Essays on Social Networks and Time Series with Structural Breaks

Thèse

Elysee Aristide Houndetoungan

Doctorat en économique
Philosophiæ doctor (Ph. D.)

Québec, Canada

Essays on Social Networks and Time Series with Structural Breaks

Thèse

Elysée Aristide Houndetoungan

Sous la direction de:

Vincent Boucher, directeur de recherche
Bernard Fortin, codirecteur de recherche

Résumé

Cette thèse, structurée en trois (03) essais, développe de nouveaux modèles économétriques pour l'analyse des interactions sociales et des séries temporelles.

Le premier chapitre (coécrit avec le Professeur Vincent Boucher) étudie une méthode d'estimation des effets de pairs à travers les réseaux sociaux lorsque la structure du réseau n'est pas observée. Nous supposons que nous connaissons (avons une estimation convergente de) la *distribution* du réseau. Nous montrons que cette hypothèse est suffisante pour l'estimation des effets de pairs en utilisant un modèle linéaire en moyennes. Nous proposons un estimateur de variables instrumentales et un estimateur bayésien. Nous présentons et discutons des exemples importants où notre méthodologie peut être appliquée. Nous présentons également une application avec la base de données Add Health largement utilisée et qui comporte de nombreux liens non observés. Nous estimons un modèle des effets de pairs sur la réussite scolaire des élèves. Nous montrons que notre estimateur bayésien reconstruit les liens manquants et permet d'obtenir une estimation valide des effets de pairs. En particulier, nous montrons qu'ignorer les liens manquants sous-estime l'effet endogène des pairs sur la réussite scolaire.

Dans le deuxième chapitre, je présente un modèle structurel des effets de pairs dans lequel la variable dépendante est de type comptage (nombre de cigarettes fumées, fréquence des visites au restaurant, fréquence de participation aux activités). Le modèle est basé sur un jeu statique à information incomplète dans lequel, les individus interagissent à travers un réseau dirigé et sont influencés par leur croyance sur la décision de leurs pairs. Je présente des conditions suffisantes sous lesquelles l'équilibre du jeu est unique. Je montre que l'utilisation du modèle spatial autorégressif (SAR) linéaire-en-moyennes ou du modèle Tobit SAR pour estimer les effets de pairs sur des variables de comptage générées à partir du jeu sous-estime asymptotiquement les effets de pairs. Le biais d'estimation diminue lorsque la dispersion de la variable de comptage augmente. Je propose également une application empirique. J'estime les effets de pairs sur le nombre d'activités parascolaires auxquelles les étudiants sont inscrits. En contrôlant l'endogénéité du réseau, je trouve que l'augmentation du nombre d'activités dans lesquelles les amis d'un étudiant sont inscrits d'une unité implique une augmentation du nombre d'activités dans lesquelles l'étudiant est inscrit de 0,295. Je montre également que les effets de pairs sont sous-estimés à 0,150 lorsqu'on ignore la nature de comptage de la variable

dépendante.

Le troisième chapitre (coécrit avec le Professeur Arnaud Dufays et le Professeur Alain Coen) présente une approche de modélisation de séries temporelles. Les processus avec changements structurels sont une approche flexible pour modéliser des longues séries chronologiques. En considérant un modèle linéaire en moyennes, nous proposons une méthode qui relâche l'hypothèse selon laquelle une cassure structurelle dans une série temporelle implique un changement de tous les paramètres du modèle. Pour ce faire, nous estimons d'abord les dates de cassures potentielles présentées par la série, puis nous utilisons une régression pénalisée pour détecter les paramètres du modèle qui changent à chaque date de cassure. Étant donné que certains segments de la régression peuvent être courts, nous optons pour une fonction de pénalité (presque) non biaisée, appelée fonction de pénalité *seamless-L0* (SELO). Nous montrons que l'estimateur SELO détecte de manière convergente les paramètres qui varient à chaque cassure et nous proposons d'utiliser un algorithme de maximisation d'espérance de recuit déterministe (DAEM) pour traiter la multimodalité de la fonction objectif. Étant donné que la fonction de pénalité SELO dépend de deux paramètres, nous utilisons un critère pour choisir les meilleurs paramètres et par conséquent le meilleur modèle. Ce nouveau critère présente une interprétation bayésienne qui permet d'évaluer l'incertitude des paramètres ainsi que l'incertitude du modèle. Les simulations de Monte Carlo montrent que la méthode fonctionne bien pour de nombreux modèles de séries temporelles, y compris des processus hétéroscédastiques. Pour un échantillon de 14 stratégies de hedge funds (HF), utilisant un modèle de tarification basé sur l'actif, nous mettons en exergue la capacité prometteuse de notre méthode à détecter la dynamique temporelle des expositions au risque ainsi qu'à prévoir les rendements HF.

Abstract

This dissertation, composed of three (03) separate chapters, develops new econometric models for peer effects analysis and time series modelling.

The first chapter (a joint work with Professor Vicent Boucher) studies a method for estimating peer effects through social networks when researchers do not observe the network structure. We assume that researchers know (a consistent estimate of) the *distribution* of the network. We show that this assumption is sufficient for the estimation of peer effects using a linear-in-means model. We propose an instrumental variables estimator and a Bayesian estimator. We present and discuss important examples where our methodology can be applied. We also present an application with the widely used Add Health database which presents many missing links. We estimate a model of peer effects on students' academic achievement. We show that our Bayesian estimator reconstructs these missing links and leads to a valid estimate of peer effects. In particular, we show that disregarding missing links underestimates the endogenous peer effect on academic achievement.

In the second chapter, I present a structural model of peer effects in which the dependent variable is counting (Number of cigarettes smoked, frequency of restaurant visits, frequency of participation in activities). The model is based on a static game with incomplete information in which individuals interact through a directed network and are influenced by their belief over the choice of their peers. I provide sufficient conditions under which the equilibrium of the game is unique. I show that using the standard linear-in-means spatial autoregressive (SAR) model or the SAR Tobit model to estimate peer effects on counting variables generated from the game asymptotically underestimates the peer effects. The estimation bias decreases when the range of the dependent counting variable increases. I estimate peer effects on the number of extracurricular activities in which students are enrolled. I find that increasing the number of activities in which a student's friends are enrolled by one implies an increase in the number of activities in which the student is enrolled by 0.295, controlling for the endogeneity of the network. I also show that the peer effects are underestimated at 0.150 when ignoring the counting nature of the dependent variable.

The third chapter (a joint work with Professor Arnaud Dufays and Professor Alain Coen) presents an approach for time series modelling. Change-point (CP) processes are one flexible

approach to model long time series. Considering a linear-in-means models, we propose a method to relax the assumption that a break triggers a change in all the model parameters. To do so, we first estimate the potential break dates exhibited by the series and then we use a penalized likelihood approach to detect which parameters change. Because some segments in the CP regression can be small, we opt for a (nearly) unbiased penalty function, called the seamless-L0 (SELO) penalty function. We prove the consistency of the SELO estimator in detecting which parameters indeed vary over time and we suggest using a deterministic annealing expectation-maximisation (DAEM) algorithm to deal with the multimodality of the objective function. Since the SELO penalty function depends on two tuning parameters, we use a criterion to choose the best tuning parameters and as a result the best model. This new criterion exhibits a Bayesian interpretation which makes possible to assess the parameters' uncertainty as well as the model's uncertainty. Monte Carlo simulations highlight that the method works well for many time series models including heteroskedastic processes. For a sample of 14 Hedge funds (HF) strategies, using an asset based style pricing model, we shed light on the promising ability of our method to detect the time-varying dynamics of risk exposures as well as to forecast HF returns.

Contents

List of Tables

List of Figures

To my parents
To Daphne

A rational expectations equilibrium is a likelihood function. Maximize it.

Thomas J. Sargent

Acknowledgements

First and foremost, I would like to express my deep gratitude to my advisors, Professor Vincent Boucher and Professor Bernard Fortin, for their continuous support during my doctoral studies, for their guidance, patience, and encouragement. I have learned a lot from working with Professor Vincent Boucher and Professor Bernard Fortin, not only in terms of intellectual knowledge, but also how to live, work and think as a researcher. I could not have imagined having such great advisors and mentors.

I would like to thank Professor Arnaud Dufays, Professor Luc Bissonnette, Professor Maripier Isabelle, and Professor Ismael Mourifié for their great support on my research and on the job market. I could not achieve what I have at this moment without them.

Special thanks to Marion Goussé and Luc Bissonnette for accepting to be examiners for my dissertation defense. I am also grateful to Yann Bramoullé for acting as external examiner.

My acknowledgement also goes to all the faculty members of the Department of Economics at Laval University for their support and their presence at our PhD student seminars during which, I was pleased to present my research on multiple occasions. Their suggestions and criticisms are invaluable in helping me complete the three chapters in this dissertation.
I would also like to thank other staff members of the Department of Economics at Laval University for their logistical support during my doctoral studies.

Thank you to my fellow doctoral students, Guy Morel Kossivi Amouzou Agbe, Ibrahima Diallo, and David Zoundi for their constructive feedback and comments.
I thank other current and former fellow colleagues in the PhD program in economics at Laval University, Finagnon Antoine Dedewanou, Marius Sossou, Rolande Kpekou Tossou, Koffi Akakpo, Jean-Louis Bago, Gilles Koumou, Morvan Nongni Donfack, Bodel Aymele Gnintedem, Horace Gninafon, Elfried Faton, Mélissa Huguet, Ibrahima Sarr, François Seyler, James Wabenga Yango, Ichola Soulé, and Josette Gbeto for creating a positive atmosphere in the department.
I would also like to thank Marius Adom, Abdoul Haki Maoude, Firmin Ayivodji, Lucien Chaffa, Romel Degboe, Bérénice Nagbo, Josué Awonon, Sulpice Amonle, and Ismael Assani for their support while writing this dissertation.

Special thanks to Dr. Murray Hay for taking the time to proofread important parts of this dissertation.

Furthermore, I would like to express my deep and sincere gratitude to my family. I am grateful to my mother Honorine Aledji, my father Hilaire Houndetoungan, my brother Dr. Gilles David Houndetoungan, and my sisters Odette and Dorine Houndetoungan for their support from afar. Special thanks to my beloved Daphne Kamanan for her support, encouragement, and patience during the completion of this dissertation and my job market. I could not achieve what I have without her love.

I also want to thank my friends in Quebec for their support and for always being with me when I need to have fun after hard work. Especially, I thank Cyr Loïc Afoudah, Amen Tchenagni, Rodrigue Daassi, Antoine Minko, Kevin Some, Line Kinkonda, Daniel Diakite, and Jeff-Teddy Papoin.

Foreword

The three chapters of this dissertation are separate articles published or in preparation for submission to peer-reviewed scientific journals. The first chapter was written jointly with my thesis director, Vincent Boucher, an associate professor within the Department of Economics at Laval University. The second chapter is my job market paper. The third chapter is a joint work with Arnaud Dufay, a former assistant professor within the Department of Economics at Laval University, and Alain Coen, a full Professor of Finance at the Graduate School of Business of the University of Quebec in Montreal. The two first chapters, on social networks, are in preparation for submission to peer-reviewed scientific journals whereas the third chapter has recently been accepted for publication in the *Journal of Financial Econometrics*.

I have also developed two easy-to-use R packages that implement the methods developed in the two first chapters. The package `PartialNetwork`, joint with Professor Vincent Boucher, offers a routine to replicate all the results in the first chapter. The package `CDatanet` can be used to replicate all the results in the second chapter. These packages are available on GitHub and on the CRAN website.

Introduction

The design and implementation of economic policies are often based on econometric and statistical models. To be convenient, these models are expected to replicate fairly faithfully the real world. In most cases, the models must depart from simplified and strong assumptions to suit the real world, and this can be very challenging. Many econometric models become inefficient when they are built on strong theoretical assumptions which are violated in practice. To illustrate this inefficiency, let us consider the following important examples.

1. In economics of social interactions, most models assume that the econometricians observe the entire network data; that is, they observe the friends of every individual in the studied population (see Bramoullé et al., 2009). However, network data are very expensive to collect because they require to survey the entire population instead of a sample. Therefore such an assumption is not realistic when dealing with a large population. In addition, most developed and studied models are for continuous dependent variables (e.g, Lee, 2004; Lee et al., 2010), whereas survey data contain most often discrete variables. Even if the estimator of the model parameters could still be consistent in some cases when the distribution of the interest variable is misspecified, the estimation bias is not always negligible in finite samples, as well as for policy implications.
2. In time series modelling, many methods are based on the assumption that the series are generated from a non-invariant data generation process (DGP) over time. However, long time series are likely to be affected by structural breaks due to changes in the government policies, occurrence of unusual events, and economic agents' expectations (see Fryzlewicz et al., 2014). Depending on the time series, the structural breaks can change the nature of the DGP or only affect a small set of its parameters. The assumption of non-invariant DGP does not suit long times series.

This dissertation, composed of three (03) separate chapters, develops new econometrics models for estimating peer effects and analysing times series by relaxing strong assumptions often made in the literature. In doing so, it offers tools to eliminate bias that often plague estimates presented in the empirical literature. Especially, it develops a method for estimating peer effects with partial network data. It also develops a structural model of social interactions which deals with counting dependent variables. This dissertation also presents a time series

modelling approach through the linear-in-means specification by relaxing the assumption that a break triggers a change in all the model parameters.

The first chapter, *Estimating Peer Effects Using Partial Network Data*, develops a new method to estimate peer effects when the network data is not (or partially) observed. This chapter is coauthored work with Professor Vincent Boucher. Peer effects estimation is based on the assumption that the entire network data is available. However, eliciting network data is expensive (Breza et al., 2020), and since networks must be sampled completely (Chandrasekhar and Lewis, 2011), there are few existing data sets that contain detailed network information. We explore the estimation of the widely used linear-in-means model (e.g. Manski (1993), Bramoullé et al. (2009)) when the researcher observes the entire network structure. Specifically, we assume that the researcher knows the *distribution* of the network but not necessarily the network itself. An important example is when a researcher is able to estimate a network formation model using some partial information about the network structure (e.g. Breza et al. (2020)). We present an instrumental variable estimator and show that we can adapt the strategy proposed by Bramoullé et al. (2009).[1] We also present a Bayesian estimator. The assumed distribution of the network acts as a prior distribution, and the inferred network structure is updated through the Markov chain Monte Carlo (MCMC) algorithm. We also present an empirical application. We explore the impact of errors in the observed networks using data on adolescents' friendship networks.

In the second chapter, *Count Data Models with Social Interactions under Rational Expectations*, I develop a structural model for peer effects analysis in which the dependent variable is counting.[2] Recent contributions to the literature of peer effects estimation include many models for limited dependent variables. However, there are no existing structural models dealing with count variables, despite these variables being prevalent in survey data (e.g., Liu et al., 2014; Fortin and Yazbeck, 2015; Lee et al., 2020a). I present a static game with incomplete information (see Harsanyi, 1967; Osborne and Rubinstein, 1994) to rationalize the model. Individuals in the game interact through a directed network and are influenced by their belief over the choice of their peers. I provide sufficient conditions under which the model game has a unique Bayesian Nash Equilibrium (BNE). I show that using the linear-in-means spatial autoregressive (SAR) model (Lee, 2004) or the SAR Tobit (SART) model (Xu and Lee, 2015b) to estimate peer effects on counting variables generated from the model asymptotically underestimates the peer effects. The estimation bias asymptotically decreases when the dependent variable takes its values from a large range. I also provide an empirical application. I estimate peer effects on the number of extracurricular activities in which students are enrolled. I find that increasing the number of activities in which a student's friends are enrolled by one implies an increase in the number of activities in which the student is enrolled by 0.295, controlling for

[1] This strategy constructs instruments using the powers of the interaction matrix.

[2] E.g, number of cigarettes smoked, frequency of restaurant visits, frequency of participation in activities.

the endogeneity of the network. I also find that the SART and the SAR models underestimate peer effects at 0.141 and 0.166, respectively.

The third chapter, *Selective linear segmentation for detecting relevant parameter changes*, is a joint work with Professor Arnaud Dufays and Professor Alain Coen. Long time series are standard in this period of large publicly available data sets. Care is required when modeling such time series, as many of them span over critical events that may change the series dynamic. Considering a linear-in-means models, we propose a method to relax the assumption that a break triggers a change in all the model parameters. To do so, we first estimate the potential break dates exhibited by the series and then we use a penalized likelihood approach to detect which parameters change. Since some segments in the CP regression can be small, we opt for a (nearly) unbiased penalty function, called the seamless-L0 (SELO) penalty function, recently proposed by Dicker et al. (2013). We prove the consistency of the SELO estimator in detecting which parameters indeed vary over time and we suggest using a deterministic annealing expectation-maximisation (DAEM) algorithm to deal with the multimodality of the objective function (see Ueda and Nakano, 1998). Since the SELO penalty function depends on two tuning parameters, we use a criterion (new in this literature) to choose the best tuning parameters and as a result the best model. This new criterion exhibits a Bayesian interpretation which makes possible to assess the parameters' uncertainty as well as the model's uncertainty. This last feature is determinant when predicting a time series since the Bayesian model averaging technique, that typically improves forecast accuracy, is readily applicable (see, e.g., Raftery et al., 2010; Koop and Korobilis, 2012). Monte Carlo simulations highlight that the method works well for many time series models including heteroskedastic processes. For a sample of 14 Hedge funds (HF) strategies, using an asset based style pricing model, we shed light on the promising ability of our method to detect the time-varying dynamics of risk exposures as well as to forecast HF returns.

Aside, I also develop two easy-to-use R packages which implement the methods developed in social interactions. For instance, the package `PartialNetwork`, joint with Professor Vincent Boucher, offers a routine to replicate all the results of the first chapter. Moreover, the package `CDatanet` can be used to replicate all the results of the second chapter. These packages are available on my GitHub page or website.

Chapter 1

Estimating Peer Effects Using Partial Network Data

Résumé

Dans ce chapitre, nous étudions l'estimation des effets de pairs à travers les réseaux sociaux lorsque la structure du réseau n'est pas observée. Nous supposons que nous connaissons (avons une estimation convergente de) la *distribution* du réseau. Nous montrons que cette hypothèse est suffisante pour l'estimation des effets de pairs en utilisant un modèle linéaire en moyennes. Nous proposons une stratégie d'estimation qui adapte la procédure de variables instrumentales utilisée pour estimer ce modèle dans le cas où la structure du réseau est observée. Nous présentons également un estimateur bayésien en supposant que la distribution du réseau est une distribution a priori. Nous inférons ensuite la structure du réseau en utilisant un algorithme de Monte Carlo par Chaîne de Markov (MCMC). Nous présentons et discutons également des exemples importants où notre méthodologie peut être appliquée. Nous montrons que la base de données « Add Health » largement utilisée en économie des interactions sociales comporte de nombreux liens non observés : seulement 70% du nombre total de liens sont observés. Nous estimons un modèle des effets de pairs sur la réussite scolaire des élèves. Nous montrons que notre estimateur bayésien reconstruit les liens manquants et permet d'obtenir une estimation valide des effets de pairs. En particulier, nous montrons qu'ignorer les liens manquants sous-estime l'effet endogène des pairs sur la réussite scolaire.

Abstract

We study the estimation of peer effects through social networks when researchers do not observe the network structure. Instead, we assume that researchers know (have a consistent estimate of) the *distribution* of the network. We show that this assumption is sufficient for the estimation of peer effects using a linear-in-means model. We present an estimation strategy

that adapts the instrumental variables procedure used to estimate this model when the network structure is observed. We also present a Bayesian estimator. The assumed distribution for the network acts as a prior distribution, and the inferred network structure is updated through the Markov chain Monte Carlo (MCMC) algorithm. We also present and discuss important examples where our methodology can be applied. We show that the widely used Add Health database features many missing links: only 70% of the total number of links are observed. We estimate a model of peer effects on students' academic achievement. We show that our Bayesian estimator reconstructs these missing links and obtains a valid estimate of peer effects. In particular, we show that disregarding missing links underestimates the endogenous peer effect on academic achievement.

Keywords: Social networks, Peer effects, Missing variables, Measurement errors.

JEL Classification: C31, C36, C51.

1.1 Introduction

There is a large and growing literature on the impact of peer effects in social networks.[1] However eliciting network data is expensive (Breza et al., 2020), and since networks must be sampled completely (Chandrasekhar and Lewis, 2011), there are few existing data sets that contain detailed network information.

In this paper, we explore the estimation of the widely used linear-in-means model (e.g. Manski (1993), Bramoullé et al. (2009)) when the researcher does not observe the entire network structure. Specifically, we assume that the researcher knows the *distribution* of the network but not necessarily the network itself. An important example is when a researcher is able to estimate a network formation model using some partial information about the network structure (e.g. Breza et al. (2020)). Other examples are when the researcher observes the network with noise (e.g. Hardy et al. (2019)) or only observes a subsample of the network (e.g. Chandrasekhar and Lewis (2011)).

We present an instrumental variable estimator and show that we can adapt the strategy proposed by Bramoullé et al. (2009), which uses instruments constructed using the powers of the interaction matrix. Specifically, we use two different draws from the distribution of the network. One draw is used to approximate the endogenous explanatory variable, while the other is used to construct the instruments.

We show that since the true networks and the two approximations are drawn from the same distribution, the instruments are uncorrelated with the approximation error and are therefore valid. We explore the properties of the estimator using Monte Carlo simulations. We show that the method performs well, even when the distribution of the network is diffuse and when we allow for group-level fixed effects.

We also present a Bayesian estimator. The estimator imposes more structure but allows to cover cases for which the instrumental variable strategy fails.[2] Our estimator is general enough that it can be applied to many peer-effect models having misspecified networks (e.g. Chandrasekhar and Lewis (2011), Hardy et al. (2019), or Griffith (2019)). The approach relies on data augmentation (Tanner and Wong, 1987). The assumed distribution for the network acts as a prior distribution, and the inferred network structure is updated through the Markov chain Monte Carlo (MCMC) algorithm.

We present numerous examples of settings in which our estimators are implementable. In particular, we present an implementation of our instrumental variable estimator using the network formation model developed by Breza et al. (2020). We show that the method performs

[1] For recent reviews, see Boucher and Fortin (2016), Bramoullé et al. (2020), Breza (2016), and De Paula (2017).

[2] We also provide a classical version of the estimator (using an expectation maximization algorithm) in Appendix A.10, which is similar to the strategies used by Griffith (2018) and Hardy et al. (2019).

very well. We also show that the recent estimator proposed by Alidaee et al. (2020) works well but is less precise.

We also present an empirical application. We explore the impact of errors in the observed networks using data on adolescents' friendship networks. We show that the widely used Add Health database features many missing links: only 70% of the total number of links are observed. We estimate a model of peer effects on students' academic achievement. We show that our Bayesian estimator reconstructs these missing links and obtains a valid estimate of peer effects. In particular, we show that disregarding missing links underestimates the endogenous peer effect on academic achievement.

This paper contributes to the recent literature on the estimation of peer effects when the network is either not entirely observed or observed with noise. Chandrasekhar and Lewis (2011) show that models estimated using sampled networks are generally biased. They propose an analytical correction as well as a two-step general method of moment (GMM) estimator. Liu (2013) shows that when the interaction matrix is not row-normalized, instrumental variable estimators based on an out-degree distribution are valid, even with sampled networks. Relatedly, Hsieh et al. (2018) focus on a regression model that depends on global network statistics. They propose analytical corrections to account for non-random sampling of the network (see also Chen et al. (2013)).

Hardy et al. (2019) look at the estimation of (discrete) treatment effects when the network is observed noisily. Specifically, they assume that observed links are affected by iid errors and present an expectation maximization (EM) algorithm that allows for a consistent estimate of the treatment effect. Griffith (2018) also presents an EM algorithm to impute missing network data. Griffith (2019) explores the impact of imposing an upper bound to the number of links when eliciting network data. He shows, analytically and through simulations, that these bounds may bias the estimates significantly.

Relatedly, some papers derive conditions under which peer effects can be identified even without any network data. De Paula et al. (2018a) and Manresa (2016) use panel data and present models of peer effect having an unknown network structure. Both approaches require observing a large number of periods and some degree of *sparsity* for the interaction network. De Paula et al. (2018a) prove a global identification result and estimate their model using an adaptive elastic net estimator, while Manresa (2016) uses a lasso estimator, while assuming no endogenous effect and deriving its explicit asymptotic properties.

Souza (2014) studies the estimation of a linear-in-means model when the network is not known. He presents a pseudo-likelihood model in which the true (unobserved) network is replaced by its expected value, given a parametric network formation model. He formally derives the identified set and applies his methodology to study the spillover effects of a randomized intervention.

Thirkettle (2019) focuses on the estimation of a given network statistic (e.g. some centrality measure), assuming that the researcher only observes a random sample of links. Using a structural network formation model, he derives bounds on the identified set for both the network formation model and the network statistic of interest. Lewbel et al. (2019) use a similar strategy but focus on the estimation of a linear-in-means model and assume a network formation model having conditionally independent linking probabilities. They show that their estimator is point-identified given some exclusion restrictions.

We contribute to the literature by proposing two estimators for the linear-in-means model, in a cross-sectional setting, when the econometrician does not know the true social network but rather knows the *distribution* of true network. Our estimators are both simple to implement and flexible. In particular, they can be used when network formation models can be estimated given only limited network information (e.g. Breza et al. (2020) or Graham (2017)) or when networks are observed imperfectly (e.g. Chandrasekhar and Lewis (2011), Griffith (2019), or Hardy et al. (2019)). We show that having partial information about network structure (as opposed to no information) allows the development of flexible and easily implementable estimators. Finally, we also present an easy-to-use R package—named `PartialNetwork`—for implementing our estimators and examples, including the estimator proposed by Breza et al. (2020). The package is available online at: `https://github.com/ahoundetoungan/PartialNetwork`.

The remainder of the paper is organized as follows. In Section 2, we present the econometric model as well as the main assumptions. In Section 3, we present an instrumental variable estimator. In Section 4, we present our Bayesian estimation strategy. In Section 5, we present important economic contexts in which our method is implementable. In Section 6, we present an empirical application in which the network is only partly observed. Section 7 concludes with a discussion of the main results, limits, and challenges for future research.

1.2 The Linear-in-Means Model

Let $\mathbf{A}$ represent the $N \times N$ *adjacency matrix* of the network. We assume a directed network: $a_{ij} \in \{0, 1\}$, where $a_{ij} = 1$ if i is linked to j. We normalize $a_{ii} = 0$ for all i and let $n_i = \sum_j a_{ij}$ denote the number of links of i. Let $\mathbf{G} = f(\mathbf{A})$, the $N \times N$ *interaction matrix* for some function f. Unless otherwise stated, we assume that $\mathbf{G}$ is a row-normalization of the adjacency matrix $\mathbf{A}$.[3] Our results extend to alternative specifications of f.

We focus on the following model:

$$\mathbf{y} = c\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \alpha\mathbf{G}\mathbf{y} + \mathbf{G}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{1.1}$$

[3] In such a case, $g_{ij} = a_{ij}/n_i$ whenever $n_i > 0$, while $g_{ij} = 0$ otherwise.

where $\mathbf{y}$ is a vector of an outcome of interest (e.g. academic achievement), c is a constant, $\mathbf{X}$ is a matrix of observable characteristics (e.g. age, gender...), and $\boldsymbol{\varepsilon}$ is a vector of errors. The parameter α therefore captures the impact of the average outcome of one's peers on their behaviour (the endogenous effect). The parameter $\boldsymbol{\beta}$ captures the impact of one's characteristics on their behaviour (the individual effects). The parameter $\boldsymbol{\gamma}$ captures the impact of the average characteristics of one's peers on their behaviour (the contextual effects).

This *linear-in-means* model (Manski, 1993) is perhaps the most widely used model for studying peer effects in networks (see Bramoullé et al. (2020) for a recent review). In this paper, we contrast with the literature by assuming that the researcher does not know the interaction matrix $\mathbf{G}$. Specifically, we assume instead that the researcher knows the distribution of the interaction matrix.

The next assumption summarizes our set-up.

Assumption A. *We maintain the following assumptions:*

A.1. $|\alpha| < 1/\|\mathbf{G}\|$ *for some submultiplicative norm* $\|\cdot\|$.

A.2. The distribution $P(\mathbf{A})$ *of the true network* $\mathbf{A}$ *(which potentially depends on* $\mathbf{X}$*) is known.*

A.3. The population is partitioned in $M > 1$ *groups, where the size* N_r *of each group* $r = 1, ..., M$ *is bounded. The probability of a link between individuals of different groups is equal to* 0.

A.4. For each group, the outcome and individual characteristics are observed, i.e. $(\mathbf{y}_r, \mathbf{X}_r)$, $r = 1, ..., M$, *are observed.*

A.5. The network is exogenous in the sense that $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{G}] = \mathbf{0}$.

Assumption A.1. ensures that the model is coherent and that there exists a unique vector $\mathbf{y}$ compatible with (1.1). When $\mathbf{G}$ is row-normalized, $|\alpha| < 1$ is sufficient.

Assumption A.2. states that the researcher knows the distribution of the true network $\mathbf{A}$. Of course, knowledge of $P(\mathbf{A})$ is sufficient for $P(\mathbf{G})$, since $\mathbf{G} = f(\mathbf{A})$ for some known function f. Assumption A.2. is weaker than assuming that the econometrician observes the entire network structure. In Section 1.5, we discuss some important examples where Assumption A.2. is reasonable for important economic contexts. In particular, we present examples from the literature on network formation models that allow for a consistent estimation of $P(\mathbf{A})$ using only partial network information.

As will be made clear, our estimation strategy requires that the econometrician be able to draw iid samples from $P(\mathbf{A})$. As such, and for the sake of simplicity, all of our examples will be based on network distributions that are conditionally independent across links (i.e. $P(a_{ij}|\mathbf{A}_{-ij}) = P(a_{ij})$), although this is not formally required.[4]

[4] A prime example of a network distribution that is not conditionally independent is the distribution for

Assumption A.3. is by no means necessary; however, it simplifies the exposition and ensures a law of large numbers (LLN) in our context. We refer the reader to Lee (2004) and Lee et al. (2010) for more general, alternative sufficient conditions.

Assumption A.4. implies that the data is composed of a subset of fully sampled groups.[5] A similar assumption is made by Breza et al. (2020). Note that we assume that the network is exogenous (Assumption A.5.) mostly to clarify the presentation of the estimators. In Section 1.7, we discuss how recent advances for the estimation of peer effects in endogenous networks can be adapted to our context.

Finally, note that Assumption A *does not* imply that one can simply proxy $\mathbf{G}$ in (1.1) using a draw $\hat{\mathbf{G}}$ from $P(\mathbf{G})$. The reason is that for any vector $\mathbf{w}$, $\hat{\mathbf{G}}\mathbf{w}$ generally does not converge to $\mathbf{G}\mathbf{w}$ as N goes to infinity. In other words, knowledge of $P(\mathbf{G})$ and $\mathbf{w}$ is not sufficient to obtain a consistent estimate of $\mathbf{G}\mathbf{w}$. We discuss some exceptions in Section 1.7.

1.3 Estimation Using Instrumental Variables

As discussed in the introduction, we show that it is possible to estimate (1.1) given only partial information on network structure. To understand the intuition, note that it is not necessary to observe the complete network structure to observe $\mathbf{y}$, $\mathbf{X}$, $\mathbf{G}\mathbf{X}$, and $\mathbf{G}\mathbf{y}$. For example, one could simply obtain $\mathbf{G}\mathbf{y}$ from survey data: "What is the average value of your friends' y?"

However, the observation of $\mathbf{y}$, $\mathbf{X}$, $\mathbf{G}\mathbf{X}$, and $\mathbf{G}\mathbf{y}$ is not sufficient for the estimation of (1.1). The reason is that $\mathbf{G}\mathbf{y}$ is endogenous; thus, a simple linear regression would produce biased estimates. (e.g. Manski (1993), Bramoullé et al. (2009)).

