UNIVERSITÉ PARIS II PANTHÉON-ASSAS

Synthèse des travaux
En vue de l'obtention de l'habilitation à diriger des recherches

MICROÉCONOMÉTRIE ET STATISTIQUE BAYÉSIENNE : ESSAIS SUR L'INNOVATION, LA CROISSANCE, LA SANTÉ ET LE MARCHÉ DU TRAVAIL

Jean-Michel ETIENNE

Soutenue le 11 décembre 2020

Directeur de recherches : Professeur Georges BRESSON
Université Paris II Panthéon-Assas

Jury :

Georges BRESSON	Professeur, Université Paris II Panthéon-Assas
Guy LACROIX	Professeur, Université Laval Québec, Canada, rapporteur
Pierre MOHNEN	Professeur, Université de Maastricht, Pays-Bas, rapporteur
Patrick SEVESTRE	Professeur, Ecole d'Economie d'Aix-Marseille, rapporteur
Michael VISSER	Directeur de Recherche, CNRS, président

1 Introduction

L'économétrie est actuellement l'une des branches très actives de la recherche économique. Elle est devenue un outil indispensable en révélant des problèmes nouveaux nécessitant des développements appropriés. L'objectif de l'économétrie est multiple et varié. Il est multiple car elle cherche non seulement à tester les théories, mais aussi à mettre en évidence des relations causales entre phénomènes économiques. Il est varié car, en fonction de la complexité du phénomène étudié, sa mise en oeuvre pratique repose principalement et alternativement sur deux approches d'inférence statistique : l'inférence fréquentiste et l'inférence bayésienne.

Schématiquement, les méthodes d'inférence statistique ont connu deux grandes phases de développement. A la fin du XIXème siècle et au début du XXème siècle, Pearson, Fisher, Neyman, et Wald, entre autres, définissent les notions fondamentales de vraisemblance, de puissance des tests d'hypothèse et d'intervalle de confiance. Ensuite, à partir de la fin des années 1950, la puissance de calcul des ordinateurs et la banalisation de l'outil informatique permettent de dépasser les hypothèses traditionnelles d'indépendance et de normalité, commodes du point de vue mathématique mais souvent considérées comme simplistes [1], pour donner toute leur fécondité à des concepts plus anciens tels que l'hypothèse bayésienne. L'informatique a permis aussi l'explosion des méthodes de simulation par application des techniques de ré-échantillonnage (méthode de Monte Carlo, bootstrap, jackknife, ...).

Au cours de ces 50 dernières années, et grâce à la capacité croissante des ordinateurs, les économètres et plus particulièrement les micro-économètres ont accompli d'énormes progrès dans tous les domaines de la science économique. L'utilisation intensive des inférences fréquentiste et bayésienne a permis de développer de nouvelles techniques d'estimation et de simulation pour modéliser toutes sortes de données (séries temporelles, données en coupe, données de panel, variables qualitatives, censurées, variables de durée et de comptage, données hiérarchiques ou emboîtées, données multi-dimensionnelles, ...).

Cependant, les deux approches d'inférence statistique ont des objectifs différents quoique complémentaires. L'inférence bayésienne, qualifiée de déductive, porte sur la crédibilité d'une hypothèse et permet de combiner l'information apportée par les données avec les connaissances *a priori* dans le but d'obtenir une information *a posteriori*. L'inférence fréquentiste, qualifiée d'inductive, repose sur la loi des observations et la révision des croyances est remplacée par le test d'hypothèse. Il ne s'agit pas ici pour moi de participer aux débats pour ou contre telle ou telle approche [2] mais simplement de présenter très sommairement dans cette introduction quelques avantages et inconvénients respectifs qui ont guidés mon choix d'une inférence plutôt que d'une autre dans mes travaux [3].

1. Je pense entre autres aux travaux des statisticiens Box, Geary ou Tukey sur la statistique robuste et à la célèbre citation de Geary reprise dans Baltagi and Bresson (2015) : "*Normality is a myth ; there never was and never will be a normal distribution*" (Geary (1947) p. 241).

2. Je renvoie le lecteur aux synthèses de Royall (1997), Gelman and Shalizi (2013) et Sprenger (2016) sur les débats relatifs aux approches fréquentiste et bayésienne.

