles et alternatives, et du log ratio des vraisemblances associées, présentée dans la section 3.4.2, est appliquée sur un jeu de données simulé.

Les deux hypothèses considérées correspondent à celles du test de nullité des variances des paramètres **LA_RERmax** et **R_RGRmax** i.e. sous l'hypothèse nulle $\lambda_{R\_RGRmax} = 0$, et $\lambda_{LA\_RERmax} = 0$.

Le log ratio d'intérêt est défini de la manière suivante :

$$g^* = \sum_{i=1}^{N} g_i^* = \sum_{i=1}^{N} \log \left(\frac{\int_{\mathbb{R}^{nindiv}} f_i(y_i, \xi_i; \tilde{\theta}_N) d\xi_i}{\int_{\mathbb{R}^{nindiv}} f_i(y_i, \xi_i; \hat{\theta}_N) d\xi_i} \right) \tag{4.4}$$

où $\tilde{\theta}_N$ et $\hat{\theta}_N$ sont respectivement les maxima de vraisemblance sous l'hypothèse nulle et sous l'hypothèse alternative.

La procédure jointe, directement dérivée de l'algorithme 3.4, permet de générer à chaque itération $k \geq 0$:

- un estimateur du maximum de vraisemblance sous l'hypothèse nulle $\theta_{0,k}$,
- un estimateur du maximum de vraisemblance sous l'hypothèse alternative $\theta_{1,k}$,
- un estimateur du log ratio des vraisemblances individuelles associées $(g_i^{(k)})_{i=1,...,N}$.

À chaque itération $k = 0, 1, ...$, la procédure est la suivante :

1. Obtenir $\left((\xi_{i,0}^{(k+1)})_{i=1,...N}, \theta_{0,k+1} \right)$ et $\left((\xi_{i,1}^{(k+1)})_{i=1,...N}, \theta_{1,k+1} \right)$ comme sorties d'une étape de la procédure d'inférence présentée en section 4.3.2 sous l'hypothèse nulle et sous l'hypothèse alternative. On rappelle que les variables latentes $\xi_{i,0}^{(k+1)}$ et $\xi_{i,1}^{(k+1)}$ sont simulées selon leur distribution *a posteriori*, conditionnellement aux données.
2. Pour tout $i = 1, ..., N$ simuler : $\tilde{\xi}_i^{(k+1)} \sim \text{Unif} \left\{ \xi_{i,0}^{(k+1)}; \xi_{i,1}^{(k+1)} \right\}$
3. Pour tout $i = 1, ..., N$ mettre à jour :

$$g_i^{(k+1)} = g_i^{(k)} + \frac{f_i(y_i, \tilde{\xi}_i^{(k+1)}, \theta_{0,k+1}) - e^{g_i^{(k)}} f_i(y_i, \tilde{\xi}_i^{(k+1)}, \theta_{1,k+1})}{f_i(y_i, \tilde{\xi}_i^{(k+1)}, \theta_{0,k+1}) + e^{g_i^{(k)}} f_i(y_i, \tilde{\xi}_i^{(k+1)}, \theta_{1,k+1})}$$

4. $k \leftarrow k + 1$

D'un point de vue computationnel, l'hypothèse nulle se réfère au cadre d'estimation présenté dans la section 4.4.1, où les paramètres **LA_RERmax** et **R_RGRmax** sont considérés comme des paramètres de population, c'est-à-dire sans variabilité interindividuelle ($\Omega^0_{indiv} = \{\textbf{sca}, \textbf{snu}\}$). En revanche, d'un point de vue théorique, les deux hypothèses ne sont pas définies sur des espaces latents identiques. La méthode SARIS, discutée dans le chapitre 3, traite de l'estimation de ratios de constantes de normalisation de densités de probabilité, lesquelles sont supposées être définies sur le même espace.

En pratique, si l'on fixe les paramètres **LA_RERmax** et **R_RGRmax** en tant que paramètres de population (pour des raisons de simplification computationnelle, par exemple), la dimension de l'espace des variables latentes sous l'hypothèse nulle devient $nindiv - 2$. Cela empêche l'application des points 2. et 3. évoqués précédemment. En réalité, les deux composantes latentes "manquantes", correspondant aux paramètres **LA_RERmax** et **R_RGRmax**, sont indépendantes des données. Par conséquent, leur distribution *a posteriori*, conditionnellement aux données, suit une loi normale centrée réduite.

Une solution consiste à ajuster artificiellement la dimension de $\tilde{\xi}_i^{(k)}$ afin de rendre possible le calcul des vraisemblances sous les deux hypothèses de manière cohérente :

- Si $\tilde{\xi}_i^{(k)} = \xi_{i,0}^{(k)}$, alors $\tilde{\xi}_i^{(k)} \in \mathbb{R}^{nindiv-2}$; on peut alors ajouter deux colonnes supplémentaires correspondant à des réalisations d'une loi $\mathcal{N}(0, 1)$. Cela permet de rendre la dimension des variables latentes compatible avec celle du modèle théorique et ainsi de rendre possible le calcul de la vraisemblance complète sous l'hypothèse alternative.
- Si $\tilde{\xi}_i^{(k)} = \xi_{i,1}^{(k)}$, les composantes associées aux variables latentes pour **LA_RERmax** et **R_RGRmax** doivent être retirées pour pouvoir appliquer la vraisemblance sous l'hypothèse nulle.

Il est important de noter que cette approche doit être utilisée avec précaution. En effet, la vraisemblance théorique complète, notée $f_i(y_i, \tilde{\xi}_i^{(k)}; \tilde{\theta}_N)$, calculée comme précédemment décrit, diffère de celle obtenue après modification artificielle des dimensions. Cette différence provient du fait que les espaces latents sous les deux hypothèses n'ont pas les mêmes dimensions. Pour clarifier, notons $\tilde{f}_i(y_i, \tilde{\xi}_i^{(k)}; \tilde{\theta}_N)$ la vraisemblance sous l'hypothèse nulle dans le modèle où les paramètres **LA_RERmax** et **R_RGRmax** sont traités comme des paramètres de population.

On a alors :

$$f_i(y_i, \tilde{\xi}_i^{(k)}; \tilde{\theta}_N) = \sqrt{2\pi}^{-2} \exp\left\{-\frac{\tilde{\xi}_i^1 + \tilde{\xi}_i^2}{2}\right\} \tilde{f}_i(y_i, \tilde{\xi}_i^{(k)}; \tilde{\theta}_N)$$

où $\tilde{\xi}_i^1$ et $\tilde{\xi}_i^2$ représentent les deux composantes manquantes dans les variables latentes. Leur définition est la suivante :

- Si $\tilde{\xi}_i^{(k)} = \xi_{i,1}^{(k)}$, alors $\tilde{\xi}_i^1$ et $\tilde{\xi}_i^2$ sont les deux composantes supprimées de $\xi_{i,1}^{(k)}$.
- Si $\tilde{\xi}_i^{(k)} = \xi_{i,0}^{(k)}$, alors $\tilde{\xi}_i^1$ et $\tilde{\xi}_i^2$ sont des réalisations d'une loi $\mathcal{N}(0,1)$, correspondant à la distribution *a posteriori* des variables latentes sous l'hypothèse nulle, comme expliqué plus haut.

La figure 4.11 montre 6 courbes d'estimation des log ratio individuels $(g_i^{(k)})_k$ au cours des itérations. En annexe, la figure 4.27 montre l'ensemble des trajectoires individuelles. Les paramètres d'inférences sont identiques à ceux présentés dans la section 4.4.2.

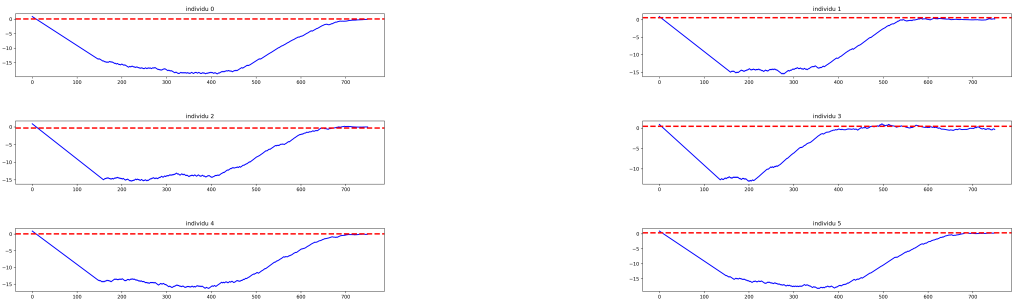


Figure 4.11 – Estimation du log ratio des vraisemblances individuelles g_i^*, pour 6 individus, sous les hypothèses H_0 et H_1, à partir de la procédure jointe d'inférence et d'estimation des paramètres dans les modèles à variables latentes. Les courbes bleues correspondent aux trajectoires d'estimation $(g_i^{(k)})_{k\geq 0}$, les pointillés rouges correspondent à une valeur estimée par échantillonnage préférentiel intensif.

Dans ces figures, les courbes bleues représentent les trajectoires individuelles $(g_i^{(k)})_{k\geq 0}$ au cours des itérations. Les pointillés rouges sont des valeurs de référence estimées par échantillonnage préférentiel intensif. Les lois de proposition utilisées sont des lois normales dont la moyenne et la variance ont été estimées sur des échantillons de la loi *a posteriori* des variables latentes, conditionnellement aux données. Pour chaque intégrale, les paramètres de la distribution de proposition ont été estimés sur 2000 échantillons de variables latentes, et 15000 échantillons ont été utilisés pour calculer les espérances empiriques.

Ces figures montrent que la procédure jointe d'inférence parvient à correctement estimer les ratios de vraisemblance individuelle sur cet exemple. Comme cela avait été expliqué dans le chapitre 3, cette approche donne une première estimation satisfaisante, à coût réduit, qui peut avoir différentes utilités. Tout d'abord elle peut servir d'initialisation à une procédure SARIS plus précise. Ensuite, dans le cas d'une procédure de test, une valeur approchée de la statistique du rapport de vraisemblance peut être suffisante pour être comparée à un seuil de référence (par exemple celui d'une région de rejet) si la différence est suffisamment conséquente. Concernant le coût computationnel de cette procédure, les avantages sont doubles. Dans un premier temps il est directement réduit, relativement à une procédure équivalente basée sur les mêmes échantillons, en profitant d'évaluations des vraisemblances et de simulations nécessaires à l'inférence des paramètres. Mais en plus, en s'intégrant dans la procédure d'inférence des paramètres, elle profite du coût temporel des boucles informatiques, ce qui réduit grandement sont temps d'exécution.

4.5 . Conclusion et perspectives

Dans ce chapitre, la problématique d'étude de la variabilité génotypique chez *Arabidopsis thaliana* est présentée. Comprendre les mécanismes porteurs de cette variabilité permettrait d'identifier des leviers pertinents pour la sélection de variétés plus performantes.

Les modèles mécanistes permettent de décrire au cours du temps des caractéristiques d'intérêt au travers de paramètres ayant un sens biologique précis. La variabilité phénotypique observée dans la population d'étude est donc également présente dans les paramètres du modèle. Identifier les paramètres variables d'un génotype à l'autre permet donc de mieux comprendre cette variabilité observée.

Le modèle ARNICA (Richard-Molard et al., 2007) décrit au cours du temps, à l'échelle d'un génotype, les échanges de carbone et d'azote entre la plante et son environnement responsables de sa croissance. Ce modèle dépend de nombreux paramètres. Afin d'identifier les paramètres porteurs de la variabilité observée, une approche populationnelle a été proposée. Cette approche consiste à intégrer le modèle mécaniste ARNICA dans un modèle à effet mixtes. La modélisation à effets mixtes permet de modéliser les paramètres du modèle mécaniste ayant

un sens biologique précis, par une composante fixe commune à l'ensemble des individus statistiques (les génotypes) de la population, et par une composante aléatoire porteuse de la variabilité génotypique.

Une procédure d'inférence de ces paramètres a été mise en place, et testée sur des données simulées, dans un modèle statistique restreint. La procédure a également été appliquée sur le jeu de données réelles, sur lesquelles les paramètres estimés semblent correctement décrire certains traits phénotypiques observés, à l'échelle populationnelle, mais également à l'échelle individuelle. Cet aspect semble prometteur quant aux capacités descriptives et prédictives futures de l'approche, lorsqu'un plus grand nombre de paramètres sera considéré.