The typical instrumental approach to deal with this endogeneity is to use instruments based on the structural model, i.e. instruments constructed using second-degree peers (e.g. $\mathbf{G}^2\mathbf{X}$, see Bramoullé et al. (2009)). These are less likely to be found in survey data. Indeed, we could doubt the informativeness of questions such as: "What is the average value of your friends' average value of their friends' x?"

Under the assumption that the network is observed, the literature has focused mostly on efficiency: that is, how to construct the optimal set of instruments (e.g. Kelejian and Prucha (1998) or Lee et al. (2010)). Here, we are interested in a different question. We would like to understand how much information on the network structure is needed to construct relatively "good" instruments for $\mathbf{G}\mathbf{y}$? As we will discuss, it turns out that even very imprecise estimates of $\mathbf{G}$ allow for constructing valid instruments.

We present valid instruments in Proposition 1.1 and Proposition 1.2 below. We also study the

an exponential random graph model (ERGM), e.g. Mele (2017). See also our discussion in Section 1.7.

[5] Contrary to Liu et al. (2017) or Wang and Lee (2013), for example.

properties of the implied estimators using Monte Carlo simulations. Unless otherwise stated, these simulations are performed as follows: we simulate 100 groups of 50 individuals each. Within each group, each link (i, j) is drawn from a Bernoulli distribution with probability:

$$p_{ij} = \frac{\exp\{c_{ij}/\lambda\}}{1 + \exp\{c_{ij}/\lambda\}}, \tag{1.2}$$

where $c_{ij} \sim N(0, 1)$, and $\lambda > 0$.

This approach is convenient since it allows for some heterogeneity among linking probabilities. Moreover, λ can easily control the spread of the distribution, and hence the quality of the approximation of the true network.[6] Indeed, when $\lambda \to 0$, $p_{ij} \to 1$ whenever $c_{ij} > 0$, while $p_{ij} \to 0$ whenever $c_{ij} < 0$. Similarly, as $\lambda \to \infty$, $p_{ij} \to 1/2$. Then, simulations are very precise for $\lambda \to 0$ and very imprecise (and homogeneous) for $\lambda \to \infty$.

We also let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2]$, where $x_i^1 \sim N(0, 5^2)$ and $x_i^2 \sim Poisson(6)$. We set the true value of the parameters to: $\alpha = 0.4$, $c = 2$, $\beta_1 = 1$, $\beta_2 = 1.5$, $\gamma_1 = 5$, and $\gamma_2 = -3$. Finally, we let $\varepsilon_i \sim N(0, 1)$.

We now present our formal results. To clearly expose the argument, we first start by discussing the special case where there are no contextual effects: $\boldsymbol{\gamma} = \mathbf{0}$. The model in (1.1) can therefore be rewritten as:

$$\mathbf{y} = c\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \alpha\mathbf{G}\mathbf{y} + \boldsymbol{\varepsilon}.$$

The following proposition holds.

Proposition 1.1. *Assume that* $\boldsymbol{\gamma} = \mathbf{0}$. *There are two cases:*

1. *Suppose that* $\mathbf{G}\mathbf{y}$ *is observed and let* $\boldsymbol{H}$ *be an interaction matrix, correlated with* $\mathbf{G}$*, and such that* $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{H}] = \mathbf{0}$. *Then,* $\mathbf{H}\mathbf{X}$, $\mathbf{H}^2\mathbf{X}$,... *are valid instruments.*
2. *Suppose that* $\mathbf{G}\mathbf{y}$ is not *observed and let* $\tilde{\mathbf{G}}$ *and* $\hat{\mathbf{G}}$ *be two draws from the distribution* $P(\mathbf{G})$. *Then,* $\hat{\mathbf{G}}\mathbf{X}$, $\hat{\mathbf{G}}^2\mathbf{X}$,... *are valid instruments when* $\tilde{\mathbf{G}}\mathbf{y}$ *is used as a proxy for* $\mathbf{G}\mathbf{y}$.

First, suppose that $\mathbf{G}\mathbf{y}$ is observed directly from the data; then, any instrument correlated with the usual instruments $\mathbf{G}\mathbf{X}, \mathbf{G}^2\mathbf{X}$,... while being exogenous are valid. Note that a special case of the first part of Proposition 1.1 is when $\mathbf{H}$ is drawn from $P(\mathbf{G})$. However, the instrument remains valid if the researcher uses the *wrong* distribution $P(\mathbf{G})$.[7] A similar strategy is used by Kelejian and Piras (2014) and Lee et al. (2020b) in a different context. An example, presented in Section 1.5.2, is when $P(\mathbf{G})$ is estimated imprecisely in small samples.

Of course, the specification error on $P(\mathbf{G})$ must be independent of $\boldsymbol{\varepsilon}$. Note also that if the specification error is too large, the correlation between $\mathbf{G}\mathbf{y}$ and $\mathbf{H}\mathbf{X}$ will likely be weak. It is

[6] The true network and the approximations are drawn from the same distribution.

[7] We would like to thank Chih-Sheng Hsieh and Arthur Lewbel for discussions on this important point.

also worth noting that the first part of Proposition 1.1 does not depend on the assumption that groups are entirely sampled (i.e. Assumption A.4.).

When $\mathbf{Gy}$ is *not* observed directly, however, specification errors typically produce invalid instruments. Note also that the estimation requires two draws from $P(\mathbf{G})$ instead of just one. To see why, let us rewrite the model as:

$$\mathbf{y} = c\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \alpha\tilde{\mathbf{G}}\mathbf{y} + [\boldsymbol{\eta} + \boldsymbol{\varepsilon}],$$

where $\boldsymbol{\eta} = \alpha[\mathbf{Gy} - \tilde{\mathbf{G}}\mathbf{y}]$ is the approximation error for $\mathbf{Gy}$. Suppose also that $\hat{\mathbf{G}}\mathbf{X}$ is used as an instrument for $\mathbf{Gy}$.

The validity of the instrument therefore requires $\mathbb{E}[\boldsymbol{\eta} + \boldsymbol{\varepsilon}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}] = 0$, and in particular:

$$\mathbb{E}[\mathbf{Gy}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}] = \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}],$$

which is true since $\mathbf{G}$ and $\tilde{\mathbf{G}}$ are drawn from the same distribution.

Table 1.1 presents the results of the Monte Carlo simulations, which are in line with the above discussion. Figure 1.1 shows that the estimator is still centred and precise, even when the constructed networks are really imprecise estimates of the true network. Finally, note that this also implies a non-intuitive property: if $\boldsymbol{\gamma} = \mathbf{0}$, and if $\mathbf{GX}$ is observed, but not $\mathbf{Gy}$, then $\mathbf{GX}$ is not a valid instrument since it is correlated with the approximation error $\boldsymbol{\eta}$.

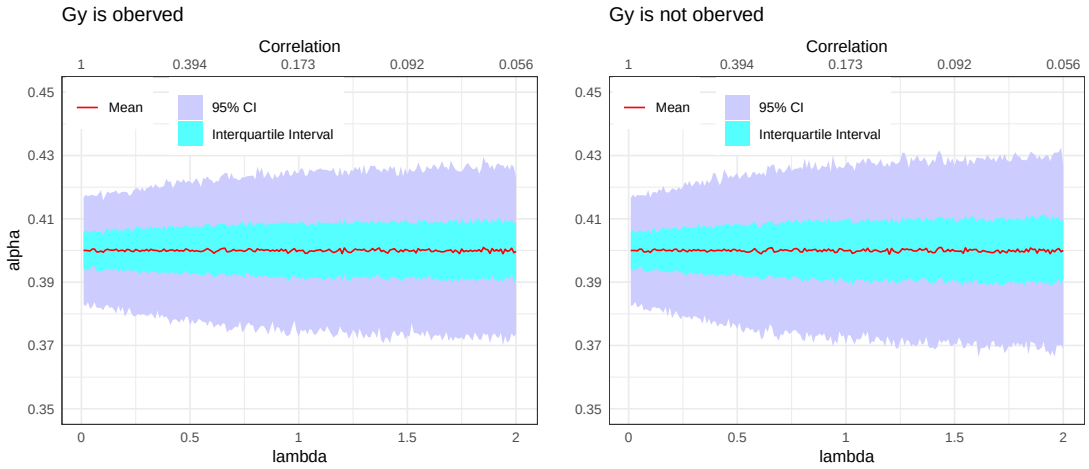


Figure 1.1 – Peer effect without contextual effects

The graph shows estimates of α for 1,000 replications of the model without contextual effects for various values of λ. The upper x-axis reports the average correlation between two independent network draws using the distribution given by equation (1.2).

Of course, Proposition 1.1 assumes that there are no contextual effects. We show that a similar result holds when $\boldsymbol{\gamma} \neq \mathbf{0}$. However, to estimate (1.1) using an instrumental variable approach, we must assume that $\mathbf{GX}$ is *observed*. The reason is that there are no natural instruments for

Table 1.1 – Simulation results without contextual effects

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - **Gy is Observed**					
Estimation results					
$Intercept = 2$	2.000	0.242	1.830	1.992	2.162
$\alpha = 0.4$	0.400	0.013	0.391	0.401	0.409
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	1,816.759	281.478	1,623.922	1,800.054	1,995.179
Hausman	1.198	1.607	0.120	0.566	1.669
Sargan	0.905	1.315	0.088	0.402	1.168
$N = 50$, $M = 100$ - **Gy is not observed - same draw**					
Estimation results					
$Intercept = 2$	4.348	0.287	4.156	4.332	4.535
$\alpha = 0.4$	0.271	0.015	0.261	0.272	0.282
$\beta_1 = 1$	1.002	0.003	0.999	1.001	1.004
$\beta_2 = 1.5$	1.503	0.006	1.498	1.503	1.507
Tests					
F-test	26,656.064	2,108.805	25,237.919	26,492.586	27,972.810
Hausman	245.060	36.134	220.376	242.230	267.029
Sargan	1.939	2.768	0.208	0.910	2.452
Validity					
$cor(\eta_i, \hat{x}_{i,1})$	−0.367	0.018	−0.380	−0.367	−0.355
$cor(\eta_i, \hat{x}_{i,2})$	−0.269	0.017	−0.280	−0.269	−0.257
$N = 50$, $M = 100$ - **Gy is not observed - different draws**					
Estimation results					
$Intercept = 2$	2.001	0.264	1.809	1.994	2.175
$\alpha = 0.4$	0.400	0.014	0.390	0.400	0.410
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	1,824.774	280.901	1,623.689	1812.479	2,014.936
Hausman	69.842	17.204	57.169	69.691	81.438
Sargan	0.891	1.245	0.082	0.431	1.143
Validity					
$cor(\eta_i, \hat{x}_{i,1})$	0.000	0.014	−0.010	0.000	0.010
$cor(\eta_i, \hat{x}_{i,2})$	0.000	0.014	−0.010	0.000	0.010

Number of simulations: 1,000, $\lambda = 1$. Instruments: $\mathbf{GX}$ if $\mathbf{Gy}$ is observed, $\mathbf{G}^c\mathbf{X}$ if $\mathbf{Gy}$ is not observed and approximated by $\mathbf{G}^c\mathbf{y}$. Additional results for alternative instruments and $\lambda = +\infty$ are available in Table A.1, Table A.2, and Table A.3 of Appendix A.3.

$\mathbf{GX}$. In Section 1.4, we present an alternative estimation strategy that does not require the observation of $\mathbf{GX}$.

We have the following:

Proposition 1.2. *Assume that* $\mathbf{GX}$ *is observed. There are two cases:*

1. *Suppose that* $\mathbf{Gy}$ *is observed and let* $\boldsymbol{H}$ *be an interaction matrix, correlated with* $\mathbf{G}$*, and such that* $\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{H}] = \mathbf{0}$*. Then,* $\mathbf{H}^2\mathbf{X}$*,* $\mathbf{H}^3\mathbf{X}$*,... are valid instruments.*
2. *Suppose that* $\mathbf{Gy}$ is not *observed and let* $\tilde{\mathbf{G}}$ *and* $\hat{\mathbf{G}}$ *be two draws from the distribution* $P(\mathbf{G})$*. Then,* $\hat{\mathbf{G}}^2\mathbf{X}$*,* $\hat{\mathbf{G}}^3\mathbf{X}$*,... are valid instruments when* $\tilde{\mathbf{G}}\mathbf{y}$ *is used as a proxy for* $\mathbf{Gy}$*, if* $\tilde{\mathbf{G}}\mathbf{X}$ *is added as additional explanatory variables.*

The first part of Proposition 1.2 is a simple extension of the first part of Proposition 1.1. The second part of Proposition 1.2 requires more discussion. Essentially, it states that $\hat{\mathbf{G}}^2\mathbf{X}$, $\hat{\mathbf{G}}^3\mathbf{X}$, ... are valid instruments when the following *expanded model* is estimated:

$$\mathbf{y} = c\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \alpha\tilde{\mathbf{G}}\mathbf{y} + \mathbf{GX}\boldsymbol{\gamma} + \tilde{\mathbf{G}}\mathbf{X}\check{\boldsymbol{\gamma}} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{1.3}$$

where the true value of $\check{\boldsymbol{\gamma}}$ is $\mathbf{0}$.

To understand why the introduction of $\tilde{\mathbf{G}}\mathbf{X}\check{\boldsymbol{\gamma}}$ is needed, recall that the constructed instrument must be uncorrelated with the approximation error $\boldsymbol{\eta}$. This correlation is conditional on the explanatory variables, that contain $\mathbf{G}$. In particular, it implies that generically,

$$\mathbb{E}[\mathbf{Gy}|\mathbf{X}, \mathbf{GX}, \hat{\mathbf{G}}^2\mathbf{X}] \neq \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X}, \mathbf{GX}, \hat{\mathbf{G}}^2\mathbf{X}].$$

It turns out that adding the auxiliary variable $\tilde{\mathbf{G}}\mathbf{X}$ as a covariate is sufficient to restore the result, i.e.

$$\mathbb{E}[\mathbf{Gy}|\mathbf{X}, \mathbf{GX}, \hat{\mathbf{G}}^2\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}] = \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X}, \mathbf{GX}, \hat{\mathbf{G}}^2\mathbf{X}, \tilde{\mathbf{G}}\mathbf{X}].$$

Table 1.2 presents the simulations' results. We see that most of the estimated parameters are not biased. However, we also see that estimating the expanded model, instead of the true one, comes at a cost. Due to multicollinearity, the estimation of $\boldsymbol{\gamma}$ is contaminated by $\tilde{\mathbf{G}}\mathbf{X}$, and the parameters are biased. Figure 1.2 also shows that the estimation of α remains precise, even as the value of λ increases.

Proposition 1.1 and Proposition 1.2 therefore show that the estimation of (1.1) is possible, even with very limited information about the network structure. We conclude this section by discussing how one can adapt this estimation strategy while allowing for group-level unobservables.

Table 1.2 – Simulation results with contextual effects

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - Instrument: $(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is observed					
Estimation results					
$Intercept = 2$	1.996	0.177	1.879	1.997	2.115
$\alpha = 0.4$	0.400	0.003	0.398	0.400	0.402
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
$\gamma_1 = 5$	5.000	0.021	4.985	5.000	5.015
$\gamma_2 = -3$	−2.999	0.029	−3.018	−2.999	−2.980
Tests					
F-test	18295.381	2049.380	16864.174	18258.774	19581.640
Hausman	1.202	1.624	0.127	0.568	1.593
Sargan	1.046	1.559	0.103	0.448	1.321
$N = 50$, $M = 100$ - Instrument: $(\hat{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is not observed					
Estimation results					
$Intercept = 2$	1.987	0.207	1.844	1.983	2.128
$\alpha = 0.4$	0.400	0.004	0.397	0.400	0.402
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
$\gamma_1 = 5$	5.357	0.021	5.342	5.356	5.370
$\gamma_2 = -3$	−2.381	0.038	−2.408	−2.378	−2.355
$\hat{\gamma}_1 = 0$	−0.356	0.024	−0.372	−0.356	−0.339
$\hat{\gamma}_2 = 0$	−0.617	0.038	−0.643	−0.618	−0.592
Tests					
F-test	13562.892	1402.029	12583.175	13547.357	14445.031
Hausman	17.051	8.277	11.093	15.779	22.061
Sargan	1.003	1.425	0.125	0.470	1.267

Number of simulations: 1,000, $\lambda = 1$. Additional results for $\lambda = +\infty$ are available in Table A.4 of Appendix A.3.

1.3.1 Group-Level Unobservables

A common assumption is that each group in the population is affected by a common shock, unobserved by the econometrician (e.g. Bramoullé et al. (2009)). As such, for each group $r = 1, ..., M$, we have:

$$\mathbf{y}_r = c_r\mathbf{1}_r + \mathbf{X}_r\boldsymbol{\beta} + \alpha\mathbf{G}_r\mathbf{y}_r + \mathbf{G}_r\mathbf{X}_r\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_r,$$

where c_r is not observed, $\mathbf{1}_r$ is a N_r-dimensional vector of ones, N_r is the size of the group r, $\mathbf{G}_r$ is the sub-interaction matrix in the group r, and $\boldsymbol{\varepsilon}_r$ is the vector of error terms in the group r

Under Assumption A.3., it is not possible to obtain a consistent estimate of $\{c_r\}_{r=1}^m$ since the number of observations used to estimate each c_r is bounded. This is known as the *incidental*

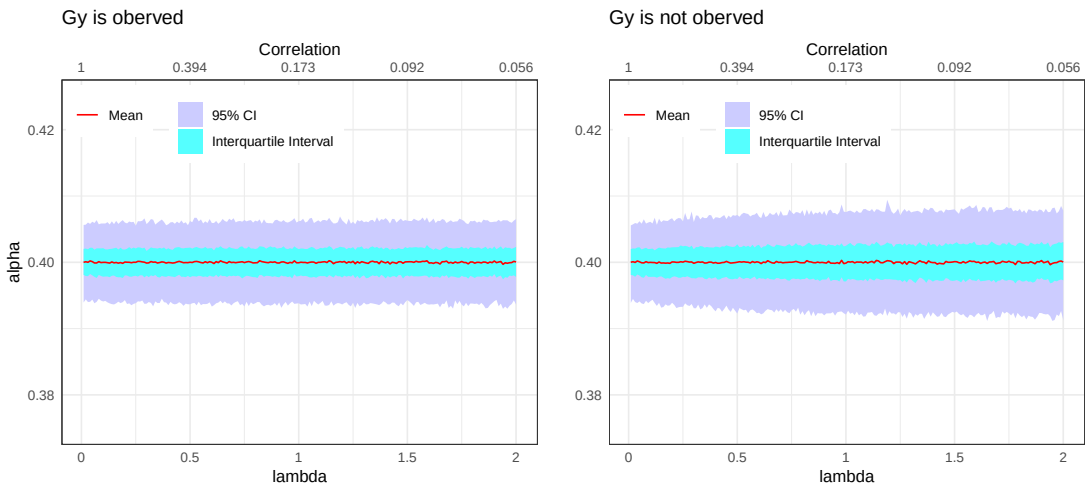


Figure 1.2 – Peer effect with contextual effects

The graph shows estimates of α for 1,000 replications of the model with contextual effects for various values of λ. The upper x-axis reports the average correlation between two independent network draws using the distribution given by equation (1.2).

parameter problem.[8] A common strategy is to use deviations from the group average and to estimate the model in deviations (e.g. Bramoullé et al. (2009)).

Let $\mathbf{J} = diag\{\mathbf{I}_{N_r} - \frac{1}{N_r}\mathbf{1}_r\mathbf{1}_r'\}$ be the group-differentiating matrix, where $\mathbf{I}_{N_r}$ is the identified matrix of dimension N_r. The operator *diag* generates a blog-diagonal matrix in which each group is a block.[9] We can rewrite:

$$\mathbf{Jy} = \mathbf{JX}\boldsymbol{\beta} + \alpha\mathbf{JGy} + \mathbf{JGX}\boldsymbol{\gamma} + \boldsymbol{J\varepsilon}.$$

Note that the results of Propositions 1.1 and 1.2 extend directly. Figure 1.3 shows that the estimation performs well; however, the loss of information can be large. Indeed, as λ increases, not only does the correlation between the true network and the constructed network decrease, but the linking probabilities become homogeneous. Then, it becomes hard to distinguish between the (almost uniform) network effects and the group effects. In practice, we therefore expect our approach to perform well when the distribution of the true network exhibits heterogeneous linking probabilities.

For example, Table 1.3 presents the estimation results when we assume that the network formation process is a function of the observed characteristics. Specifically:

$$p(a_{ij} = 1|\mathbf{x}_i, \mathbf{x}_j) = \Phi(-4.5 + |x_{i,1} - x_{j,1}| - 2|x_{i,2} - x_{j,2}|),$$

where Φ is the cumulative distribution for the standardized normal distribution. As such, the network features *heterophily* with respect to the first variable and *homophily* with respect

[8] See Lancaster (2000) for a review.

[9] Then, $\mathbf{Jw}$ gives $\mathbf{w}$ minus the group average of $\mathbf{w}$.

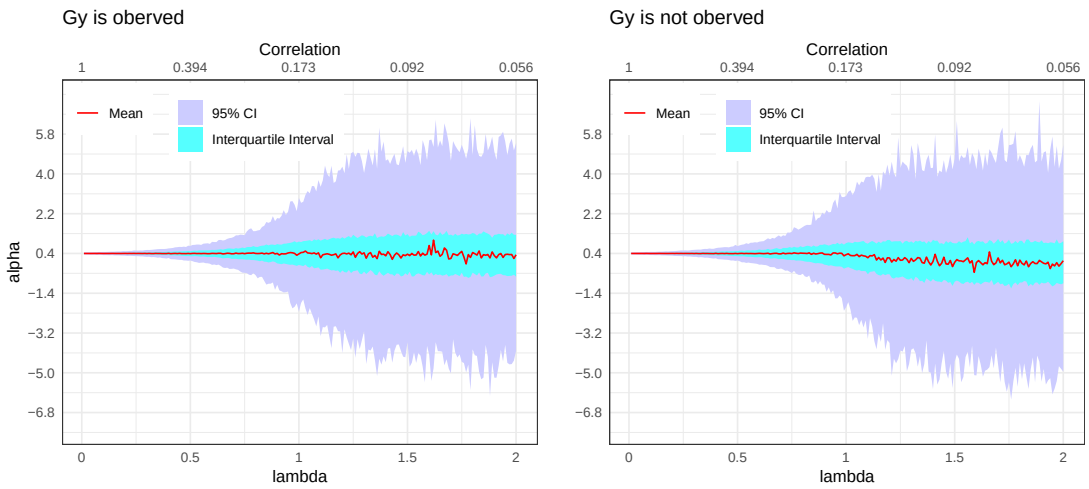


Figure 1.3 – Peer effect with fixed effects

The graph shows α estimates for 1,000 replications of the model with fixed effects for various values of λ. The above x-axis reports the average correlation between two independent draws of graphs from the distribution given by equation (1.2). Complete estimates for $\lambda \in \{1, +\infty\}$ are presented in Tables A.5 and A.6 of the Appendix A.3.

to the second variable.[10] As anticipated, the estimation performs well. We now present our likelihood-based estimator.

1.4 Likelihood-Based Estimators

The approach developed in the previous section assumes that $\mathbf{GX}$ is observed. When it is not, the instrumental variable estimators fail. We therefore present a likelihood-based estimator. Accordingly, more structure must be imposed on the errors $\boldsymbol{\varepsilon}$.[11]

To clarify the exposition, we will focus on the network adjacency matrix $\mathbf{A}$ instead of the interaction matrix $\mathbf{G}$. Of course, this is without any loss of generality. Given parametric assumptions for $\boldsymbol{\varepsilon}$, one can write the log-likelihood of the outcome as:[12]

$$\ln \mathcal{P}(\mathbf{y}|\mathbf{A}, \boldsymbol{\theta}), \tag{1.4}$$

where $\boldsymbol{\theta} = [\alpha, \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\sigma}']'$, $\boldsymbol{\sigma}$ are unknown parameters from the distribution of $\boldsymbol{\varepsilon}$. Note that $\mathbf{y} = (\mathbf{I}_N - \alpha\mathbf{G})^{-1}(c\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{GX}\boldsymbol{\gamma} + \boldsymbol{\varepsilon})$ and $(\mathbf{I}_N - \alpha\mathbf{G})^{-1}$ exist under our Assumption A.1..

If the adjacency matrix $\mathbf{A}$ was observed, then (1.4) could be estimated using a simple maximum likelihood estimator (as in Lee et al. (2010)) or using Bayesian inference (as in Goldsmith-Pinkham and Imbens (2013)).

[10] That is, individuals with different values of x_1 and similar values of x_2 are more likely to be linked.

[11] Lee (2004) presents a quasi maximum-likelihood estimator that does not require such a specific assumption for the distribution of the error term. His estimator could be used alternatively. As well, as will be made clear, our approach can be used for a large class of extremum estimators, following Chernozhukov and Hong (2003), and in particular for GMM estimators, as in Chandrasekhar and Lewis (2011).

[12] Note that under Assumption A.3., the likelihood can be factorized across groups.

Table 1.3 – Simulation results with subpopulation unobserved fixed effects (3)

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - Instrument: $\mathbf{J}(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\mathbf{Gy}$ is observed					
Estimation results					
$\alpha = 0.4$	0.400	0.006	0.396	0.400	0.404
$\beta_1 = 1$	1.000	0.007	0.995	1.000	1.005
$\beta_2 = 1.5$	1.500	0.020	1.486	1.499	1.514
$\gamma_1 = 5$	5.000	0.008	4.995	5.000	5.005
$\gamma_2 = -3$	−2.999	0.030	−3.021	−2.998	−2.979
Tests					
F-test	1123.431	178.101	999.270	1116.900	1242.319
Hausman	1.039	1.503	0.114	0.472	1.289
Sargan	1.037	1.370	0.111	0.509	1.458
$N = 50$, $M = 100$ - Instrument: $\mathbf{J}(\hat{\mathbf{G}})^2\mathbf{X}$ - $\mathbf{Gy}$ is not observed					
Estimation results					
$\alpha = 0.4$	0.399	0.015	0.389	0.398	0.408
$\beta_1 = 1$	1.002	0.013	0.994	1.002	1.011
$\beta_2 = 1.5$	1.418	0.054	1.380	1.419	1.453
$\gamma_1 = 5$	4.743	0.046	4.713	4.743	4.775
$\gamma_2 = -3$	−3.655	0.252	−3.843	−3.669	−3.490
$\hat{\gamma}_1 = 0$	0.256	0.046	0.224	0.255	0.286
$\hat{\gamma}_2 = 0$	0.788	0.280	0.609	0.794	0.987
Tests					
F-test	1153.330	200.889	1003.857	1147.411	1277.991
Hausman	161.862	60.319	117.502	154.631	197.888
Sargan	9.257	13.256	0.825	4.052	12.405

Number of simulations: 1,000. In each group, the fixed effect is generated as $0.3x_{1,1} + 0.3x_{3,2} - 1.8x_{50,2}$. The network's true distribution follows the network formation model, such that $p_{ij} = \Phi\left(-4.5 + |x_{i,1} - x_{j,1}| - 2|x_{i,2} - x_{j,2}|\right)$, where Φ represents the cumulative distribution function of $\mathcal{N}(0, 1)$.

Since $\mathbf{A}$ is not observed, an alternative would be to focus on the unconditional likelihood, i.e.

$$\ln \mathcal{P}(\mathbf{y}|\boldsymbol{\theta}) = \ln \sum_{\mathbf{A}} \mathcal{P}(\mathbf{y}|\mathbf{A}, \boldsymbol{\theta}) P(\mathbf{A}).$$

A similar strategy is proposed by Chandrasekhar and Lewis (2011) using a GMM estimator.

One particular issue with estimating $\ln \mathcal{P}(\mathbf{y}|\boldsymbol{\theta})$ is that the summation is not tractable. Indeed, the sum is over the set of possible adjacency matrices, which contain $2^{N(N-1)}$ elements. Then, simply simulating networks from $P(\mathbf{A})$ and taking the average is likely to lead to poor approximations.[13] A classical way to address this issue is to use an EM algorithm (Dempster

[13] That is: $\ln \mathcal{P}(\mathbf{y}|, \boldsymbol{\theta}) \approx \ln \frac{1}{S} \sum_{s=1}^{S} \mathcal{P}(\mathbf{y}|\mathbf{A}_s, \boldsymbol{\theta})$, where $\mathbf{A}_s$ is drawn from $P(\mathbf{A})$. This is the approximation

et al., 1977). The interested reader can consult Appendix A.10 for a presentation of such an estimator. Although valid, we found that the Bayesian estimator proposed in this section is less restrictive and numerically outperforms its classical counterpart.

For concreteness, we will assume that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_N)$; however, it should be noted that our approach is valid for a number of alternative assumptions. We have for $\mathbf{G} = f(\mathbf{A})$,

$$\begin{aligned} \ln \mathcal{P}(\mathbf{y}|\mathbf{A}, \boldsymbol{\theta}) &= -N \ln(\sigma) + \ln |\mathbf{I}_N - \alpha \mathbf{G}| - \frac{N}{2} \ln(\pi) \\ &\quad - \frac{[(\mathbf{I}_N - \alpha \mathbf{G})\mathbf{y} - c\mathbf{1}_N - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\mathbf{X}\boldsymbol{\gamma}]'[(\mathbf{I}_N - \alpha \mathbf{G})\mathbf{y} - c\mathbf{1}_N - \mathbf{X}\boldsymbol{\beta} - \mathbf{G}\mathbf{X}\boldsymbol{\gamma}]}{2\sigma^2}. \end{aligned}$$

Since $\mathbf{A}$ is not observed, we follow Tanner and Wong (1987) and Albert and Chib (1993), and we use data augmentation to evaluate the posterior distribution of $\boldsymbol{\theta}$. That is, instead of focusing on the posterior $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{A}, \mathbf{X})$, we focus on the posterior $p(\boldsymbol{\theta}, \mathbf{A}|\mathbf{y}, \mathbf{X})$, treating $\mathbf{A}$ as another set of unknown parameters.

Indeed, it is possible to obtain draws from $p(\boldsymbol{\theta}, \mathbf{A}|\mathbf{y}, \mathbf{X})$ using Algorithm 1.1..

Algorithm 1.1. MCMC to draw from $p(\boldsymbol{\theta}, \mathbf{A}|\mathbf{y}, \mathbf{X})$

Step 0: Initialize $\mathbf{A}, \boldsymbol{\theta}$ to $\mathbf{A}_0, \boldsymbol{\theta}_0$, respectively;
for $t = 1, ..., T$, *where* T *is the number of simulations* **do**
- **Step 1.1**: Propose $\mathbf{A}^*$ from the proposal distribution $q_A(\mathbf{A}^*|\mathbf{A}_{t-1})$ and accept $\mathbf{A}^*$ with probability $\min\left\{1, \frac{\mathcal{P}(\mathbf{y}|\boldsymbol{\theta}_{t-1}, \mathbf{A}^*) q_A(\mathbf{A}_{t-1}|\mathbf{A}^*) P(\mathbf{A}^*)}{\mathcal{P}(\mathbf{y}|\boldsymbol{\theta}_{t-1}, \mathbf{A}_{t-1}) q_A(\mathbf{A}^*|\mathbf{A}_{t-1}) P(\mathbf{A}_{t-1})}\right\}$;
- **Step 1.2**: Draw α^* from the proposal $q_\alpha(\cdot|\alpha_{t-1})$ and accept α^* with probability $\min\left\{1, \frac{\mathcal{P}(\mathbf{y}|\mathbf{A}_t; \boldsymbol{\beta}_{t-1}, \boldsymbol{\gamma}_{t-1}, \alpha^*) q_\alpha(\alpha_{t-1}|\alpha^*) P(\alpha^*)}{\mathcal{P}(\mathbf{y}|\mathbf{A}_t; \boldsymbol{\theta}_{t-1}) q_\alpha(\alpha^*|\alpha_{t-1}) P(\alpha_{t-1})}\right\}$;
- **Step 1.3**: Draw $[\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma]$ from their conditional distributions.