3. Je serai plus précis dans les sections suivantes lorsque j'aborderai des méthodes spécifiques telles que la méthode de Monte Carlo par Chaîne de Markov (MCMC), l'échantillonnage de Gibbs ou

L'analyse bayésienne repose sur l'hypothèse que tous les paramètres du modèle sont des quantités aléatoires et peuvent donc incorporer des connaissances antérieures, *a priori*. Cette hypothèse contraste fortement avec l'inférence fréquentiste, selon laquelle tous les paramètres sont considérés comme des quantités inconnues mais fixes. L'analyse bayésienne suit une règle simple de probabilité, la règle de Bayes, utilisée pour former la distribution jointe *a posteriori* des paramètres du modèle, proportionnelle au produit de la vraisemblance et des distributions *a priori*. Cette distribution *a posteriori* résulte donc de la mise à jour des connaissances *a priori* sur les paramètres du modèle étant données les observations et fournit des statistiques sur les paramètres telles que moyennes, médianes, quantiles, intervalle de crédibilité, ... Bien que les distributions *a posteriori* exactes ne soient connues que dans un petit nombre de cas, les distributions jointes *a posteriori*, peuvent généralement être estimées à partir de méthodes d'approximation (MCMC, méthode variationnelle, méthode de Laplace, ...) sans aucune hypothèse sur la taille de l'échantillon. Tous les tests statistiques sur les paramètres du modèle peuvent alors être exprimés sous forme d'énoncés de probabilité basés sur la distribution *a posteriori* estimée. Au contraire, l'inférence fréquentiste est basée sur les distributions d'échantillonnage des estimateurs des paramètres et fournit des estimations ponctuelles des paramètres, leurs écarts-types ainsi que leurs intervalles de confiance. Les distributions d'échantillonnage exactes sont rarement connues et sont souvent approchées par une distribution normale (lorsque la taille de l'échantillon est grande).

Pourquoi utiliser l'analyse bayésienne ou plus précisement, quand utiliser l'analyse bayésienne ou l'analyse fréquentiste ? La réponse à cette question réside principalement dans la problématique de recherche. L'analyse bayésienne répond aux questions basées sur la distribution des paramètres conditionnelle à l'échantillon observé. Au contraire, l'analyse fréquentiste répond aux questions basées sur la distribution des statistiques obtenues à partir d'échantillons hypothétiques répétitifs, qui seraient générés par le même processus qui a produit l'échantillon observé. Cependant, l'analyse bayésienne présente des avantages et des inconvénients qui vont avoir plus ou moins d'importance selon le problème à résoudre. Elle est particulièrement utile lorsqu'il n'y a pas de méthode fréquentiste disponible ou lorsque les méthodes fréquentistes existantes permettent difficilement de résoudre le problème.

L'inférence bayésienne est exacte, dans le sens où l'estimation et la prédiction sont basées sur la distribution *a posteriori* qui est, soit connue analytiquement, soit estimée numériquement. En revanche, de nombreuses procédures d'estimation fréquentistes, telles que le maximum de vraisemblance, reposent sur l'hypothèse d'une normalité asymptotique pour l'inférence.

L'inférence bayésienne fournit une interprétation directe et peut-être plus intuitive des résultats en termes de probabilités. Par exemple, dans l'approche fréquentiste, l'intervalle de confiance au niveau 95% n'est interprétable qu'en référence à l'ensemble de tous les intervalles qu'on aurait pu observer si l'expérience avait été reproduite dans les mêmes conditions. En revanche, l'intervalle de crédibilité bayésienne s'interprète comme un intervalle qui a 95% de chance de contenir le paramètre, mais cet intervalle

l'approximation variationnelle.

dépend de la loi *a priori*[4].
L'approche bayésienne respecte le principe de vraisemblance selon lequel l'information apportée par une observation de x sur θ est entièrement contenue dans la fonction de vraisemblance $f(\theta|x)$. Birnbaum (1962) a montré que ce principe de vraisemblance est équivalent à la conjonction des principes d'exhaustivité et de conditionnement[5] (voir Berger and Wolpert (1988)). L'approche bayésienne satisfait ces trois principes, contrairement à l'approche fréquentiste, car, dans le cadre bayésien, toute inférence est faite à partir de la loi *a posteriori*, et satisfait donc le principe de vraisemblance (voir Birnbaum (1962), Mayo (2010)) .
Enfin, comme je l'ai brièvement mentionné, la précision de l'estimation dans l'analyse bayésienne n'est pas limitée par la taille de l'échantillon, les méthodes de simulation bayésienne pouvant fournir un degré arbitraire de précision.
Malgré les avantages conceptuels et méthodologiques de l'approche bayésienne, son application dans la pratique est encore parfois considérée comme controversée. Il y a deux raisons principales à cela : la subjectivité présumée dans la spécification des informations *a priori* et les difficultés de calcul dans la mise en œuvre des méthodes bayésiennes. Parallèlement à l'objectivité qui provient des données, l'approche bayésienne utilise des distributions *a priori* potentiellement subjectives, différents individus pouvant spécifier différentes distributions *a priori*. Pour les partisans de l'inférence fréquentiste, les méthodes bayésiennes manquent d'objectivité et devraient être évitées. Cependant, une approche bayésienne équilibrée et fiable est possible. L'utilisation des *a priori* non informatifs[6] ou peu informatifs comme les *g-priors* de Zellner[7] est un moyen de régler la question de la subjectivité dans les modèles bayésiens.
Enfin, l'un des principaux inconvénients de l'analyse bayésienne est le coût de calcul. Généralement, l'analyse bayésienne implique des intégrales multiples qui ne peuvent être