Ce travail en cours, ouvre sur différentes perspectives. En particulier un plus grand nombre de paramètres doit être pris en compte dans le modèle statistique, afin de décrire au mieux les données observées. Cet aspect permettra d'établir des descriptions précises des traits phénotypiques mesurés, et d'obtenir de bonnes prédictions des comportements individuels. Ces résultats permettraient aux biologistes de mieux comprendre les processus étudiés, et de diminuer les coûts expérimentaux. Ensuite, la procédure d'inférence devra être améliorée afin de mieux estimer les maxima de vraisemblance, et d'intégrer de manière plus optimisée l'inférence de la statistique du rapport de vraisemblance, grâce à l'estimateur SARIS, dans l'objectif de mettre en place des tests statistiques qui permettraient d'identifier les paramètres porteurs de la variabilité génotypique.

4.6 . Annexes

Trajectoires des estimations présentées en section 4.4

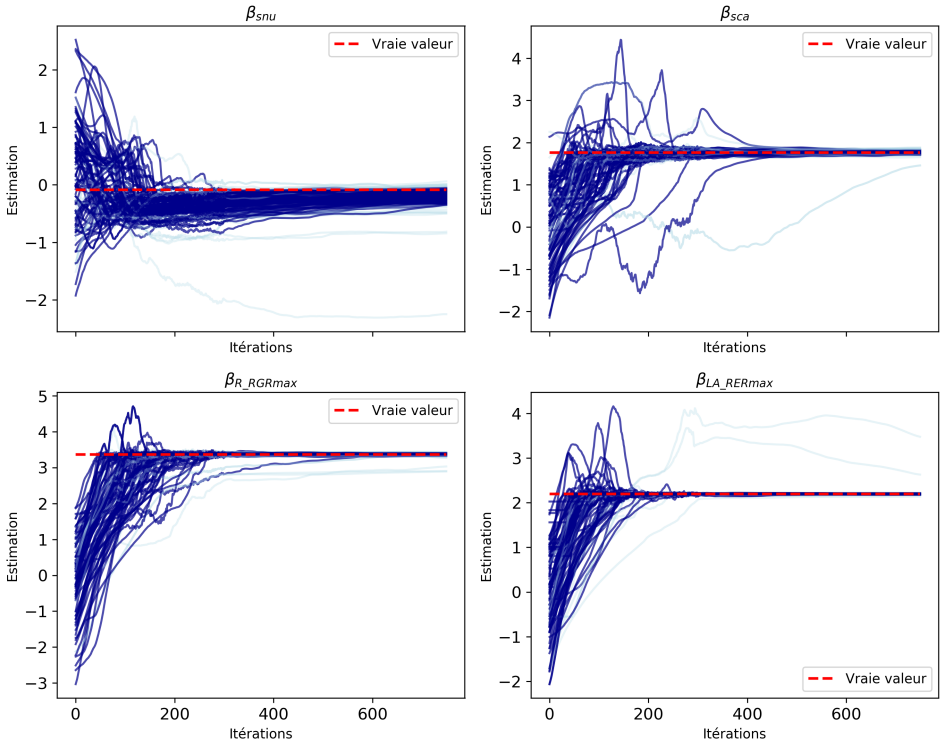


Figure 4.12 – Trajectoires complètes décrivant l'évolution de la valeur des paramètres de localisation $\beta_\nu, \nu \in \Omega_{indiv}$, pendant la procédure d'inférence, au cours des itérations. Les lignes bleu foncé correspondent aux courbes dont la valeur finale se situent en les quantiles d'ordre 10 et 90%, les courbes bleu clair correspondent au reste. La ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions.

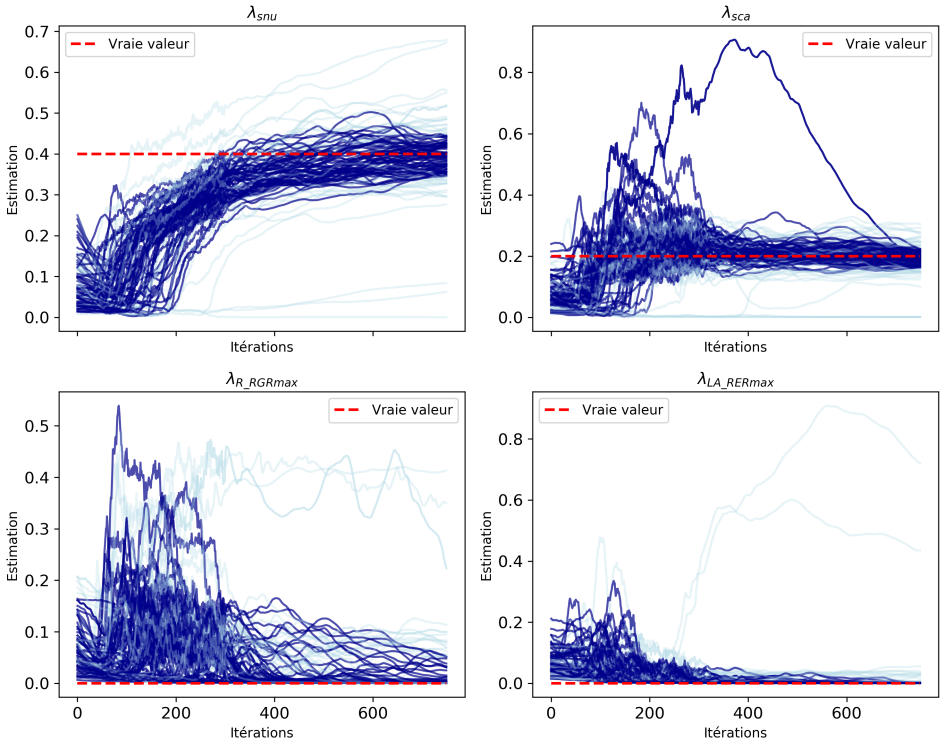


Figure 4.13 – Trajectoires complètes décrivant l'évolution de la valeur des paramètres d'échelle $\lambda_\nu, \nu \in \Omega_{indiv}$ pendant la procédure d'inférence, au cours des itérations. La ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions.

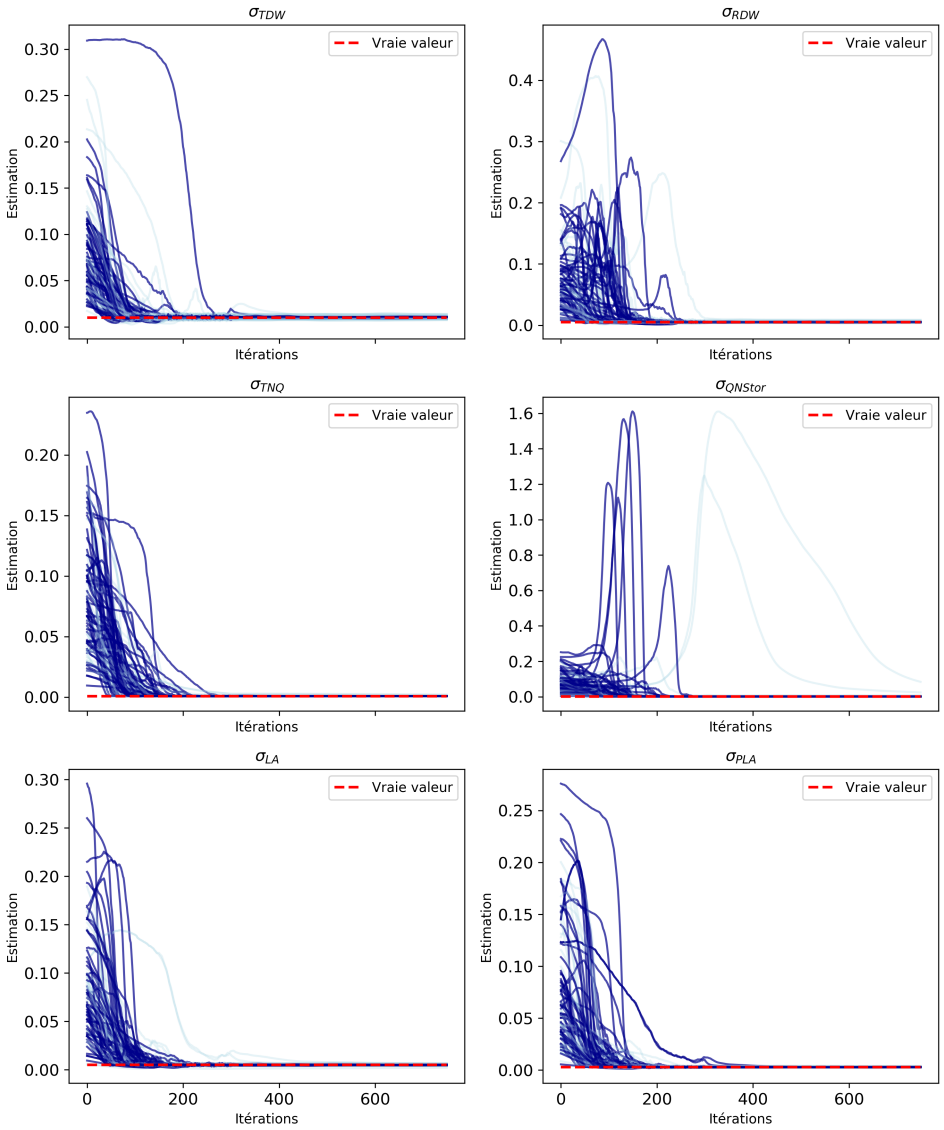


Figure 4.14 – Trajectoires complètes décrivant l'évolution de la valeur des paramètres de variance résiduelle $\sigma_\nu, \nu \in \Omega_{output}$ pendant la procédure d'inférence, au cours des itérations. Les lignes bleu foncé correspondent aux courbes dont la valeur finale se situent en les quantiles d'ordre 10 et 90%, les courbes bleu clair correspondent au reste. La ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions.

Estimation sous H_0

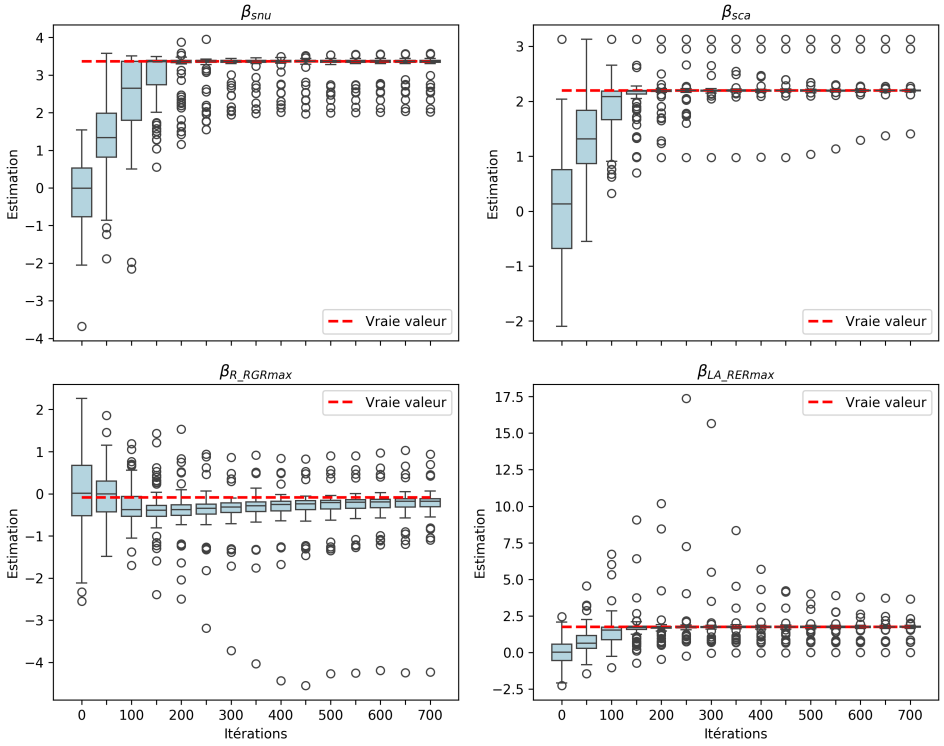


Figure 4.15 – Evolution de la valeur des paramètres de localisation $\beta_\nu, \nu \in \Omega_{indiv}$ pendant la procédure d'inférence, sous l'hypothèse nulle $\lambda_{R\_RGRmax}, \lambda_{LA\_RERmax} = 0$, au cours des itérations. Les boxplots sont affichés toutes les 50 itérations, la ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions.