Detailed distributions for Steps 2 and 3 can be found Appendix A.8. Step 1, however, involves some additional complexities. Indeed, the idea is the following: starting from a given network formation model (i.e. $P(\mathbf{A})$), one has to be able to draw samples from the posterior distribution of $\mathbf{A}$, given $\mathbf{y}$. This is not a trivial task. The strategy used here is to rely on a Metropolis–Hastings algorithm, a strategy that has also been used in the related literature on ERGMs (e.g. Snijders (2002), Mele (2017)).

The acceptance probability in Step 1 of Algorithm 1.1. clearly exposes the role of the assumed distribution for the true network $P(\mathbf{A})$, i.e. the prior distribution of $\mathbf{A}$. This highlights the importance of $P(\mathbf{A})$ for the identification of the model. Since $\boldsymbol{\theta}$ and $\mathbf{A}$ are unobserved, we have $N(N-1) + k$ parameters to estimate, where k is the number of dimensions of $\boldsymbol{\theta}$. In particular, if $P(a_{ij} = 1) = 1/2$ for all i, j, then the probability of acceptance in Step 1 of

suggested by Chandrasekhar and Lewis (2011) (see their Section 4.3). In their case, they only need to integrate over the $m < N(N-1)$ pairs that are not sampled. Still, the number of compatible adjacency matrices is 2^m. As such, the approach is likely to produce bad approximations.

Algorithm 1.1. reduces to:

$$\min\left\{1, \frac{\mathcal{P}(\mathbf{y}|\boldsymbol{\theta}_{t-1}, \mathbf{A}^*)q_A(\mathbf{A}_{t-1}|\mathbf{A}^*)}{\mathcal{P}(\mathbf{y}|\boldsymbol{\theta}_{t-1}, \mathbf{A}_{t-1})q_A(\mathbf{A}^*|\mathbf{A}_{t-1})}\right\},$$

which only depends on the likelihood of the model and $q_A(\cdot|\cdot)$.[14] We explore the impact of the information encoded in $P(\mathbf{A})$ on the identification of $\boldsymbol{\theta}$ using Monte Carlo simulations later in this section.

One issue, however, is that there is no general rule for selecting the network proposal distribution $q_A(\cdot|\cdot)$. A natural candidate is a Gibbs sampling algorithm for each link, i.e. change only one link ij at every step t and propose a_{ij} according to its marginal distribution:

$$a_{ij} \sim P(\cdot|\mathbf{A}_{-ij}, \mathbf{y}) = \frac{\mathcal{P}(\mathbf{y}|a_{ij}, \mathbf{A}_{-ij})P(a_{ij}|\mathbf{A}_{-ij})}{\mathcal{P}(\mathbf{y}|1, \mathbf{A}_{-ij})P(a_{ij}=1|\mathbf{A}_{-ij}) + \mathcal{P}(\mathbf{y}|0, \mathbf{A}_{-ij})P(a_{ij}=0|\mathbf{A}_{-ij})},$$

where $\mathbf{A}_{-ij} = \{a_{kl}; k \neq i, l \neq j\}$. In this case, the proposal is always accepted.

However, it has been argued that Gibbs sampling could lead to slow convergence (e.g. Snijders (2002), Chatterjee et al. (2013)), especially when the network is *sparse* or exhibits a high level of *clustering*. For example, Mele (2017) and Bhamidi et al. (2008) propose different blocking techniques that are meant to improve convergence.

Here, however, the realization of Step 1 involves an additional computational issue since evaluating the likelihood ratio in Step 1 requires comparing the determinants $|\mathbf{I} - \alpha f(\mathbf{A}^*)|$ for each proposed $\mathbf{A}^*$, which is computationally intensive. In particular, taking $\mathbf{G}^* = f(\mathbf{A}^*)$ to be a row-normalization of $\mathbf{A}^*$, changing a single element of $\mathbf{A}^*$ results in a change in the entire corresponding row of $\mathbf{G}^*$. Still, comparing the determinant of two matrices that differ only in a single row is relatively fast. Moreover, when $\mathbf{G} = \mathbf{A}$, Hsieh et al. (2019) propose a blocking technique that facilitates the computation of the determinant.

Since the appropriate blocking technique depends strongly on $P(\mathbf{A})$ and the assumed distribution for $\boldsymbol{\varepsilon}$, we use the Gibbs sampling algorithm for each link of the simulations and estimations presented in this paper, adapting the strategy proposed by Hsieh et al. (2019) to our setting (see Proposition A.1 in Appendix A.7). This can be viewed as a *worse-case* scenario. We encourage researchers to try other updating schemes if Gibbs sampling performs poorly in their specific contexts. In particular, we present a blocking technique in Appendix A.7 that is also implemented in our R package `PartialNetwork`.[15]

Table 1.4 presents the Monte Carlo simulations using Algorithm 1.1.. The simulated population is the same as in Section 1.3; however, for computational reasons, we limit ourselves to $M = 50$ groups of $N = 30$ individuals each. As expected, the average of the means of the

[14] In this case, the model would not be identified since there would be more parameters to estimate than there are observations.

[15] Available at: https://github.com/ahoundetoungan/PartialNetwork.

posterior distributions are centred relatively on the parameters' true values. Note, however, that due to the smaller number of groups and the fact that we performed only 200 simulations, the results in Table 1.4 may exhibit small sample as well as simulation biases.

Table 1.4 – Simulation results with a Bayesian method

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 30$, $M = 50$					
Estimation results					
$Intercept = 2$	1.873	0.893	1.312	1.815	2.486
$\alpha = 0.4$	0.398	0.025	0.383	0.398	0.414
$\beta_1 = 1$	1.003	0.027	0.982	1.002	1.019
$\beta_2 = 1.5$	1.500	0.019	1.489	1.501	1.512
$\gamma_1 = 5$	5.011	0.167	4.909	5.009	5.117
$\gamma_2 = -3$	−2.987	0.135	−3.084	−2.983	−2.887
$\sigma^2 = 1$	1.018	0.113	0.946	1.016	1.089

Simulation results for 200 replications of the model with unobserved exogenous effects estimated by a Bayesian method where the graph precision parameter λ is set to 1.

1.5 Network Formation Models

Our main assumption (Assumption A.2.) is that the researcher has access to the true distribution of the observed network. An important special case is when the researcher has access to a consistent estimate of this distribution. For concreteness, in this section we assume that links are generated as follows:

$$\mathbb{P}(a_{ij} = 1) \propto \exp\{Q(\boldsymbol{\theta}, \mathbf{w}_{ij})\}, \tag{1.5}$$

where Q is some known function, $\mathbf{w}_{ij}$ is a vector of (not necessarily observed) characteristics for the pair ij, and $\boldsymbol{\theta}$ is a vector of parameters to be estimated.

An important feature of such models is that their estimation may not necessarily require the observation of the entire network structure. To understand the intuition, assume a simple logistic regression framework:

$$\mathbb{P}(a_{ij} = 1) = \frac{\exp\{\mathbf{x}_{ij}\boldsymbol{\theta}\}}{1 + \exp\{\mathbf{x}_{ij}\boldsymbol{\theta}\}},$$

where $\mathbf{x}_{ij}$ is a vector of *observed* characteristics of the pair ij. Here, note that $\mathbf{s} = \sum_{ij} a_{ij}\mathbf{x}_{ij}$ is a vector of sufficient statistics. In practice, this therefore means that the estimation of $\boldsymbol{\theta}$ only requires the observation of such sufficient statistics.

To clarify this point, consider a simple example where individuals are only characterized by their gender and age. Specifically, assume that $\mathbf{x}_{ij} = [1, \mathbb{1}\{gender_i = gender_j\}, |age_i - age_j|]$.

Then, the set of sufficient statistics is resumed by (1) the number of links, (2) the number of same-gender links, and (3) average age difference between linked individuals.

Note that these statistics are much easier (and cheaper) to obtain than the entire network structure; however, they nonetheless allow for estimating the distribution of the true network.

Of course, in general, the simple logistic regression above might be unrealistically simple as the probability of linking might depend on unobserved variables.

In this section, we discuss some examples of network formation models that can be estimated using only partial information about the network. We subdivide such models into two categories: models that can be estimated using sampled network data and latent surface models.

1.5.1 Sampled Network

As discussed in Chandrasekhar and Lewis (2011), sampled data can be used to estimate a network formation model under the assumptions that (1) the sampling is exogenous and (2) links are conditionally independent, i.e. $P(a_{ij}|\mathbf{A}_{-ij}) = P(a_{ij})$, as in (1.5).

Indeed, if the sampling was done, for example, as a function of the network structure, the estimation of the network formation model would likely be biased. Also, if the network formation model is such that links are *not* conditionally independent, then consistent estimation usually requires the observation of the entire network structure.[16]

An excellent illustration of a compatible sampling scheme is presented in Conley and Udry (2010). Rather than collecting the entire network structure, the authors asked the respondents about their relationship with a random sample of the other respondents: "Have you ever gone to ____ for advice about your farm?" If the answer is "Yes," then a link is assumed between the respondents.

Since the pairs of respondents for which the "Yes/No" question is asked are random, the estimation of a network formation model with conditionally independent links gives consistent estimates. If, in addition, the individual characteristics of the sampled pair of respondents cover the set of observable characteristics for the entire set of respondents, one can compute the predicted probability that any two respondents are linked.

For concreteness, consider the simple model presented above, such that:

$$\mathbb{P}(a_{ij} = 1) = \frac{\exp\{\mathbf{x}_{ij}\boldsymbol{\theta}\}}{1 + \exp\{\mathbf{x}_{ij}\boldsymbol{\theta}\}},$$

where $\mathbf{x}_{ij} = [1, \mathbb{1}\{gender_i = gender_j\}, |age_i - age_j|]$.

[16] Or at least requires additional network summary statistics, such as individual degree or clustering coefficients; see Boucher and Mourifié (2017) or Mele (2017).

Then, as long as the random sample of pairs for which the "Yes/No" question is asked includes both men and women and includes individuals of different ages, then these sampled pairs allow for a consistent estimation of $\boldsymbol{\theta}$. As such, for any two respondents the (predicted) probability of a link is given by $\hat{p}_{ij} = \exp\{\mathbf{x}_{ij}\hat{\boldsymbol{\theta}}\}/(1+\exp\{\mathbf{x}_{ij}\hat{\boldsymbol{\theta}}\})$, where $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$.

The argument can be extended to models featuring an unobserved degree of heterogeneity. Specifically, Graham (2017) studies the following *undirected* network formation model:

$$\mathbb{P}(a_{ij}=1) = \frac{\exp\{\mathbf{x}_{ij}\boldsymbol{\theta}+\nu_i+\nu_j\}}{1+\exp\{\mathbf{x}_{ij}\boldsymbol{\theta}+\nu_i+\nu_j\}},$$

where ν_i and ν_j are unobserved. He presents a *tetrad logit* estimator based on the assumption that only a random sample of links are observed (his Assumption 2), as in Conley and Udry (2010).

Graham (2017) shows that $\boldsymbol{\theta}$ can be recovered consistently given some regularity conditions on the asymptotic behaviour of the model (his Assumption 4, which is implied by our Assumption A.3.). Once the consistent estimator $\hat{\boldsymbol{\theta}}$ is recovered, the predicted probabilities are given by:

$$\hat{P}(\mathbf{A}|\mathbf{n}) = \frac{\exp\{\sum_{ij:j<i} a_{ij}\mathbf{x}_{ij}\hat{\boldsymbol{\theta}}\}}{\sum_{B\in\mathcal{A}}\exp\{\sum_{ij:j<i} b_{ij}\mathbf{x}_{ij}\hat{\boldsymbol{\theta}}\}}, \tag{1.6}$$

where $\mathbf{n} = [n_1, ..., n_n]$ is the *degree sequence*, and $\mathcal{A}$ is the set of adjacency matrices that have the same degree sequence as $\mathbf{A}$, i.e. $n_i = \sum_{j\neq i} a_{ij} = \sum_{j\neq i} b_{ij}$ for all i and all $\mathbf{B}\in\mathcal{A}$ (see Graham (2017), equation (3)).

Note that computing $\hat{P}(\mathbf{A})$ therefore requires knowledge of the degree sequence, but this information can easily be incorporated as a survey question: "How many people have you gone to for advice about your farm?" Also, as noted by Graham (2017), the computation of the normalizing term in (1.6) is not tractable for networks of moderate size. As such, the predicted probabilities cannot be computed directly and must be simulated, for example using the sequential importance sampling algorithm proposed by Blitzstein and Diaconis (2011).

1.5.2 Latent Surface Models

Recently, McCormick and Zheng (2015) and Breza et al. (2020) have proposed a novel approach for the estimation of network formation models represented by:

$$\mathbb{P}(a_{ij}=1) \propto \exp\{\nu_i+\nu_j+\zeta\mathbf{z}_i'\mathbf{z}_j\}, \tag{1.7}$$

where ν_i, ν_j, ζ, $\mathbf{z}_i$, and $\mathbf{z}_j$ are not observed by the econometrician but follow parametric distributions. As in Graham (2017), ν_i and ν_j can be interpreted as i and j's propensity to create links, irrespective of the identity of the other individual involved. The other component, $\zeta\mathbf{z}_i'\mathbf{z}_j$, is meant to capture homophily on an abstract latent space (e.g. Hoff et al. (2002)).

Breza et al. (2020) show that it is possible to use aggregate relational data (ARD) to recover the values of the variables in (1.7) and therefore obtain an estimate of $\mathbb{P}(a_{ij} = 1)$. ARD are obtained from survey questions such as: "How many friends with trait 'X' do you have?" We refer the interested reader to McCormick and Zheng (2015) and Breza et al. (2020) for a formal discussion of the model. Here, we discuss the intuition using a simple analogy.

Suppose that individuals are located according to their geographical position on Earth. Suppose also that there are a fixed number of cities on Earth in which individuals can live. The econometrician does not know the individuals' location on Earth nor do they know the location of the cities. In model (1.7), $\mathbf{z}_i$ represent i's position on Earth.

Suppose that the researcher has data on ARD for a subset of the population. In the context of our example, ARD data are count variables of the type: "How many of your friends live in city A?"[17] Given (1.7) and parametric assumptions for the distribution of ν' and $\mathbf{z}$'s, the goal is to use ARD responses to infer the positions and sizes of the cities on Earth, as well as the values for ν_i and $\mathbf{z}_i$.[18]

To understand the intuition behind the identification of the model, consider the following example: suppose that individual i has many friends living in city A. Then, city A is likely located close to i's location. Similarly, if many individuals have many friends living in city A, then city A is likely a large city. Finally, if i has many friends from many cities, i likely has a large ν_i.

As mentioned above, we refer the interested reader to McCormick and Zheng (2015) and Breza et al. (2020) for a formal description of the method as well as formal identification conditions. Here, we provide Monte Carlo simulations for the estimators developed in Section 1.3, assuming that the true network follows (1.7). The details of the Monte Carlo simulations can be found in Appendix A.4.

We simulate 20 groups of 250 individuals each. Within each subpopulation, we simulate the ARD responses as well as a series of observable characteristics (e.g. cities). We then estimate the model in (1.7) and compute the implied probabilities, $\mathbb{P}(a_{ij} = 1)$, which we used as the distribution of our true network.[19] We estimate peer effects using the instrumental variable strategy presented in Section 1.3. Results are presented in Tables 1.5 and 1.6.

Results show that the method performs relatively well when $\mathbf{Gy}$ is observed but slightly less well when $\mathbf{Gy}$ is not observed and when the model allows for group-level unobservables. Note, however, that one potential issue with this specific network formation model is that it is based

[17] The general approach works for any discrete characteristic.

[18] One also needs the ARD traits of the entire population, which is similar to our Assumption A.4.. See Section C.I, Step II in Breza et al. (2020) for details.

[19] We fix $\zeta = 1.5$ (i.e. ζ is not estimated) to mitigate part of the small sample bias. See our discussion below.

on a single population setting (i.e. there is only one Earth). The researcher should keep in mind that the method should only be used on medium- to large-sized groups.

If the method proposed by Breza et al. (2020) performs well, note that our instrumental variable estimator does not require the identification of the structural parameters in (1.7). Indeed, the procedure only requires a consistent estimate of the linking probabilities.

As such, we could alternatively use the approach recently proposed by Alidaee et al. (2020). They present an alternative estimation procedure for models with ARD that does not rely on the parametric assumption in equation (1.7). They propose a penalized regression based on a low-rank assumption. One main advantage of their estimator is that it allows for a wider class of model and ensures that the estimation is fast and easily implementable.[20]

As for most penalized regressions, the estimation requires the user to select a tuning parameter, which effectively controls the weight of the penalty. We found that the value recommended by the authors is too large in the context of (1.7), using our simulated values. Since the choice of this tuning parameter is obviously context dependent, we recommend choosing it using a cross-validation procedure.

To explore the properties of their estimator in our context, we do the following. First, we simulate data using (1.7), using the same specification as for Tables 1.5 and 1.6. Second, we estimate the linking probabilities using their penalized regression under different tuning parameters, including the optimal (obtained through cross-validation) and the recommended parameter (taken from Alidaee et al. (2020)). Third, we estimate the peer-effect model using our instrumental variable estimator.

Table A.8 of Appendix A.3 presents the results under alternative tuning parameters when $\mathbf{Gy}$ is observed and $\boldsymbol{\gamma} = 0$. We see that the procedure performs well but is less precise than when using the parametric estimation procedure. This is intuitive since the estimation procedure in Breza et al. (2020) imposes more structure (and is specified correctly in our context). The procedure proposed by Alidaee et al. (2020) is valid for a large class of models but is less precise.

Tables A.9 and A.10 also exemplify the results of Propositions 1.1 and 1.2. When $\mathbf{Gy}$ is observed, the estimation is precise, even if the network formation model is not estimated precisely. However, when $\mathbf{Gy}$ is not observed, small sample bias strongly affects the performance of the estimator.

Results from this section imply that, when $\mathbf{Gy}$ is observed, the estimator proposed by Alidaee et al. (2020) is the most attractive since it is less likely to be misspecified (and correlated with $\boldsymbol{\varepsilon}$). However, when $\mathbf{Gy}$ is not observed, the estimator from Breza et al. (2020) should be privileged. Of course, in the latter case, the validity of the results are based on the assumption

[20] The authors developed user-friendly packages in R and Python. See their paper for links and details.

that (1.7) is correctly specified.

Table 1.5 – Simulation results using ARD with contextual effects (1,000 replications)

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 250$, $M = 20$ - Instrument: $(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is observed					
Estimation results					
$Intercept = 2$	1.991	0.222	1.845	1.992	2.141
$\alpha = 0.4$	0.400	0.006	0.396	0.400	0.404
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.005	1.496	1.500	1.504
$\gamma_1 = 5$	5.000	0.020	4.986	4.999	5.013
$\gamma_2 = -3$	−2.999	0.032	−3.020	−2.998	−2.977
Tests					
F-test	5473.171	1735.035	4232.103	5325.774	6528.537
Hausman	0.986	1.346	0.106	0.475	1.291
Sargan	1.045	1.461	0.108	0.465	1.353
$N = 250$, $M = 20$ - Instrument: $(\hat{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is not observed					
$Intercept = 2$	2.065	0.327	1.852	2.051	2.275
$\alpha = 0.4$	0.399	0.008	0.394	0.400	0.405
$\beta_1 = 1$	1.002	0.003	1.000	1.002	1.004
$\beta_2 = 1.5$	1.499	0.006	1.495	1.499	1.503
$\gamma_1 = 5$	5.411	0.020	5.397	5.411	5.425
$\gamma_2 = -3$	−2.403	0.040	−2.429	−2.402	−2.375
$\hat{\gamma}_1 = 0$	−0.383	0.023	−0.399	−0.384	−0.367
$\hat{\gamma}_2 = 0$	−0.608	0.038	−0.635	−0.609	−0.583
Tests					
F-test	4790.020	1407.596	3760.700	4682.825	5686.841
Hausman	70.940	19.503	57.143	70.430	82.175
Sargan	1.167	1.615	0.103	0.523	1.534
F-test	3,867.077	1,093.165	3,037.776	3,855.692	4,588.458
Hausman	228.290	49.002	194.110	227.981	261.617
Sargan	26.953	13.515	17.184	25.380	34.583

Results without contextual effects are presented in Table A.7 of Appendix A.3.

1.6 Imperfectly Measured Networks

In this section, we assume that the econometrician has access to network data but that the data may contain errors. For example, Hardy et al. (2019) assume that some links are missing with some probability, while others are included falsely with some other probability.

To show how our method can be used to address these issues, we consider a simple example where we are interested in estimating peer effects on adolescents' academic achievements. We assume that we observe the network but that some links are missing.

Table 1.6 – Simulation results with ARD and subpopulation unobserved fixed effects (1,000 replications)

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 250$, $M = 20$ - Instrument: $\mathbf{J}(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\mathbf{Gy}$ is observed					
Estimation results					
$\alpha = 0.4$	0.401	0.054	0.366	0.400	0.436
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
$\gamma_1 = 5$	4.999	0.067	4.956	4.999	5.043
$\gamma_2 = -3$	−3.001	0.082	−3.051	−2.999	−2.949
Tests					
F-test	169.973	55.243	130.603	165.188	205.152
Hausman	0.978	1.306	0.096	0.438	1.360
Sargan	0.919	1.393	0.084	0.402	1.228
$N = 250$, $M = 20$ - Instrument: $\mathbf{J}(\hat{\mathbf{G}})^2\mathbf{X}$ - $\mathbf{Gy}$ is not observed					
Estimation results					
$\alpha = 0.4$	0.477	0.077	0.426	0.474	0.528
$\beta_1 = 1$	1.001	0.003	0.999	1.001	1.003
$\beta_2 = 1.5$	1.499	0.007	1.495	1.499	1.504
$\gamma_1 = 5$	5.401	0.024	5.386	5.400	5.416
$\gamma_2 = -3$	−2.399	0.040	−2.425	−2.401	−2.374
$\hat{\gamma}_1 = 0$	−0.467	0.087	−0.524	−0.465	−0.408
$\hat{\gamma}_2 = 0$	−0.721	0.119	−0.795	−0.721	−0.641
Tests					
F-test	125.898	43.107	94.650	122.580	150.436
Hausman	23.096	11.431	15.100	21.238	29.881
Sargan	1.096	1.611	0.131	0.479	1.369

To estimate this model, we use the AddHealth database. Specifically, we focus on a subset of schools from the "In School" sample that each have less than 200 students. Table 1.7 displays the summary statistics.

Most of the papers estimating peer effects that use this particular database have taken the network structure as given. One notable exception is Griffith (2019), looking at the issue of top coding.[21] In practice, even if the schools are meant to be entirely sampled, we are still potentially missing many links. To understand why, we discuss the organization of the data.

Each adolescent is assigned to a unique identifier. The data includes ten variables for the ten potential friendships (maximum of 5 male and 5 female friends). These variables can contain missing values (no friendship was reported), an error code (the named friend could not be found in the database), or an identifier for the reported friends. This data is then used to

[21] Although we are not exploring this issue here, our method can also be applied to analyse the impact of top coding.

Table 1.7 – Summary statistics

Statistic	Mean	Std. Dev.	Pctl(25)	Pctl(75)
Female	0.540	0.498	0	1
Hispanic	0.157	0.364	0	0
Race				
White	0.612	0.487	0	1
Black	0.246	0.431	0	0
Asian	0.022	0.147	0	0
Other	0.088	0.283	0	0
Mother education				
High	0.310	0.462	0	1
$<$ High	0.193	0.395	0	0
$>$ High	0.358	0.480	0	1
Missing	0.139	0.346	0	0
Mother job				
Stay-at-home	0.225	0.417	0	0
Professional	0.175	0.380	0	0
Other	0.401	0.490	0	1
Missing	0.199	0.399	0	0
Age	13.620	1.526	13	14
GPA	2.088	0.794	1.5	2.667

generate the network's adjacency matrix $\mathbf{A}$.

Of course, error codes cannot be matched to any particular adolescent; as well, even in the case where the friendship variable refers to a valid identifier, however, the referred adolescent may still be absent from the database. A prime example is when the referred adolescent has been removed from the database by the researcher, perhaps due to other missing variables for these particular individuals. These missing links are quantitatively important as they account for roughly 30% of the total number of links (7,830 missing for 17,993 observed links). Figure 1.4 displays the distribution of the number of "unmatched named friends".[22]

Given that we observe the number of missing links for each individual, we can use the general estimator proposed in Section 1.4 to estimate the model. For this, however, we need one additional assumption.

Let N_s be the size of school s, n_i be the number of observed (matched) friends of i, and $\check{n}_i$ be the number of missing (unmatched) friends of i. We make the following assumption for the network formation process:

$$\begin{aligned} \mathbb{P}(a_{ij}=1) &= 1 \text{ if the link is observed in the data,} \\ \mathbb{P}(a_{ij}=1) &= \frac{\check{n}_i}{N_s-1-n_i} \text{ otherwise.} \end{aligned}$$

In large schools, this is equivalent to assuming that missing friendships are drawn at random

[22] We focus on within-school friendships; thus, nominations outside of school are discarded.

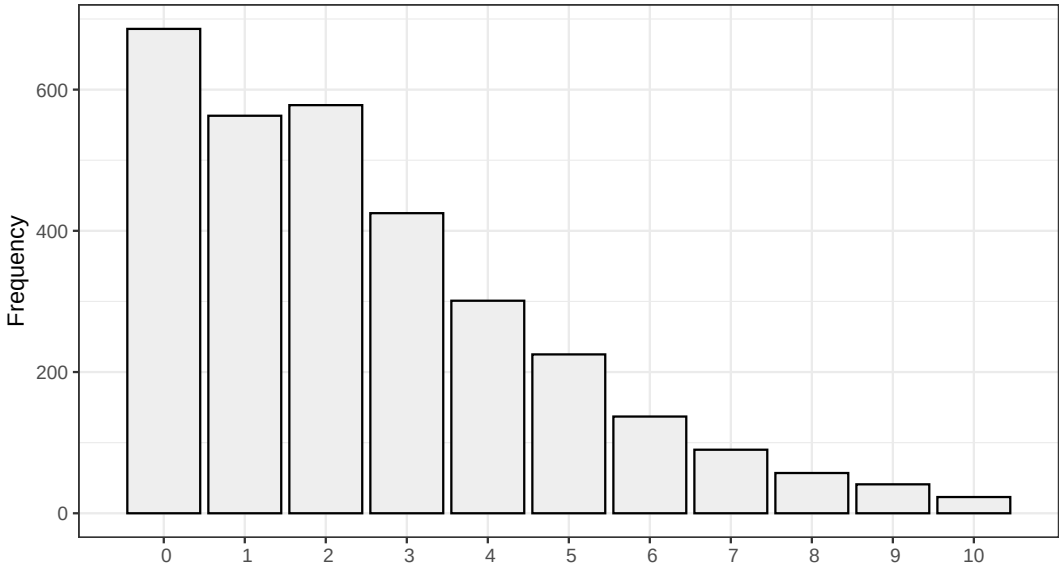


Figure 1.4 – Frequencies of the number of missing links per adolescent

among the remaining adolescents. In small schools, however, this approach disregards the dependence between links.

To understand why we cannot directly assume that the $\check{n}_i$ missing links are assigned randomly to the adolescents with whom i is not a friend, consider the following simple example. Suppose that there are only three adolescents, i, j, and k, and that i has no matched (observed) friends, but we know that they have one friend (i.e. $\check{n}_i = 1$). This friend can either be j or k. Selecting friendship relations at random therefore implies that a_{ij} and a_{ik} are perfectly (negatively) correlated. This will lead many Metropolis–Hastings algorithms to fail. In particular, this is the case of the Gibbs sampling procedure.[23]

Assuming $\mathbb{P}(a_{ij} = 1) = \mathbb{P}(a_{ik} = 1) = \dfrac{\check{n}_i}{N_s - 1 - n_i} = 1/2$ for unobserved links circumvents this issue. Moreover, since this assumption only affects the prior distribution of our Bayesian inference procedure, it need not have important consequences on the posterior distribution. Table 1.8 presents the estimation results.[24] Importantly, we see that the estimated value for α is significantly larger when the network is reconstructed. Also notable is the fact that the reconstructed network captures an additional contextual peer effect: having a larger fraction of Hispanic friends significantly reduces academic achievement.[25] The remainder of the estimated parameters are roughly the same for both specifications.

[23] This issue could be solved by updating an entire line of $\mathbf{A}$ for each step of the algorithm. However, this proves to be computationally intensive for networks of moderate size.

[24] Trace plots and posterior distributions are presented in Figures A.1, A.2, and A.3 of Appendix A.9.

[25] Note that one has to be careful in discussing the structural interpretation of the contextual effects. See Boucher and Fortin (2016) for a discussion.

Table 1.8 – Posterior distribution

	Observed network			Reconstructed network		
Statistic	Mean	Std. Dev.	t-stat	Mean	Std. Dev.	t-stat
Peer effect	0.350^{***}	0.022	15.519	0.526^{***}	0.036	14.468
Intercept	1.196^{***}	0.132	9.086	1.153^{***}	0.139	8.293
Own effects						
Female	-0.144^{***}	0.029	-5.013	-0.131^{***}	0.028	-4.690
Hispanic	0.084^{**}	0.042	1.999	0.122^{***}	0.044	2.745
Race						
Black	0.231^{***}	0.045	5.084	0.233^{***}	0.047	4.913
Asian	0.090	0.090	1.008	0.106	0.087	1.212
Other	-0.056	0.051	-1.085	-0.044	0.052	-0.862
Mother education						
$<$ High	0.122^{***}	0.039	3.125	0.120^{***}	0.038	3.136
$>$ High	-0.139^{***}	0.033	-4.165	-0.111^{***}	0.034	-3.317
Missing	0.060	0.051	1.179	0.058	0.052	1.131
Mother job						
Professional	-0.081^{*}	0.044	-1.833	-0.068	0.044	-1.551
Other	-0.003	0.035	-0.078	0.009	0.034	0.253
Missing	0.066	0.047	1.411	0.067	0.047	1.426
Age	0.073^{***}	0.009	7.749	0.076^{***}	0.010	7.824
Contextual effects						
Female	-0.012	0.049	-0.234	-0.048	0.071	-0.679
Hispanic	-0.061	0.069	-0.876	-0.275^{***}	0.091	-3.042
Race						
Black	-0.051	0.058	-0.868	-0.097	0.065	-1.478
Asian	-0.212	0.184	-1.150	-0.131	0.324	-0.404
Other	0.138	0.090	1.539	0.154	0.140	1.099
Mother education						
$<$ High	0.269^{***}	0.072	3.733	0.337^{***}	0.109	3.093
$>$ High	-0.071	0.060	-1.196	-0.118	0.086	-1.377
Missing	0.078	0.094	0.828	0.013	0.146	0.089
Mother job						
Professional	0.109	0.081	1.352	0.060	0.115	0.519
Other	0.101^{*}	0.060	1.684	0.057	0.088	0.647
Missing	0.093	0.087	1.080	0.055	0.136	0.400
Age	-0.066^{***}	0.006	-11.583	-0.087^{***}	0.008	-11.178
SE^2	0.523			0.493		

N = 3,126. Observed links = 17,993. Missing links = 7,830. Significance levels: *** = 1%, ** = 5%, * = 10%.

1.7 Discussions

In this paper, we proposed two types of estimators that can estimate peer effects given that the researcher only has access to the distribution of the true network. In doing so, we abstracted from many important considerations. In this section, we discuss some limits, some areas for future research, and some general implications of our results.

1.7.1 Endogenous Networks

In this paper, we assumed away any endogeneity of the network structure (Assumption A.5.). As discussed in Section 1.2, this is done for the purposes of presentation. Indeed, there are multiple ways to introduce, and correct for, endogenous networks, which lead to many possible models. In this section, we discuss how existing endogenous network corrections can be adapted to our setting. Specifically, we assume that there exists some unobserved variable correlated with both the network $\mathbf{A}$ and the outcome $\mathbf{y}$. This violates Assumption A.5..