4. Pour reprendre l'exemple de Lecoutre (2005), l'intervalle de confiance à 95% de sécurité pour un paramètre π s'interprète de la façon suivante : "95% des intervalles calculés sur l'ensemble des échantillons possibles (tous ceux qu'il est possible de tirer) contiennent la vraie valeur π". Cette interprétation est conditionnelle à π, les valeurs possibles du paramètre ne peuvent pas être probabilisées mais seules les probabilités d'échantillonnage conditionnelles à π peuvent l'être. En revanche, dans l'approche bayésienne, π est une variable aléatoire définie par sa loi *a priori* et l'intervalle de crédibilité à 95% de π est calculé à partir de la loi *a posteriori* de π. L'intervalle de crédibilité de π dépend donc de la loi *a priori*.

5. Principe d'exhaustivité : deux observations x et y donnant la même valeur d'une statistique exhaustive T doivent conduire à la même inférence sur θ. Principe de conditionnement : si deux expériences E_1 et E_2 sur le paramètre θ sont possibles et si on choisit une de ces expériences au hasard avec une probabilité p, alors l'inférence sur θ ne doit dépendre que de l'expérience choisie.

6. La règle de Jeffreys (Jeffreys (1961)) permet de définir des lois *a priori* non informatives. L'argument de Jeffreys est le suivant : l'information de Fisher $I(\theta)$ représente une mesure de la quantité d'information sur θ contenue dans l'observation. Plus $I(\theta)$ est grande, plus l'observation apporte de l'information. Il semble alors naturel de favoriser (au sens rendre plus probable) les valeurs de θ pour lesquelles $I(\theta)$ est grande ; ce qui minimise l'influence de la loi *a priori* au profit de l'observation.

7. Dans un modèle simple du type : $y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \tau^{-1}I)$, la loi *a priori* de β est $\beta \sim N(\beta_0, \Sigma_\beta)$. Au lieu de proposer une distribution de Wishart pour Σ_β^{-1}, Zellner a proposé d'utiliser l'information contenue dans les données X en posant : $\Sigma_\beta = g\tau^{-1}\left(X'X\right)^{-1}$ où g est une constante (voir Baltagi et al. (2018)).

calculées qu'à l'aide de méthodes d'intégration numérique [8] qui nécessitent des temps de calcul relativement long. Des méthodes de simulation doivent donc être envisagées pour la réalisation effective de l'inférence bayésienne des modèles multi-paramétriques complexes (voir par exemple Robert (2007), Chib (2008), Berger (2013), Chan et al. (2019)). L'utilisation de ces méthodes de simulation ne compromet pas les avantages discutés de l'approche bayésienne, mais ajoute incontestablement à la complexité de son application dans la pratique.

Dans mes travaux de recherche, réalisés avec Miriam Abdelmoula, Badi Baltagi, Marc Beltempo, Georges Bresson, Guy Lacroix, Pierre Mohnen, Mathieu Narcy, Ali Skalli, Ioannis Théodossiou et Annick Vignes, j'utilise les deux approches d'inférence statistique selon la complexité de la question étudiée et la difficulté de sa mise en oeuvre. Il n'y a pas à proprement parlé de fil directeur unique dans l'ensemble des travaux présentés ici. Après des études de Licence, Maîtrise et DEA d'économétrie à l'Université de Strasbourg durant lesquelles le Professeur François Laisney m'a initié à l'inférence bayésienne, j'ai entrepris des études de doctorat à l'Université Paris II sous la direction d'Ali Skalli (MCF-HDR) dans le cadre d'un contrat de recherche européen obtenu par l'ERMES (UMR 7181 - CNRS) dirigée par le Professeur Georges Bresson. Le thème de ma thèse portait sur les inégalités sociales de santé. Les méthodes utilisées étaient d'optique fréquentiste. Ensuite, en fonction des différents contrats de recherche auxquels j'ai participé et des différentes collaborations scientifiques que j'ai eues le privilège d'avoir (voir mon CV *supra*), les thèmes de recherche ont alterné. J'ai étudié les inégalités sur le marché du travail, la formation des prix sur le marché aux poissons, le marché immobilier parisien, l'innovation et la croissance inclusive, les dépenses de R&D dans les régions françaises, les modèles de productivité, la pollution atmosphérique et les activités humaines, les probabilités d'accidents et d'infections dans un service de néonatalogie, ... Les méthodes d'estimation ont également alterné entre les approches fréquentistes et bayésiennes selon les spécifications retenues dans ces études. Cependant, depuis les trois dernières années, l'essentiel de mes travaux de recherche a concerné des études faisant appel à l'approche bayésienne, étant donnée la complexité des spécifications retenues. Mes travaux s'inscrivent principalement dans le champ de la micro-économétrie appliquée et se répartissent de manière équilibrée entre approches fréquentiste (5 articles) et bayésienne (4 articles et une soumission). Cette dichotomie est plus logique que la classification thématique car mes travaux concernant l'innovation, la santé et la productivité totale des facteurs sont abordés alternativement sous des angles fréquentistes ou bayésiens.