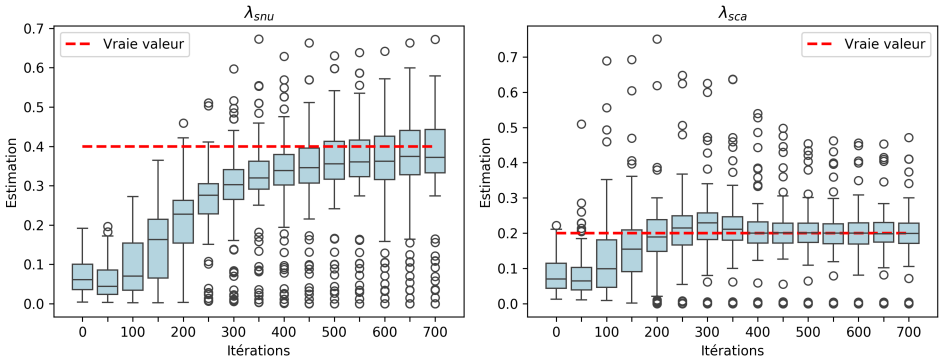


Figure 4.16 – Evolution de la valeur des paramètres d'échelle $\lambda_\nu, \nu \in \Omega_{indiv}$ pendant la procédure d'inférence, sous l'hypothèse nulle $\lambda_{R\_RGRmax}, \lambda_{LA\_RERmax} = 0$, au cours des itérations. Les boxplots sont affichés toutes les 50 itérations, la ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions.

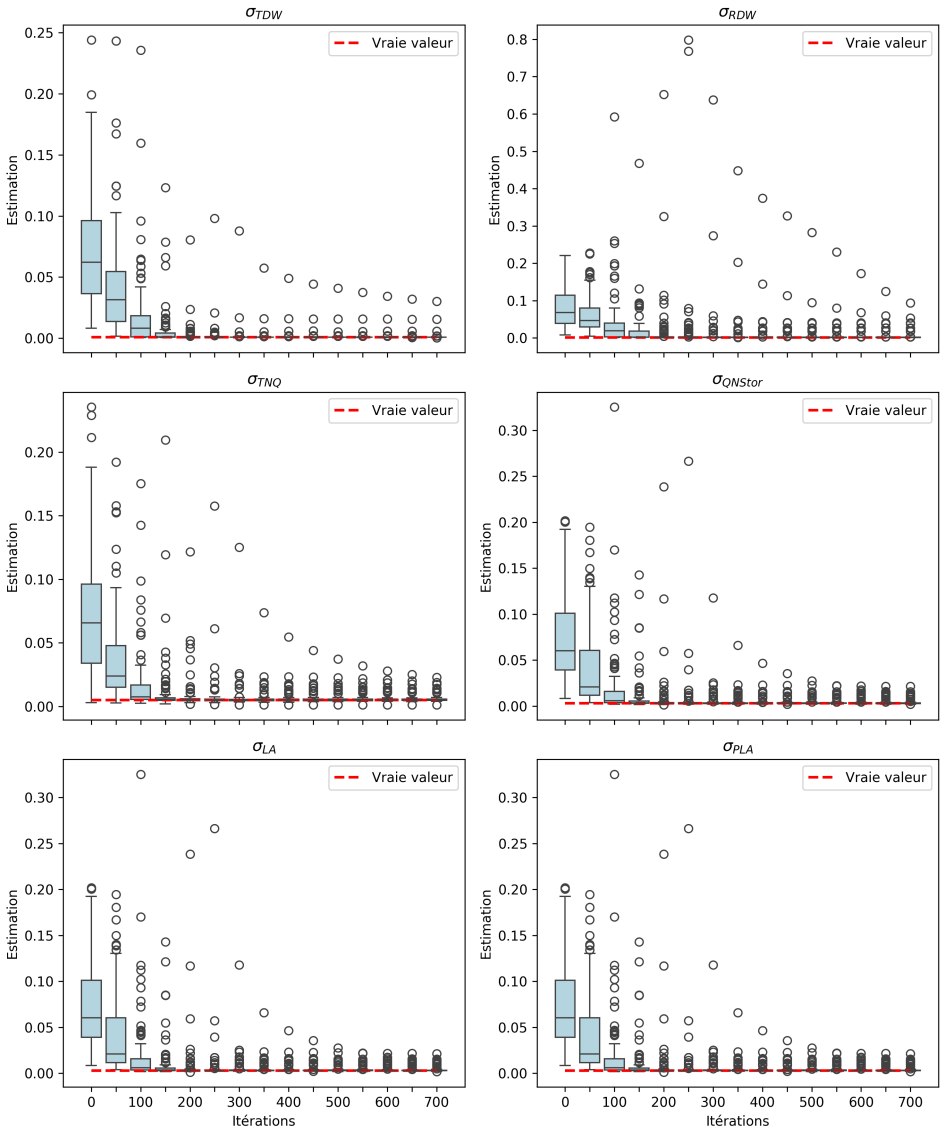


Figure 4.17 – Evolution de la valeur des paramètres de variance résiduelle $\sigma_\nu, \nu \in \Omega_{output}$ pendant la procédure d'inférence, sous l'hypothèse nulle $\lambda_{R\_RGRmax}, \lambda_{LA\_RERmax} = 0$, au cours des itérations. Les boxplots sont affichés toutes les 50 itérations, la ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions.

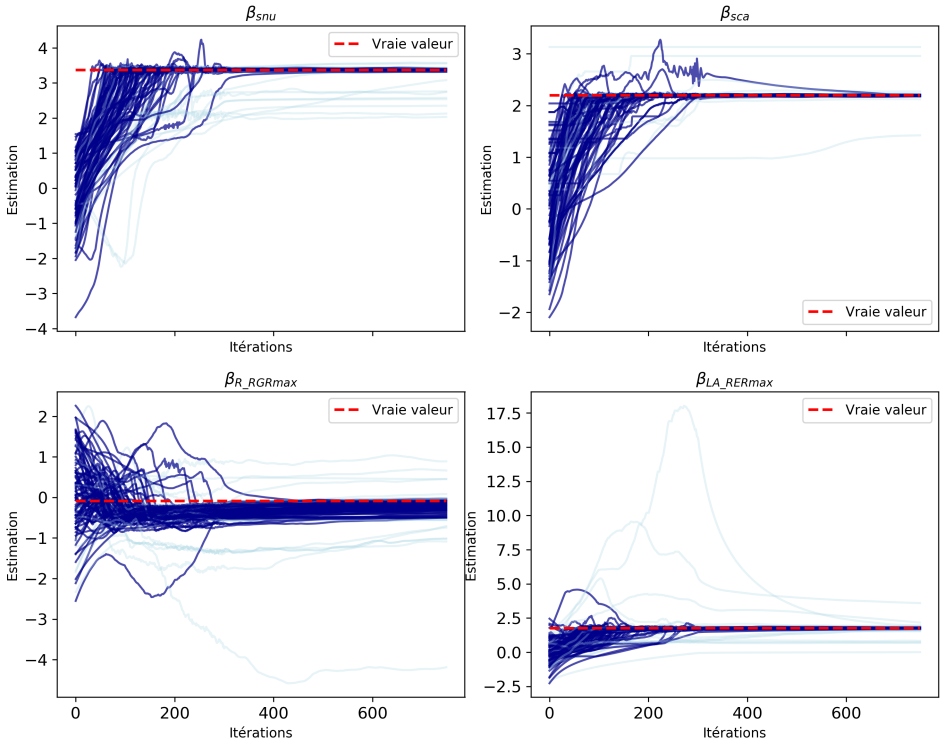


Figure 4.18 – Trajectoires complètes décrivant l'évolution de la valeur des paramètres de localisation $\beta_\nu, \nu \in \Omega_{indiv}$ pendant la procédure d'inférence, sous l'hypothèse nulle $\lambda_{R\_RGRmax}, \lambda_{LA\_RERmax} = 0$, au cours des itérations. Les lignes bleu foncé correspondent aux courbes dont la valeur finale se situent en les quantiles d'ordre 10 et 90%, les courbes bleu clair correspondent au reste. La ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions.

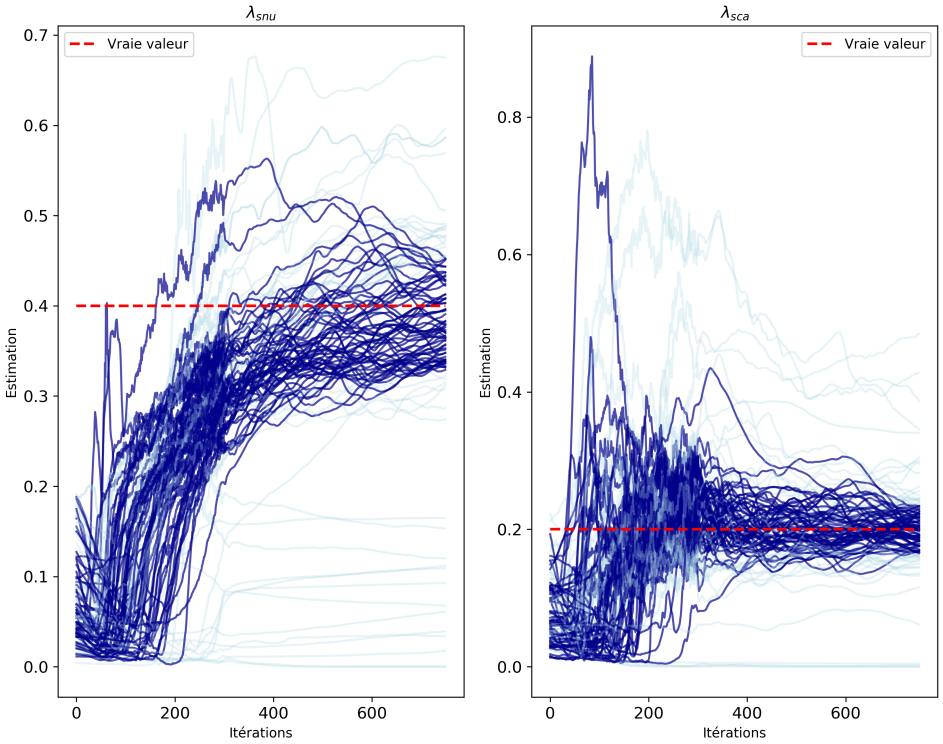


Figure 4.19 – Trajectoires complètes décrivant l'évolution de la valeur des paramètres d'échelle $\lambda_\nu, \nu \in \Omega_{indiv}$ pendant la procédure d'inférence, sous l'hypothèse nulle $\lambda_{R\_RGRmax}, \lambda_{LA\_RERmax} = 0$, au cours des itérations. Les lignes bleu foncé correspondent aux courbes dont la valeur finale se situent en les quantiles d'ordre 10 et 90%, les courbes bleu clair correspondent au reste. La ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions.

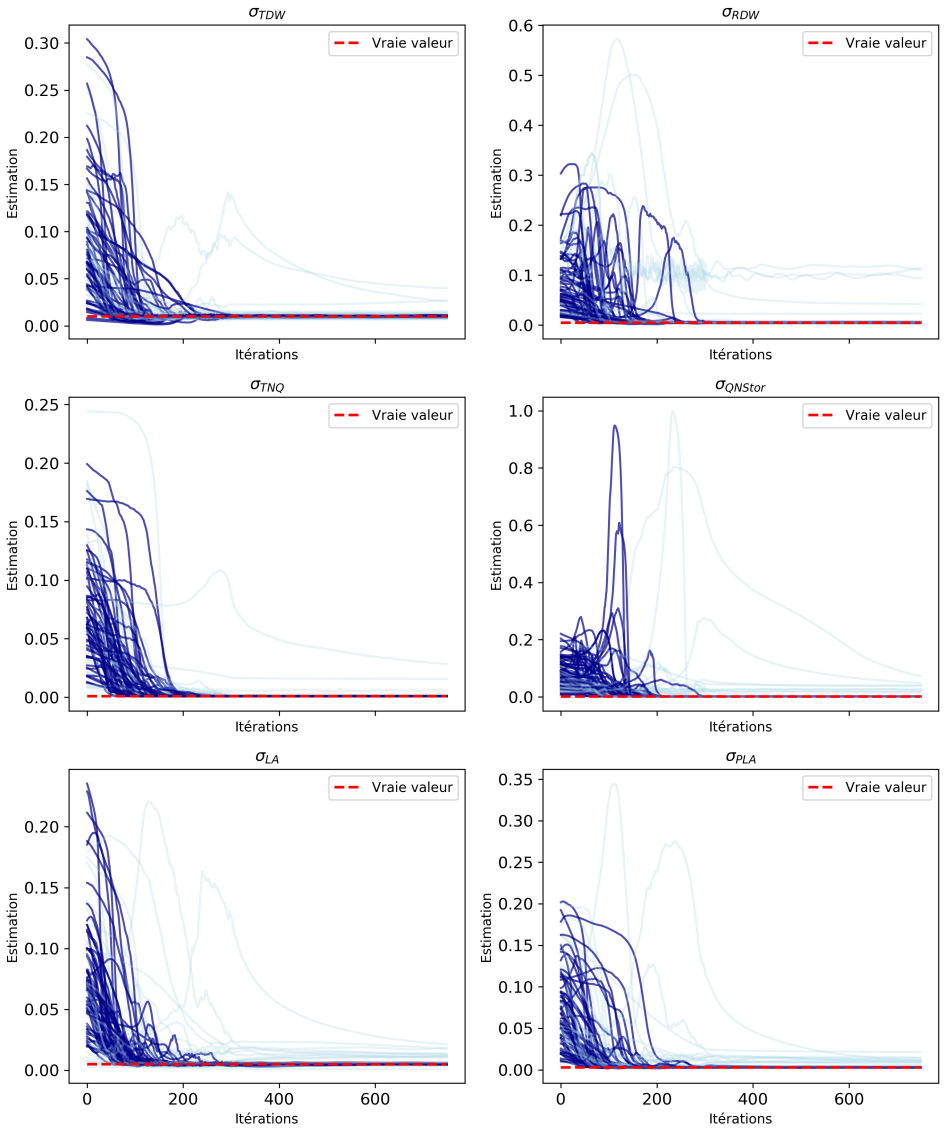


Figure 4.20 – Trajectoires complètes décrivant l'évolution de la valeur des paramètres de variance résiduelle $\sigma_\nu, \nu \in \Omega_{output}$ pendant la procédure d'inférence, sous l'hypothèse nulle $\lambda_{R\_RGRmax}, \lambda_{LA\_RERmax} = 0$, au cours des itérations. Les lignes bleu foncé correspondent aux courbes dont la valeur finale se situent en les quantiles d'ordre 10 et 90%, les courbes bleu clair correspondent au reste. La ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions.