A first remark is that our ability to accommodate such an unobserved variable depends on the flexibility of the used network formation model. Indeed, some models can obtain estimates of the unobserved heterogeneity (e.g. Breza et al. (2020) or Graham (2017)). In such cases, one could simply include the estimated unobserved variables as additional explanatory variables. Johnsson and Moon (2015) discuss this approach as well as other control-function approaches in detail. Since their instrumental variable estimator is based on the higher-order link relations (i.e. $\mathbf{G}^2\mathbf{X}, \mathbf{G}^3\mathbf{X}, ...$), it is also valid for the instrumental variable estimator proposed in Section 1.3.[26]

However, this is not entirely satisfying since it assumes that the network formation is estimated consistently, independent of the peer-effect model.[27] The Bayesian estimator presented in Section 1.4 allows for more flexibility. Indeed, one could expand Algorithm 1.1. and perform the estimation of the network formation model jointly with the peer-effect model, as has been done, for instance, by Goldsmith-Pinkham and Imbens (2013), Hsieh and Van Kippersluis (2018), and Hsieh et al. (2020)) in contexts where the network is observed. Essentially, instead of relying on the known distribution $P(\mathbf{A})$, one would need to rely on $P(\mathbf{A}|\mathbf{S}, \boldsymbol{\kappa}, \mathbf{z})$, where $\mathbf{S}$ is a matrix (possibly a vector) of observed statistics about $\mathbf{A}$ (e.g. a sample, a vector of summary statistics, or ARD), $\mathbf{z}$ is an unobserved latent variable (correlated with $\boldsymbol{\varepsilon}$), and $\boldsymbol{\kappa}$ is a vector of parameters to be estimated.

This contrasts, for example, with the network formation model presented in Section 1.5.2 where $\mathbf{S}$ (i.e. ARD) is sufficient for the estimation of all of the models' parameters. Here, the estimation of $\mathbf{z}$ and $\boldsymbol{\kappa}$ also requires knowledge about the likelihood of $\mathbf{y}$. The exploration of

[26] Note that estimated unobserved variables can also be added as explanatory variables in the context of the estimator presented in Section 1.4.

[27] See, for example, Assumption 1 and Assumptions 6–11 in Johnsson and Moon (2015).

such models (especially their identification) goes far beyond the scope of the current paper and is left for future (exciting) research.

1.7.2 Large Populations and Partial Sampling

In this section, we discuss the estimation of (1.1) when the population cannot be partitioned into groups of bounded sizes (i.e. when Assumption A.3. does not hold). Note that doing so will also most likely violate Assumption A.4.. Indeed, in large populations (e.g. cities, countries...), assuming that the individuals' characteristics $(\mathbf{y}, \mathbf{X})$ are observed for the entire population is unrealistic (that is, except for census data). This implies that strategies such as the one presented in Section 1.4 would also be unfeasible, irrespective of the network formation model.[28]

One would therefore have to rely on an instrumental strategy, such as the one presented in Section 1.3. In this section, we discuss the properties of the estimator in Section 1.3 in the context of a single, large sample.

To fix the discussion, assume for now that $\mathbf{x}_i = x_i$ has only one dimension and only takes on a finite number of distinct values. Assume also that for any two i, j, $P(a_{ij} = 1) = \phi(x_i, x_j)$ for some known function ϕ. We have:

$$(\mathbf{Gx})_i = \sum_{j=1}^{N} g_{ij} x_j,$$

which we can rewrite as:

$$\sum_{j=1}^{N} g_{ij} x_j = \frac{\sum_x x (n_x/N) \frac{1}{n_x} \sum_{j:x_j=x} a_{ij}}{\sum_x (n_x/N) \frac{1}{n_x} \sum_{j:x_j=x} a_{ij}},$$

where n_x is the number of individuals having the trait x. Note that since we assumed that the support of x only takes a finite number of values, n_x goes to infinity with N. We therefore have (by a strong law of large numbers):

$$\frac{1}{n_x} \sum_{j:x_j=x} a_{ij} \to \phi(x_i, x), \tag{1.8}$$

and $n_x/N \to p(x)$, where $p(x)$ is the fraction of individuals with trait x in the population. We therefore have:

$$(\mathbf{Gx})_i \to \frac{\sum_x x p(x) \phi(x_i, x)}{\sum_x p(x) \phi(x_i, x)} \equiv \hat{x}(x_i).$$

For example, for an Erdös–Rényi network (i.e. $\phi(x, x') = \bar{\phi}$ for all x, x'), we have $\hat{x}(x_i) = \mathbb{E}x$.

This means that the knowledge of $\phi(\cdot, \cdot)$ is sufficient to construct a *consistent* estimate of $\mathbf{Gx}$. Note also that a similar argument allows constructing a consistent estimate for $\mathbf{Gy}$ or $\mathbf{G}^2\mathbf{X}$.

[28] Also, in terms of computational cost, the estimator presented in Section 1.4 is likely to be very costly for very large populations.

As such, the instrumental variable strategy proposed in Section 1.3 can be applied even if $\mathbf{GX}$ is not observed.

Unfortunately, this approach relies on the (perhaps unrealistic) assumption that $P(a_{ij} = 1) = \phi(x_i, x_j)$, which here implies that individuals form an (asymptotically) infinite number of links.[29] When the number of links is bounded (e.g. De Paula et al. (2018b)), the average in (1.8) does not converge for each i. Note, however, that the first part of Proposition 1.2 still applies: if $\mathbf{Gy}$ and $\mathbf{GX}$ are observed, then the constructed (biased) instrument $\mathbf{H}^2\mathbf{X}$ drawn from a network formation model with bounded degrees is valid.

Finally, note that the argument presented here can be generalized. In particular, Parise and Ozdaglar (2019) recently proposed a means of approximating games on large networks using *graphon games*, i.e. games played directly on the network formation model. If the approach is promising, its implications for the estimation of peer effects go far beyond the scope of this paper and are left for future research.

1.7.3 Survey Design

As discussed in Section 1.3, instrumental variable estimators are only valid if the researcher observes $\mathbf{GX}$. Also, Breza et al. (2020) and Alidaee et al. (2020) propose using ARD to estimate network formation models. Importantly, although ARD responses and $\mathbf{GX}$ are similar, they are not equivalent. For example, consider a binary variable (e.g. gender). One can obtain $\mathbf{GX}$ by asking questions such as "What fraction of your friends are female?" For ARD, the question would be "How many of your friends are female?" This suggests asking two questions. One related to the number of female friends and one related to the number of friends.[30]

For continuous variables (e.g. age), this creates additional issues. One can obtain $\mathbf{GX}$ by asking about the average age of one's friends, but ARD questions must be discrete: "How many of your friends are in the same age group as you?" Then, in practice, an approach could be to ask individuals about the number of friends they have, as well as the number of friends they have from multiple age groups: "How many of your friends are between X and Y years old?" Using this strategy allows construction of both the ARD and $\mathbf{GX}$.

Finally, an implication of Propositions 1.1 and 1.2 is that asking directly for $\mathbf{Gy}$ in the survey leads to a more robust estimation strategy. Indeed, the constructed instruments are valid even if the network formation model is misspecified.

[29] A special case, when the network is complete, is presented in Brock and Durlauf (2001).

[30] Breza et al. (2020) do not require information on the number of friends, although this significantly helps the estimation.

1.7.4 Next Steps

In this paper, we proposed two estimators where peer effects can be estimated without having knowledge of the entire network structure. We found that, perhaps surprisingly, even very partial information on network structure is sufficient. However, there remains many important challenges, in particular with respect to the study of compatible models of network formation.

Chapter 2

Count Data Models with Social Interactions under Rational Expectations

Résumé

Dans ce chapitre, je présente un modèle structurel des effets de pairs pour analyser une variable de comptage (nombre de cigarettes fumées, fréquence des visites au restaurant, fréquence de participation aux activités). Le modèle est basé sur un jeu statique à information incomplète dans lequel les joueurs interagissent à travers un réseau dirigé et sont influencés par leur croyance sur le choix de leurs amis. Je présente des conditions suffisantes sous lesquelles l'équilibre du jeu est unique. J'estime les paramètres du modèle en utilisant une approche the *vraisemblance partielle imbriquée*. Je montre que l'utilisation du modèle spatial autorégressif (SAR) linéaire-en-moyennes ou du modèle Tobit SAR pour estimer les effets de pairs sur des variables de comptage générées à partir du jeu sous-estime asymptotiquement les effets de pairs. Le biais d'estimation diminue lorsque la dispersion de la variable de comptage augmente. Je propose également une application empirique. J'estime les effets de pairs sur le nombre d'activités parascolaires auxquelles les étudiants sont inscrits. Je trouve que l'augmentation d'une unité du nombre d'activités dans lesquelles les amis d'un étudiant sont inscrits implique une augmentation du nombre d'activités dans lesquelles l'étudiant est inscrit de 0,295, en contrôlant l'endogénéité du réseau. Je montre également que les effets de pairs sont sous-estimés à 0,150 lorsqu'on ignore la nature de comptage de la variable dépendante.

Abstract

I present a structural model of peer effects to analyze count data (Number of cigarettes smoked, frequency of restaurant visits, frequency of participation in activities). The model is based

on a static game with incomplete information in which individuals' outcome is counting. In addition, individuals interact through a directed network and are influenced by their belief over the choice of their peers. I provide sufficient conditions under which the equilibrium of the game is unique. I estimate the model's parameters using the Nested Partial Likelihood method. I show that using the standard linear-in-means spatial autoregressive (SAR) model or the SAR Tobit model to estimate peer effects on counting variables generated from the game asymptotically underestimates the peer effects. The estimation bias decreases when the range of the dependent variable increases. I estimate peer effects on the number of extracurricular activities in which students are enrolled. I find that increasing the number of activities in which student's friends are enrolled by one implies an increase in the number of activities in which the student is enrolled by 0.295, controlling for the endogeneity of the network. I also show that the peer effects are underestimated at 0.150 when ignoring the counting nature of the dependent variable.

Keywords: Discrete model, Social networks, Bayesian game, Rational expectations, Network formation.

JEL Classification: C25, C31, C73, D84, D85.

2.1 Introduction

There is a large and growing literature on peer effects in economics.[1] Recent contributions include, among others, models for limited dependent variables, including binary (e.g., Brock and Durlauf, 2001; Lee et al., 2014; Liu, 2019), ordered (e.g., Boucher et al., 2018), multinomial (e.g., Guerra and Mohnen, 2020), and censored (e.g., Xu and Lee, 2015b) variables. To my knowledge, however, there are no existing models for count variables with microeconomic foundations, despite these variables being prevalent in survey data (e.g., Liu et al., 2012; Patacchini and Zenou, 2012; Fujimoto and Valente, 2013; Liu et al., 2014; Fortin and Yazbeck, 2015; Boucher, 2016; Lee et al., 2020a).

In this paper, I propose a network model in which the dependent variable is the number of occurrences of an event in a constant period.[2] The model generalizes the rational expectations model of Lee et al. (2014), which is used to study peer effects on binary data. I show that the model's parameters can be estimated using the Nested Partial Likelihood (NPL) method (Aguirregabiria and Mira, 2007). I show that using the linear-in-means spatial autoregressive (SAR) model (Lee, 2004) or the SAR Tobit (SART) model (Xu and Lee, 2015b) to estimate peer effects on counting variables generated from the model asymptotically underestimates the peer effects. The estimation bias decreases when the range of the dependent variable increases. I estimate peer effects on the number of extracurricular activities in which students are enrolled using the data set provided by the National Longitudinal Study of Adolescent Health (Add Health). I control for network endogeneity. I find that ignoring the endogeneity of the network overestimates the peer effects. Finally, I provide an easy-to-use R package—named `CDatanet`—for implementing the model.[3]

I present a static game with incomplete information (see Harsanyi, 1967; Osborne and Rubinstein, 1994) to rationalize the model. Individuals in the game interact through a directed network, simultaneously choose their strategy, and are influenced by their belief over the choice of their peers. As in many discrete games (e.g., Xu and Lee, 2015a; Liu, 2019), I assume that individuals do not directly choose the observed integer outcome. Instead, they choose a latent variable that can be interpreted as an intention. This latent variable determines the observed integer outcome (see also Maddala, 1986; Cameron and Trivedi, 2013).
I provide sufficient conditions under which the model game has a unique Bayesian Nash Equilibrium (BNE). To estimate the model parameters, I rely on the NPL algorithm proposed by Aguirregabiria and Mira (2007). The estimation process is straightforward and can be readily implemented. Moreover, it does not require computing the game equilibrium. I show that the estimator is consistent, and I study its limiting distribution.

[1] For recent reviews, see Boucher and Fortin (2016), De Paula (2017), and Bramoullé et al. (2020).

[2] Examples are number of cigarettes smoked, frequency of restaurant visits, frequency of participation in activities.

[3] The package is available at CRAN.R-project.org/package=CDatanet.

I show that modeling the counting dependent variable generated from the game through use of a misspecified continuous model, such as the SART model or the SAR, asymptotically underestimates the peer effects. The estimation bias decreases when the range of the dependent variable increases. In practice, the bias could almost disappear if the range of the dependent variable is sufficiently large. This result is also confirmed through Monte Carlo simulations.

I provide an empirical application. I use the Add Health data to estimate peer effects on the number of extracurricular activities in which students are enrolled. I find that increasing the number of activities in which a student's friends are enrolled by one implies an increase in the number of activities in which the student is enrolled by 0.295. As in the Monte Carlo study, I find that the SART and the SAR models underestimate peer effects at 0.141 and 0.166, respectively.

I control for the endogeneity of the network in the empirical application. Endogeneity is due to unobservable individual characteristics, such as the gregariousness or degree of sociability, which influence both link formation in the network and participation in extracurricular activities (see Johnsson and Moon, 2015; Graham, 2017). To deal with the endogeneity, I use a two-stage estimation strategy. In the first stage, I consider a dyadic linking model in which the probability of link formation between two students depends, among others, on their gregariousness (see Graham, 2017; Breza et al., 2020). Using a Markov Chain Monte Carlo (MCMC) approach, I simulate the posterior distribution of this gregariousness. In the second stage, the estimator of gregariousness is included in the count data model as a supplementary explanatory variable.[4] I find that the network is endogenous and that ignoring the endogeneity overestimates peer effects.

This paper contributes to the literature on social interaction models for limited dependent variables. The existing models deal with binary (e.g., Brock and Durlauf, 2001; Soetevent and Kooreman, 2007; Lee et al., 2014; Xu and Lee, 2015a; Liu, 2019), censored (e.g., Xu and Lee, 2015b), ordered (e.g., Boucher et al., 2018), and multinomial outcomes (e.g., Guerra and Mohnen, 2020). My model fits between the rational expectations model for binary data developed by Lee et al. (2014) and the SAR model used to study continuous outcomes. When the distribution of the outcome is almost degenerated, such that the outcome takes only two values, I show that the structure of my model game and the BNE are similar to those of Lee et al. (2014). In addition, when the outcome is not left-censored and its range is sufficiently large, I show that the model is similar to the SAR model.

The paper contributes to the extensive empirical literature on social interactions by being the first to deal with the counting nature of count data. Existing papers studying peer effects using count data rely on linear-in-means models estimated by the maximum likelihood approach

[4] I use the posterior distribution of the estimator of gregariousness to account for the uncertainty related to first-stage estimation in the second stage.

of Lee (2004) or the two-stage least squares method of Kelejian and Prucha (1998), which ignores the counting nature of the outcome (e.g., Liu et al., 2012; Patacchini and Zenou, 2012; Fujimoto and Valente, 2013; Liu et al., 2014; Fortin and Yazbeck, 2015; Boucher, 2016; Lee et al., 2020a). I show that peer effects estimated in this way are potentially biased downward. In my empirical application on students' participation in extracurricular activities, I account for the counting nature of the outcome.

Importantly, in the literature on spatial autoregressive models for limited dependent variables, cases of count data have been studied (e.g, Karlis, 2003; Liesenfeld et al., 2016; Inouye et al., 2017; Glaser, 2017). These papers consider reduced form equations in which the dependent count variable is spatially autocorrelated. However, the models are not based on any process (game) that explains how the individuals choose their strategy, and thus how they are influenced by their peers. Therefore, the reduced form cannot be interpreted as a best-response function, and the spatial dependence parameter cannot be interpreted as peer effects.

The paper also contributes to the literature on peer effects models with endogenous networks. Goldsmith-Pinkham and Imbens (2013) as well as Hsieh and Lee (2016) consider a Bayesian hierarchical model to control for endogeneity. They use a MCMC approach to jointly simulate from the posterior distribution of the network formation model parameters and the outcome model parameters. While this method is efficient as the estimation is done in a single step, it can be cumbersome to implement with a discrete data model. Johnsson and Moon (2015) also develop a strategy to estimate the linear-in-means peer effects model by controlling for the endogeneity of the network. Their estimation method is semiparametric and relies on a control function approach. My method to control for endogeneity is similar in spirit to that of Johnsson and Moon (2015) and can be readily implemented with discrete outcome models. The network formation model is estimated, in a first stage, separately from the outcome model estimation. Moreover, I provide a way to estimate the variance of the estimator of the outcome model, which takes into account the uncertainty of the estimation in the first stage.

The remainder of the paper is organized as follows. Section 2.2 presents the microeconomic foundation of the model based on an incomplete information network game. Section 2.3 addresses the identification and the estimation of the model parameters. It also presents the link between the model and the linear-in-means model. Section 2.4 documents the Monte Carlo experiments. Section 2.5 presents the empirical results and the method used to control for the endogeneity of the network. Section 2.6 discusses some limits and some general implications of the results. Section 2.7 concludes this paper.

2.2 Incomplete Information Network Game

I present a game of incomplete information with social interactions. Let $\mathcal{V} = \{1, \ldots, n\}$ be a set of n players indexed by i and y_i, the observed integer outcome of player i (e.g., the

number of cigarettes smoked per day or per week). The integer variable y_i is considered as a generalization of a binary variable (see Lee et al., 2014; Liu, 2019).[5] As in Xu and Lee (2015a) and Liu (2019), I assume that the players do not directly choose y_i. Instead, they choose y_i^*, a latent variable that determines the observed outcome y_i. This latent variable can be interpreted as an intention that leads to the observed choice y_i (see Maddala, 1986).
I assume that y_i^* and y_i are linked as follows:

Assumption B. *Let $(a_q)_{q\in\mathbb{N}}$ be a sequence given by $a_0 = -\infty$, $a_1 \in \mathbb{R}$, and $a_q = a_1 + \gamma(q-1)$ for $q \in \mathbb{N}^*$ and $\gamma \in \mathbb{R}_+^*$. If $y_i^* \in (a_q, a_{q+1}]$, then $y_i = q$.*

The outcome y_i is called the *count variable* or *count data*. As in a binary game (e.g., Liu, 2019), Assumption B sets $y_i = 0$ if y_i^* is not greater than some real value a_1. When $y_i^* > a_1$, Assumption B implies that there are increasing boundaries a_1, a_2, ..., such that $y_i = q$ if $y_i^* \in (a_q, a_{q+1}]$. A similar assumption is also set to link a polytomous ordered variable to a latent variable (e.g., Amemiya, 1981; Baetschmann et al., 2015; Boucher et al., 2018).
Assumption B restricts the boundaries to be equally spaced from a_1; that is, a_1, $a_1+\gamma$, $a_1+2\gamma$, and so on. This is stronger than the usual assumption for an ordered model, which allows the boundary increment to vary (see Amemiya, 1981). However, two important points motivate such simplification. First, it is intuitively natural to set that the boundaries increase uniformly by γ as the count variable y_i increases uniformly by 1. This allows to interpret y_i^* as a *ratio* variable and the model a as linear model.[6] Second, if the increment varies, then the number of unknown parameters increases with the number of values taken by y_i. In practice, estimating the model can be cumbersome when the outcome takes many values. As the count variable y_i is unbounded, Assumption B fixes in particular the incidental parameter issues, which could appear in the econometric model.
However, I also show that the proof of the Bayesian Nash Equilibrium (BNE) of the game could be readily generalized when one assumes a sequence with varying increments over i and q.[7]
Interestingly, Assumption B also generalizes the binary outcome game of Lee et al. (2014). Indeed, if $\gamma = \infty$, then $a_r = \infty$ for $r \geq 2$. In that case, y_i can only take 2 values: $y_i = 0$ if $y_i^* \leq a_1$, and $y_i = 1$ otherwise.

Individuals interact through a directed network. Let $\mathbf{G} = [g_{ij}]$ be an $n \times n$ adjacency matrix, where the (i,j)-th element is non-negative and captures the proximity of the individuals i and

[5] For example, when the binary variable is coded 0, y_i also takes 0, and when the binary variable is coded 1, y_i could take any strictly positive value.

[6] Unlike the case of an ordered variable, the latent variable increases at the same rate as the exposure time. For instance, $y_i^* = \alpha$ in a week is supposed to be equivalent to $y_i^* = \frac{\alpha}{7}$ in a day. Using a constant increment allows dealing with time-varying exposure (see Section 2.6.2).

[7] In that case, $\left(a_q^i\right)_{q\in\mathbb{N}}$ would be any strictly increasing sequence, such that if $y_i^* \in (a_q^i, a_{q+1}^i]$, then $y_i = q$. To generalize the equilibrium results of the game, I assume that $\lim_{q\to\infty} a_{q+1}^i - a_q^i > 0$, $\forall\, i \in \mathcal{V}$ (see Appendix B.1.3).

j in the network. I define the peers of individual i as the set of individuals $\mathcal{V}_i = \{j, g_{ij} > 0\}$. By convention, nobody interacts with himself/herself, that is $g_{ii} = 0 \; \forall \; i \in \mathcal{V}$.

I assume that the individuals' preferences can be characterized by the following linear-quadratic utility function:[8]

$$\mathcal{U}_i = \underbrace{(\psi_i + \varepsilon_i)\, y_i^* - \frac{y_i^{*2}}{2}}_{\text{private sub-utility}} + \underbrace{\lambda y_i^* \sum_{j \neq i} g_{ij} y_j}_{\text{social sub-utility}}, \tag{2.1}$$

where ψ_i, $\lambda \in \mathbb{R}$, and ε_i is an idiosyncratic shock that can be interpreted as the player's type. The term ψ_i captures observed characteristics of i.[9] I assume that the idiosyncratic shock ε_i is identically and independently distributed over i. The player i observed their own type ε_i but not that of the others. All players know the common distribution of ε_i.

The first two terms of the utility function (2.1) are the private subutility, in which $-\frac{1}{2} y_i^{*2}$ is the intention cost, and $\psi_i + \varepsilon_i$ is the own marginal benefit. The third term is a social sub-utility. It depends on the intention y_i^*, the average of the peers' outcomes $\sum_{j \neq i} g_{ij} y_j$, and the peer effects parameter λ. Importantly, each individual i chooses the intention y_i^*, but each is affected by their peers' outcomes y_j, $j \in \mathcal{V}$. As argued by Fortin and Boucher (2015), the utility function (2.1) describes *complementarity* in social interactions if $\lambda > 0$ and *substitutability* in social interactions if $\lambda < 0$. A similar utility function is used by Liu (2019) to model bivariate binary outcomes with social interactions.

Individuals observe neither the private information ε_j of their peers, nor do they then observe the outcome y_j of their peers. The utility function (2.1) characterizes a game of incomplete information (Bayesian game) in which the players form beliefs regarding their peers' outcomes. Moreover, as the players know the common distribution of their type ε_i, they form rational beliefs (see Lee et al., 2014; Liu, 2019). This implies that for any player $j \in \mathcal{V}$, any player $i \neq j$ puts the same probability on the event $\{y_j = q\}$, $q \in \mathbb{N}$. In addition, this probability is consistent with the common distribution of ε_j. Let p_{jq} be this probability; that is, $\forall \; j \in \mathcal{V}$, $q \in \mathbb{N}$, $p_{jq} = Prob\,(y_j = q | \boldsymbol{\psi}, \mathbf{G})$, where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$.

Individuals simultaneously choose their strategy y_i^* as to maximize their expected utilities.

$$\mathbf{E}\,(\mathcal{U}_i | y_i^*, \varepsilon_i, \lambda, \boldsymbol{\psi}, \mathbf{G}) = (\psi_i + \varepsilon_i)\, y_i^* - \frac{y_i^{*2}}{2} + \lambda y_i^* \sum_{j \neq i} g_{ij} \bar{y}_j, \tag{2.2}$$

where $\bar{y}_j = \sum_{r=0}^{\infty} r p_{jr}$ is the expectation of y_j with respect to the rational beliefs. For $\bar{y}_i$ to exist and be finite, I assume that the distribution of ε_i belongs to a specific class of distributions.

[8] The linear-quadratic specification of the utility function is common for network games (e.g., Ballester et al., 2006; Calvó-Armengol et al., 2009).

[9] For example, $\psi_i = \mathbf{x}_i' \boldsymbol{\beta}$, where $\mathbf{x}_i$ is a vector of observed characteristics and $\boldsymbol{\beta}$ is a vector of parameters.

Assumption C. *ε_i follows a continuous symmetric distribution having a cumulative distribution function (cdf) F_ε and a probability density function (pdf) $f_\varepsilon = o(1/x^\alpha)$ at ∞ for some $\alpha > 3$.*

The assumption of continuity is necessary so that ε_i has a continuous density function. The symmetry of this density function simplifies many equations. The condition $f_\varepsilon = o(1/x^\alpha)$ at ∞ for some $\alpha > 3$ implies that the probability of $y_i = q$ should decrease at some rate when q grows to infinity.[10] This condition plays an important role. It implies that $\bar{y}_i = \sum_{r=1}^{\infty} r q_{ir}$ exists and is finite. Many usual distributions suit Assumption C, such as normal, logistic, and student with a degree of freedom greater than 2, ...

The first-order conditions (f.o.cs) of the expected utility maximization imply that

$$y_i^* = \lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i + \varepsilon_i, \tag{2.3}$$

where $\forall\ i \in \mathcal{V}$, $\mathbf{g}_i = (g_{i1} \dots g_{in})$, $\bar{\mathbf{y}} = (\bar{y}_1\ \dots \bar{y}_n)'$. Equation (2.3) shows that an individual's intention is explained linearly by the average of their peers' expected outcomes.

Let $\mathbf{y}^* = (y_1^*\ \dots y_n^*)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1\ \dots\ \varepsilon_n)'$. The f.o.cs (2.3) is also equivalent to

$$\mathbf{y}^* = \lambda \mathbf{G} \bar{\mathbf{y}} + \boldsymbol{\psi} + \boldsymbol{\varepsilon}. \tag{2.4}$$

For any $q \in \mathbb{N}$, I denote by $\mathbf{p}_\mathrm{q} = (p_{1q}, \dots, p_{nq})'$, an n-dimensional vector of the probabilities that $y_1 = q$, ..., $y_n = q$. Let also $\mathbf{p} = \left(\mathbf{p}_0', \mathbf{p}_1', \mathbf{p}_2', \mathbf{p}_3', \dots\right)'$, an infinite-dimensional vector of beliefs. The f.o.cs (2.3) imply that any vector of beliefs $\mathbf{p}$ characterizes a BNE (see Osborne and Rubinstein, 1994) of the game with the utility (2.1) if

$$\forall\ i \in \mathcal{V}, q \in \mathbb{N},\ p_{iq} = F_\varepsilon\left(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_q\right) - F_\varepsilon\left(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_{q+1}\right). \tag{2.5}$$

Equation (2.5) can also be expressed as $\mathbf{p} = \mathbf{H}(\mathbf{p})$, where $\mathbf{H}$ is some mapping that depends on λ, $\boldsymbol{\psi}$, $\mathbf{G}$, and F_ε. Finding belief systems that verify this equation amounts to computing the fixed points of $\mathbf{H}$. However, since $\mathbf{H}$ is defined from an infinite space to itself, establishing the conditions for the existence of a unique fixed point is challenging. In addition, computing the fixed points would be cumbersome in practice.

Equation (2.5) also implies that the knowledge of the expected outcome $\bar{\mathbf{y}}$ at the equilibrium is sufficient to compute the equilibrium beliefs $\mathbf{p}$ and vice versa. This result has a very useful implication: to prove the uniqueness of the equilibrium beliefs, it is sufficient to prove that the expected equilibrium outcome is unique.[11] Moreover, as the expected outcome $\bar{\mathbf{y}}$ is an n-dimensional vector, this simplifies the establishment of uniqueness conditions.

Importantly, the expected outcome $\bar{\mathbf{y}}$ at equilibrium also verifies a fixed-point equation as stated by the following proposition.

[10] Note that this condition does not imply that the probability of $y_i = q$ is null for some $q \in \mathbb{N}$.

[11] I show that the vector of equilibrium beliefs $\mathbf{p}$ exists (which implies the existence of an expected outcome $\bar{\mathbf{y}}$ at equilibrium) and that there is at most one expected equilibrium outcome $\bar{\mathbf{y}}$.

Proposition 2.1. *Let* $\mathbf{L}(\bar{\mathbf{y}}) = (\ell_1(\bar{\mathbf{y}}) \ \ldots \ell_n(\bar{\mathbf{y}}))'$, *where* $\ell_i(\bar{\mathbf{y}}) = \sum_{r=1}^{\infty} F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_r)$ *for all* $i \in \mathcal{V}$. *Under Assumptions B and C, the expected outcome* $\bar{\mathbf{y}}$ *at the equilibrium verifies* $\bar{\mathbf{y}} = \mathbf{L}(\bar{\mathbf{y}})$.

Proof. See Appendix B.1.1. □

Proposition 2.1 states that any n-dimensional vector $\bar{\mathbf{y}}^e$, which is an expected outcome at equilibrium, is also a fixed point of the mapping $\mathbf{L}$. To find sufficient conditions for $\mathbf{L}$ to have a unique fixed point, I show that $\mathbf{L}$ is a contracting mapping under the following assumption.

Assumption D. $|\lambda| < \dfrac{C_\gamma}{||\mathbf{G}||_\infty}$, *where* $C_\gamma = \left(\max_{u \in \mathbb{R}} \sum_{k=-\infty}^{\infty} f_\varepsilon(u + \gamma k) \right)^{-1}$.

Assumption D sets a maximal value that the peer effects parameter cannot exceed. This assumption also generalizes the restriction imposed on $|\lambda|$ in other models. For instance, in the case of the binary model ($\gamma = \infty$), Assumption D implies that $|\lambda| < \dfrac{1}{||\mathbf{G}||_\infty f_\varepsilon(0)}$, which is the restriction set on $|\lambda|$ in the rational expectation models for binary data developed by Lee et al. (2014) and Liu (2019).

In the case of the binary model, if $f_\varepsilon(0) < 1$ and $\mathbf{G}$ is row-normalized ($||\mathbf{G}||_\infty = 1$), Assumption D is not too restrictive in practice because it is weaker than $|\lambda| < 1$.[12] In the general case, the upper bound of $|\lambda|$ depends on the assumed distribution of ε_i. In Section 2.3.1, I discuss the implication of Assumption D when $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$.

The following theorem establishes the existence and uniqueness of the pure strategy BNE of the incomplete information network game.

Theorem 2.1. *Under Assumptions B, C, and D, the incomplete information network game with the utility (2.1) has a unique pure strategy BNE with the equilibrium strategy profile* $\mathbf{y}^{e*}$, *given by* $\mathbf{y}^{e*} = \lambda \mathbf{G} \bar{\mathbf{y}}^e + \boldsymbol{\psi} + \boldsymbol{\varepsilon}$, *where* $\bar{\mathbf{y}}^e = (\bar{y}_1^e \ldots \bar{y}_n^e)$ *is the unique solution of* $\bar{\mathbf{y}} = \mathbf{L}(\bar{\mathbf{y}})$.