Mes premières recherches d'optique fréquentiste (durant les années suivant la soutenance de thèse) ont porté sur les inégalités sur le marché du travail et sur les inégalités sociales de santé. La décomposition de l'écart de salaire entre les femmes et les hommes au sein des secteurs associatif et privé, à différents niveaux de la distribution des salaires, a été étudiée dans un modèle de régression quantile sur des données en coupe transversale durant la période $1994-2001$. Cette décomposition a été réalisée à l'aide de la méthode de décomposition de Machado and Mata (2005) qui présente l'avantage d'étudier le

8. telles que l'intégration par la méthode de Monte Carlo ou la méthode d'approximation analytique de Laplace (voir Robert (1996)).

phénomème de discrimination tout le long de la distribution des salaires. La relation entre inégalités de revenu et santé des individus pour 14 pays de l'UE sur 7 années (1994-2001) a été explorée dans un modèle probit à effets aléatoires. Les procédures d'intégration numérique pour approcher la vraisemblance ont été utilisées pour estimer le modèle. Mes recherches se sont ensuite orientées vers le processus de formation des prix sur le marché aux poissons en tenant compte des caractérisques non-marchandes de ce marché particulier (réseaux acheteurs-vendeurs et vendeurs-vendeurs). La relation entre interactions sociales et prix sur le marché aux poissons de Marseille pour 15473 transactions sur des données mensuelles (1988-1990) a été étudiée dans un modèle dynamique à l'aide de la méthode System GMM de Blundell and Bond (1998) et d'Arellano and Bover (1995). J'ai ensuite étudié le marché privatif immobilier parisien en appliquant la méthode des prix hédoniques sur 156896 transactions immobilières couvrant la période $1990-2003$ et structurées en 5 niveaux emboîtés dans le temps et spatialement. Les déterminants des prix des logements parisiens ont été étudiés dans un modèle spatial autorégressif (*spatial lag*) sur un pseudo-panel de données non cylindrées, à poids spatiaux variant dans le temps et à effets aléatoires emboîtés. Généralisant l'approche d'Antweiler (2001) au cas *spatial lag*, le modèle a été estimé par la méthode du maximum de vraisemblance. Mes derniers travaux utilisant l'approche fréquentiste concernent l'influence de la R&D et de l'innovation sur la croissance inclusive, définie comme la combinaison de la croissance du PIB par tête et celle de l'équité dans la répartition des revenus. La spécification d'un système dynamique d'équations simultanées à correction d'erreur concerne 63 pays sur la période $1990-2013$ et a été estimée par la méthode GMM d'Arellano and Bond (1991) en deux étapes.
Les travaux d'optique bayésienne ont débuté avec l'étude des déterminants des dépenses de R&D à l'aide d'un modèle "*two-part*" hiérarchique à effets aléatoires corrélés. Le modèle "*two-part*" est composé de deux équations : la décision de R&D des entreprises françaises, réparties dans 19 régions sur la période $1990-2007$, et spécifiée à l'aide d'un modèle probit et l'intensité de leur R&D, analysée dans une relation log-linéaire des dépenses de R&D, en tenant compte des effets spécifiques corrélés (firmes, régions). L'approche bayésienne a été choisie, car cette spécification est plus facile à estimer dans un cadre bayésien que dans un cadre fréquentiste. J'ai ensuite montré l'importance de l'innovation dans un modèle de productivité totale des facteurs à composantes communes inobservables. L'importance de la technologie, des infrastructures et des institutions dans l'explication des différences de productivité totale des facteurs entre 82 pays sur 19 ans ($1990-2008$) a été étudiée dans un modèle *FAR* (*Factor Augmented Regression*). Une approche bayésienne, plus aisée que l'approche multi-étapes fréquentiste, a été proposée. Les liens entre émissions de CO_2 et activités économiques pour 81 pays sur 25 années (1991-2015) ont été étudiés dans un modèle de croissance à l'aide d'estimations semi-paramétriques à coefficients aléatoires. Du fait de la complexité des liens entre émissions de CO_2 et activités économiques, et de la spécification retenue, une approche bayésienne d'approximation variationnelle à champ moyen a été choisie. Cette approche bayésienne très puissante a été également utilisée pour une estimation semi-paramétrique à coefficients aléatoires entre le taux de croissance du PIB par tête, les taux de croissance des capitaux physiques et humain, de la force de travail, d'autres