Robustesse de l'algorithme d'estimation sur un même jeu de données

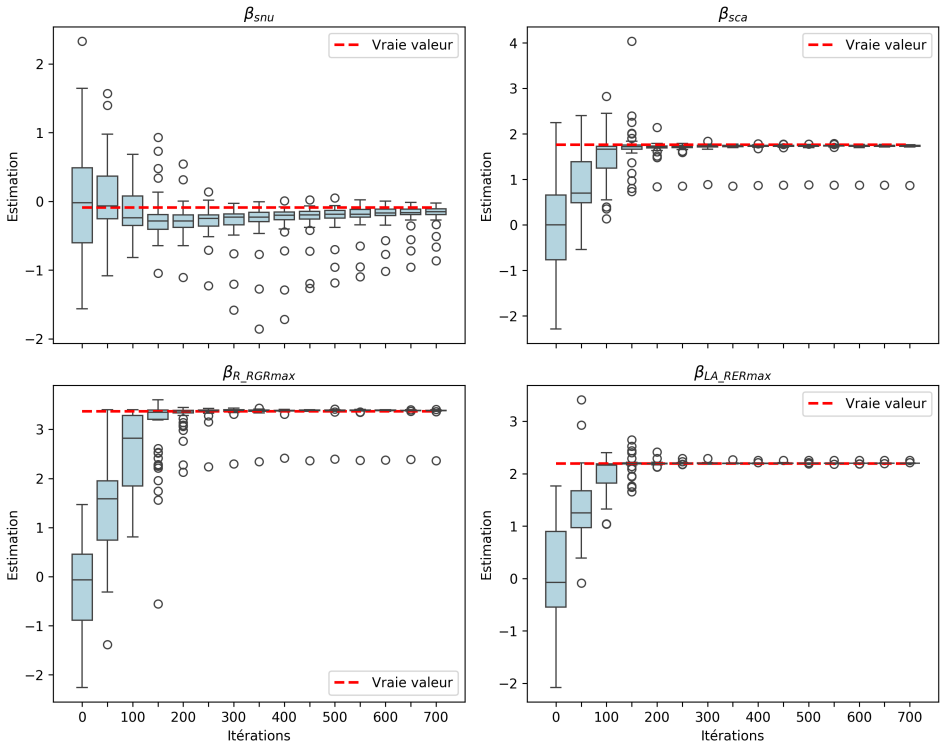


Figure 4.21 – Evolution de la valeur des paramètres de localisation $\beta_\nu, \nu \in \Omega_{indiv}$ pendant la procédure d'inférence, au cours des itérations. Les boxplots sont affichés toutes les 50 itérations, la ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions, réalisées sur un même jeu de données.

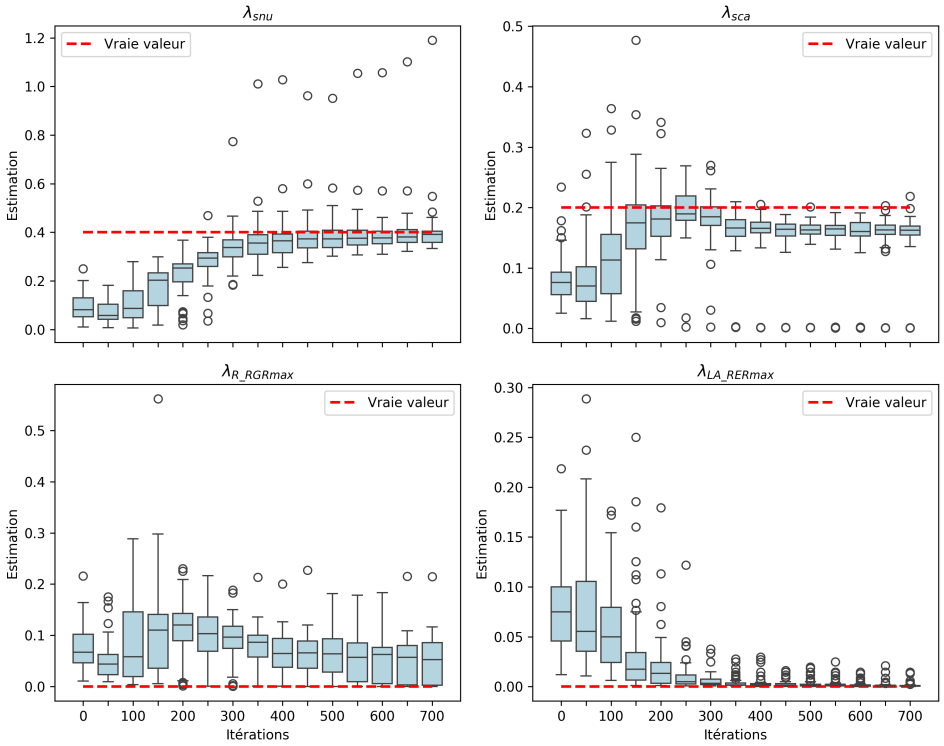


Figure 4.22 – Evolution de la valeur des paramètres d'échelle $\lambda_\nu, \nu \in \Omega_{indiv}$ pendant la procédure d'inférence, au cours des itérations. Les boxplots sont affichés toutes les 50 itérations, la ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions, réalisées sur un même jeu de données.

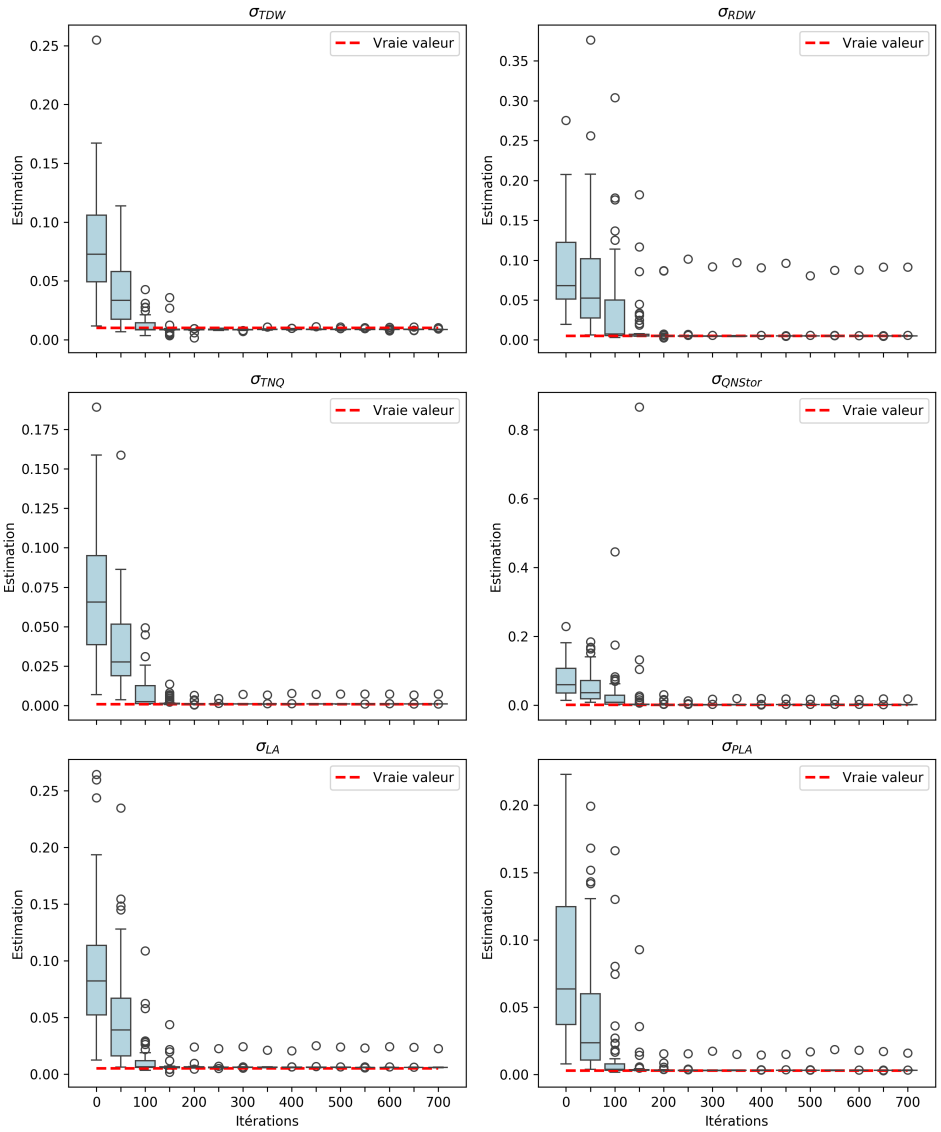


Figure 4.23 – Evolution de la valeur des paramètres de variance résiduelle $\sigma_\nu, \nu \in \Omega_{output}$ pendant la procédure d'inférence, au cours des itérations. Les boxplots sont affichés toutes les 50 itérations, la ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions, réalisées sur un même jeu de données.

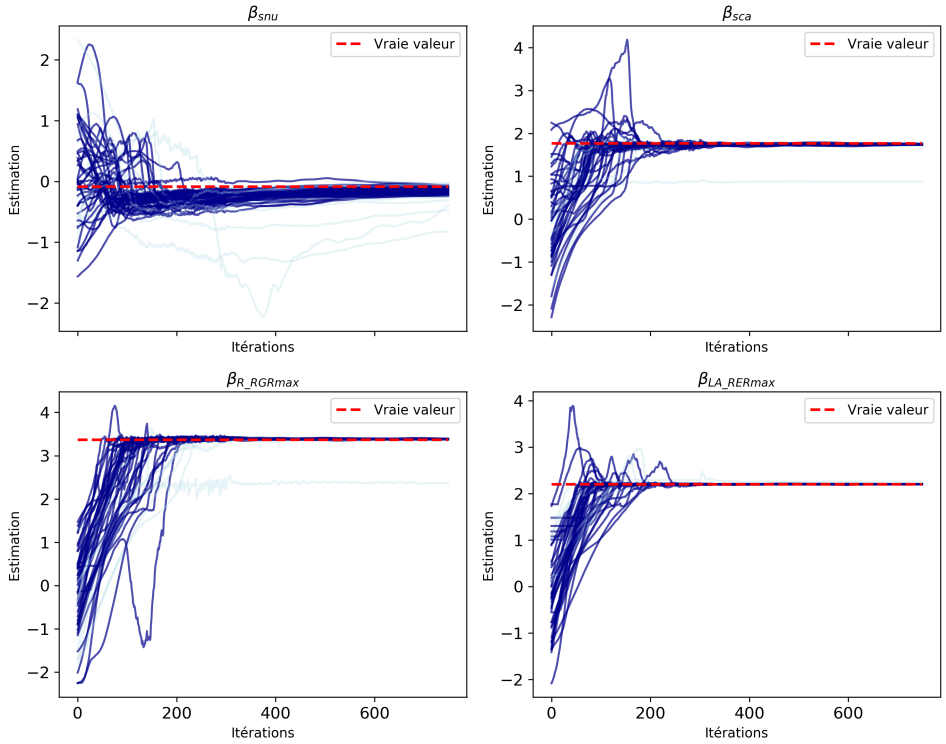


Figure 4.24 – Trajectoires complètes décrivant l'évolution de la valeur des paramètres de localisation $\beta_\nu, \nu \in \Omega_{indiv}$ pendant la procédure d'inférence, au cours des itérations. Les lignes bleu foncé correspondent aux courbes dont la valeur finale se situent en les quantiles d'ordre 10 et 90%, les courbes bleu clair correspondent au reste. La ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions, réalisées sur un même jeu de données.