Proof. See Appendix B.1.2. □

There are two important remarks concerning Theorem 2.1. First, the model generalizes the rational expectations model proposed by Lee et al. (2014) for discrete binary outcomes. Indeed, if $\gamma = \infty$, then $p_{ir} = 0$ for $r \geq 2$ and $i \in \mathcal{V}$. As a result, $\bar{y}_i = \sum_{r=0}^{\infty} r p_{ir} = p_{i1} \ \forall \ i \in \mathcal{V}$, and $\bar{\mathbf{y}} = \mathbf{p}_1$, where $\mathbf{p}_1 = (p_{11} \ \ldots \ p_{n1})$. Under these considerations, Assumptions C and D still ensure that the game has a unique BNE with the equilibrium strategy $\mathbf{y}^{e*}$, given by

[12] In practice, it is generally assumed that $|\lambda| < 1$, as individuals will not experience an increase in their intention/outcome greater than the increase in their peers' outcomes.

$\mathbf{y}^{e*} = \lambda \mathbf{G}\mathbf{p}_1^e + \boldsymbol{\psi} + \boldsymbol{\varepsilon}$, where $p_{i1}^e = f_\varepsilon\left(\lambda \mathbf{g}_i \mathbf{p}_1^e + \psi_i - a_1\right)$ for all $i \in \mathcal{V}$. This characterization of the equilibrium is the same as that of Lee et al. (2014).

Second, the equilibrium belief is not necessary to compute the equilibrium strategy. The knowledge of $\bar{\mathbf{y}}^e$, the expected outcome at equilibrium, is sufficient to compute $\mathbf{y}^{e*}$, the equilibrium strategy , and $\mathbf{p}^e$, the equilibrium belief. This result is important in practice as it simplifies the model estimation.

I also generalize the uniqueness of the BNE when the increment of the sequence $(a_q)_{q\in\mathbb{N}}$ varies (see Appendix B.1.3). However, this raises an important issue in practice, as it implies an infinite number of parameters to estimate. Additional assumptions must be considered for a consistent estimate of the model.

Theorem 2.1 guarantees that the mapping $\mathbf{L}$ has a unique fixed point, which is sufficient to compute the BNE. This also suggests using the Nested Pseudo Likelihood (NPL) algorithm proposed by Aguirregabiria and Mira (2007) to estimate the model. In the next section, I study the parameter identification and present the model estimation strategy.

2.3 Econometric Model

This section presents the identification and estimation strategy of the model. It also studies the link between the model and the SAR and SART models.

My strategy to estimate the model parameters relies on the likelihood approach. This requires being specific about the distribution of ε_i, as in Lee et al. (2014), Xu and Lee (2015b), Liu (2019), ... Given that the expected outcome at equilibrium depends on the cdf F_ε, it is very challenging to obtain a consistent estimator of the model parameters without specifying this cdf. Later, in Section 2.3.3, I discuss a particular case where a General Method of Moment (GMM) could be used as alternative estimation strategy that does not require specifying a distribution.

Assumption E. $\varepsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$.

The choice of the normal distribution is natural since this facilitates the comparison of this model with the SART and SAR models, which also consider a normal distribution. In addition, the normal distribution allows dealing with the endogeneity of the network (see also Hsieh and Lee, 2016).

2.3.1 Identification

In this section, I describe restrictions on the model parameters that are necessary to ensure identifiability. Let $\boldsymbol{\psi} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X} = (\mathbf{x}_1 \ \dots \ \mathbf{x}_n)'$ is an $n \times K$-dimensional matrix of explanatory variables, and $\boldsymbol{\beta}$ is a K-dimensional vector of unknown parameters. The matrix $\mathbf{X}$ may also include the average of the explanatory variables among peers; that is, $\boldsymbol{\psi} = \tilde{\mathbf{X}}\boldsymbol{\beta}$, where $\tilde{\mathbf{X}} = [\mathbf{X}, \ \mathbf{G}\mathbf{X}]$. The coefficients of $\mathbf{G}\mathbf{X}$ represent the contextual effects (Manski, 1993).

To identify the model parameters, I assume that the matrix of the explanatory variables is a full rank matrix.

Assumption F. *Let* $\mathbf{Z} = [\mathbf{G}\bar{\mathbf{y}},\ \mathbf{X}]$. $\mathbf{Z}$ *is a full rank matrix.*

The BNE characterization (2.5) becomes

$$p_{iq} = \Phi\left(\frac{\lambda \mathbf{g}_i \bar{\mathbf{y}} + \mathbf{x}_i'\boldsymbol{\beta} - a_q}{\sigma_\varepsilon}\right) - \Phi\left(\frac{\lambda \mathbf{g}_i \bar{\mathbf{y}} + \mathbf{x}_i'\boldsymbol{\beta} - a_{q+1}}{\sigma_\varepsilon}\right), \tag{2.6}$$

where Φ is the cdf of $\mathcal{N}(0,1)$.
As $a_0 = -\infty$, and $a_q = a_1 + \gamma(q-1)$ for $q \in \mathbb{N}^*$,

$$p_{iq} = \begin{cases} 1 - \Phi\left(\dfrac{\lambda \mathbf{g}_i \bar{\mathbf{y}} + \mathbf{x}_i'\boldsymbol{\beta} - a_1}{\sigma_\varepsilon}\right) & \text{if } q = 0, \\ \Phi\left(\dfrac{\lambda \mathbf{g}_i \bar{\mathbf{y}} + \mathbf{x}_i'\boldsymbol{\beta} - a_1 - \gamma(q-1)}{\sigma_\varepsilon}\right) - \Phi\left(\dfrac{\lambda \mathbf{g}_i \bar{\mathbf{y}} + \mathbf{x}_i'\boldsymbol{\beta} - a_1 - \gamma q}{\sigma_\varepsilon}\right) & \text{if } q \in \mathbb{N}^*. \end{cases} \tag{2.7}$$

Estimating the model requires additional restrictions on the parameters. Equation (2.7) poses two identification issues. First, Equation (2.7) does not change when λ, $\boldsymbol{\beta}$, a_1, γ, and σ_ε are multiplied by any positive number. To fix this identification issue, I set γ to one.[13] Second, if the explanatory variables include a constant, such that $\mathbf{x}_i'\boldsymbol{\beta} = \beta_1 + x_{2i}\beta_2 + \dots x_{Ki}\beta_K$, the parameters β_1 and a_1 cannot be identified because they enter the equation only through their difference. Therefore, I also set $a_1 = 0$. Following these restrictions, Assumption B can be simplified.

Assumption B'. *Let* $(a_q)_{q\in\mathbb{N}}$ *be a sequence given by* $a_0 = -\infty$, $a_q = q - 1$ *for* $q \in \mathbb{N}^*$. *If* $y_i^* \in (a_q, a_{q+1}]$, *then* $y_i = q$.

Under Assumptions B', D, E, and F, the parameters $\boldsymbol{\theta} = \left(\lambda, \boldsymbol{\beta}', \sigma_\varepsilon\right)'$ are identified. Indeed, given the adjacency matrix $\mathbf{G}$ and the exogenous variable $\mathbf{X}$, the parameters $\boldsymbol{\theta} = \left(\lambda, \boldsymbol{\beta}', \sigma_\varepsilon\right)'$ and the alternative parameters $\tilde{\boldsymbol{\theta}} = \left(\tilde{\lambda}, \tilde{\boldsymbol{\beta}}', \tilde{\sigma}_\varepsilon\right)'$ are equivalent if they lead to the same BNE equilibrium; that is $\bar{\mathbf{y}} = \tilde{\bar{\mathbf{y}}}$, where $\bar{\mathbf{y}}$ and $\tilde{\bar{\mathbf{y}}}$ are the expected outcomes associated with $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$, respectively. In addition, Theorem 2.1 ensures that $\bar{\mathbf{y}}$ and $\tilde{\bar{\mathbf{y}}}$ are uniquely determined by the fixed point mappings. Then,

$$\begin{aligned} \bar{\mathbf{y}} = \sum_{r=1}^{\infty} \Phi\left(\frac{\lambda \mathbf{g}_i \bar{\mathbf{y}} + \mathbf{x}_i'\boldsymbol{\beta} - a_r}{\sigma_\varepsilon}\right) = \sum_{r=1}^{\infty} \Phi\left(\frac{\tilde{\lambda} \mathbf{g}_i \tilde{\bar{\mathbf{y}}} + \mathbf{x}_i'\tilde{\boldsymbol{\beta}} - a_r}{\tilde{\sigma}_\varepsilon}\right),\ \forall\ i \in \mathcal{V}, \\ \left(\frac{\lambda}{\sigma_\varepsilon} - \frac{\tilde{\lambda}}{\tilde{\sigma}_\varepsilon}\right) \mathbf{g}_i \bar{\mathbf{y}} + \mathbf{x}_i' \left(\frac{\boldsymbol{\beta}}{\sigma_\varepsilon} - \frac{\tilde{\boldsymbol{\beta}}}{\tilde{\sigma}_\varepsilon}\right) + q\left(\frac{1}{\sigma_\varepsilon} - \frac{1}{\tilde{\sigma}_\varepsilon}\right) = 0,\ \forall\ i \in \mathcal{V}, q \in \mathbb{N}. \end{aligned} \tag{2.8}$$

As $\mathbf{Z}$ is a full rank matrix, it follows from Equation (2.8) that $\sigma_\varepsilon = \tilde{\sigma}_\varepsilon$, $\lambda = \tilde{\lambda}$, and $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$. Therefore, $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

[13] Alternatively, I could also set σ_ε to one. However, this complicates the comparison of the model with the SAR and SART models. Moreover, the restriction $\gamma = 1$ excludes the binary cases for which $\gamma = \infty$.

With the assumed distribution of ε_i, one can quantify the upper bound of $|\lambda|$. Assume that $\mathbf{G}$ is row-normalized ($||\mathbf{G}||_\infty = 1$). Under Assumptions B' and E, the upper bound of $|\lambda|$ set in Assumption D is

$$C_{1,\sigma_\varepsilon} = \frac{\sigma_\varepsilon}{\phi(0) + 2\sum_{k=1}^{\infty} \phi\left(\frac{k}{\sigma_\varepsilon}\right)}, \tag{2.9}$$

where ϕ is the pdf of $\mathcal{N}(0,1)$.[14]

Figure 2.1 plots C_{1,σ_ε} as a function of σ_ε. One can notice that $C_{1,\sigma_\varepsilon} \approx 1$ if $\sigma_\varepsilon > 0.5$. In that case, Assumption D is not much stronger than $|\lambda| < 1$. In contrast, when $\sigma_\varepsilon < 0.5$, Assumption D implies a stronger restriction. However, the condition $\sigma_\varepsilon < 0.5$ is likely violated in practice when $\gamma = 1$. Indeed, σ_ε is the standard deviation of y_i^* conditional on $\mathbf{Z}$. As y_i^* takes values in disjoint intervals of range γ, the standard deviation must be sufficiently large for y_i^* to span several intervals. If σ_ε is too low, y_i will be likely constant given $\mathbf{Z}$.

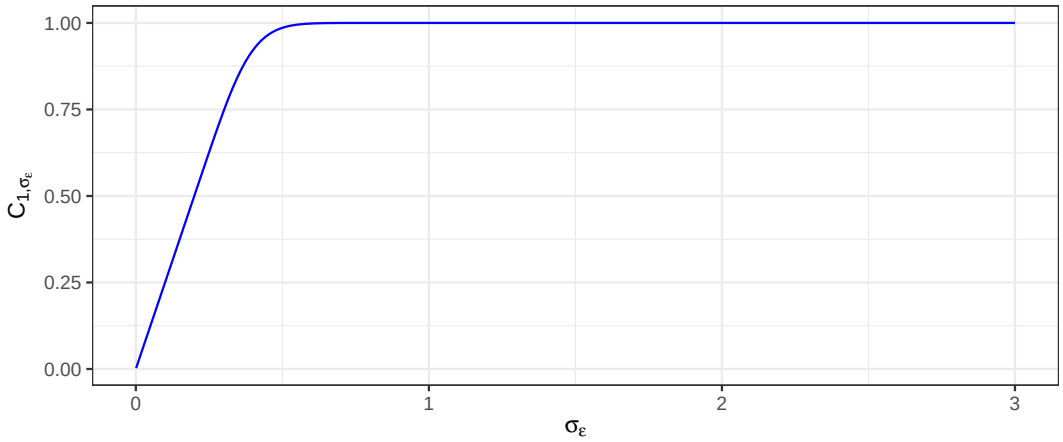


Figure 2.1 – C_{1,σ_ε}, upper bound of λ when $\gamma = 1$ as a function of σ_ε

In the next section, I present the strategy used to estimate $\boldsymbol{\theta}$, and I study the limiting distribution of the estimator.

2.3.2 Estimation

The estimation strategy is based on the NPL algorithm proposed by Aguirregabiria and Mira (2007) and recently used by Lin and Xu (2017) and Liu (2019). If $\bar{\mathbf{y}}$ were observed, estimating the model would result in a simple *probit* estimation by the maximum likelihood (ML) method. As $\bar{\mathbf{y}}$ is not observed, the ML estimation requires computing $\bar{\mathbf{y}}$; that is, solve a fixed point problem in $\mathbb{R}^n$ for each proposal of $\boldsymbol{\theta}$. This may be computationally cumbersome for large samples. The NPL algorithm uses an iterative process and does not require solving a fixed point problem.

[14] I also show that C_{1,σ_ε} can be evaluated using the third Theta function (see Section 2 in Bellman, 2013) available in most software (see Appendix B.1.4).

Let $\mathcal{L}$ be the pseudo likelihood[15] function in $(\boldsymbol{\theta}, \bar{\mathbf{y}})$, defined as

$$\mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}}) = \sum_{i=1}^{n} \sum_{r=0}^{\infty} d_{ir} \log(p_{ir}), \tag{2.10}$$

where $p_{iq} = \Phi\left(\frac{\lambda \mathbf{g}_i \bar{\mathbf{y}} + \mathbf{x}_i'\boldsymbol{\beta} - a_q}{\sigma_\varepsilon}\right) - \Phi\left(\frac{\lambda \mathbf{g}_i \bar{\mathbf{y}} + \mathbf{x}_i'\boldsymbol{\beta} - a_{q+1}}{\sigma_\varepsilon}\right) \; \forall \; i \in \mathcal{V}$, $q \in \mathbb{N}$, and $d_{ir} = 1$ if $y_i = r$, and $d_{ir} = 0$ otherwise. As I set above that $\boldsymbol{\psi} = \mathbf{X}\boldsymbol{\beta}$, the mapping $\mathbf{L}$ can be redefined as $\mathbf{L}(\bar{\mathbf{y}}, \boldsymbol{\theta}) = (\ell_1(\bar{\mathbf{y}}, \boldsymbol{\theta}) \; \ldots \ell_n(\bar{\mathbf{y}}, \boldsymbol{\theta}))'$, where

$$\ell_i(\bar{\mathbf{y}}, \boldsymbol{\theta}) = \sum_{r=1}^{\infty} \Phi\left(\frac{\lambda \mathbf{g}_i \bar{\mathbf{y}} + \mathbf{x}_i'\boldsymbol{\beta} - a_r}{\sigma_\varepsilon}\right) \quad \text{for all } i \in \mathcal{V}. \tag{2.11}$$

The NPL algorithm consists of starting with a proposal $\bar{\mathbf{y}}_0$ for $\bar{\mathbf{y}}$ and constructing a sequence of estimators $(\mathcal{Q}_m)_{m \geq 1}$, defined as $\mathcal{Q}_m = \{\boldsymbol{\theta}_m, \bar{\mathbf{y}}_m\}$ for $m \geq 1$, where $\boldsymbol{\theta}_m = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}}_{m-1})$ is the estimator of $\boldsymbol{\theta}$ at the m-th stage, and $\mathbf{y}_m = \mathbf{L}(\bar{\mathbf{y}}_{m-1}, \boldsymbol{\theta}_m)$ is the estimator of $\bar{\mathbf{y}}$ at the m-th stage. In other words, given the guess $\bar{\mathbf{y}}_0$, $\boldsymbol{\theta}_1 = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}}_0)$ and $\mathbf{y}_1 = \mathbf{L}(\bar{\mathbf{y}}_0, \boldsymbol{\theta}_1)$; then $\boldsymbol{\theta}_2 = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}}_1)$, $\mathbf{y}_2 = \mathbf{L}(\bar{\mathbf{y}}_1, \boldsymbol{\theta}_2)$, ...
The sequence $\mathcal{Q}_m$ is well defined for any $m > 1$. Notice that each value of $\mathcal{Q}_m$ requires evaluating the mapping $\mathbf{L}$ only once. If $(\mathcal{Q}_m)_{m \geq 1}$ converges, regardless of the initial guess $\bar{\mathbf{y}}_0$, its limit $\{\hat{\boldsymbol{\theta}}, \hat{\bar{\mathbf{y}}}\}$ satisfies the following two properties: $\hat{\boldsymbol{\theta}}$ maximizes the pseudo likelihood $\mathcal{L}(\boldsymbol{\theta}, \hat{\bar{\mathbf{y}}})$ and $\hat{\bar{\mathbf{y}}} = \mathbf{L}(\hat{\boldsymbol{\theta}}, \hat{\bar{\mathbf{y}}})$.
As shown by Kasahara and Shimotsu (2012), a key determinant of the convergence of the NPL algorithm is the contraction property of the fixed point mapping $\mathbf{L}$ guaranteed by Theorem 2.1. In practice, when $||\hat{\boldsymbol{\theta}}_M - \hat{\boldsymbol{\theta}}_{M-1}||_1$ and $||\hat{\bar{\mathbf{y}}}_M - \hat{\bar{\mathbf{y}}}_{M-1}||_1$ are less than some tolerance values (for example 10^{-6}), I set $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_M$ and $\hat{\bar{\mathbf{y}}} = \hat{\bar{\mathbf{y}}}_M$. Aguirregabiria and Mira (2007) prove that the NPL estimator is root-n consistent and asymptotically normal. I adapt their proof to my framework. The convergence and the limiting distribution of $\hat{\boldsymbol{\theta}}$ are given by the following proposition.

Proposition 2.2. *Under regularity conditions (see Proposition 2 of Aguirregabiria and Mira, 2007), the NPL estimator $\hat{\boldsymbol{\theta}}$ is consistent, and*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(0, \; (\boldsymbol{\Sigma}_0 + \boldsymbol{\Omega}_0)^{-1} \boldsymbol{\Sigma}_0 \left(\boldsymbol{\Sigma}_0' + \boldsymbol{\Omega}_0'\right)^{-1}\right), \tag{2.12}$$

where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$; $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Omega}_0$ are given in Appendix B.2.1.

Proof. See Appendix B.2.1. □

Some numerical aspects about the NPL estimator must be pointed out. First, the pseudo likelihood (2.10) involves an infinite sum. However, as $d_{iq} = 0$ for any $q \neq y_i$, this pseudo

[15] This is a pseudo likelihood because it is defined for any $\boldsymbol{\theta}$ and $\bar{\mathbf{y}}$, where $\bar{\mathbf{y}}$ is not necessary, the equilibrium expected outcome associated with $\boldsymbol{\theta}$ (see Aguirregabiria and Mira, 2007).

likelihood can also be expressed as $\mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}}) = \sum_{i=1}^{n} \log(p_{iy_i})$. Second, the mapping $\mathbf{L}$, which is used to compute the sequence $(\mathcal{Q}_m)$ and the asymptotic variance of $\hat{\boldsymbol{\theta}}$, also involves an infinite sum. However, note that the summed elements decrease exponentially. A very good approximation of these sums can be readily reached by only summing a few elements.

2.3.3 Comparison with other models

In this section, I compare the reduced form of the expected outcome to that of the Poisson model. I also make the link between the count variable model and the SAR and SART models, which are often used in empirical studies to estimate peer effects with count data.

Reduced form of the expected outcome

One of the most commonly used models to study count data is the Poisson model, in which the expected outcome has an exponential form with respect to the explanatory variables (see Cameron and Trivedi, 2013). Note that the exponential form of the Poisson model essentially prevents negative expected outcomes and is not based on microeconomic foundations.
From Proposition 2.1, the expected outcome of the new count variable model is given by

$$\bar{y}_i = \sum_{r=1}^{\infty} \Phi\left(\frac{\lambda \mathbf{g}_i \bar{\mathbf{y}} + \mathbf{x}_i' \boldsymbol{\beta} - a_r}{\sigma_\varepsilon}\right). \tag{2.13}$$

The reduced form (2.13) also prevents negative values in the expected outcome $\bar{\mathbf{y}}$. However, this specification is different from that of the Poisson model. I show in Section 2.6.1 that the new count variable model is flexible in terms of dispersion fitting, as it allows equidispersion, overdispersion, and underdispersion.

My specification is different from that of the Poisson model because motivating a network game with an outcome that follows a Poisson distribution is challenging. For example, this would require specifying a utility function with an exponential form in the game. Such a utility function is not common in network games. Moreover, another specification of the sequence $(a_q)_{q\in\mathbb{N}}$ having increments that decrease exponentially leads to an expected outcome with an exponential form. However, as discussed in Section 2.2, a uniform increment is more appropriate when the model is compared with a linear model. Indeed, if $\sigma_\varepsilon > 0.5$, the reduced form (2.13) is *nearly* linear and similar to the expected outcome of the Tobit model (see Figure 2.2).

Links with the SAR and SART models

Assume that the researcher estimates a SAR or SART model, whereas the counting variable is generated from the game described by the utility function (2.1). Unlike the SAR model,

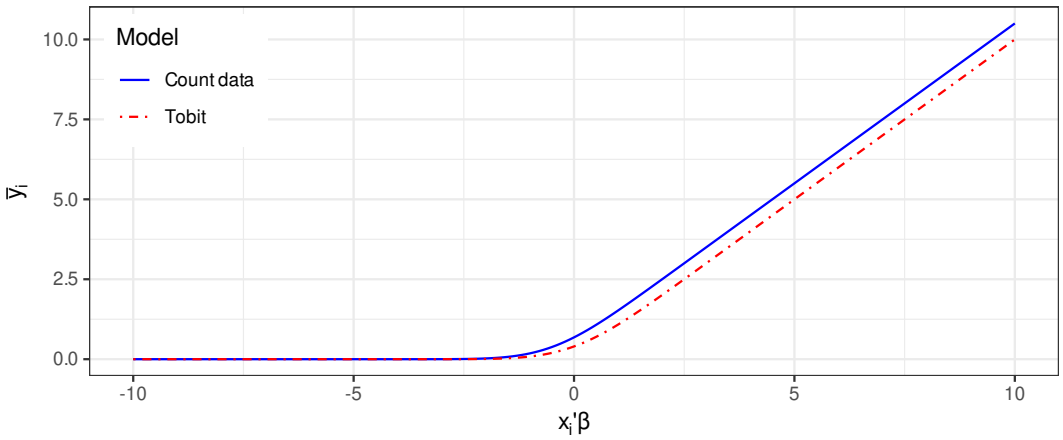


Figure 2.2 – Expected outcome at $\lambda = 0$ and $\sigma_\varepsilon = 1$

the SART model controls for the left-censoring nature of the dependent variable (see Xu and Lee, 2015a). I assume that y_i takes values as large as possible so that one can consider that y_i is non-censored. In this case, the SAR and the SART are almost equivalent, and the results of the comparison of the counting variable model to the SAR model could also be generalized to the SART model.
Let us recall the f.o.cs (2.3).

$$y_i^* = \lambda \mathbf{g}_i \bar{\mathbf{y}} + \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i. \tag{2.14}$$

The reduced form of the linear SAR model is given by

$$y_i = \tilde{\lambda} \mathbf{g}_i \mathbf{y} + \mathbf{x}_i' \tilde{\boldsymbol{\beta}} + \nu_i. \tag{2.15}$$

When y_i in Equation (2.15) is generated from the game described by the utility function (2.1), I show that the standard MLE of $\tilde{\lambda}$ is generally asymptotically biased.

Proposition 2.3. *The MLE of the parameter $\tilde{\lambda}$, based on the assumption that $\nu_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\nu^2\right)$, where σ_ν^2 is an unknown parameter, is inconsistent.*

Proof. See Appendix B.2.2. □

The inconsistency of the MLE is due to a heteroskedasticity in (2.15) that is not taken into account. Indeed,

$$\nu_i = \varepsilon_i + \lambda \mathbf{g}_i \boldsymbol{\eta} - \zeta_i, \tag{2.16}$$

where $\boldsymbol{\eta} = \bar{\mathbf{y}} - \mathbf{y}$ and $\zeta_i = y_i^* - y_i$. The heteroskedasticity is caused by the term $\mathbf{g}_i \boldsymbol{\eta}$, which comes from the approximation of the expected outcome $\bar{\mathbf{y}}$ on the right side of Equation (2.14) by the actual outcome $\mathbf{y}$ in Equation (2.15). Because the individuals have private information in the counting variable model, they are influenced by the expected choice and not by the actual choice of their friends.

As the covariance structure of $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)'$ is not known, the maximum likelihood (ML) method cannot be used. However, $\tilde{\lambda}$ can be estimated consistently using the General Method of Moment (GMM) with unknown heteroskedasticity as developed by Lin and Lee (2010). This approach takes into account the unobserved covariance structure of $\boldsymbol{\nu}$ in the case of the SAR model. Nevertheless, the GMM estimator does not account for the left-censoring nature of y_i. This estimator may be significantly biased in finite sample when data contain many zeros.

The two-stage least square (2SLS) estimator (see Kelejian and Prucha, 1998) of the model (2.15) also leads to biased estimations. Importantly, I show that the bias is downward and decreases when the range of the dependent variable increases.

Assume for simplicity that $\mathbf{X}$ is a column vector of ones.[16] In this case, the 2SLS estimator of $\tilde{\lambda}$ is

$$\hat{\tilde{\lambda}}_{2SLS} = \frac{\frac{1}{n}\sum_{i=1}^{n} \tilde{y}_i(\mathbf{g}_i\tilde{\mathbf{y}}) - \hat{\tilde{y}}(\hat{\mathbf{g}}\tilde{\mathbf{y}})}{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{g}_i\tilde{\mathbf{y}})^2 - (\hat{\mathbf{g}}\tilde{\mathbf{y}})^2},$$

where $\tilde{y}_i = \mathbf{P}_{\mathbf{Z}i}\mathbf{y}$, $\mathbf{g}_i\tilde{\mathbf{y}} = \mathbf{P}_{\mathbf{Z}i}\mathbf{G}\mathbf{y}$, $\mathbf{P}_{\mathbf{Z}i}$ is the i-th row of $\mathbf{P}_{\mathbf{Z}}$, $\hat{\tilde{y}} = \frac{1}{n}\sum_{i=1}^{n}\tilde{y}_i$, $\hat{\mathbf{g}}\tilde{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i\tilde{\mathbf{y}}$, and $\mathbf{P}_{\mathbf{Z}} = \mathbf{Z}\left(\mathbf{Z}\mathbf{Z}\right)^{-1}\mathbf{Z}'$.

Proposition 2.4. *The probability limit of* $\hat{\tilde{\lambda}}_{2SLS}$ *is*

$$\operatorname{plim}\hat{\tilde{\lambda}}_{2SLS} = \lambda - \lambda \operatorname{plim}\frac{\frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}\left(\mathbf{g}_i\tilde{\mathbf{y}}|\mathbf{X},\mathbf{G},\mathbf{Z}\right)}{\frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}\left(\mathbf{g}_i\tilde{\mathbf{y}}\right)}. \tag{2.17}$$

Proof. See Appendix B.2.3. □

The estimator $\hat{\tilde{\lambda}}_{2SLS}$ is biased downward. Proposition 2.4 also implies that the bias of the 2SLS estimator decreases when $\mathbf{Var}\left(\mathbf{g}_i\tilde{\mathbf{y}}\right)$ increases and $\mathbf{Var}\left(\mathbf{g}_i\tilde{\mathbf{y}}|\mathbf{X},\mathbf{G},\mathbf{Z}\right)$ is fixed. Note that the conditional variance $\mathbf{Var}\left(\mathbf{g}_i\tilde{\mathbf{y}}|\mathbf{X},\mathbf{G},\mathbf{Z}\right)$ does not increase with the range of y_i if σ_ε^2 is constant. Indeed, $\mathbf{Var}\left(\mathbf{g}_i\tilde{\mathbf{y}}|\mathbf{X},\mathbf{G},\mathbf{Z}\right) = \mathbf{P}_{\mathbf{Z}i}\mathbf{G}\,\mathbf{Var}\left(\mathbf{y}|\mathbf{X},\mathbf{G}\right)\mathbf{G}'\mathbf{P}'_{\mathbf{Z}i}$, where $\mathbf{Var}\left(\mathbf{y}|\mathbf{X},\mathbf{G}\right)$ is only function of σ_ε^2 and the sequence $(a_q)_{q\in\mathbb{N}}$. However, the term $\mathbf{Var}\left(\mathbf{g}_i\tilde{\mathbf{y}}\right)$ at the denominator of the bias increases with the range of y_i. This result has an important implication in practice. The bias of the 2SLS estimator decreases if the dependent variable takes its values from a large range and σ_ε^2 is constant. An example is when the counting variable is observed over a long period compared with the case where the same variable is observed over a short period. The bias of the SAR model is expected to be smaller when the variable is observed over a long period.[17] This result is confirmed by Monte Carlo simulations (see Section 2.4).

[16] The 2SLS approach requires instruments that can be computed from $\mathbf{X}$ and $\mathbf{G}$ (see Kelejian and Prucha, 1998). If $\mathbf{X}$ is a column vector of ones, then this implies that I have other valid instruments to compute the estimator.

[17] This result can also be generalized to the MLE because under some moment conditions, the 2SLS, as a GMM estimator, and the MLE have the same limiting distribution (see Kelejian and Prucha, 1998).

2.4 Monte Carlo Experiments

In this section, I conduct a Monte Carlo study to assess the performance of the NPL estimator in a finite sample. I also compare the model to the spatial autoregressive Tobit (SART) and the standard linear-in-mean spatial autoregressive (SAR) models.

I consider two types of data generating processes (DGP). The DGP of type A simulates many *zeros*,[18] whereas the DGP of type B simulates few *zeros*. In both cases, the latent variables y_i^* are defined as follows:

$$y_i^* = \lambda \mathbf{g}_i \bar{\mathbf{y}} + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \gamma_1 \mathbf{g}_i \mathbf{x}_1 + \gamma_2 \mathbf{g}_i \mathbf{x}_2 + \varepsilon_i,$$

where $\bar{\mathbf{y}} = \mathbf{L}(\bar{\mathbf{y}}, \boldsymbol{\theta})$. The explanatory variables $\mathbf{g}_i\mathbf{x}_1$ and $\mathbf{g}_i\mathbf{x}_2$ are the averages x_1 and x_2, respectively, among friends. Once $\mathbf{y}^*$ is generated, I compute the count outcome $\mathbf{y}$ following Assumption B'.

As pointed out in Section 2.3.3, the estimator of $\boldsymbol{\theta}$ from the count data model may be close to that of the SAR and SART models if the dependent variable has a large dispersion. To illustrate this through the Monte Carlo study, I set two values for the parameter $\tilde{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}',\ \boldsymbol{\gamma}'\right)'$ by type of DGP. This allows simulating a count dependent variable with either low or high dispersion, depending on $\tilde{\boldsymbol{\beta}}$. The values used for $\tilde{\boldsymbol{\beta}}$ are presented in Table 2.1.

Table 2.1 – Slope of the observed explanatory variables

	Low dispersion	High dispersion
Type A	(-2, -2.5, 2.1, 1.5, -1.2)	(-1, -6.8, 2.3, -2.5, 2.5)
Type B	(1, 0.4, 0.5, 0.5, 0.6)	(3, -1.8, 2.3, 2.5, 2.5)

This table presents the values of $\tilde{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}',\ \boldsymbol{\gamma}'\right)'$ by type of DGP to simulate count data having either low or high dispersion. For instance, to simulate data from the DGP of type B with a low dispersion, I set $\boldsymbol{\beta} = (1, 0.4, 0.5)$ and $\boldsymbol{\gamma} = (0.5, 0.6)$.