variables de contrôle ainsi que des tendances communes pour 23 pays de l'OCDE sur la période 1971-2015. Enfin, je me suis intéressé aux probabilités d'infection nosocomiale et d'accident en liens avec les surcharges de travail dans un service de néonatalogie au Canada. Pour cela, j'ai estimé des modèles logit bayésiens semi-paramétriques à coefficients aléatoires sur données de panel avec la même approche variationnelle.
Cette synthèse présente dans un premier temps mes activités de recherche antérieures, puis, dans un second temps, précise le programme de recherche que je souhaite développer dans le futur proche. Il s'agira de développements relatifs à l'économétrie bayésienne variationnelle de modèles linéaires et non linéaires semi-paramétriques sur données de panel avec trois thèmes d'application qui m'intéressent particulièrement : l'estimation de frontières de production, l'estimation de modèles logit emboîtés et l'économétrie du changement climatique.

2 Références citées

Antweiler, W. (2001). Nested random effects estimation in unbalanced panel data. *Journal of Econometrics*, 101 :295–313.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data : Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58 :277–297.

Arellano, M. and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics*, 68 :29–51.

Baltagi, B. H. and Bresson, G. (2015). Robust panel data methods and influential observations. In Baltagi, B. H., editor, *The Oxford Handbook of Panel Data*, pages 418–450. Oxford University Press.

Baltagi, B. H., Bresson, G., Chaturvedi, A., and Lacroix, G. (2018). Robust linear static panel data models using ε-contamination. *Journal of Econometrics*, 202(1) :108–123.

Berger, J. O. (2013). *Statistical Decision Theory and Bayesian Analysis*. Springer Science.

Berger, J. O. and Wolpert, R. L., editors (1988). *The Likelihood Principle : A Review, Generalizations, and Statistical Implications*. Hayward, CA : Institute of Mathematical Statistics.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of American Statistical Association*, 57 :269–306.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87 :115–143.

Chan, J., Koop, G., Poirier, D. J., and Tobias, J. L. (2019). *Bayesian Econometric Methods*. Cambridge University Press.

Chib, S. (2008). Panel data modeling and inference : a bayesian primer. In Sevestre, P. and Mátyás, L., editors, *The Econometrics of Panel Data : Fundamentals and Recent Developments in Theory and Practice*, pages 479–515. Springer.

Geary, R. C. (1947). Testing for normality. *Biometrika*, 34(3/4) :209–242.

Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1) :8–38.

Jeffreys, H. (1961). *Theory of Probability*. Oxford Classic Texts in the Physical Sciences, Oxford Unversity Press, 3rd edition.

Lecoutre, B. (2005). Et si vous étiez un bayésien qui s'ignore ? *Revue MODULAD, Monde des Utilisateurs de l'Analyse des Données*, 32 :92–105.

Machado, J. A. F. and Mata, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*, 20(4) :445–465.

Mayo, D. G. (2010). An error in the argument from conditionality and sufficiency to the likelihood principle. In Mayo, D. G. and Spanos, A., editors, *Error and Inference : Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*, pages 305–314. Cambridge University Press.

Robert, C. (1996). *Méthodes de Monte Carlo par chaînes de Markov*. Economica.

Robert, C. (2007). *The Bayesian Choice : from Decision-Theoretic Foundations to Computational Implementation*. Springer Science.

Royall, R. (1997). *Statistical Evidence : a Likelihood Paradigm*. Chapman & Hall/CRC press.

Sprenger, J. (2016). Bayesianism vs. frequentism in statistical inference. In Hàjek, A. and Hitchcock, C., editors, *Oxford Handbook of the Philosophy of Probability*, pages 382–405. Oxford University Press.