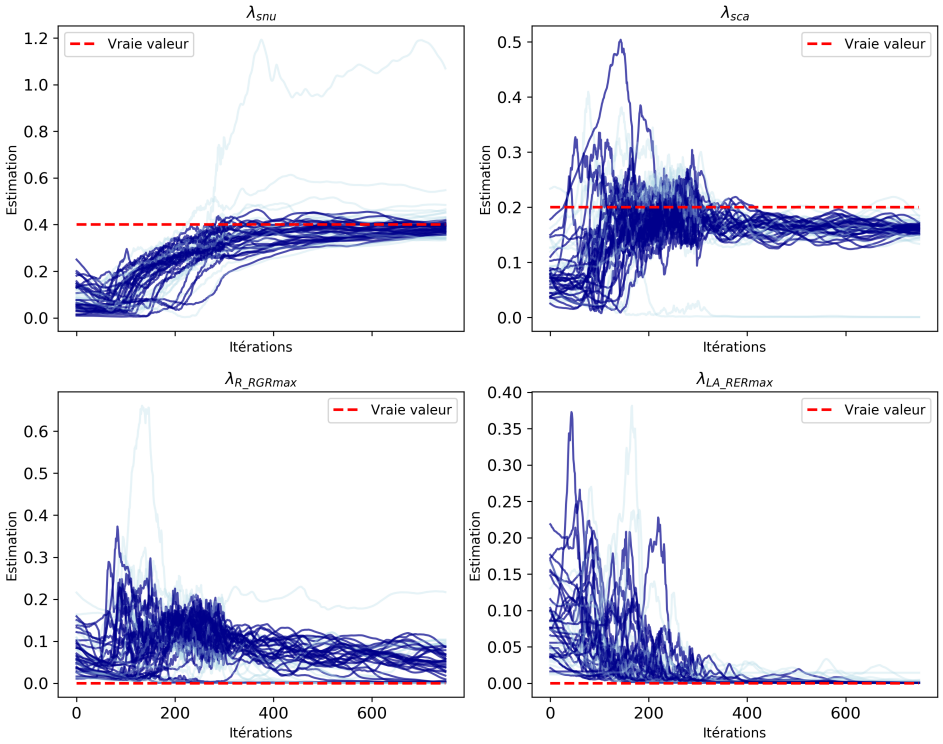


Figure 4.25 – Trajectoires complètes décrivant l'évolution de la valeur des paramètres d'échelle $\beta_\nu, \nu \in \Omega_{indiv}$ pendant la procédure d'inférence, au cours des itérations. Les lignes bleu foncé correspondent aux courbes dont la valeur finale se situent en les quantiles d'ordre 10 et 90%, les courbes bleu clair correspondent au reste. La ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions, réalisées sur un même jeu de données.

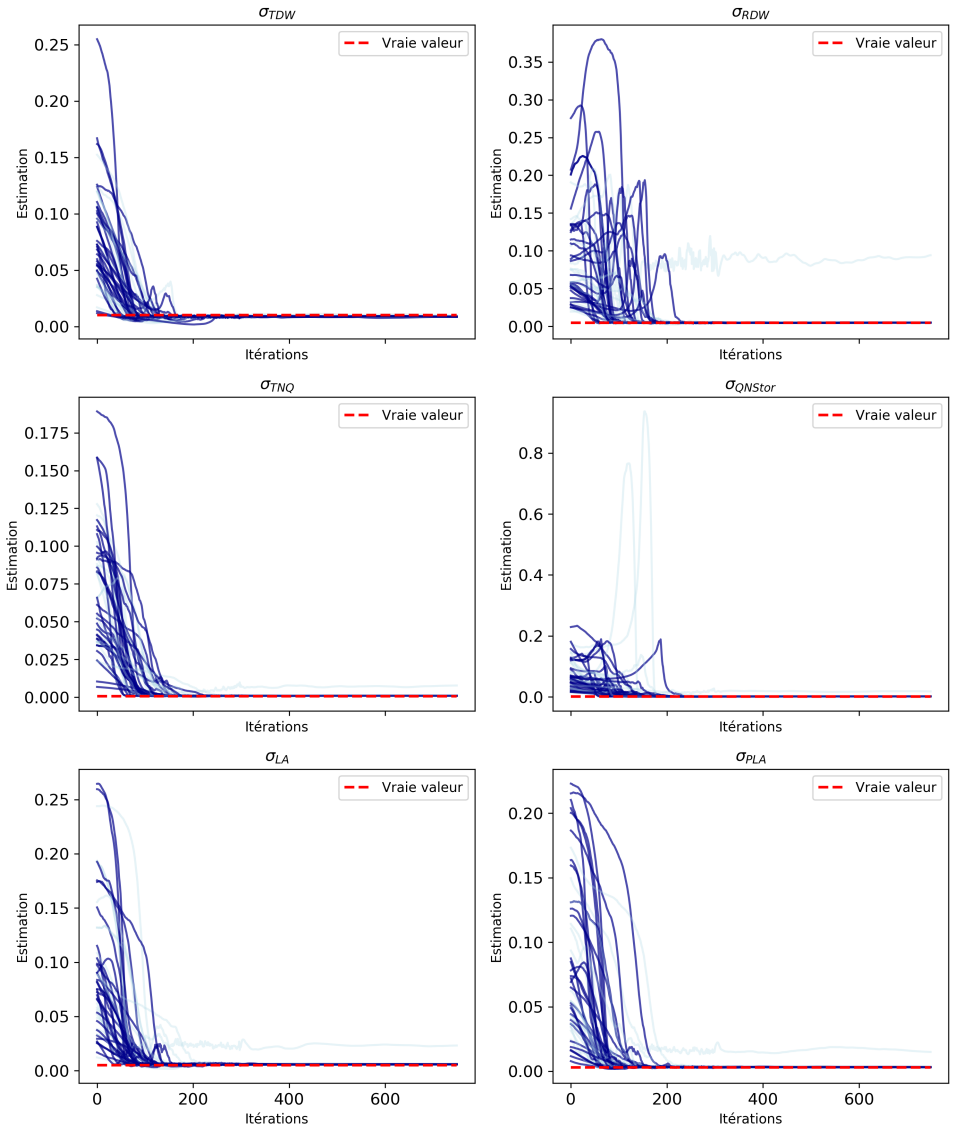


Figure 4.26 – Trajectoires complètes décrivant l'évolution de la valeur des paramètres de variance résiduelle $\sigma_\nu, \nu \in \Omega_{output}$ pendant la procédure d'inférence, au cours des itérations. Les lignes bleu foncé correspondent aux courbes dont la valeur finale se situent en les quantiles d'ordre 10 et 90%, les courbes bleu clair correspondent au reste. La ligne rouge en pointillés représente la vraie valeur utilisée pour simuler les données. Les résultats ont été obtenus sur 100 répétitions, réalisées sur un même jeu de données.

Annexes : trajectoires individuelles des 48 estimations de log ratios de vraisemblance

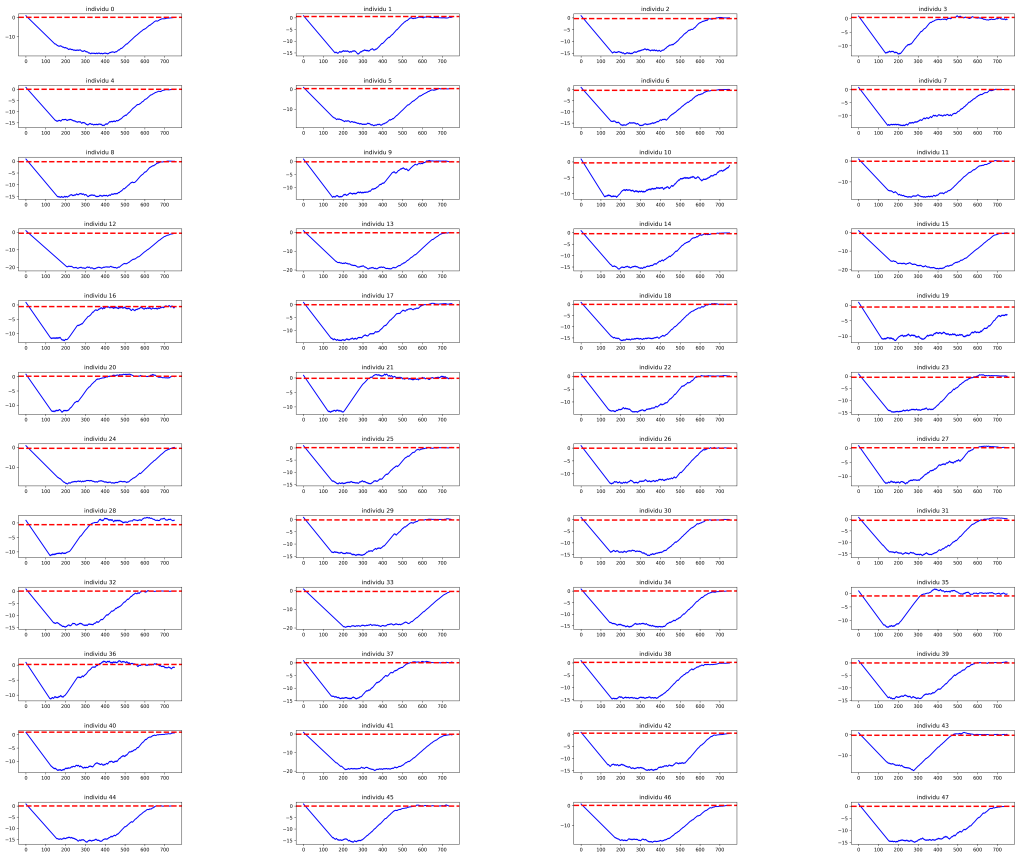


Figure 4.27 – Estimation du log ratio des vraisemblances individuelles, sous les hypothèses H_0 et H_1, à partir de la procédure jointe d'inférence et d'estimation des paramètres dans les modèles à variables latentes.

Chapitre 5

Conclusion et perspectives de recherche

5.1 . Conclusion générale

Ce travail de doctorat s'inscrit dans le projet *Stat4Plant* qui vise à développer des méthodes statistiques pour caractériser les interactions entre la plante et son environnement. Plus précisément, la problématique appliquée ayant motivé ce travail de thèse est celle de l'amélioration des plantes par sélection variétale.

La sélection variétale a pour but de sélectionner, pour une espèce donnée, les génotypes présentant les meilleurs caractéristiques phénotypiques au sein d'une population (rendement, assimilation des ressources, etc.). Cette approche est donc basée sur la variabilité de ces traits phénotypiques d'intérêt, observée au sein de la population. La modélisation mathématique permet de décrire ces traits au cours de la croissance de la plante, grâce à des paramètres ayant un sens biologique précis, tels que des paramètres d'absorption de ressources extérieures, des taux de croissance maximaux, etc. Ces paramètres permettent donc de décrire à l'échelle d'un génotype, les traits phénotypiques d'intérêt, au cours de la croissance de la plante. L'environnement considéré étant fixé, la variabilité observée devrait donc majoritairement être due à la variabilité génétique. Cette variabilité devrait donc être portée par ces paramètres biologiques. Identifier les paramètres porteurs de cette variabilité permettrait d'identifier les leviers d'action pertinents pour la sélection variétale.

L'approche classique pour étudier cette variabilité consiste à estimer les paramètres de manière individuelle, génotype par génotype, puis de les comparer. Cette méthode présente des limites liées à la quantité de données disponibles, souvent coûteuses et longues à recueillir, et aux méthodes utilisées pour l'estimation des paramètres.