The exogenous variables x_1 and x_2 are simulated from $\mathcal{N}(0, 4)$ and $\mathcal{P}oisson(3)$, respectively. I also consider several sample sizes, N $\in \{250, 750, 1500\}$. The adjacency matrix $\mathbf{G}$ is such that $g_{ij} = \dfrac{1}{n_i}$ if i is connected to j, and $g_{ij} = 0$ otherwise, where n_i is the degree of i randomly chosen between 0 and 20 for $N = 250$, 0 and 35 for $N = 750$, and 0 and 50 for $N = 1500$. Figure 2.3 presents the histogram of the simulated data for $N = 1500$. Data from a DGP of type A exhibit excess zeros (e.g., *number of cigarettes smoked daily* for low dispersion data or *weekly* for high dispersion data), whereas data from a DGP of type B concern frequent events (e.g., *number of recreational activities in which students participated in the last school year* for low dispersion data or *the last two school years* for large dispersion data).

[18] When the proportion of zeros is very high, one may need zero-inflated or hurdle specifications (see Jones, 1989; Lambert, 1992). I discuss this point in Section 2.6.3.

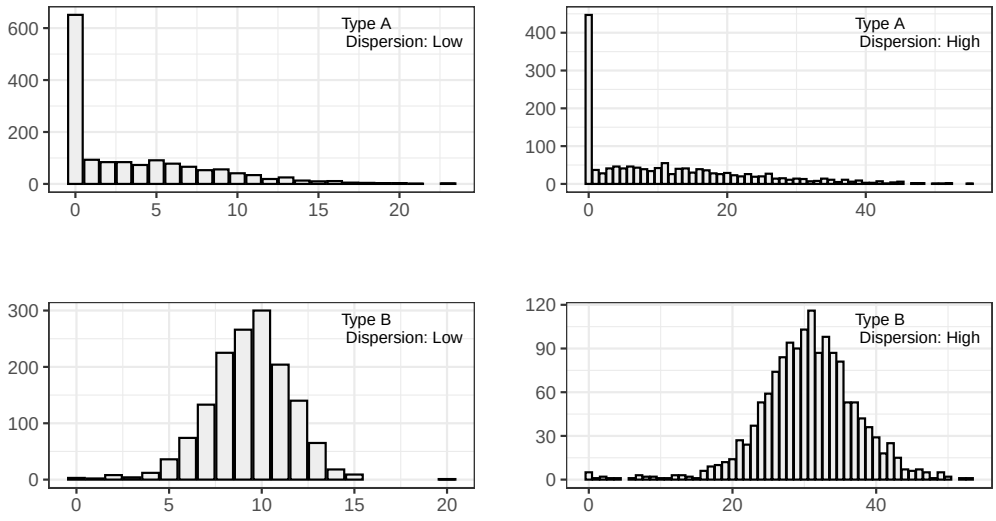


Figure 2.3 – Simulated data using the count data model with social interactions

I simulate each DG 1,000 times. The results can be replicated using my R package `CDatanet` and the replication code.[19]

The Monte Carlo results show that the NPL estimator of the count data model performs well in finite samples regardless of the type of DGP (see Tables 2.2 and 2.3). The estimator seems consistent. Moreover, the model performs better when the dependent variable has a higher dispersion.

When comparing the count data model to the SART and SAR models, it stands out that the SART and SAR models bias the peer effects downward. The bias remains substantial in a large sample for both types of DGP when the dependent variable has a lower dispersion. In contrast, when the dependent variable has a large dispersion, the SART model estimator is close to that of the count data model (see Propositions 2.3 and 2.4). However, the bias of the SAR model is still large for the DGP of type A. Indeed, the SAR model does not control for the left-censoring nature of the dependent variable.

2.5 Effect of Social Interactions on Participation in Extracurricular Activities

In this section, I present an empirical illustration of the model using a unique and now widely used data set provided by the National Longitudinal Study of Adolescent Health (Add Health).

[19]The package and the replication code are located at CRAN.R-project.org/package=CDatanet.

Table 2.2 – Monte Carlo simulations with low dispersion

	CDSI[(1)]		SART		SAR	
Statistic	Mean	Sd.	Mean	Sd.	Mean	Sd.
	$N = 250$					
	Type A					
$\lambda = 0.4$	0.399	0.171	0.270	0.141	0.193	0.139
$\beta_0 = -2$	−2.009	0.441	−1.698	0.455	0.946	0.488
$\beta_1 = -2.5$	−2.500	0.075	−2.543	0.076	−1.689	0.078
$\beta_2 = 2.1$	2.100	0.072	2.133	0.073	1.534	0.084
$\gamma_1 = 1.5$	1.499	0.313	1.300	0.281	0.887	0.286
$\gamma_2 = -1.2$	−1.196	0.280	−1.016	0.247	−0.707	0.252
$\sigma_\varepsilon = 1.5$	1.469	0.085	1.546	0.087	2.013	0.106
	Type B					
$\lambda = 0.4$	0.407	0.088	0.303	0.076	0.283	0.104
$\beta_0 = 1$	0.984	0.454	1.806	0.451	1.911	0.492
$\beta_1 = 0.4$	0.400	0.049	0.400	0.049	0.399	0.049
$\beta_2 = 0.5$	0.500	0.057	0.501	0.058	0.500	0.058
$\gamma_1 = 0.5$	0.496	0.127	0.537	0.126	0.545	0.130
$\gamma_2 = 0.6$	0.588	0.164	0.738	0.148	0.754	0.178
$\sigma_\varepsilon = 1.5$	1.480	0.071	1.528	0.072	1.523	0.071
	$N = 750$					
	Type A					
$\lambda = 0.4$	0.394	0.112	0.263	0.096	0.171	0.118
$\beta_0 = -2$	−1.991	0.284	−1.685	0.298	0.945	0.334
$\beta_1 = -2.5$	−2.500	0.042	−2.543	0.043	−1.684	0.047
$\beta_2 = 2.1$	2.099	0.041	2.132	0.042	1.534	0.048
$\gamma_1 = 1.5$	1.489	0.206	1.288	0.190	0.854	0.235
$\gamma_2 = -1.2$	−1.193	0.181	−1.012	0.164	−0.679	0.201
$\sigma_\varepsilon = 1.5$	1.490	0.049	1.564	0.050	2.028	0.062
	Type B					
$\lambda = 0.4$	0.399	0.064	0.292	0.057	0.275	0.085
$\beta_0 = 1$	1.002	0.323	1.874	0.317	1.971	0.389
$\beta_1 = 0.4$	0.401	0.028	0.401	0.028	0.400	0.028
$\beta_2 = 0.5$	0.501	0.032	0.502	0.032	0.501	0.032
$\gamma_1 = 0.5$	0.500	0.088	0.543	0.088	0.550	0.091
$\gamma_2 = 0.6$	0.601	0.118	0.749	0.109	0.762	0.139
$\sigma_\varepsilon = 1.5$	1.494	0.040	1.533	0.040	1.531	0.040
	$N = 1500$					
	Type A					
$\lambda = 0.4$	0.402	0.088	0.268	0.078	0.143	0.132
$\beta_0 = -2$	−2.009	0.225	−1.705	0.234	0.930	0.271
$\beta_1 = -2.5$	−2.500	0.029	−2.543	0.029	−1.682	0.030
$\beta_2 = 2.1$	2.101	0.028	2.135	0.028	1.532	0.031
$\gamma_1 = 1.5$	1.502	0.162	1.296	0.149	0.804	0.238
$\gamma_2 = -1.2$	−1.200	0.141	−1.015	0.132	−0.632	0.217
$\sigma_\varepsilon = 1.5$	1.496	0.035	1.569	0.036	2.030	0.042
	Type B					
$\lambda = 0.4$	0.401	0.056	0.288	0.050	0.272	0.074
$\beta_0 = 1$	0.995	0.280	1.915	0.278	2.006	0.343
$\beta_1 = 0.4$	0.401	0.020	0.401	0.020	0.400	0.020
$\beta_2 = 0.5$	0.499	0.023	0.500	0.023	0.499	0.023
$\gamma_1 = 0.5$	0.503	0.072	0.549	0.072	0.555	0.076
$\gamma_2 = 0.6$	0.599	0.101	0.753	0.093	0.764	0.118
$\sigma_\varepsilon = 1.5$	1.497	0.028	1.533	0.028	1.531	0.028

(1): CDSI stands for count data model with social interactions. The count data model is estimated using the NPL method as described in Section 2.3.2, whereas the SART and the SAR models are estimated using the ML method. The "Mean" column reports the average of the 1,000 estimations, and the "Sd." column reports the standard deviation.

Table 2.3 – Monte Carlo simulations with high dispersion

	CDSI[(1)]		SART		SAR	
Statistic	Mean	Sd.	Mean	Sd.	Mean	Sd.
	$N = 250$					
	Type A					
$\lambda = 0.4$	0.401	0.032	0.387	0.033	0.306	0.092
$\beta_0 = -1$	−1.007	0.492	−0.406	0.498	2.812	1.558
$\beta_1 = -6.8$	−6.801	0.058	−6.807	0.058	−6.289	0.178
$\beta_2 = 2.3$	2.298	0.061	2.300	0.061	2.146	0.100
$\gamma_1 = -2.5$	−2.499	0.251	−2.591	0.256	−2.725	0.665
$\gamma_2 = 2.5$	2.497	0.185	2.559	0.188	2.395	0.376
$\sigma_\varepsilon = 1.5$	1.481	0.071	1.533	0.073	2.585	0.395
	Type B					
$\lambda = 0.4$	0.401	0.025	0.389	0.025	0.388	0.025
$\beta_0 = 3$	2.986	0.439	3.610	0.443	3.663	0.436
$\beta_1 = -1.8$	−1.800	0.048	−1.801	0.048	−1.798	0.049
$\beta_2 = 2.3$	2.300	0.056	2.301	0.056	2.299	0.056
$\gamma_1 = 2.5$	2.505	0.132	2.485	0.135	2.481	0.135
$\gamma_2 = 2.5$	2.497	0.178	2.560	0.180	2.562	0.180
$\sigma_\varepsilon = 1.5$	1.477	0.069	1.528	0.071	1.528	0.070
	$N = 750$					
	Type A					
$\lambda = 0.4$	0.400	0.024	0.384	0.023	0.299	0.078
$\beta_0 = 1$	−0.999	0.356	−0.359	0.358	2.751	1.219
$\beta_1 = -6.8$	−6.801	0.031	−6.807	0.031	−6.354	0.102
$\beta_2 = 2.3$	2.300	0.034	2.302	0.034	2.169	0.056
$\gamma_1 = -2.5$	−2.500	0.180	−2.607	0.179	−2.793	0.545
$\gamma_2 = 2.5$	2.502	0.133	2.571	0.133	2.447	0.297
$\sigma_\varepsilon = 1.5$	1.494	0.041	1.538	0.041	2.454	0.229
	Type B					
$\lambda = 0.4$	0.400	0.019	0.389	0.019	0.387	0.019
$\beta_0 = 3$	2.991	0.314	3.632	0.316	3.681	0.316
$\beta_1 = -1.8$	−1.801	0.028	−1.801	0.028	−1.800	0.028
$\beta_2 = 2.3$	2.301	0.034	2.301	0.034	2.300	0.034
$\gamma_1 = 2.5$	2.508	0.091	2.487	0.094	2.484	0.094
$\gamma_2 = 2.5$	2.499	0.133	2.560	0.134	2.563	0.134
$\sigma_\varepsilon = 1.5$	1.494	0.042	1.535	0.042	1.535	0.042
	$N = 1500$					
	Type A					
$\lambda = 0.4$	0.400	0.020	0.383	0.020	0.296	0.063
$\beta_0 = -1$	−1.006	0.298	−0.339	0.299	2.717	1.006
$\beta_1 = -6.8$	−6.801	0.023	−6.806	0.023	−6.381	0.072
$\beta_2 = 2.3$	2.301	0.023	2.302	0.023	2.180	0.038
$\gamma_1 = -2.5$	−2.501	0.148	−2.615	0.149	−2.828	0.441
$\gamma_2 = 2.5$	2.504	0.106	2.576	0.107	2.475	0.231
$\sigma_\varepsilon = 1.5$	1.496	0.029	1.536	0.029	2.391	0.158
	Type B					
$\lambda = 0.4$	0.400	0.016	0.387	0.016	0.385	0.016
$\beta_0 = 3$	3.012	0.269	3.672	0.272	3.721	0.272
$\beta_1 = -1.8$	−1.800	0.020	−1.800	0.020	−1.799	0.020
$\beta_2 = 2.3$	2.300	0.023	2.301	0.023	2.300	0.023
$\gamma_1 = 2.5$	2.500	0.074	2.477	0.075	2.474	0.076
$\gamma_2 = 2.5$	2.498	0.106	2.563	0.107	2.566	0.107
$\sigma_\varepsilon = 1.5$	1.224	0.012	1.239	0.011	1.239	0.011
$\sigma_\varepsilon = 1.5$	1.498	0.029	1.536	0.028	1.536	0.028

(1): CDSI stands for count data model with social interactions. The count data model is estimated using the NPL method as described in the Section 2.3.2, whereas the SART and the SAR models are estimated using the ML method. The "Mean" column reports the average of the 1,000 estimations, and the "Sd." column reports the standard deviation.

2.5.1 Data

The Add Health data provides national representative information on 7th–12th graders in the United States (US). I use the Wave I in-school data, which were collected between September 1994 and April 1995. The surveyed sample is made up of 80 high schools and 52 middle schools. In particular, the data provides information on the social and demographic characteristics of students as well as their friendship links (i.e., best friends, up to 5 females and up to 5 males), education level, occupation of parents, etc.

I remove self-friendships and friendships between two students from different schools. Moreover, an important number of listed friend identifiers are missing or associated with "error codes."[20] I therefore remove from the study sample schools having many missing links and those having less than 100 students. I end up with 72,291 students from 120 schools. The largest school has 2,156 students, and about 50% of the schools have more than 500 students. The average number of friends per student is 3.8 (1.8 male friends and 2.0 female friends).

The studied counting dependent variable is the number of extracurricular activities in which students are enrolled. Students were presented with a list of clubs, organizations, and teams found in many schools. The students were asked to identify any of these activities in which they participated during the current school year or in which they planned to participate later in the school year. The students do not observe the activities in which their peers plan to participate. Therefore, the studied dependent variable is a good example for illustrating the model because the outcome is suited to a Bayesian game used to address the model. Throughout the paper, I write "*the number of extracurricular activities in which students are enrolled*" to mean the number of extracurricular activities in which the students participate during the year or in which they plan to participate.

Table B.1 provides the data summary. Figure B.1 in Appendix B.3 presents the distribution of the number of extracurricular activities in which the students are enrolled. It varies from 0 to 33 with an average of 2.4. Most students are enrolled in fewer than 10 activities. As observable characteristics, I consider age, sex, being Hispanic, race, number of years spent at their current school, living with both parents, mother's education, and mother's profession.

2.5.2 Empirical estimation

I estimate the count data model as well as the SART and the SAR models by controlling for contextual effects and school heterogeneity as fixed effects. It is well known that controlling for fixed effects in a non-linear model leads to an inconsistent estimation because of the accidental parameter issue (see Neyman and Scott, 1948; Lancaster, 2000). However, as argued by Lee et al. (2014) and Liu (2019), school fixed effects can be included as *dummy* variables because the number of schools in the Add Health data is low relative to sample size. Moreover,

[20] In the recent literature, numerous papers have developed methods for estimating peer effects using partial network data (e.g., Boucher and Houndetoungan, 2020). To focus on the main purpose of this paper, I do not address that issue here.

I remove schools having fewer than 100 students from the data.

The estimation results without school heterogeneity are reported in Table 2.4, whereas those with school heterogeneity are reported in Table 2.5. The comparison of log-likelihoods of both estimations confirms that there is a school heterogeneity effect.[21] As stated by Propositions 2.3 and 2.4 and highlighted through the Monte Carlo simulations, the SART and SAR models significantly underestimate the peer effects. Moreover, the estimation results of the SART and the SAR models are quite similar. This is because the DGP of the number of extracurricular activities in which students are enrolled is similar to the DGP of type B (see Section 2.4). As a result, the left-censoring nature of the dependent variable is not too important.

The coefficients of the count data model cannot be interpreted directly. Policy makers may be interested in the marginal effect of the explanatory variables on the expected number of extracurricular activities in which students are enrolled.[22] I present how to derive the marginal effects and the corresponding standard errors for the count data model in Appendix B.2.4.

The results confirm that an increase by one in the number of activities in which friends are enrolled implies an increased number of activities in which the students are enrolled of 0.363 (when controlling for school fixed effects). However, the SART and the SAR models underestimate this effect at 0.157 and 0.185, respectively.

Moreover, the own control variables are also significant. For instance, older students participate less in extracurricular activities, whereas Black and Asian students as well as students who have spent a greater number of years at their current school participate more. It is also found that many contextual effects are significant; for example, being a friend with male students increases one's participation, whereas being a friend with a student who has spent a greater number of years at their current school decreases ones's participation.

2.5.3 Endogeneity of the network

The estimation results above are based on the exogeneity of the network; that is, link formation does not depend on the error term ε_i in Equation (2.3). This assumption is strong and may imply inconsistent estimations (see Hsieh and Lee, 2016). To release this assumption, I consider a dyadic linking model in which the probability of link formation between two students i and j is specified with degree heterogeneity (e.g., Graham, 2017).

Let be $\mathbf{A} = [a_{ij}]$, the network data, such that $a_{ij} = 1$ if i knows j, and $a_{ij} = 0$ otherwise. Let also the latent variable a^*_{ij}, given by $a^*_{ij} = \Delta\mathbf{x}'_{ij}\bar{\boldsymbol{\beta}} + \mu_i + \mu_j + \varepsilon^*_{ij}$, where $\Delta\mathbf{x}_{ij}$ is a vector of observed dyad-specific variables, $\bar{\boldsymbol{\beta}}$ contains the parameters associated with the dyad-specific variables, μ_i is an unobserved individual-level attribute (gregariousness) that captures the *degree heterogeneity*, and $\varepsilon^*_{ij} \overset{iid}{\sim} logistic$. The latent variable a^*_{ij} can be interpreted as a link formation utility. I assume that $a_{ij} = 1$ if $a^*_{ij} > 0$. Therefore, the probability of link formation

[21] This result is found using the likelihood ratio test. The test statistic is compared with the value of the Chi-squared distribution table for 119 degrees of freedom.

[22] This is also the case for the SART model. Only the estimators of the SAR model's parameters can be interpreted as marginal effects.

Table 2.4 – Application results without fixed effects

Parameters	CDSI[(1)] Coef.	CDSI[(1)] Marginal effects		SART Coef.	SART Marginal effects		SAR	
λ	0.668	0.549	(0.025)***	0.249	0.203	(0.004)***	0.237	(0.006)***
Own effects								
Intercept	1.061	0.870	(0.096)***	2.415	1.963	(0.07)***	2.597	(0.094)***
Age	−0.019	−0.016	(0.006)**	−0.077	−0.063	(0.004)***	−0.075	(0.006)***
Male	−0.237	−0.195	(0.017)***	−0.243	−0.198	(0.017)***	−0.208	(0.019)***
Hispanic	0.036	0.029	(0.027)	0.012	0.010	(0.02)	0.052	(0.029)*
Race								
Black	0.250	0.205	(0.031)***	0.210	0.170	(0.023)***	0.235	(0.034)***
Asian	0.670	0.550	(0.035)***	0.651	0.529	(0.023)***	0.639	(0.039)***
Other	0.211	0.173	(0.029)***	0.197	0.160	(0.023)***	0.192	(0.033)***
Years at school	0.122	0.100	(0.008)***	0.132	0.107	(0.005)***	0.127	(0.008)***
With both par.	0.160	0.131	(0.020)***	0.158	0.129	(0.019)***	0.150	(0.022)***
Mother Educ.								
<High	−0.065	−0.054	(0.024)**	−0.068	−0.055	(0.024)**	−0.054	(0.027)**
>High	0.376	0.309	(0.02)***	0.381	0.310	(0.021)***	0.359	(0.022)***
Missing	0.222	0.182	(0.033)***	0.206	0.167	(0.028)***	0.240	(0.037)***
Mother job								
Professional	0.211	0.174	(0.025)***	0.219	0.178	(0.026)***	0.197	(0.029)***
Other	0.058	0.047	(0.021)**	0.055	0.045	(0.021)**	0.041	(0.024)*
Missing	−0.081	−0.066	(0.03)**	−0.080	−0.065	(0.027)**	−0.061	(0.033)*
Contextual effects								
Age	−0.078	−0.064	(0.004)***	−0.035	−0.028	(0.004)***	−0.042	(0.004)***
Male	0.108	0.088	(0.029)***	0.013	0.010	(0.031)	0.051	(0.034)
Hispanic	−0.153	−0.126	(0.039)***	−0.241	−0.196	(0.042)***	−0.217	(0.046)***
Race								
Black	−0.169	−0.139	(0.037)***	−0.095	−0.077	(0.035)**	−0.102	(0.043)**
Asian	−0.589	−0.484	(0.046)***	−0.447	−0.363	(0.047)***	−0.440	(0.058)***
Other	−0.279	−0.229	(0.05)***	−0.229	−0.186	(0.061)***	−0.220	(0.061)***
Years at school	−0.028	−0.023	(0.010)**	0.021	0.017	(0.01)*	0.021	(0.011)*
With both par.	0.069	0.057	(0.037)	0.244	0.198	(0.039)***	0.226	(0.041)***
Mother Educ.								
<High	−0.222	−0.182	(0.042)***	−0.204	−0.166	(0.049)***	−0.175	(0.05)***
>High	0.019	0.016	(0.036)	0.250	0.203	(0.038)***	0.239	(0.040)***
Missing	−0.247	−0.203	(0.060)***	−0.152	−0.123	(0.064)*	−0.099	(0.071)
Mother job								
Professional	0.094	0.078	(0.045)*	0.272	0.221	(0.051)***	0.252	(0.054)***
Other	−0.006	−0.005	(0.036)	0.107	0.087	(0.041)**	0.093	(0.044)**
Missing	−0.030	−0.024	(0.053)	0.067	0.055	(0.056)	0.054	(0.064)
σ_ε	2.426			2.447			2.315	
N	72,291			72,291			72,291	
log-likelihood	−159,923.7			−160,606.6			−163,430.3	
Fixed effects	No			No			No	

(1): CDSI stands for count data model with social interactions. The count data model is estimated using the NPL method as described in Section 2.3.2, whereas the SART and the SAR models are estimated using the ML method. Under the CSDI and the SART models, the column Coef. refers to the parameter values, while both columns of marginal effects refer to the marginal effects with their corresponding standard errors reported in parentheses. The columns under SAR report the parameter values (equal to the marginal effects) of the SAR model, with their standard error reported in parentheses. The codes ***, **, * mean that the corresponding parameter is significant at 1%, 5%, and 10%, respectively.

Table 2.5 – Application results with fixed effects

Parameters	CDSI[(1)] Coef.	CDSI[(1)] Marginal effects		SART Coef.	SART Marginal effects		SAR	
λ	0.443	0.363	(0.028)***	0.194	0.157	(0.005)***	0.185	(0.006)***
Own effects								
Age	−0.049	−0.040	(0.008)***	−0.073	−0.059	(0.006)***	−0.061	(0.009)***
Male	−0.253	−0.207	(0.017)***	−0.261	−0.212	(0.018)***	−0.225	(0.019)***
Hispanic	0.123	0.101	(0.026)***	0.128	0.104	(0.021)***	0.158	(0.03)***
Race								
Black	0.309	0.253	(0.031)***	0.308	0.250	(0.025)***	0.312	(0.035)***
Asian	0.701	0.576	(0.035)***	0.704	0.572	(0.025)***	0.689	(0.04)***
Other	0.220	0.181	(0.028)***	0.217	0.176	(0.024)***	0.209	(0.033)***
Years at school	0.120	0.099	(0.007)***	0.120	0.097	(0.006)***	0.112	(0.009)***
With both par.	0.158	0.129	(0.019)***	0.153	0.124	(0.019)***	0.149	(0.022)***
Mother Educ.								
<High	−0.044	−0.036	(0.024)	−0.045	−0.036	(0.025)	−0.033	(0.027)
>High	0.392	0.321	(0.019)***	0.389	0.316	(0.021)***	0.369	(0.022)***
Missing	0.231	0.190	(0.032)***	0.214	0.174	(0.029)***	0.246	(0.037)***
Mother job								
Professional	0.236	0.193	(0.025)***	0.238	0.193	(0.026)***	0.217	(0.029)***
Other	0.069	0.057	(0.02)***	0.069	0.056	(0.022)***	0.057	(0.024)**
Missing	−0.064	−0.052	(0.029)*	−0.063	−0.051	(0.028)*	−0.042	(0.033)
Contextual effects								
Age	−0.064	−0.052	(0.005)***	−0.032	−0.026	(0.004)***	−0.039	(0.005)***
Male	0.032	0.026	(0.030)	−0.034	−0.027	(0.032)	0.011	(0.034)
Hispanic	−0.048	−0.039	(0.042)	−0.071	−0.057	(0.046)	−0.059	(0.049)
Race								
Black	−0.085	−0.070	(0.039)*	−0.028	−0.023	(0.038)	−0.045	(0.045)
Asian	−0.331	−0.272	(0.052)***	−0.219	−0.178	(0.054)***	−0.229	(0.062)***
Other	−0.245	−0.201	(0.052)***	−0.208	−0.169	(0.063)***	−0.203	(0.061)***
Years at school	−0.015	−0.012	(0.011)	−0.002	−0.001	(0.011)	−0.004	(0.013)
With both par.	0.165	0.135	(0.037)***	0.239	0.194	(0.040)***	0.228	(0.041)***
Mother Educ.								
<High	−0.180	−0.148	(0.043)***	−0.173	−0.141	(0.050)***	−0.147	(0.051)***
>High	0.190	0.156	(0.038)***	0.299	0.243	(0.040)***	0.286	(0.041)***
Missing	−0.178	−0.146	(0.061)**	−0.145	−0.118	(0.066)*	−0.095	(0.072)
Mother job								
Professional	0.257	0.211	(0.047)***	0.341	0.277	(0.053)***	0.321	(0.055)***
Other	0.076	0.062	(0.038)*	0.133	0.108	(0.043)**	0.124	(0.045)***
Missing	0.055	0.045	(0.054)	0.105	0.085	(0.059)	0.091	(0.064)
σ_ε	2.394			2.425			2.295	
N	72,291			72,291			72,291	
log-likelihood	−158,963.9			−159,881.0			−162,744.4	
Fixed effects	Yes			Yes			Yes	

(1): CDSI stands for count data model with social interactions. The count data model is estimated using the NPL method as described in Section 2.3.2, whereas the SART and the SAR models are estimated using the ML method. Under the CSDI and the SART models, the column Coef. refers to the parameter values, while both columns of marginal effects refer to the marginal effects with their corresponding standard errors reported in parentheses. The columns under SAR report the parameter values (equal to the marginal effects) of the SAR model, with their standard error reported in parentheses. The codes ***, **, * mean that the corresponding parameter is significant at 1%, 5%, and 10%, respectively.

between i and j, denoted P_{ij}, is defined as

$$P_{ij} = \frac{\exp\left(\Delta \mathbf{x}'_{ij}\bar{\boldsymbol{\beta}} + \mu_i + \mu_j\right)}{1 + \exp\left(\Delta \mathbf{x}'_{ij}\bar{\boldsymbol{\beta}} + \mu_i + \mu_j\right)}. \tag{2.18}$$

By convention, I set $P_{ii} = 0$ and $P_{ij} = 0$ if i and j come from different schools. A similar network formation model can be found in McCormick and Zheng (2015) and Breza et al. (2020), where the term $\Delta \mathbf{x}'_{ij}\bar{\boldsymbol{\beta}}$ is replaced by the distance between the individuals on a latent space.

As dyad-specific variables, I choose the absolute value of age difference, the absolute value of the difference in the number of years spent at the current school, whether both students are of the same sex, Hispanic, White, Black, Asian, and whether the mother's job for both students is professional. Importantly, the probability of link formation (2.18) is symmetric ($P_{ij} = P_{ji}$ for any $i, j \in \mathcal{V}$), but it allows the network to be directed because $\varepsilon^*_{ij} \neq \varepsilon^*_{ji}$. This specification is different from that of Graham (2017) in which $\varepsilon^*_{ij} = \varepsilon^*_{ji}$ and $a_{ij} = a_{ji}$ for all $i, j \in \mathcal{V}$.

Let $s(i)$ be the school of the individual i. I assume that the unobserved attribute μ_i is random and distributed according to $\mathcal{N}\left(u_{\mu s(i)}, \sigma^2_{\mu s(i)}\right)$. It is important to notice that the mean and the variance of μ_i vary across schools. Such a specification enables the capturing of school heterogeneity (as fixed effects) in the probability of link formation.

As pointed out in Hsieh and Lee (2016), the unobserved attributes μ_i may be correlated to the error term ε_i. This violates the exogeneity condition on $\mathbf{G}$.

For any i, let $\boldsymbol{v}_i = (\varepsilon_i, \mu_i)'$. The variable $\boldsymbol{v}_i$ is distributed according to a bivariate normal distribution. Let $\boldsymbol{\Sigma}^i_{\mu\varepsilon}$ be the covariance matrix of $\boldsymbol{v}_i$.

$$\boldsymbol{\Sigma}_{\mu\varepsilon} = \begin{pmatrix} \sigma^2_\varepsilon & \rho\sigma_\varepsilon\sigma_{\mu s(i)} \\ \rho\sigma_\varepsilon\sigma_{\mu s(i)} & \sigma^2_{\mu s(i)} \end{pmatrix}, \tag{2.19}$$

where ρ is the partial correlation between μ_i and ε_i. The error term ε_i can be rewritten as $\varepsilon_i = \rho\sigma_\varepsilon \dfrac{\mu_i - u_{\mu s(i)}}{\sigma_{\mu s(i)}} + \nu_i$, where $\nu_i \sim \mathcal{N}\left(0, (1-\rho^2)\sigma^2_\varepsilon\right)$ and $\mathbf{Cov}(\mu_i, \nu_i) = 0$. Let $\tilde{\mu}_i = \dfrac{\mu_i - u_{\mu s(i)}}{\sigma_{\mu s(i)}}$.

By looking for more evidence of endogeneity, one can also control for the contextual effect of $\tilde{\mu}_i$. In that case, $\varepsilon_i = \rho\sigma_\varepsilon\tilde{\mu}_i + \bar{\rho}\sigma_\varepsilon\bar{\tilde{\mu}}_i + \tilde{\nu}_i$, where $\bar{\tilde{\mu}}_i$ is the average of $\tilde{\mu}_i$ among i's friends, $\bar{\rho}$ is the partial correlation between $\bar{\tilde{\mu}}_i$ and ε_i and $\tilde{\nu}_i \sim \mathcal{N}\left(0, \bar{\sigma}^2_\varepsilon\right)$. If μ_i or μ_j is correlated to ε_i, that is $\rho \neq 0$ or $\bar{\rho} \neq 0$, then the network is endogenous. To control for endogeneity, $\tilde{\mu}_i$ and $\bar{\tilde{\mu}}_i$ may simply be included in the count data model as additional explanatory variables (see Johnsson and Moon, 2015; Boucher and Houndetoungan, 2020). In that case, the BNE characterization (2.6) becomes

$$p_{iq} = \Phi\left(\frac{\lambda \mathbf{g}_i\bar{\mathbf{y}} + \mathbf{x}'_i\boldsymbol{\beta} + \rho\sigma_\varepsilon\tilde{\mu}_i + \bar{\rho}\sigma_\varepsilon\bar{\tilde{\mu}}_i - a_q}{\bar{\sigma}_\varepsilon}\right) - \Phi\left(\frac{\lambda \mathbf{g}_i\bar{\mathbf{y}} + \mathbf{x}'_i\boldsymbol{\beta} + \rho\sigma_\varepsilon\tilde{\mu}_i + \bar{\rho}\sigma_\varepsilon\bar{\tilde{\mu}}_i - a_{q+1}}{\bar{\sigma}_\varepsilon}\right).$$

My estimation strategy is in two stages. The first stage is based on a Bayesian approach. Using MCMC, I simulate $\bar{\boldsymbol{\beta}}$, μ_i, $u_{\mu s(i)}$, and $\sigma^2_{\mu s(i)}$ from their posterior distributions (see details

in Appendix B.4.1). The simulations from the posterior distribution are then used to draw $\tilde{\mu}_i$ and $\bar{\tilde{\mu}}_i$. At the second stage, the draws of $\tilde{\mu}_i$ and $\bar{\tilde{\mu}}_i$ are used as additional explanatory variables to estimate the count data model.
I take into account the uncertainty of estimation of the first stage. By replicating drawings of μ_i, $u_{\mu s(i)}$, and $\sigma^2_{\mu s(i)}$ from the posterior distribution, I correct the asymptotic variance of the estimator at the second stage. The approach I use is similar in spirit to that of Krinsky and Robb (1986). The new variance accounts for the variability of $\tilde{\mu}_i$ (see details in Appendix B.4.2).

The estimation results (controlling for schools' heterogeneity and network endogeneity) are presented in Table 2.6. The results are significantly different to those of Table 2.5. The parameters of the additional explanatory variables are significantly different to zero at 1%. This confirms that the network is endogenous.
Although friends incite participation in extracurricular activities, the sociability degree (gregariousness) of the students also plays an important role. Students with high μ_i are more *extroverted* (more likely to form links) and also participate in more extracurricular activities.[23] In contrast, *introverted* students participate less in extracurricular activities. Similar evidence has been found in sociology studies, which highlight that an individual's gregariousness determines their participation in activities.[24] As well, being friends of a highly gregarious student also increases one's participation in extracurricular activities.[25]
Peer effects are reduced when controlling for network endogeneity but remain significant. An increase by one in the number of activities in which friends are enrolled implies an increase in the number of activities in which students are enrolled of 0.295. The endogeneity of the network is also confirmed with the models SART and SAR. However, they still underestimate peer effects at 0.141 and 0.166, respectively.
To understand the decrease in peer effects, notice that λ could capture other effects if students' gregariousness is not included in the count data model. For example, λ can capture the effect of an exogenous shock that increases students' and peers' gregariousness because students and their friends will experiment and increase in their participation in extracurricular activities. This is similar to the correlated effects (see Manski, 1993).

2.6 Discussions

In this section, I discuss some general implications of the model, some limits, and some areas for future research.

[23] Because $\rho\sigma_\varepsilon$, the sign of $\tilde{\mu}_i$ is positive in the count data model.

[24] For example, specific personality traits are associated with activity participation (e.g., Newton et al., 2018); extroverted people work more often in jobs having more social interactions (e.g., Pfeiffer and Schulz, 2012), and highly gregarious individuals are more likely to be a member of a group (e.g., Erbe, 1962).

[25] Because $\bar{\rho}\sigma_\varepsilon$, the sign of $\mathbf{g}_i\tilde{\mu}_i$ is positive in the count data model.

Table 2.6 – Application results controlling for fixed effects and network endogeneity

Parameters	CDSI[(1)]			SART			SAR	
	Coef.	Marginal effects		Coef.	Marginal effects			
λ	0.359	0.294	(0.028)***	0.173	0.141	(0.005)***	0.166	(0.006)***
$\rho\sigma_\varepsilon$	0.246	0.202	(0.011)***	0.253	0.205	(0.010)***	0.240	(0.013)***
$\bar{\rho}\sigma_\varepsilon$	0.202	0.166	(0.019)***	0.240	0.195	(0.018)***	0.218	(0.020)***
Own effects								
Age	−0.049	−0.040	(0.008)***	−0.066	−0.053	(0.006)***	−0.061	(0.009)***
Male	−0.241	−0.198	(0.017)***	−0.249	−0.202	(0.018)***	−0.213	(0.019)***
Hispanic	0.179	0.147	(0.027)***	0.184	0.150	(0.022)***	0.211	(0.031)***
Race								
Black	0.557	0.457	(0.033)***	0.564	0.458	(0.027)***	0.552	(0.038)***
Asian	0.848	0.696	(0.035)***	0.847	0.687	(0.026)***	0.827	(0.041)***
Other	0.281	0.231	(0.028)***	0.281	0.228	(0.024)***	0.269	(0.033)***
Years at school	0.099	0.081	(0.007)***	0.097	0.079	(0.006)***	0.092	(0.009)***
With both par.	0.145	0.119	(0.019)***	0.142	0.115	(0.019)***	0.135	(0.022)***
Mother Educ.								
<High	−0.021	−0.017	(0.024)	−0.021	−0.017	(0.025)	−0.012	(0.027)
>High	0.377	0.309	(0.019)***	0.376	0.305	(0.021)***	0.354	(0.022)***
Missing	0.226	0.185	(0.032)***	0.210	0.170	(0.029)***	0.242	(0.036)***
Mother job								
Professional	0.209	0.171	(0.024)***	0.209	0.170	(0.026)***	0.191	(0.029)***
Other	0.054	0.044	(0.020)**	0.056	0.045	(0.022)**	0.043	(0.023)*
Missing	−0.060	−0.050	(0.029)*	−0.058	−0.047	(0.028)*	−0.041	(0.033)
Contextual effects								
Age	−0.075	−0.061	(0.005)***	−0.051	−0.041	(0.004)***	−0.056	(0.005)***
Male	−0.002	−0.002	(0.029)	−0.042	−0.034	(0.032)	0.002	(0.034)
Hispanic	0.002	0.001	(0.042)	−0.009	−0.007	(0.047)	−0.001	(0.049)
Race								
Black	0.171	0.140	(0.043)***	0.241	0.196	(0.042)***	0.205	(0.048)***
Asian	−0.114	−0.094	(0.055)*	−0.013	−0.011	(0.055)	−0.039	(0.064)
Other	−0.157	−0.129	(0.053)**	−0.122	−0.099	(0.063)	−0.127	(0.061)**
Years at school	−0.016	−0.013	(0.011)	−0.010	−0.008	(0.011)	−0.009	(0.013)
With both par.	0.153	0.126	(0.037)***	0.207	0.168	(0.04)***	0.193	(0.041)***
Mother Educ.								
<High	−0.152	−0.125	(0.043)***	−0.143	−0.116	(0.050)**	−0.122	(0.051)**
>High	0.169	0.139	(0.038)***	0.246	0.200	(0.040)***	0.236	(0.041)***
Missing	−0.147	−0.120	(0.062)*	−0.124	−0.101	(0.065)	−0.081	(0.071)
Mother job								
Professional	0.205	0.168	(0.047)***	0.269	0.218	(0.053)***	0.246	(0.055)***
Other	0.034	0.028	(0.038)	0.083	0.067	(0.043)	0.072	(0.045)
Missing	0.037	0.030	(0.055)	0.083	0.067	(0.059)	0.065	(0.064)
$\bar{\sigma}_\varepsilon$	2.377			2.412			2.283	
N	72,291			72,291			72,291	
log-likelihood	−158,467.7			−159,462.2			−162,328.3	
Fixed effects	Yes			Yes			Yes	

(1): CDSI stands for count data model with social interactions. The count data model is estimated using the NPL method as described in Section 2.3.2, whereas the SART and the SAR models are estimated using the ML method. Under the CSDI and the SART models, the column Coef. refers to the parameter values, while both columns of marginal effects refer to the marginal effects with their corresponding standard errors reported in parentheses. The columns under SAR report the parameter values (equal to the marginal effects) of the SAR model, with their standard error reported in parentheses. The codes ***, **, * mean that the corresponding parameter is significant at 1%, 5%, and 10%, respectively.

2.6.1 Flexibly dispersed count variable model

The most commonly used models to study count data (without social interactions) are the Poisson model and related models, such as the generalized Poisson (Consul and Jain, 1973) and Negative Binomial (Hilbe, 2011). The main difference between these models is in the way they fit the dispersion of the dependent variable.

The fundamental feature of Poisson models is the mean-variance equality conditional on the explanatory variables (equidispersion), whereas Negative Binomial models allow the variance to be greater than the mean (overdispersion). In addition to the overdispersion, the generalized Poisson allows the variance to be smaller than the mean (underdispersion)

The count data model of this paper is flexible in terms of dispersion fitting. The conditional variance of y_i can be expressed as

$$\mathbf{Var}\left(y_i|\mathbf{X},\mathbf{G}\right) = \bar{y}_i + \underbrace{2\sum_{r=1}^{\infty} r\Phi\left(\hat{\psi}_{ir}\right)}_{\Delta(\sigma_\varepsilon)} - \bar{y}_i^2, \tag{2.20}$$

where $\forall\ i \in \mathcal{V},\ q \in \mathbb{N}^*$, and $\hat{\psi}_{iq} = \dfrac{\lambda \mathbf{g}_i\bar{\mathbf{y}} + \mathbf{x}_i'\boldsymbol{\beta} - a_q}{\sigma_\varepsilon}$. The equation $\Delta(\sigma_\varepsilon) = 0$ does not have a closed form, but $\Delta(\sigma_\varepsilon)$ is increasing in σ_ε. Depending on σ_ε, the term $\Delta(\sigma_\varepsilon)$ may be null, negative, or positive. The new count variable model is flexible in terms of dispersion fitting. It allows equidispersion, overdispersion, and underdispersion as the Generalized Poisson model.

2.6.2 Time-varying exposure

Data from "*How many times do you smoke a day?*" are not the same as those of "*How many times do you smoke a week?*" When individuals are not followed for the same amount of time, it is more relevant to model rates instead of counts.

Let e_i be the exposure time of i. In the traditional count data models (Poisson and Negative Binomial), the time-varying exposure issue can be fixed using an offset (see Hakim et al., 1991; Winkelmann and Zimmermann, 1995). This consists of adding $log(e_i)$ as a supplementary explanatory variable and constraining its coefficient to one. In doing so, $\dfrac{\bar{y}_i}{e_i}$ does not depend on e_i because $\bar{y}_i$ is a log-linear function of explanatory variables. The rate $\dfrac{\bar{y}_i}{e_i}$ can be compared between individuals having different exposure times. Since the reduced form of the expected outcome $\bar{y}_i$ in the new count variable model has a more complex form, this offset approach cannot be used.

To control for time-varying exposure, the sequence $(a_q)_{q\in\mathbb{N}}$ of Assumption B' may be redefined as $a_0 = -\infty$ and $a_q = e_i(q-1)\ \forall\ q \in \mathbb{N}^*$. Under this specification, the distribution of $\dfrac{y_i^*}{e_i}$ does not depend on the exposure time. Note that this result holds because the increment of the sequence $(a_q)_{q\in\mathbb{N}}$ is constant.

2.6.3 Zero-inflated and Hurdle specifications

In applications with excess zeros, zero-inflated (see Lambert, 1992) or Hurdle (see Jones, 1989) specifications are suggested for modeling count data. These specifications assume that "zeros" could be generated by processes other than those of the positive values. For instance, for the question "*How many times did you smoke during the last week*?" smokers may report zero because they did not smoke during that *specific* week. However, other individuals may report zero because they are non-smokers. The first type of zeros are *sampling*, whereas the second type of zeros are *structural*. It may be important to distinguish both processes because they do not have the same policy implications (see Tüzen and Erbaş, 2018).
The zero-inflated model assumes that there is a mix of sampling of structural zeros in the data, whereas the Hurdle specification allows only structural zeros. I refer the reader to Jones (1989) and (Lambert, 1992) for more details. However, these specifications are not compatible with the microeconomic foundation of my model. This could be investigated in future research.

2.7 Conclusion

In this paper, I study a social network model for count data using a static Bayesian game. I provide sufficient conditions under which the game has a unique Nash Bayesian equilibrium. I show that the model parameter can be estimated using the Nested Partial Likelihood (NPL) method. I also show that the counting nature of the dependent variable is important, especially when the variable has a small range. Indeed, modeling data that are generated from the game using the standard linear-in-means peer effects model, which incorrectly assumes that the dependent variable is normally distributed, lead to asymptotically inconsistent estimations. The estimation bias decreases when the range of the dependent variable increases. This result is also confirmed through Monte Carlo simulations.
I also provide an empirical application. I estimate peer effects on the number of extracurricular activities in which a student is enrolled. By controlling for the endogeneity of the network, I find that an increase by one in the number of activities in which friends are enrolled implies an increase in the number of activities in which students are enrolled by 0.295. However, the SART and SAR models underestimate this effect at 0.141 and 0.166, respectively. I also find that ignoring the endogeneity overestimates the peer effects.
The model implementation is simple and not computational. I provide an easy to use R package that implements all the methods used in this paper.[26] Nevertheless, the model also has limits. In particular, it does not consider zero-inflated specifications for data having excess zeros.

[26]The package is available at CRAN.R-project.org/package=CDatanet.

Chapter 3

Selective linear segmentation for detecting relevant parameter changes

Résumé

Les processus avec changements structurels sont une approche flexible pour modéliser des longues séries chronologiques. En considérant un modèle linéaire en moyennes, nous proposons une méthode qui relâche l'hypothèse selon laquelle une cassure structurelle dans une série temporelle implique un changement de tous les paramètres du modèle. Pour ce faire, nous estimons d'abord les dates de cassures potentielles présentées par la série, puis nous utilisons une régression pénalisée pour détecter les paramètres du modèle qui changent à chaque date de cassure. Étant donné que certains segments de la régression peuvent être courts, nous optons pour une fonction de pénalité (presque) non biaisée, appelée fonction de pénalité *seamless-L0* (SELO). Nous montrons que l'estimateur SELO détecte de manière convergente les paramètres qui varient à chaque cassure et nous proposons d'utiliser un algorithme de maximisation d'espérance de recuit déterministe (DAEM) pour traiter la multimodalité de la fonction objectif. Étant donné que la fonction de pénalité SELO dépend de deux paramètres, nous utilisons un critère pour choisir les meilleurs paramètres et par conséquent le meilleur modèle. Ce nouveau critère présente une interprétation bayésienne qui permet d'évaluer l'incertitude des paramètres ainsi que l'incertitude du modèle. Les simulations de Monte Carlo montrent que la méthode fonctionne bien pour de nombreux modèles de séries temporelles, y compris les processus hétéroscédastiques. Pour un échantillon de 14 stratégies de hedge funds (HF), utilisant un modèle de tarification basé sur l'actif, nous mettons en exergue la capacité prometteuse de notre méthode à détecter la dynamique temporelle des expositions au risque ainsi qu'à prévoir les rendements de HF.

Abstract

Change-point (CP) processes are one flexible approach to model long time series. Considering a linear-in-means model, we propose a method to relax the assumption that a break triggers a change in all the model parameters. To do so, we first estimate the potential break dates exhibited by the series and we use a penalized likelihood approach to detect which parameters change. Since some segments in the CP regression can be small, we opt for a (nearly) unbiased penalty function, called the seamless-L0 (SELO) penalty function. We prove the consistency of the SELO estimator in detecting which parameters indeed vary over time and we suggest using a deterministic annealing expectation-maximisation (DAEM) algorithm to deal with the multimodality of the objective function. Since the SELO penalty function depends on two tuning parameters, we use a criterion to choose the best tuning parameters and as a result the best model. This new criterion exhibits a Bayesian interpretation that makes possible to assess the parameters' uncertainty as well as the model's uncertainty. Monte Carlo simulations highlight that the method works well for many time series models including heteroskedastic processes. For a sample of 14 Hedge funds (HF) strategies, using an asset based style pricing model, we shed light on the promising ability of our method to detect the time-varying dynamics of risk exposures as well as to forecast HF returns.

Keywords: Change-point, Structural change, Time-varying parameter, Model selection, Hedge funds.

JEL Classification: C11, C12, C22, C32, C52, C53.

3.1 Introduction

Long time series are standard in this period of large publicly available datasets. Care is required when modeling such a time series since many of them span over critical events that may change the series dynamic. At least two statistical solutions exist to take into account these changes. On the one hand, a process with fixed parameters can be used but it needs to exhibit a rich and complex dynamic. This complexity often makes the model difficult to estimate and to interpret (see, for instance, long memory processes such as Geweke and Porter-Hudak (1983)). On the other hand, one can rely on time-varying parameter (TVP) models and in particular Markov-switching and change-point (CP) processes since they allow for abrupt changes in the model parameters when a critical event affects the series dynamic (see Hamilton, 1989; Bauwens et al., 2015). This paper deals with CP linear regression models where we allow the mean parameters to change over time.

The CP literature dates back to Chernoff and Zacks (1964) and is nowadays vast. Just focusing on linear regressions, Andrews (1993), Bai and Perron (1998), Killick et al. (2012), Fryzlewicz et al. (2014) and Yau and Zhao (2016) develop prominent procedures to detect breakpoints. On the Bayesian side, there also exist many ways to estimate structural breaks and important contributions can be found in Stephens (1994), Chib (1998), Fearnhead and Liu (2007), Rigaill et al. (2012) and Maheu and Song (2013). While all these methods differ in the criterion or in the algorithm used to detect the changes, most of them rely on the assumption that, when a break is detected (that may be triggered by the change in only one model parameter), a new segment is created and a new set of parameters needs to be estimated. Although the assumption seems harmless, it creates two important drawbacks:

1. From an interpretation perspective, if all the parameters have to change when a break is detected, it is difficult to assess which parameters have indeed abruptly varied and so it complicates the economic interpretation of the structural break.
2. Forecasting wise, when a parameter does not vary from one regime to another, its estimation is more accurate than if two parameters were considered over these two regimes. This feature could improve the predictions of the model.

In this paper, we propose a method to relax the assumption that a break triggers a change in all the model parameters. To do so, we first estimate the potential break dates exhibited by the series and then we use a penalized likelihood approach to detect which parameters change. Because some segments in the CP regression can be small, we opt for a (nearly) unbiased penalty function, called the seamless-L0 (SELO) penalty function, recently proposed by Dicker et al. (2013). We prove the consistency of the SELO estimator in detecting which parameters indeed vary over time and we suggest using a deterministic annealing expectation-maximisation (DAEM) algorithm to deal with the multimodality of the objective function (see Ueda and Nakano, 1998). Since the SELO penalty function depends on two tuning parameters, we use a criterion (new in this literature) to choose the best tuning parameters and as a result

the best model. This new criterion exhibits a Bayesian interpretation which makes possible to assess the parameters' uncertainty as well as the model's uncertainty. This last feature is determinant when predicting a time series since the Bayesian model averaging technique, that typically improves forecast accuracy, is readily applicable (see, e.g., Raftery et al., 2010; Koop and Korobilis, 2012).

We are aware of five other papers that also relax the assumption on the number of parameters that changes when a break is detected. In the frequentist literature, the influential paper of Bai and Perron (1998) proposes a method that also operates when only a subset of parameters can break. However, the number of possibilities grows exponentially with the number of breaks as well as with the number of parameters that can break. From a Bayesian perspective, Giordani and Kohn (2008), Eo (2016), Huber et al. (2019) and Dufays and Rombouts (2020) propose flexible state-space models to capture which parameters vary over time. However, all these estimation procedures break down when the number of parameters is large (see Supplementary Appendix C.5 for more details).

We believe that our method exhibits several advantages over the existing alternatives. Firstly, it operates for small and large dimensions. Secondly, the estimation is fast compared to the Bayesian alternatives. As a final advantage, we relax the assumption on breakpoints *once* the structural breaks have been detected which makes our approach operating in combination with any existing CP methods. In this paper, we illustrate our approach with the CP procedure of Yau and Zhao (2016) but any other CP method could have been used.

A final reference close to our framework is Chan et al. (2014) who propose a penalized regression for segmenting time series in piecewise linear models. The paper uses a group Lasso penalty function (see Yuan and Lin, 2006) to get an overestimated number of segments and in a second phase, an information criterion is used to improve the estimation. Nevertheless, we differ from their methods in many aspects. First, we use an almost unbiased penalty function and from a theoretical perspective, as we use the penalized regression on a potential break date set, our assumptions for a consistent estimator are different and in line with the standard penalized regression literature. We also use a Bayesian criterion to select among the promising models uncovered by the penalty function which allows for model uncertainty and for Bayesian model averaging. Also, our estimation procedure is fast compared to Chan et al. (2014) since we iterate on closed-form expressions and because our model exhibits fewer parameters. As a final difference, we provide break uncertainty.

Eventually, we apply our method on Hedge funds (HF) returns. As highlighted by Fung et al. (2008), by Meligkotsidou and Vrontos (2008), by Bollen and Whaley (2009), and more recently by Patton et al. (2015), the dynamics of HF risk exposures and the nonlinear generating process of HF returns should be associated with market events and structural breaks. For a sample of 14 monthly Credit Suisse HF indices spanning from March 1994 to March 2016, and

using the asset based style pricing model introduced by Fung and Hsieh (2001), we show that our modeling is particularly appealing to detect time-varying exposures in HF tradings. In particular, our results report the relative role played by static and dynamic parameters and factors in the decomposition of HF returns. We also investigate the prediction performance of our approach and it turns out that the selective segmentation approach compares favorably in terms of root mean squared forecast errors and cumulative log-predictive densities with respect to other CP processes. In particular, it almost systematically dominates the CP model which assumes that all the parameters vary when a break is detected.

The paper is organized as follows. Section 3.2 documents the model specification and the SELO penalty function. Section 3.3 explains how the DAEM algorithm is applied to our framework. In Section 3.4, we detail the criterion used to select the SELO tuning parameters and we relate it to the Bayesian paradigm. Section 3.5 documents the CP method of Yau and Zhao (2016) and discusses how it can be slightly improved. An extensive Monte Carlo study is proposed in Section 3.6. We end the paper by applying the method on HF returns in Section 3.7.

3.2 Model specification

We consider a standard linear regression specified as

$$\begin{aligned} y_t &= \beta_1 + \beta_2 x_{t,2} + \ldots + \beta_K x_{t,K} + \epsilon_t, \\ &= \mathbf{x}_t' \boldsymbol{\beta}_1 + \epsilon_t \,, \end{aligned} \tag{3.1}$$

where $\epsilon_t \sim MDS(0, \sigma^2)$ (in which MDS stands for the martingale difference sequence), $\mathbf{x}_t = (1, x_{t,2}, \ldots, x_{t,K})'$ and $\boldsymbol{\beta}_1 = (\beta_1, \beta_2, \ldots, \beta_K)'$. Typically, if a linear model is estimated over a long period, the parameters are subject to abrupt changes over time. To take this time-varying dynamic into account, we allow for $m-1$ structural breaks in the model parameters as follows,

$$y_t = \mathbf{x}_t' \boldsymbol{\beta}_i^* + \epsilon_t \,, \text{ for } \tau_{i-1} < t \leq \tau_i, \tag{3.2}$$

in which $\boldsymbol{\beta}_i^*$, is the true parameter of the explanatory variables over the regime i, $\boldsymbol{\tau}_0 = \{\tau_0, \ldots, \tau_m\} \in \mathbb{N}^{m+1}$ where $\tau_0 = 0$, $\tau_m = T$ and $\tau_i < \tau_{i+1}$ $\forall i \in [0, m-1]$. In this paper, we are interested in capturing which parameters are subject to breaks and which do not vary over time. To do so, we reframe the model (3.2) as follows,

$$\begin{aligned} y_t &= \mathbf{x}_t' \boldsymbol{\beta}_1^* + \mathbf{x}_t' \Big(\sum_{j=2}^{m} \Delta \boldsymbol{\beta}_j^* \mathbf{1}_{\{t > \tau_{j-1}\}} \Big) + \epsilon_t \,, \\ \mathbf{y} &= \mathbf{X}_{\boldsymbol{\tau}} \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \end{aligned} \tag{3.3}$$

where $\mathbf{1}_{\{x>a\}} = 1$ if $x > a$ and zero otherwise, $\Delta \boldsymbol{\beta}_j^* = \boldsymbol{\beta}_j^* - \boldsymbol{\beta}_{j-1}^*$, for $j \in [2, m]$, stands for the model parameters in first-difference, $\mathbf{y} = (y_1, \ldots, y_T)'$, $\mathbf{X}_{\boldsymbol{\tau}} = (\tilde{\mathbf{X}}_{\tau_0}, \tilde{\mathbf{X}}_{\tau_1}, \ldots, \tilde{\mathbf{X}}_{\tau_{m-1}})$

with $\tilde{\mathbf{X}}_{\tau_i} = (\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0}, \mathbf{x}_{\tau_i+1}, \ldots, \mathbf{x}_T)'$, $\boldsymbol{\varepsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_T)'$ and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*\prime}, \Delta\boldsymbol{\beta}_2^{*\prime}, \ldots, \Delta\boldsymbol{\beta}_m^{*\prime})' \in \Re^{Km\times 1}$. Note that the matrix $\tilde{\mathbf{X}}_{\tau_0}$ stands for the standard regressors since $\tau_0 = 0$. Regarding the notations, the first-difference parameter in regime j is a K-dimensional vector $\Delta\boldsymbol{\beta}_j^*$ such that $\Delta\boldsymbol{\beta}_j^* = (\Delta\beta_{j1}^*, \ldots, \Delta\beta_{jK}^*)'$. Let us also denote $A = \{(j,k); \Delta\beta_{jk}^* \neq 0, \text{ for } j \in [2,m] \text{ and for } k \in [1,K]\}$, the set of indices defining the true model.

Our strategy to uncover which parameters truly vary over time consists in first finding where are the potential break dates $\boldsymbol{\tau}$, then, in a second phase, in detecting which parameters evolve. Note that even when we know the true break dates $\boldsymbol{\tau}$, the problem of finding which parameters vary when a break occurs is not straightforward as the number of models to consider amounts to $2^{(m-1)K}$. Consequently, it is infeasible to carry out an exhaustive model selection when K or m is large. We propose a penalized likelihood approach to explore this large model space and to select which parameters experience breaks. To focus on our selective segmentation approach, we shall first assume that we have obtained a set of potential break dates $\boldsymbol{\tau}$. We discuss how we estimate this set in Section 3.5.

Remark 3.1. *In the situation where all the models can be considered (i.e., $(m-1)K \leq 10$), we do not need to rely on the penalized likelihood approach explained in Section 3.2.1. In particular, we could directly estimate all the model combinations and select the best one according to the marginal likelihood criterion given in Section 3.4.*

3.2.1 Penalized likelihood and choice of the penalty function

As emphasized by Equation (3.3), given a set of break dates $\boldsymbol{\tau}$, the problem of finding which parameters abruptly change when a break occurs boils down to a penalized linear regression problem. Specifically, one can solve the following optimization problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + T\sum_{j=2}^{m}\sum_{k=1}^{K} \text{pen}(\Delta\beta_{jk}), \tag{3.4}$$

where $||\,.\,||_p$ denotes the L_p norm and $\text{pen}(\Delta\beta_{jk})$ stands for a penalty function. Popular choices of $\text{pen}(\Delta\beta_{jk})$ are the Lasso penalty function (i.e., $\text{pen}(\Delta\beta_{jk}) = \lambda||\Delta\boldsymbol{\beta}_{jk}||_1$, see Tibshirani (1994)) or the rigde function (i.e., $\text{pen}(\Delta\beta_{jk}) = \lambda\Delta||\beta_{jk}||_2^2$, see, for instance, Ishwaran and Rao (2005)).

Following Fan and Li (2001), standard desirable properties induced by a penalty function are i) unbiasdness, ii) sparsity and iii) continuity. For instance, the ridge function is only continuous while the Lasso penalty function achieves sparsity and continuity (beside at zero). However one standard issue with these popular penalty functions is that they provide biased (but typically consistent) estimators. In our framework, this drawback is problematic since a segment can sometimes contain a small amount of observations that makes consistency results not sufficient. Recently, Dicker et al. (2013) propose a penalty function, called seamless-L_0

(SELO), that exhibits all the desirable properties. For a model parameter denoted ω, the penalty function reads as

$$\mathcal{P}_{\mathbf{SELO}}(\omega|\zeta,\lambda) \quad = \quad \frac{\lambda}{\ln 2}\ln(\frac{2|\omega|+\zeta}{|\omega|+\zeta}),$$

where the parameter ζ controls for the concavity of the function and λ stands for the penalty imposed when $\omega \neq 0$. We slightly modify their function to end up with parameters that are directly interpretable. In fact, we use the following penalty function,

$$\mathcal{P}_{\mathbf{SELO}}(\omega|a,\lambda) \quad = \quad \frac{\lambda}{\ln 2}\ln(\frac{2(\frac{|\omega|}{a})+\zeta}{(\frac{|\omega|}{a})+\zeta}), \tag{3.5}$$

where $\zeta = \dfrac{2^y-2}{1-2^y}$ with $y \in (0,1)$ and we set $y = 0.99$. In most cases, the parameter a can be interpreted as an interval $\omega \in [-a,a]$ in which ω will be biased with respect to the OLS estimate since $\mathcal{P}_{\mathbf{SELO}}(a) = \lambda y$. Intuitively, when $|\omega| > a$, we have $\mathcal{P}_{\mathbf{SELO}}(\omega) \approx \lambda$ and $\dfrac{d\mathcal{P}_{\mathbf{SELO}}(\omega)}{d\omega}|_{|\omega|\geq a} \approx 0$ for large values of a. Figure 3.1 shows the SELO penalty function with $\{a,\lambda\} = \{1, 0.9\}$ and illustrates that the function is almost flat for absolute values greater than a. To be more precise about how large a must be, when $|\bar{\beta}| \geq a$ with $a \geq \dfrac{\zeta}{\ln 2[\zeta^2+3\zeta+2]} = 0.0099$, we have that $\dfrac{d\mathcal{P}_{\mathbf{SELO}}(\omega)}{d\omega}|_{|\omega|\geq a} \leq \lambda$ which implies that the bias imposed by the SELO penalty function is smaller than the one of the Lasso function (i.e. $\lambda|\omega|$).

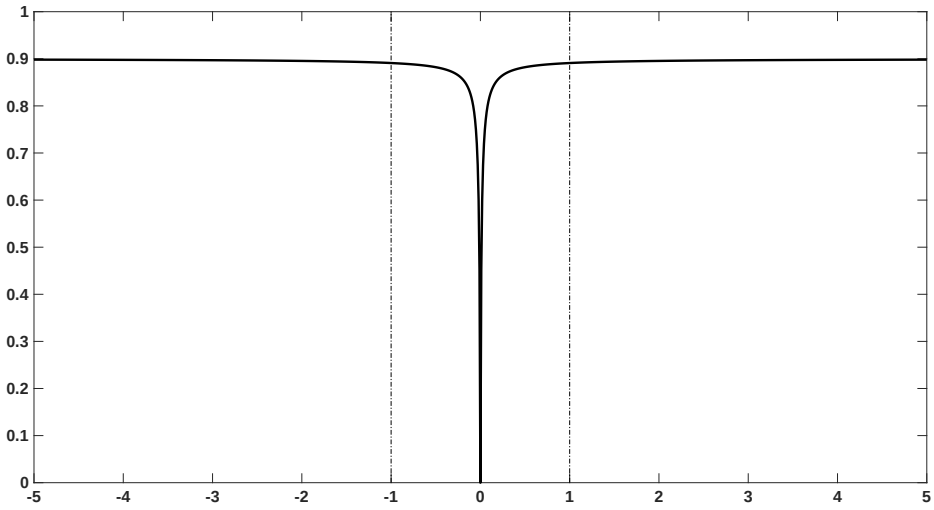


Figure 3.1 – SELO penalty function

Penalty function is shown in solid black lines while vertical dotted lines highlight the interval [-a,a]. The SELO parameters are set to $\lambda = 0.9$ and $a = 1$.

3.2.2 Consistency of the SELO estimator

The interval $[-a, a]$ in which a parameter is biased is likely to change with the variable to which it refers. Furthermore, if we assume that this interval is fixed over time, we should set a new parameter a for each variable on the m segments. Unlike Dicker et al. (2013) who define a single parameter for all the variables, we use K parameters $a_1, \ldots, a_K$, that is, one per explanatory variable. Thus, the objective function to minimize is given by

$$f(\boldsymbol{\beta}) = ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + T\frac{\lambda}{\ln(2)}\sum_{j=2}^{m}\sum_{k=1}^{K}\ln\left(\frac{2(\frac{|\Delta\beta_{jk}|}{a_k}) + \zeta}{(\frac{|\Delta\beta_{jk}|}{a_k}) + \zeta}\right). \tag{3.6}$$

Before discussing how to maximize the objective function, we present the main results about the modified SELO estimator. As highlighted in Dicker et al. (2013), the SELO estimator is consistent under reasonable conditions. Proposition 3.1 shows that this consistency result also applies in our framework. To do so, we consider the following assumptions (in which a sequence $\omega_T \to \omega$ is understood as $\lim_{T\to\infty}\omega_T = \omega$).