L'approche considérée dans ce travail est une approche statistique, qui considère chaque génotype comme un individu statistique, sur lequel des caractéristiques sont observées au cours du temps. Les modèles à effets mixtes permettent de traiter ce type de données, appelées données longitudinales, en modélisant à la fois la variabilité intra-individuelle, décrivant les caractéristiques liées à chaque individu, et la variabilité inter-individuelle existant au sein de la population. La variabilité intra-individuelle décrit la tendance générale observée, commune à tous les individus, par l'intermédiaire d'un modèle mathématique dépendant de paramètres biologiques inconnus, comme décrit plus haut. Ces paramètres biologiques sont modélisés par

deux types d'effets : des effets fixes communs à tous les individus, et des effets aléatoires représentés par des variables aléatoires non observées, appelées variables latentes. Ces derniers modélisent la variabilité inter-individuelle. Afin d'identifier les paramètres porteurs de la variabilité observée au sein de la population, l'approche considérée est celle d'un test statistique portant sur la variance des effets aléatoires. Les principales contributions de ce travail de doctorat sont maintenant résumées.

Dans un premier temps, cette thèse considère la problématique de tester la nullité des variances dans les modèles à effets mixtes. Cette question pose deux enjeux théoriques principaux. Le premier porte sur l'aspect contraint de l'inférence, un paramètre positif égal à 0 se situe sur la frontière de l'espace des paramètres. Le second porte sur la singularité systématique de la matrice d'Information de Fisher dans ce contexte. Ces deux aspects peuvent rendre les tests asymptotiques usuels difficiles, ou même impossibles, à appliquer correctement. Afin de surmonter ces enjeux une procédure de test par Bootstrap paramétrique est proposée. Dans cette dernière, des hypothèses suffisantes sur le paramètre utilisé pour générer les données Bootstrap sont proposées, assurant que la procédure de test soit asymptotiquement du niveau souhaité. Les principales contributions de cette procédure par rapport à la littérature sont les suivantes :

- la procédure proposée s'applique indifféremment aux modèles non linéaires et aux modèles linéaires quelle que soit la forme des hypothèses testées ,
- la procédure prend en compte la présence possible de variances non testées égales à 0, i.e. la présence de paramètres de nuisance,
- la procédure présente de bonnes performances dans le cadre de petits échantillons, ce qui est souvent le cas dans les applications où les données peuvent être coûteuses en temps et/ou financièrement.

Dans un second temps, une contribution computationnelle a été réalisée. Les modèles à effets mixtes sont des modèles à variables latentes, dans lesquels la vraisemblance n'est souvent pas calculable de manière exacte. Le test proposé étant basé sur la statistique du rapport de vraisemblance, la problématique d'estimation de la vraisemblance marginale est centrale. De manière plus spécifique une nouvelle méthode appelée SARIS pour *Stochastic Approximation*

of Ratio Importance Sampling est proposée, permettant d'estimer des ratios de constantes de normalisation de densités de probabilité, tel que le rapport de vraisemblance. Cette nouvelle approche présente de très bonnes propriétés théoriques vérifiées en simulations. L'estimateur est fortement consistant et asymptotiquement Gaussien. L'algorithme proposé itératif permet une estimation en ligne, et présente donc une plus grande flexibilité que les estimateurs par Monte-Carlo usuels. Dans le cadre de l'inférence dans les modèles à variables latentes, une procédure jointe est proposée permettant d'utiliser l'effort computationnel fourni pour l'estimation des paramètres dans l'estimation de la vraisemblance marginale (ou du ratio de vraisemblance). En résumé, les contributions de ce travail sont les suivantes. L'algorithme SARIS proposé :

- permet d'obtenir un estimateur fortement consistant et asymptotiquement Gaussien du ratio de constantes de normalisation de densités de probabilité. La variance asymptotique de l'estimateur est compétitive avec la littérature et un schéma applicable atteignant cette variance est proposé,
- permet une inférence en ligne, et offre donc la possibilité de bénéficier des avantages computationnels correspondants : règles d'arrêt, économie de stockage,
- permet, de par sa définition, de s'intégrer à une procédure d'inférence dans des modèles à variables latentes, pour économiser de l'effort computationnel.

Le troisième axe de travail de ce doctorat, réalisé en collaboration avec Céline Richard-Molard, porte sur l'étude la variabilité génotypique observée chez *Arabidopsis thaliana* au cours de son processus de croissance dans un environnement faiblement azoté. L'approche considérée consiste à intégrer un modèle mécaniste déterministe existant, dans une modélisation statistique à effets mixtes, et à mettre en place une procédure d'inférence. Les contributions liées à cette étude sont les suivantes :

- la ré-implémentation complète du modèle mécaniste, et son intégration dans un modèle statistique, permettant une approche populationnelle généralisant l'approche déterministe individuelle existante,
- la mise en place d'une procédure d'inférence des paramètres du modèle, permettant d'estimer de manière jointe les paramètres biologiques, et donc de mieux comprendre les mécanismes mis en jeu lors de la croissance d'*Arabidopsis thaliana*, tout en prenant en compte la variabilité inter individuelle existant au sein de la population.

Après avoir présenté les résultats obtenus durant ce doctorat, cette section finale présente les perspectives de recherche associées aux différents travaux réalisés selon le plan du manuscrit.

5.2 . Perspectives de recherche

Perspectives sur le chapitre 2

Le chapitre 2 présente une procédure de tests de nullité des composantes de la variance dans les modèles à effets mixtes, basée sur le Bootstrap paramétrique.

Généralisation du modèle considéré :

Dans ce chapitre le modèle considéré est le suivant :

$$\begin{cases} y_{ij} & = g(x_{ij}, \beta, \Lambda\xi_i) + \varepsilon_{ij} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \xi_i & \sim \mathcal{N}(0, I_p) \end{cases},$$

Cette définition pourrait être étendue et généralisée selon les aspects suivants :

Structure hiérarchique du modèle :

Un seul niveau hiérarchique a été considéré ici. Les modèles à effets mixtes peuvent être étendus à plusieurs niveaux de variabilité. Dans le cadre de l'amélioration des plantes, un autre niveau hiérarchique pourrait être celui des conditions environnementales. Cette modélisation permettrait d'étudier en plus de la variabilité génotypique, l'interaction génotype x environnement (GxE). Si l'on indexe par k le contexte environnemental, le modèle peut se réécrire comme suit :

$$\begin{cases} y_{kij} & = g(x_{kij}, \beta, \Lambda\xi_{ik}, D\zeta_k) + \varepsilon_{kij}, \quad \varepsilon_{kij} \sim \mathcal{N}(0, \sigma^2) \\ \zeta_k & \sim \mathcal{N}(0, I_{n_k}) \\ \xi_{ik} & \sim \mathcal{N}(0, I_p) \end{cases}$$

où

- y_{kij} est la j-ème mesure de l'individu i dans la condition k

- ζ_k un effet aléatoire spécifique à la condition environnementale k
- ξ_{ik} un effet aléatoire spécifique à l'individu i dans la condition k

Homoscédasticité du modèle :

Le modèle considéré dans ce manuscrit est homoscédastique. Un modèle plus général prenant en compte le caractère potentiellement hétéroscédastique des données peut être considéré de la manière suivante :

$$\begin{cases} y_{ij} & = g(x_{ij}, \beta, \Lambda\xi_i) + h(x_{ij}, \beta, \Lambda\xi_i)\varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \\ \xi_i & \sim \mathcal{N}(0, I_p) \end{cases},$$

où la fonction h est une fonction connue. La modélisation considérée dans le chapitre 2 est un cas particulier de celle-ci.

Modélisation multivariée :

Les résultats de ce chapitre pourraient être étendus au cadre multivarié : $y_{ij} \in \mathbb{R}^q$ où $q > 1$.

Modification de la procédure Bootstrap :

La procédure Bootstrap considérée ici est entièrement paramétrique et repose donc très fortement sur les hypothèses du modèle. L'approche non paramétrique ne semble pas considérable. En effet afin de faire des tests par Bootstrap il faut imposer que l'hypothèse nulle soit vérifiée dans les données Bootstrap, ce qui n'est pas toujours aisé. Dans un cadre simple comme un test de nullité de moyenne d'un échantillon $\mathbf{X} = (X_i)_i$, on peut rééchantillonner dans $\tilde{X} = (X_i - \bar{X})_i$ qui vérifie l'hypothèse nulle Allison (2008). Cette approche est intéressante mais malheureusement semble très difficile à appliquer dans des cadres plus complexes comme celui des modèles non linéaires. Une approche permettant d'imiter la distribution sous l'hypothèse nulle est celle de Lavergne and Bertail (2020), qui consiste à modifier le critère d'estimation de l'estimateur Bootstrap. Cette modification est basée sur la statistique du score, et pose donc problème dans le cadre considéré ici, qui ne vérifie pas les hypothèses de régularité imposées, à cause de la singularité de l'Information de Fisher.

Une approche semi-paramétrique est cependant envisageable, comme souvent utilisée en régression (MacKinnon, 2006; Davidson and MacKinnon, 2006). Elle consiste à rééchantillonner

les résidus estimés. Cette approche est considérée dans Comets et al. (2021) pour construire des intervalles de confiance sur les paramètres estimés dans les modèles à effets mixtes. Il est possible de proposer une procédure de test basée sur cette approche de la manière suivante :

1. Entrées : $B \in \mathbb{N}^*$, $0 < \alpha < 1$
2. Estimer $\tilde{\theta}_N = (\tilde{\beta}_N, \tilde{\Lambda}_N, \tilde{\sigma}_N)$ l'EMV restreint
3. Estimer $(\hat{\xi}_i)_{i=1,...,N}$ les effets individuels
4. Estimer $(y_i - g(x_i, \tilde{\beta}_N, \tilde{\Lambda}_N \hat{\xi}_i))$ les résidus
5. Pour $b = 1, ..., B$
   - Pour $i = 1, ..., N$, tirer indépendamment avec remise $\varepsilon_i^{*,b}$ et $\xi_i^{*,b}$ dans les résidus et les effets individuels estimés (et correctement normalisés)
   - Construire la i-ème réponse du b-ème échantillon Bootstrap :

$$y_i^{*,b} = g(x_i, \tilde{\beta}_N, \tilde{\Lambda}_N \xi_i^{*,b}) + \varepsilon_i^{*,b}$$

   - Calculer la statistique du rapport de vraisemblance Bootstrap $\text{lrt}(y_{1:N}^{*,b})$
6. Calculer la p-valeur Bootstrap comme $p_{boot} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}_{\text{lrt}(y_{1:N}^{*,b}) > \text{lrt}(y_{1:N})}$
7. Rejeter H_0 si $p_{boot} < \alpha$

Une telle procédure permettrait de s'affranchir partiellement de l'hypothèse paramétrique faite sur les distributions des résidus et des effets aléatoires. Il serait intéressant de comparer cette procédure et celle proposée en présence de distributions d'effets aléatoires mal spécifiés (Drikvandi, 2020). Cependant il est important de mentionner que l'inférence des paramètres dans les modèles à effets mixtes repose sur l'hypothèse faite sur les distributions, ainsi une distribution mal spécifiée pollue également l'inférence pré-procédure de test. Cette approche semi-paramétrique semble également plus difficile à étudier théoriquement.

Choix du paramètre de seuillage c_N :

Le choix du paramètre de seuillage, comme présenté dans les simulations est important. Une perspective de ce chapitre est donc la mise en place d'une méthodologie pour choisir ce paramètre efficacement, de façon adaptative. Il a été montré dans les simulations qu'un choix

de c_N trop grand est très punitif en terme d'estimation de niveau empirique. Ce choix est d'autant plus sensible lorsque la variance seuillée correspond à une part importante de la variabilité des observations. Ce constat porte à considérer un choix spécifique de c_N pour chaque paramètre de variance, dépendant de la sensibilité du modèle au paramètre considéré.

Utilisation du test proposée dans une procédure de construction de modèles à effets mixtes :

Une perspective intéressante est celle de l'utilisation de la procédure de test afin de construire un modèle et d'identifier toutes les variances non nulles à prendre en compte. Une approche possible serait de réaliser des tests séquentiels, en augmentant au fur et à mesure le nombre de variances présentes dans l'hypothèse nulle. Une autre approche serait de tester successivement la nullité de chaque variance de manière indépendante. De telles procédures ne sont pas permises si la possibilité de variances non testées égales à zéro n'est pas prise en compte, ce qui est fait ici. Cependant ces approches sont délicates à étudier théoriquement, car elles posent des problèmes de tests multiples, et des problèmes de choix d'hypothèses dépendant de décisions prises à partir des données.

Perspectives sur le chapitre 3

Le chapitre 3 présente un nouvel algorithme permettant d'estimer un ratio de constante de normalisation par approximation stochastique. Les algorithmes stochastiques bénéficient d'une très riche littérature, de nombreuses perspectives s'offrent quant à l'étude de ce nouvel estimateur. Cette section présente quelques-unes de ces perspectives, et donne de premières pistes de réflexion sur l'obtention de nouveaux résultats sur la procédure SARIS.

Étude théorique de la procédure jointe

Une procédure jointe d'estimation du ratio de vraisemblance marginale et des paramètres dans les modèles à variables latentes a été proposé dans 3.4.2. Aucune garantie théorique n'a pour le moment été établie. Cette procédure permet d'estimer de manière jointe les estimateurs du maximum de vraisemblance dans 2 modèles différents, et le rapport de vraisemblance associé. En reprenant les notations de la section 3.4.2, la définition de l'estimateur du rapport de vraisemblance est rappelée :

$$r_{k+1} = r_k + \gamma_k H_k\left(\tilde{z}_{k+1}, r_k\right)$$

où $H_k\left(\tilde{z}_{k+1}, r_k\right) = \frac{f_{\theta_{0,k+1}}(y,\tilde{z}_{k+1}) - r_k f_{\theta_{1,k+1}}(y,\tilde{z}_{k+1})}{f_{\theta_{0,k+1}}(y,\tilde{z}_{k+1}) + r_k f_{\theta_{1,k+1}}(y,\tilde{z}_{k+1})}$

L'étude théorique de la convergence de cet estimateur n'est pas triviale, à cause des dépendances de l'estimateur à différentes sources d'aléas. Les densités non normalisées sont aléatoires et dépendent de $\theta_{i,k+1}$. La variable latente $\tilde{z}_{k+1}$ dépend également de ces estimateurs. Cette récursion rentre dans un cadre plus général que celui de l'étude de l'algorithme de Robbins-Monro utilisé pour l'étude de la convergence de l'algorithme SARIS. Une étude plus fine est donc nécessaire, en considérant une récursion de la forme suivante :

$$r_{k+1} = r_k + h(r_k) + \gamma_k e_k + \gamma_k u_k$$

où

- $h(r_k) = c_0 - r_k c_1$, ($h(r^*) = 0$)
- $e_k = H\left(\tilde{z}_{k+1}, r_k\right) - h(r_k)$ où $H\left(\tilde{z}_{k+1}, r_k\right)$ est l'incrément théorique inconnu si les paramètres θ_i étaient connus
- $u_k = H_k\left(\tilde{z}_{k+1}, r_k\right) - H\left(\tilde{z}_{k+1}, r_k\right)$ est l'erreur d'approximation réalisée en utilisant les estimateurs $\theta_{i,k+1}$, ($i = 0, 1$) au lieu de leurs vraies valeurs

En contrôlant les suites $(e_k)_k$ et $(u_k)_k$ on pourrait obtenir la convergence de la suite $(r_k)_k$. Ce contrôle dépend de la convergence des suites $(\theta_{0,k})_k$ et $(\theta_{i,k})_k$.

Amélioration de l'estimateur SARIS-MIXT :

L'estimateur SARIS-MIXT proposé en section 1.1.2, utilisant le mélange entre les deux distributions p_0 et p_1 comme distribution d'échantillonnage préférentiel est important pour deux raisons. Tout d'abord, il ne requiert que des échantillons issus de p_0 et p_1, qui sont nécessaires pour l'inférence dans les modèles à variables latentes et pour l'inférence bayésienne. Ensuite, il est utilisé dans la procédure jointe. Cet estimateur est une version en ligne de l'estimateur RIS-BRIDGE r_K^{ris} défini comme la solution de l'équation suivante, en r :

$$\sum_{k+1}^{K} \frac{f_0(Z_k) - r f_1(Z_k)}{f_0(Z_k) + r f_1(Z_k)} = 0, \quad Z_1, ..., Z_K \sim \frac{1}{2}(p_0 + p_1) \tag{5.1}$$

qui est la version empirique de l'identité SARIS. Comme montré dans le théorème 3.5, la variance asymptotique de cette estimateur et celle de l'estimateur SARIS-MIXT sont identiques. Cependant les simulations semblent indiquer que l'estimateur RIS-BRIDGE a de meilleures performances. Cette différence peut venir du fait que pour tout K, l'estimateur RIS impose que l'équation (5.1) soit vérifiée. Cette différence pourrait être diminuée en considérant un pas adaptatif visant à imiter la différence entre r_K^{ris} et r_{K+1}^{ris}.

Soit S_K la fonction définie sur $\mathbb{R}_+$ comme :

$$S_K(r) = \sum_{k=1}^{K} \frac{f_0(Z_k) - r f_1(Z_k)}{f_0(Z_k) + r f_1(Z_k)}$$

Par définition, $S_K(r_K^{ris}) = 0$ et :

$$S_{K+1}(r) = S_K(r) + \frac{f_0(Z_{K+1}) - r f_1(Z_{K+1})}{f_0(Z_{K+1}) + r f_1(Z_{K+1})}$$

et par un développement du premier ordre, en écrivant $H_{K+1}(r) = \frac{f_0(Z_{K+1}) - r f_1(Z_{K+1})}{f_0(Z_{K+1}) + r f_1(Z_{K+1})}$ nous avons :

$$S_{K+1}(r) \approx H_{K+1}(r_K^{ris}) + (r - r_K^{ris}) H'_{K+1}(r_K^{ris})$$

Par définition de $S_{K+1}(\hat{r}_{K+1}^{ris})$ nous avons l'approximation du premier ordre suivante :

$$\hat{r}_{K+1}^{ris} - r_K^{ris} \approx -\frac{H_{K+1}(r_K^{ris})}{H'_{K+1}(r_K^{ris})}$$

qui suggère le pas adaptatif suivant pour l'algorithme SARIS-MIXT :

$$r_{k+1} = r_k - \frac{H_{\pi^{mixt}}(Z_{k+1}, r_k)}{H'_{K+1}(r_k)}, \qquad Z_{k+1} \sim \frac{1}{2}(p_0 + p_1)$$

Cette procédure de type Newton est un autre exemple de la flexibilité de la méthodologie proposée, qui sera d'intérêt pour de futures recherches.

Au-delà de l'algorithme SARIS-MIXT (algorithme 3.3), trouver des manières d'utiliser directement les distributions p_0 et p_1 dans une procédure SARIS-EXT serait de grand intérêt, afin de profiter à la fois des réductions de variances asymptotiques et des facilités pratiques que procurent l'utilisation de ces distributions. De plus, de telles distributions permettraient d'être

utilisées dans la procédure jointe.

Autres raffinements :

D'autres perspectives basées à la fois sur la littérature de l'approximation stochastique et celle de l'estimation de vraisemblances marginales (et plus généralement de ratios de constante de normalisation) sont également à considérer.

D'un point de vue théorique, les résultats présentés ici concernent le cadre spécifique de l'algorithme de Robbins-Monro où l'échantillonnage est exact. Dans la plupart des cas pratiques, cela est impossible et il est nécessaire de recourir à l'utilisation de noyaux de transition. La consistance et la normalité asymptotique pourraient toujours être obtenues dans ce cas là (Kuhn and Lavielle, 2004; Fort, 2015) moyennant des hypothèses de régularité supplémentaires sur le noyau de transition considéré. Cependant, les interprétations de ces résultats théoriques sont délicates, car ils dépendent de quantités inconnues spécifiques aux noyaux de transitions utilisés.

Enfin, des résultats non asymptotiques pourraient être étudiés (Moulines and Bach, 2011), ce qui permettrait par exemple d'obtenir des intervalles de confiance non asymptotiques de l'estimateur.

Dans le cadre de l'estimation d'une unique constante de normalisation (cf remarque 3.4), différentes pistes sont à étudier. Par exemple le *Warp bridge sampling* (Meng and Schilling, 2002; Wang et al., 2022), est un raffinement du Bridge sampling qui modifie à la fois l'échantillon utilisé et la densité de proposition normalisée considérée afin de minimiser l'overlap entre la densité d'intérêt et celle choisie. L'aspect itératif de l'algorithme SARIS pourrait permettre d'apprendre, lors de l'estimation, une densité paramétrique proche de la densité considérée, afin de réduire la variance de l'estimation de la constante de normalisation.

Enfin, de nombreuses autres méthodes permettant de calculer des ratios de constantes de normalisation existent, qui introduisent des distributions intermédiaires entre les distributions p_0 et p_1. Parmi ces méthodes nous pouvons citer le path sampling (Gelman and Meng, 1994),

l'annealed importance sampling (Neal, 2001), l'intégration thermodynamique (Lartillot and Philippe, 2006) et le stepping stone sampling (Xie et al., 2011). Cette dernière approche basée sur l'identité $r^* = \mathbb{E}_1\left(\frac{f_0(Z)}{f_1(Z)}\right)$ décompose ce ratio, en introduisant M distributions intermédiaires non normalisées $f_{\beta_1}, ..., f_{\beta_M}$ ($0 = \beta_1 < ... < \beta_M = 1$), correspondant chacune à une constante de normalisation c_{β_i} :

$$r^* = \frac{c_0}{c_1} = \prod_{i=1}^{M-1} \frac{c_{\beta_i}}{c_{\beta_{i+1}}} = \prod_{i=1}^{M-1} r_i^*$$

Chacun de ces $M-1$ ratios peut être estimé par une des différentes approches présentées dans ce manuscrit. L'avantage étant que l'overlap entre f_{β_i} et $f_{\beta_{i+1}}$ est censé être faible. Même si le coût computationnel de ces approches est bien plus important ($M-1$ estimations à réaliser), elles présentent de très bonnes performances. De plus, simuler selon plusieurs distributions très proches les unes des autres peut être réaliser à l'aide de la méthode du *parallel tempering* de Geyer (1991).

Perspectives sur le chapitre 4

Le chapitre 4 intègre un modèle mécaniste permettant de décrire différents traits phénotypiques d'une plante au cours de sa croissance, dans un modèle statistique à effets mixtes.

Enrichissement du modèle statistique

La perspective la plus évidente est celle d'augmenter le nombre de paramètres du modèle mécaniste intégré dans le modèle statistique. Cette étape est fondamentale pour décrire précisément les traits phénotypiques à l'échelle de la population et également à l'échelle individuelle. Cet aspect sous-entend donc une optimisation plus lourde et plus complexe à mettre en place. En effet le modèle ARNICA est un modèle complexe, dont les sorties décrites de manière implicites, entraînent des compensations fortes entre les paramètres, rendant l'inférence difficile.

Réponse à la question biologique

En continuité du point précédent, sous réserve d'une inférence précise, la suite logique de

ce travail est celle de l'application du test décrit au chapitre 2, afin de comparer différentes hypothèses émises par les biologistes quant aux paramètres porteurs de la variabilité génotypique observée. La principale limite actuelle est celle de l'optimisation précise du maximum de vraisemblance, en particulier concernant les variances des effets aléatoires. En effet le modèle semble peu sensible à ces paramètres ce qui rend leur inférence plus complexe. Dans un second temps, afin d'appliquer la procédure Bootstrap il sera nécessaire de calculer de nombreuses statistiques du rapport de vraisemblance, aspect rendu encore complexe à cause du poids computationnel lié à l'évaluation répétée du modèle ARNICA. La procédure jointe d'estimation des paramètres et de la statistique du rapport de vraisemblance proposée dans le chapitre 3 pourrait permettre de réduire ces coûts.

Bibliographie

Allassonniere, S. and Kuhn, E. (2015). Convergent stochastic expectation maximization algorithm with efficient sampling in high dimension. application to deformable template model estimation. *Computational Statistics & Data Analysis*, 91 :4–19.

Allison, J. S. (2008). *Bootstrap-based hypothesis testing*. PhD thesis, North-West University.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica : Journal of the Econometric Society*, pages 821–856.

Andrews, D. W. (1999). Estimation when a parameter is on a boundary. *Econometrica*, 67(6) :1341–1383.

Andrews, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, pages 399–405.

Annis, J., Evans, N. J., Miller, B. J., and Palmeri, T. J. (2019). Thermodynamic integration and steppingstone sampling methods for estimating bayes factors : A tutorial. *Journal of mathematical psychology*, 89 :67–86.

Baey, C., Cournède, P.-H., and Kuhn, E. (2019). Asymptotic distribution of likelihood ratio test statistics for variance components in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 135 :107–122.

Baey, C., Delattre, M., Kuhn, E., Leger, J.-B., and Lemler, S. (2023). Efficient preconditioned stochastic gradient descent for estimation in latent variable models. *Proceedings of the 40th International Conference on Machine Learning*.

Baey, C. and Kuhn, E. (2020). vartestnlme : an r package for variance components testing in linear and nonlinear mixed-effects models. *arXiv preprint arXiv :2007.04791*.

Bennett, C. H. (1976). Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, 22(2) :245–268.

Benveniste, A., Métivier, M., and Priouret, P. (2012). *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science, & Business Media.

Beran, R. (1997). Diagnosing bootstrap success. *Annals of the Institute of Statistical Mathematics*, 49(1) :1–24.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The annals of statistics*, pages 1196–1217.

Bickel, P. J. and Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, pages 967–985.

Boeckmann, A., Sheiner, L., and Beal, S. (1994). Nonmem users guide. *San Francisco : University of California San Francisco*.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models : a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3) :127–135.

Bonate, P. L. (2011). Nonlinear mixed effects models : theory. *Pharmacokinetic-pharmacodynamic modeling and simulation*, pages 233–301.

Brekelmans, R., Masrani, V., Bui, T., Wood, F., Galstyan, A., Steeg, G. V., and Nielsen, F. (2020). Annealed importance sampling with q-paths. *arXiv preprint arXiv :2012.07823*.

Brown, H. and Prescott, R. (2015). *Applied mixed models in medicine*. John Wiley & Sons.

Bulinski, A. V. (2017). Conditional central limit theorem. *Theory of Probability & Its Applications*, 61(4) :613–631.

Cavaliere, G., Nielsen, H. B., Pedersen, R. S., and Rahbek, A. (2020). Bootstrap inference on the boundary of the parameter space, with application to conditional volatility models. *Journal of Econometrics*.

Chant, D. (1974). On Asymptotic Tests of Composite Hypotheses in Nonstandard Conditions. *Biometrika*, 61(2) :291–298.

Chen, M.-H. and Shao, Q.-M. (1997a). Estimating ratios of normalizing constants for densities with different dimensions. *Statistica Sinica*, pages 607–630.

Chen, M.-H. and Shao, Q.-M. (1997b). On monte carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25(4) :1563–1594.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, 59(4) :762–769.

Chernoff, H. (1954). On the Distribution of the Likelihood Ratio. *The Annals of Mathematical Statistics*, 25(3) :573–578.

Comets, E., Lavenu, A., and Lavielle, M. (2017). Parameter estimation in nonlinear mixed effect models using saemix, an r implementation of the saem algorithm. *Journal of Statistical Software*, 80(1) :1–41.

Comets, E., Rodrigues, C., Jullien, V., and Ursino, M. (2021). Conditional non-parametric bootstrap for non-linear mixed effect models. *Pharmaceutical Research*, 38 :1057–1066.

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests for goodness-of-fit of a nonlinear regression model. *Journal of Multivariate Analysis*, 91(1) :35–52.

Davidian, M. and Giltinan, D. M. (2003). Nonlinear models for repeated measurement data : an overview and update. *Journal of agricultural, biological, and environmental statistics*, 8 :387–419.

Davidian, M. and Giltinan, D. M. (2017). *Nonlinear models for repeated measurement data*. Routledge.

Davidson, R. and MacKinnon, J. G. (2006). The power of bootstrap and asymptotic tests. *Journal of Econometrics*, 133(2) :421–441.

Debavelaere, V. and Allassonnière, S. (2021). On the curved exponential family in the stochastic approximation expectation maximization algorithm. *ESAIM : Probability and Statistics*, 25 :408–432.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 68(3) :411–436.

Delattre, M. and Kuhn, E. (2019). Estimating fisher information matrix in latent variable models based on the score function. *arXiv preprint arXiv :1909.06094*.

Delattre, M., Lavielle, M., and Poursat, M.-A. (2014). A note on BIC in mixed-effects models. *Electronic Journal of Statistics*, 8(1) :456 – 475.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pages 94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society : series B (methodological)*, 39(1) :1–22.

Drikvandi, R. (2020). Nonlinear mixed-effects models with misspecified random-effects distribution. *Pharmaceutical statistics*, 19(3) :187–201.

Drikvandi, R., Verbeke, G., Khodadadi, A., and Partovi Nia, V. (2013a). Testing multiple variance components in linear mixed-effects models. *Biostatistics*, 14(1) :144–159.

Drikvandi, R., Verbeke, G., Khodadadi, A., and Partovi Nia, V. (2013b). Testing multiple variance components in linear mixed-effects models. *Biostatistics*, 14(1) :144–159.

Duflo, M. (1996). *Algorithmes stochastiques*, volume 23. Springer, .

Efron, B. (1992). Bootstrap methods : another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer.

Ekvall, K. O. and Bottai, M. (2021). Confidence regions near singular information and boundary points with applications to mixed models. *arXiv preprint arXiv :2103.10236*.

Fan, Y., Wu, R., Chen, M.-H., Kuo, L., and Lewis, P. O. (2011). Choosing among partition models in bayesian phylogenetics. *Molecular biology and evolution*, 28(1) :523–532.

Fort, G. (2015). Central limit theorems for stochastic approximation with controlled markov chain dynamics. *ESAIM : Probability and Statistics*, 19 :60–80.

Friel, N. and Wyse, J. (2012). Estimating the evidence–a review. *Statistica Neerlandica*, 66(3) :288–308.

Frühwirth-Schnatter, S. (2004). Estimating marginal likelihoods for mixture and markov switching models using bridge sampling techniques. *The Econometrics Journal*, 7(1) :143–167.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice : asymptotics and exact calculations. *Journal of the Royal Statistical Society : Series B (Methodological)*, 56(3) :501–514.

Gelman, A. and Meng, X. (1994). Path sampling for computing normalizing constants : identities and theory. *University of Chicago Department of Statistics Technical Report*, (377).

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants : From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185.

Geyer, C. J. (1991). Markov chain monte carlo maximum likelihood.

Geyer, C. J. (1994a). Estimating normalizing constants and reweighting mixtures.

Geyer, C. J. (1994b). On the asymptotics of constrained m-estimation. *The Annals of statistics*, pages 1993–2010.

Giné, E. and Nickl, R. (2021). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press.

Gordon, K. R. (2019). How mixed-effects modeling can advance our understanding of learning and memory and improve clinical and educational practice. *Journal of Speech, Language, and Hearing Research*, 62(3) :507–524.

Goymann, W., Safari, I., Muck, C., and Schwabl, I. (2016). Sex roles, parental care and offspring growth in two contrasting coucal species. *Royal Society Open Science*, 3(10) :160463.

Groll, A. and Tutz, G. (2014). Variable selection for generalized linear mixed models by l 1-penalized estimation. *Statistics and Computing*, 24 :137–154.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of mathematical psychology*, 81 :80–97.

Gronau, Q. F., Singmann, H., and Wagenmakers, E.-J. (2020). bridgesampling : An r package for estimating normalizing constants. *Journal of Statistical Software*, 92(10) :1–29.

Guédon, T., Baey, C., and Kuhn, E. (2024a). Bootstrap test procedure for variance components in nonlinear mixed effects models in the presence of nuisance parameters and a singular fisher information matrix. *Biometrika*, page asae025.

Guédon, T., Baey, C., and Kuhn, E. (2024b). Estimation of ratios of normalizing constants using stochastic approximation : the saris algorithm. *arXiv preprint arXiv :2408.13022*.

Gurka, M. J. (2006). Selecting the best linear mixed model under reml. *The American Statistician*, 60(1) :19–26.

Hall, P. (2013). *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.

Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, pages 757–762.

Hansen, B. (2022). *Econometrics*. Princeton University Press.

Higham, N. J. (1990). Analysis of the cholesky decomposition of a semi-definite matrix.

Hiroyuki, K., Katsumi, S., et al. (2012). Testing the number of components in finite mixture models. Technical report, Institute of Economic Research, Hitotsubashi University.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of mathematical statistics*, pages 1977–1991.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2) :495–503.

Kingma, D. P. and Welling, M. (2020). Auto-encoding variational bayes.

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of em with an mcmc procedure. *ESAIM : Probability and Statistics*, 8 :115–131.

Kuhn, E. and Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational statistics & data analysis*, 49(4) :1020–1038.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, pages 963–974.

Lartillot, N. and Philippe, H. (2006). Computing bayes factors using thermodynamic integration. *Systematic biology*, 55(2) :195–207.

Lavergne, P. and Bertail, P. (2020). Bootstrapping quasi likelihood ratio tests under misspecification.

Leger, J.-B. (2023). Parametrization cookbook : A set of bijective parametrizations for using machine learning methods in statistical inference. *arXiv preprint arXiv :2301.08297*.

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2) :309–326.

Llorente, F., Martino, L., Delgado, D., and Lopez-Santiago, J. (2023). Marginal likelihood computation for model selection and hypothesis testing : an extensive review. *SIAM Review*, 65(1) :3–58.

Loudet, O., Chaillou, S., Camilleri, C., Bouchez, D., and Daniel-Vedele, F. (2002). Bay-0 $\times$ shahdara recombinant inbred line population : a powerful tool for the genetic dissection of complex traits in arabidopsis. *Theoretical and Applied Genetics*, 104 :1173–1184.

MacKinnon, J. G. (2006). Bootstrap methods in econometrics. *Economic Record*, 82 :S2–S18.

Meng, X.-L. and Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3) :552–586.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity : a theoretical exploration. *Statistica Sinica*, pages 831–860.

Meteyard, L. and Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112 :104092.

Molenberghs, G. and Verbeke, G. (2007). Likelihood ratio, score, and wald tests in a constrained parameter space. *The American Statistician*, 61(1) :22–27.

Moran, P. (1971). The uniform consistency of maximum-likelihood estimators. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 70, pages 435–439. Cambridge University Press.

Moulines, E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and computing*, 11 :125–139.

Newton, M. A. and Raftery, A. E. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 56(1) :3–26.

Nie, L. (2006). Strong consistency of the maximum likelihood estimator in generalized linear and nonlinear mixed-effects models. *Metrika*, 63(2) :123–143.

Pinheiro, J. and Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer science & business media.

Qu, L., Guennel, T., and Marshall, S. L. (2013). Linear score tests for variance components in linear mixed models and applications to genetic association studies. *Biometrics*, 69(4) :883–892.

Richard-Molard, C., Brun, F., Laperche, A., Chelle, M., Pagès, L., and Ney, B. (2007). Modelling n nutrition impact on plant functioning and root architecture in various genotypes of arabidopsis thaliana. In *5. International Workshop*, page np. Citeseer.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer, New York, NY.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of computational and graphical statistics*, 18(2) :349–367.

Rotnitzky, A., Cox, D. R., Bottai, M., and Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli*, pages 243–284.

Russel, P. M., Meyer, R., Veitch, J., and Christensen, N. (2018). The stepping-stone sampling algorithm for calculating the evidence of gravitational wave models. *arXiv preprint arXiv :1810.04488*.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398) :605–610.

Shirts, M. R. and Chodera, J. D. (2008). Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics*, 129(12).

Silvapulle, M. J. and Sen, P. K. (2005). *Constrained statistical inference : Inequality, order and shape restrictions*. John Wiley & Sons.

Silvapulle, M. J. and Sen, P. K. (2011). *Constrained statistical inference : Order, inequality, and shape constraints*, volume 912. John Wiley & Sons.

Sinha, S. K. (2009). Bootstrap tests for variance components in generalized linear mixed models. *Canadian Journal of Statistics*, 37(2) :219–234.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50(4) :1171–1177.

Torrie, G. M. and Valleau, J. P. (1977). Nonphysical sampling distributions in monte carlo free-energy estimation : Umbrella sampling. *Journal of Computational Physics*, 23(2) :187–199.

Vaida, F. and Blanchard, S. (2005). Conditional akaike information for mixed effects models. *Corrado Lagazio, Marco Marchi (Eds)*, page 101.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wang, L., Jones, D. E., and Meng, X.-L. (2022). Warp bridge sampling : The next generation. *Journal of the American Statistical Association*, 117(538) :835–851.

Weigel, D. and Mott, R. (2009). The 1001 genomes project for arabidopsis thaliana. *Genome biology*, 10 :1–5.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1) :60 – 62.

Wood, S. N. (2013). A simple test for random effects in regression models. *Biometrika*, 100(4) :1005–1010.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2011). Improving marginal likelihood estimation for bayesian phylogenetic model selection. *Systematic biology*, 60(2) :150–160.

Zhang, D. and Lin, X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. *Random effect and latent variable model selection*, pages 19–36.

Zhou, X., Heuvelink, G. B., Kono, Y., Matsui, T., and Tanaka, T. S. (2022). Using linear mixed-effects modeling to evaluate the impact of edaphic factors on spatial variation in winter wheat grain yield in japanese consolidated paddy fields. *European Journal of Agronomy*, 133 :126447.