Assumption G.

G.1 $\boldsymbol{\tau} = \boldsymbol{\tau}_0$ *and* $\forall j \in [1, m]$, *we have* $\tau_j - \tau_{j-1} = T\delta_{\tau_j} \to \infty$, *with* $\sum_{j=1}^{m}\delta_{\tau_j} = 1$.

G.2 $\rho\sqrt{T} \to \infty$, *where* $\rho = \min_{r,k\in A}\left(|\Delta\beta^*_{r,k}|\right)$.

G.3 There exist r_0, $R_0 > 0$ *such that* $r_0 \leq \lambda_{T,\min} < \lambda_{T,\max} \leq R_0$, *where* $\lambda_{T,\min}$ *and* $\lambda_{T,\max}$ *are the smallest and largest eigenvalues of* $\left(T^{-1}\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}\right)$,*/. respectively.*

G.4 The process $\{\epsilon_t, \mathbf{x}_t\}_{t\in(\tau_{j-1},\tau_j]}$ *is ergodic and stationary for any* $j = 1, \ldots, m$. *Moreover,* $\forall\ t \in [1, T]$, $\mathbb{E}(\epsilon_t|\mathbf{x}_t) = 0$, $\mathbb{E}(\epsilon_t^2) = \sigma^2$ *and the process* $\{g_t\} = \{\mathbf{x}_t\epsilon_t\}$ *is a martingale difference sequence with finite second moments.*

G.5 $\lambda = \mathcal{O}_p(1)$ *and* $a_k = \mathcal{O}_p\left(T^{-\frac{3}{2}}\right)$, $\forall k \in [1, K]$.

We first discuss the assumptions before detailing our consistency result. Assumptions G.1 to G.5 are similar to those found in the variable selection literature (see Fan et al., 2004; Dicker et al., 2013) and in the CP literature (see Bai and Perron, 1998; Yau and Zhao, 2016). Condition G.1 assumes that the estimated CPs are the true locations. However, the SELO estimator maintains the same asymptotic properties with a set of potential breakpoints as long as it contains the true break dates (see the adapted assumption H below). In such case, Proposition 3.1 also ensures that the number of breakpoints is consistently estimated. Note that condition G.1 implies that the length of each segment increases linearly with T. Although unattractive, this condition is generally made in the CP literature (see, e.g., Perron et al., 2006; Yau and Zhao, 2016). For interested readers, Perron et al. (2006) motivate this assumption in details. Assumption G.2 allows the minimum break size to decrease with the

sample size but at a slower rate than $T^{-\frac{1}{2}}$. Conditions G.3 are related to the eigenvalues and are standard in the variable selection literature (see, e.g., Zhang et al., 2010). However, we show in Appendix C.1.4 that this condition is not innocuous and that it implies a fixed number of regimes as well as $\min_j\{\delta_{\tau_j}\} > 0$; that is δ_{τ_j} does not drift to 0 as $T \to \infty$. Avoiding this assumption would imply stronger conditions on the process $\{y_t, \mathbf{x}_t\}$ (see, e.g., Chan et al., 2014). The assumption G.4 refers to ergodicity and stationarity of each segment and imposes the standard exogeneity hypothesis. This assumption ensures that sampled counterparts of the first two moments of $\{\mathbf{x}_t\epsilon_t\}$ are converging to finite values. Importantly, it does not rule out conditional heteroskedasticity. Eventually, condition G.5 defines restrictions on the tuning parameters rate. The same condition applies in Dicker et al. (2013). The consistency of SELO estimator is given by the following Proposition.

Proposition 3.1. *Assume that G.1-G.5 hold and let,*

$$f_T(\boldsymbol{\beta}) = \frac{1}{T}||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + \sum_{j=2}^{m}\sum_{k=1}^{K}\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda). \tag{3.7}$$

There exists a sequence of $\sqrt{T}$-consistent local minima $\hat{\boldsymbol{\beta}}$ of $f_T(\boldsymbol{\beta})$ as defined by Equation (3.7) such that:

1. $\lim_{T\to\infty} \mathbf{P}\left(\left\{(j,k); \hat{\beta}_{jk} \neq 0\right\} = A\right) = 1$
2. $\forall\ \delta > 0$, $\lim_{T\to\infty} \mathbf{P}\left(||\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A^*|| > \delta\right) = 0$

Proof. The proof is given in Appendix C.1 □

Remark 3.2. *Proposition 3.1 also applies when the set of breakpoints contains additional spurious break dates. In particular, Proposition 3.1 holds if we relax assumption G.1 by the less restrictive assumption:*

Assumption H. *$\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_{\hat{m}}\}$ with $\hat{m} \geq m$, $\boldsymbol{\tau}_0 \subseteq \boldsymbol{\tau}$ and $\forall j \in [1, \hat{m}]$, we have $\tau_j - \tau_{j-1} = T\delta_{\tau_j} \to \infty$, with $\sum_{j=1}^{\hat{m}} \delta_{\tau_j} = 1$.*

3.3 Estimation

The objective function to minimize is given by

$$\begin{aligned} f(\boldsymbol{\beta}) &= ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + T\frac{\lambda}{\ln(2)}\sum_{j=2}^{m}\sum_{k=1}^{K}\ln\left(\frac{2(\frac{|\Delta\beta_{jk}|}{a_k}) + \zeta}{(\frac{|\Delta\beta_{jk}|}{a_k}) + \zeta}\right), \\ &= ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + \sum_{j=2}^{m}\sum_{k=1}^{K}\ln q_k(\Delta\beta_{jk}), \end{aligned} \tag{3.8}$$

in which $q_k(\Delta\beta_{jk}) = \left(\frac{2(\frac{|\Delta\beta_{jk}|}{a_k}) + \zeta}{(\frac{|\Delta\beta_{jk}|}{a_k}) + \zeta} \right)^{\left(\frac{T\lambda}{\ln(2)}\right)}$. Due to the penalty function, we cannot find any analytical expression of the minimizer. In addition to that, the function likely exhibits many local modes which complicates the optimization. We address the problem of finding the global mode by using a deterministic annealing expectation-minimization (DAEM) algorithm (see Ueda and Nakano, 1998). To do so, we first approximate the penalty function by a mixture of two Normal components (to take into account the large tail of the SELO penalty function), the details of it are given in Appendix C.1.5. Secondly, since minimizing the sum of squared residuals is identical to maximizing a likelihood function when the error term is normally distributed, we work with the following model, $\mathbf{y} = \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta} + \eta$, where $\eta \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_T)$. The modified model implied the following objective function to maximize with respect to $(\boldsymbol{\beta}, \sigma^2)$:

$$f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = -\frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 - \sum_{j=2}^{m}\sum_{k=1}^{K}\ln g_k(\Delta\beta_{jk}), \tag{3.9}$$

where $g_k(\Delta\beta_{jk}) = \sum_{i=1}^{2}\omega_i^{(k)} f_N(\Delta\beta_{jk}|\mu_i^{(k)}, s_i^{(k)})$, $f_N(x|\mu, s)$ stands for the normal density function evaluated at x with expectation and variance given by μ and s respectively and $\omega_i^{(k)} \in (0,1)$ with $\sum_{i=1}^{2}\omega_i^{(k)} = 1$. Note that the function $f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)$ in Equation (3.9) is proportional to the posterior log-density of the parameter distribution $\boldsymbol{\beta}, \sigma^2|\mathbf{y}$ from a Bayesian perspective with prior distributions given by $f(\sigma^2, \boldsymbol{\beta}_1) \propto 1$ and $f(\Delta\beta_{jk}) = g_k(\Delta\beta_{jk})$ for $j \in [2, m]$ and $k \in [1, K]$. In particular, the distribution $f(\Delta\beta_{jk})$ can be understood as a spike and slab prior (e.g. George and McCulloch, 1993) and our optimization procedure fits into the framework of Ročková and George (2014) which proposes tackling the linear variable selection problem with the EM algorithm and its DAEM variant. The optimization is therefore equivalent to finding the mode of $\boldsymbol{\beta}, \sigma^2|\mathbf{y}$. Using a data augmentation approach, we add latent variables $\mathbf{z} = (z_{21}, z_{22}, \ldots, z_{mK})'$ such that $f(z_{jk} = i) = \omega_i^{(k)}$, $\forall j \in [2, m]$, $\forall k \in [1, K]$ and $\forall i \in [1, 2]$. With these latent variables, we can write the prior distribution of $\Delta\beta_{jk}$ in a convenient hierarchical way as follows,

$$f(\Delta\beta_{jk}|z_{jk} = i) \quad = \quad f_N(\Delta\beta_{jk}|\mu_i^{(k)}, s_i^{(k)}), \text{ and } f(z_{jk} = i) = \omega_i^{(k)}.$$

By fixing $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2\}$, the EM algorithm (and its DAEM variant) solves the following optimization at iteration n,

$$\text{argmax}_{\boldsymbol{\theta}_n} Q(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{n-1}) \quad = \quad \text{argmax}_{\boldsymbol{\theta}_n}\mathbb{E}_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}_{n-1}}(\ln f(\boldsymbol{\theta}_n, \mathbf{z}|\mathbf{y})|\mathbf{y}, \boldsymbol{\theta}_{n-1}).$$

One can easily show that maximizing $Q(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{n-1})$ implies that $f(\boldsymbol{\theta}_n|\mathbf{y}) \geq f(\boldsymbol{\theta}_{n-1}|\mathbf{y})$.

3.3.1 Derivation of the DAEM algorithm

To apply the DAEM algorithm, we need to find an expression of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{n-1})$. Given a set of parameter $\boldsymbol{\theta}_{n-1}$, we have that

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{n-1}) &= \mathbb{E}_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}_{n-1}}(\ln f(\boldsymbol{\theta}|\mathbf{y},\mathbf{z})f(\mathbf{z}|\mathbf{y})|\mathbf{y},\boldsymbol{\theta}_{n-1}) \\ &\propto \ln f(\mathbf{y}|\boldsymbol{\beta},\sigma^2) + \ln f(\boldsymbol{\beta}_1,\sigma^2) - \sum_{j=2}^{m}\sum_{k=1}^{K}\sum_{i=1}^{2}\frac{(\Delta\beta_{kj}-\mu_i^{(k)})^2}{2s_i^{(k)}} f(z_{kj}=i|\mathbf{y},\boldsymbol{\theta}_{n-1}), \\ &\propto \ln f(\mathbf{y}|\boldsymbol{\beta},\sigma^2) - \frac{1}{2}\sum_{i=1}^{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i(\boldsymbol{\beta}-\boldsymbol{\mu}_i), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\mu}_i &= (\underbrace{0,0,\ldots,0}_{\text{K-dimensional}},\mu_i^{(1)},\mu_i^{(2)},\ldots,\mu_i^{(K)},\mu_i^{(1)},\ldots)' \in \Re^{mK\times 1}, \\ \boldsymbol{\Sigma}_i &= \text{diag}(\underbrace{0,0,\ldots,0}_{\text{K-dimensional}},\frac{p_{21}^{(i)}}{s_i^{(1)}},\frac{p_{22}^{(i)}}{s_i^{(2)}},\ldots,\frac{p_{2K}^{(i)}}{s_i^{(K)}},\frac{p_{31}^{(i)}}{s_i^{(1)}},\ldots,\frac{p_{mK}^{(i)}}{s_i^{(K)}}), \end{aligned}$$

with $p_{jk}^{(i)} = f(z_{jk}=i|\mathbf{y},\boldsymbol{\theta}_{n-1})$ $\forall i \in [1,2]$,$\forall j \in [2,m]$ and $\forall k \in [1,K]$. Importantly, the difference between the EM algorithm and its DA version only appears in the quantities $p_{jk}^{(i)}$. In fact, the DAEM algorithm introduces an increasing function $\phi(r) : [1,N] \to (0,1]$ such that $0 < \phi(1) \leq 1$ and $\phi(N) = 1$. For each value $r = 1,\ldots,N$, it applies recursively the EM algorithm (that starts with the final estimate of the previous EM algorithm) where the posterior probabilities $p_{jk}^{(i)}$ are denoted $p_{jk}^{(i,\phi(r))}$ and are modified as follows,

$$p_{jk}^{(i,\phi(r))} \propto (f_N(\Delta\beta_{jk}|\mu_i^{(k)},s_i^{(k)})\omega_i^{(k)})^{\phi(r)}. \tag{3.10}$$

When $r = N$, the increasing function $\phi(r) = 1$ and the standard EM algorithm is run (but with a promising starting point). To find the maximum of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{n-1})$, we sequentially maximize $\boldsymbol{\beta}$ given σ^2 and then σ^2 with respect to $\boldsymbol{\beta}$. This approach, called coordinate iterative ascent, operates in two steps:

1. Compute $\boldsymbol{\beta}_n = \arg\max_{\boldsymbol{\beta}} Q(\boldsymbol{\beta},\sigma_{n-1}^2|\boldsymbol{\theta}_{n-1})$.
2. Compute $\sigma_n^2 = \arg\max_{\sigma^2} Q(\boldsymbol{\beta}_n,\sigma^2|\boldsymbol{\theta}_{n-1})$.

At the end of the two steps, we necessarily have $Q(\boldsymbol{\beta}_{n-1},\sigma_{n-1}^2|\boldsymbol{\theta}_{n-1}) \leq Q(\boldsymbol{\beta}_n,\sigma_{n-1}^2|\boldsymbol{\theta}_{n-1}) \leq Q(\boldsymbol{\beta}_n,\sigma_n^2|\boldsymbol{\theta}_{n-1})$. The maximisation of $\boldsymbol{\beta}$ given σ_{n-1}^2 leads to

$$\boldsymbol{\beta}_n = [\sigma_{n-1}^{-2}X_{\boldsymbol{\tau}}'X_{\boldsymbol{\tau}} + \sum_{i=1}^{2}\boldsymbol{\Sigma}_i]^{-1}[\sigma_{n-1}^{-2}X_{\boldsymbol{\tau}}'\mathbf{y} + \sum_{i=1}^{2}\boldsymbol{\Sigma}_i\boldsymbol{\mu}_i].$$

The update of σ^2 conditional to $\boldsymbol{\beta}_n$ is given by

$$\sigma_n^2 = \frac{[(\mathbf{y}-X_\tau\boldsymbol{\beta}_n)'(\mathbf{y}-X_\tau\boldsymbol{\beta}_n)}{T}.$$

We summarize the DAEM procedure in Algorithm 3.1.. In practice, the minimum distance e indicating a convergence of the algorithm is set to 10^{-5} and the number of DAEM iteration N is fixed to 10.

Algorithm 3.1. DAEM algorithm

Initialize $\boldsymbol{\beta}_0$ using Algorithm;
Set $\sigma_0^2 = \dfrac{[(\mathbf{y} - \mathbf{X}_\tau \boldsymbol{\beta}_0)'(\mathbf{y} - \mathbf{X}_\tau \boldsymbol{\beta}_0)}{T}$, $\phi(1) = (\dfrac{1}{N})^2$, $r = 1$ and dist $= \infty$;
while $r <= N$ **do**
  Set $n = 0$ and $\boldsymbol{\theta}_n = (\boldsymbol{\beta}_0', \sigma_0^2)'$;
  while *dist* $> e$ **do**
    Increment $n = n + 1$;
    for $i = 1, 2$ **do**
      Compute the posterior probabilities $p_{jk}^{(i,\phi(r))}$ given in Equation (3.10);
    Compute the mean parameters $\boldsymbol{\beta}_n = [\sigma_{n-1}^{-2} \mathbf{X}_{\boldsymbol{\tau}}' \mathbf{X}_{\boldsymbol{\tau}} + \sum_{i=1}^{2} \boldsymbol{\Sigma}_i]^{-1} [\sigma_{n-1}^{-2} \mathbf{X}_{\boldsymbol{\tau}}' \mathbf{y} + \sum_{i=1}^{2} \boldsymbol{\Sigma}_i \boldsymbol{\mu}_i]$;
    Compute the variance parameter $\sigma_n^2 = \dfrac{[(\mathbf{y} - \mathbf{X}_\tau \boldsymbol{\beta}_n)'(\mathbf{y} - \mathbf{X}_\tau \boldsymbol{\beta}_n)}{T}$;
    Set $\boldsymbol{\theta}_n = (\boldsymbol{\beta}_n', \sigma_n^2)'$ and compute the distance value dist $= ||\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}||_2$;
  Increment $r = r + 1$ and set $\phi(r) = (\dfrac{r}{N})^2$;
  Set $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_n$ and $\sigma_0^2 = \sigma_n^2$;

The EM and the DAEM algorithms are sensitive to starting values. Inspired by Zhao et al. (2012), we mitigate this issue by randomly exploring the model space using a swapping approach before applying the DAEM algorithm. To be specific, we generate N_{init} values as explained in Algorithm 3.2. and we initialize the DAEM algorithm with the parameter estimates that minimize the penalized function given in Equation (3.8). In practice, we set $N_{\text{init}} = \min(2^{(m-1)K-1}, 3000)$.

Algorithm 3.2. Initialization of the DAEM algorithm

for $n = 1$ *to* N_{init} **do**
  Set $\hat{A} = \emptyset$ and sample $p \sim U[0, 1]$;
  for $j = 2, ..., m$ *and for* $k = 1, ..., K$ **do**
    $\hat{A} = \hat{A} \cup (j, k)$ with probability p ;
  $(f_n, \boldsymbol{\beta}_n) = \text{Swap}(\hat{A})$ (see Algorithm 3.3.)
return the OLS estimates $\boldsymbol{\beta}_{\hat{n}}$ such that $\hat{n} = \underset{n \in [1, N_{\text{init}}]}{\arg\min}\ f_n$;

Algorithm 3.3. Swap the set of indices - Swap($\hat{A}$)

for $j = 2, ..., m$ *and for* $k = 1, ..., K$, *and Given a set of indices* $\hat{A}$ *defining the parameters* $\Delta\boldsymbol{\beta} \neq 0$ **do**
  Build the sets $\tilde{A}_{jk} = \hat{A} \cup (j,k)$ if $\hat{A} \cap (j,k) = \emptyset$ or the set $\tilde{A}_{jk} = \hat{A} \backslash (j,k)$ otherwise;
for each set $\tilde{A}_{jk}$, compute the OLS estimates $(\hat{\boldsymbol{\beta}}_{jk})$ and the penalized function $f_{jk} = f(\hat{\boldsymbol{\beta}}_{jk})$ (see (3.8));
for the set $\hat{A}$, compute the OLS estimates $(\hat{\boldsymbol{\beta}}_{\hat{A}})$ and the penalized function $f_{\hat{A}} = f(\hat{\boldsymbol{\beta}}_{\hat{A}})$ (see (3.8));
find $(\hat{j}, \hat{k}) = \underset{j,k}{\arg\min}\, f_{jk}$;
if $f_{\hat{j}\hat{k}} < f_{\hat{A}}$ **then**
  return $\hat{\boldsymbol{\beta}}_{\hat{j}\hat{k}}$ and $f_{\hat{j}\hat{k}}$;
else
  return $\hat{\boldsymbol{\beta}}_{\hat{A}}$ and $f_{\hat{A}}$;

3.4 Selection of the penalty parameters and parameter uncertainties

The SELO penalty function exhibits two tuning parameters $\boldsymbol{a}$ and λ. The standard approach to fix them consists in considering a grid of values of these parameters and in selecting the parameters that maximize a (generally consistent) information criterion (e.g., Zhang et al., 2010). Instead of relying on a standard information criterion and select the tuning parameters $\boldsymbol{a}$ and λ that maximize it, we consider each pair $(\boldsymbol{a}, \lambda)$ as a model to take into account the model uncertainty. For a given value of $(\boldsymbol{a}, \lambda)$, the DAEM algorithm exposed in Section 3.3.1 provides an estimate $\hat{\Delta\boldsymbol{\beta}}$ of $\Delta\boldsymbol{\beta}$ which delivers an estimate of $\hat{A}$, i.e., the set of indices with $\Delta\hat{\beta}_{jk} \neq 0$ for $j \in [2, m]$ and for $k \in [1, K]$. This set tells us which covariates should be included in the linear regression and which should not. Let us denote by $\tilde{\mathbf{X}}_{\boldsymbol{\tau}}^{\hat{A}}$ the covariates related to the first-difference estimates that are different from zero. We use the following criterion for selecting $\boldsymbol{a}$ and λ:

$$f(\mathbf{y}|\mathbf{a}, \lambda, \boldsymbol{\tau}) = \left(\frac{g_{\hat{A}}}{1+g_{\hat{A}}}\right)^{k_{\hat{A}}/2} \left[\frac{g_{\hat{A}}}{1+g_{\hat{A}}} s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g_{\hat{A}})} s_{\tilde{\mathbf{X}}_{\tau_0}, \tilde{\mathbf{X}}_{\boldsymbol{\tau}}^{\hat{A}}}\right]^{-\frac{T-K}{2}}, \tag{3.11}$$

where $s_{\tilde{\mathbf{X}}_{\tau_0}}$ stands for the residual sum of squares (RSS) from the ordinary least squares (OLS) with $\mathbf{X} = \tilde{\mathbf{X}}_{\tau_0}$ (i.e., a regression without break), $s_{\tilde{\mathbf{X}}_{\tau_0}, \tilde{\mathbf{X}}_{\boldsymbol{\tau}}^{\hat{A}}}$ is the RSS from the OLS with $\mathbf{X} = (\tilde{\mathbf{X}}_{\tau_0}, \tilde{\mathbf{X}}_{\boldsymbol{\tau}}^{\hat{A}})$, the value $k_{\hat{A}} = |\hat{A}|$ denotes the number of first-difference parameters different from zero in the model and $g_{\hat{A}}$ is a user parameter. We properly derive the criterion in Appendix C.2. Fernandez et al. (2001) show that the criterion (3.11) is consistent in the sense that it selects asymptotically the true subset of regressors when $g_{\hat{A}} = w(T)^{-1}$ as stated in proposition 3.2.

Proposition 3.2. *(Adaption of Fernandez et al., 2001). Conditional on the true break dates, the criterion (3.11) is asymptotically maximized for the true subset of covariate A if the following conditions on the parameter $g_{\hat{A}} = w(T)^{-1}$ holds:*

1. $\lim_{T\to\infty} w(T) = \infty$,
2. $\lim_{T\to\infty} \frac{w'(T)}{w(T)} = 0$,
3. $\lim_{T\to\infty} \frac{T}{w(T)} \in [0, \infty)$.

Proof. See Appendix C.3. □

Remark 3.3. *Proposition 3.2 can be readily adapted when the conditioning set is a potential break date set complying with Assumption H.*

In Fernandez et al. (2001), they advocate for setting $g_{\hat{A}} = \min(T^{-1}, (k_{\hat{A}} + K)^{-2})$ as this prior empirically delivers good results for selecting the true covariates in standard linear regressions. However, we deviate from this benchmark prior by fixing $g_{\hat{A}} = \frac{1}{T^\alpha - 1}$ with $\alpha = 1$ when $k_{\hat{A}} = 0$ and $\alpha = \frac{k_{\hat{A}} + \hat{m}_{\hat{A}} - 1}{k_{\hat{A}}} > 1$ when $k_{\hat{A}} > 0$ in which $\hat{m}_{\hat{A}}$ denotes the number of active segments. When $\alpha > 1$, we show in Appendix C.3.1 that the criterion in Equation (3.11) asymptotically converges in probability to

$$\ln f(\mathbf{y}|\mathbf{a}, \lambda, \boldsymbol{\tau}) - \left(-\frac{T}{2} \ln s_{\tilde{\mathbf{X}}_{\tau_0}, \tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}} - \frac{\alpha k_{\hat{A}}}{2} \ln T \right) \xrightarrow{p} 0. \tag{3.12}$$

The asymptotic value is equivalent to the Bayesian information criterion (BIC) of a linear regression model exhibiting a number of parameters of $\alpha k_{\hat{A}}$.[1] Consequently, the model penalty takes additionally into account the number of active breakpoints when $\alpha = \frac{k_{\hat{A}} + \hat{m}_{\hat{A}} - 1}{k_{\hat{A}}}$. This stronger penalty works empirically well and is motivated by several CP papers advocating for stronger penalties than the BIC as it tends to overfit the number of regimes in finite sample (see, e.g., Liu et al., 1997; Zhang and Siegmund, 2007; Kim and Kim, 2016).

Interestingly, criterion (3.11) stands for a marginal likelihood in the Bayesian paradigm under $\epsilon \sim \mathcal{N}(0, \sigma^2 I_T)$ and the following prior,

$$\begin{aligned} f(\boldsymbol{\beta}_1, \sigma^2) &\propto \sigma^{-2}, \\ f(\Delta\boldsymbol{\beta}_{\hat{A}}|\sigma^2, \boldsymbol{\tau}) &\sim \mathcal{N}(\mathbf{0}, \sigma^2 (g_{\hat{A}} (\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}})' \mathbf{M}_{\tilde{\mathbf{X}}_{\tau_0}} \tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}})^{-1}), \text{ and } f(\Delta\boldsymbol{\beta}_{\hat{A}^c}) \sim \text{Dirac}(\mathbf{0}), \end{aligned} \tag{3.13}$$

[1] The BIC of a linear regression model with K parameters is given by $-\frac{T}{2} \ln(\frac{s_{\tilde{\mathbf{X}}_{\tau_0}, \tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}}{T}) - \frac{K}{2} \ln T$. So the marginal likelihood criterion of Equation (3.11) converges to the BIC up to an additive constant (that is $\frac{T}{2} \ln T$).

where $\mathbf{M}_{\tilde{\mathbf{X}}_{\tau_0}} = I_T - \tilde{\mathbf{X}}_{\tau_0}((\tilde{\mathbf{X}}_{\tau_0})'\tilde{\mathbf{X}}_{\tau_0})^{-1}(\tilde{\mathbf{X}}_{\tau_0})'$. The prior distributions given by Equations (3.13) lead to simple posterior inference. The posterior distribution of the model parameters are given by, see Appendix C.2.1 for derivations,

$$\begin{aligned}
\sigma^2|\mathbf{y},\boldsymbol{\tau} &\sim \mathcal{IG}\left(\frac{T-K}{2}, \frac{\frac{g_{\hat{A}}}{1+g_{\hat{A}}}s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g_{\hat{A}})}s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}}{2}\right),\\
\boldsymbol{\beta}_1|\mathbf{y},\sigma^2,\Delta\boldsymbol{\beta},\boldsymbol{\tau} &\sim \mathcal{N}\left((\tilde{\mathbf{X}}'_{\tau_0}\tilde{\mathbf{X}}_{\tau_0})^{-1}\tilde{\mathbf{X}}'_{\tau_0}(\mathbf{y}-\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}\Delta\boldsymbol{\beta}), \sigma^2(\tilde{\mathbf{X}}'_{\tau_0}\tilde{\mathbf{X}}_{\tau_0})^{-1}\right),\\
\Delta\boldsymbol{\beta}_{\hat{A}}|\mathbf{y},\sigma^2,\boldsymbol{\tau} &\sim \mathcal{N}\left(\frac{[(\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}})'\mathbf{M}_{\tilde{\mathbf{X}}_{\tau_0}}\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}]^{-1}(\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}})'\mathbf{M}_{\tilde{\mathbf{X}}_{\tau_0}}\mathbf{y}}{1+g_{\hat{A}}}, \frac{\sigma^2[(\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}})'\mathbf{M}_{\tilde{\mathbf{X}}_{\tau_0}}\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}]^{-1}}{(1+g_{\hat{A}})}\right),\\
\Delta\boldsymbol{\beta}_{\hat{A}^c}|\mathbf{y},\boldsymbol{\tau} &= \mathbf{0},
\end{aligned}$$

in which $\mathcal{IG}(-,-)$ denotes the Inverse-Gamma distribution. Consequently, we can go beyond selecting the best pair $(\boldsymbol{a}_p, \lambda_p)$ (i.e., the pair that maximizes the criterion (3.11)) and can take the uncertainty of this selection into account. Given a set of models $M_z = (\boldsymbol{a}_z, \lambda_z)$, with $z = 1, ..., Z$, we can directly assess the posterior probability of a specific model as follows

$$f(M_p|\mathbf{y},\boldsymbol{\tau}) = \frac{f(\mathbf{y}|\boldsymbol{a}_p,\lambda_p,\boldsymbol{\tau})f(M_p|\boldsymbol{\tau})}{\sum_{z=1}^{Z} f(\mathbf{y}|\boldsymbol{a}_z,\lambda_z,\boldsymbol{\tau})f(M_z|\boldsymbol{\tau})}, \forall p \in [1, Z], \tag{3.14}$$

where $f(M_z|\boldsymbol{\tau})$ denotes the prior probability of model M_z. In this paper, we assume uninformative prior, so $f(M_z|\boldsymbol{\tau}) = Z^{-1}$. The posterior probability can be used to account for uncertainty on the selected regressors. In fact, we have

$$\begin{aligned}
f(\boldsymbol{\beta}_1, \Delta\boldsymbol{\beta}, \sigma^2, M|\mathbf{y},\boldsymbol{\tau}) = f(\boldsymbol{\beta}_1|\mathbf{y},\boldsymbol{\tau},\sigma^2,\Delta\boldsymbol{\beta},M)f(\Delta\boldsymbol{\beta}|\mathbf{y},\boldsymbol{\tau},\sigma^2,M)\\
f(\sigma^2|\mathbf{y},\boldsymbol{\tau},M)f(M|\mathbf{y},\boldsymbol{\tau})
\end{aligned} \tag{3.15}$$

It is worth emphasizing that the consistent property of the criterion (3.11) does not depend on the normality assumption. Only, the posterior distribution of the model parameters does. We do not see this as a limitation since one can easily extend the model with another distributional assumption and compute the posterior distribution by numerical integrations.

3.4.1 Prediction using Bayesian model averaging

Equation (3.14) shows how to take into account the uncertainty of the model parameters with respect to the selection of the SELO parameters. The Bayesian paradigm also provides a simple tool to forecast the series taking this uncertainty into account. In particular, the predictive density $f(y_{T+1:T+h}|\mathbf{y})$, for $h \geq 1$, is related to the posterior density as follows

$$\begin{aligned}
f(y_{T+1:T+h}|\mathbf{y},\boldsymbol{\tau}) &= \sum_{z=1}^{Z}\int f(y_{T+1:T+h},\boldsymbol{\beta}_1,\Delta\boldsymbol{\beta},\sigma^2,M_z|\mathbf{y},\boldsymbol{\tau})d\boldsymbol{\beta}_1 d\Delta\boldsymbol{\beta} d\sigma^2,\\
&\approx \frac{1}{N}\sum_{i=1}^{N} f(y_{T+1:T+h}|\mathbf{y},\boldsymbol{\tau},\boldsymbol{\beta}_1^{(i)},\Delta\boldsymbol{\beta}^{(i)},(\sigma^2)^{(i)},M^{(i)}),
\end{aligned} \tag{3.16}$$

where $\{\boldsymbol{\beta}_1^{(i)}, \Delta\boldsymbol{\beta}^{(i)}, (\sigma^2)^{(i)}, M^{(i)}\}_{i=1}^N$ are independent draws from the posterior distribution (i.e., $\boldsymbol{\beta}_1, \Delta\boldsymbol{\beta}, \sigma^2, M|\mathbf{y}, \boldsymbol{\tau}$). From (3.16), it is apparent that the predictive density takes the model uncertainty into account.[2] This feature should be contrasted with the standard penalized regression literature in which forecasting is performed using one unique set of parameter estimates; i.e., the estimates given by one penalty parameter selected, for instance, by cross-validation or by an information criterion.

In practice, simulations from the posterior distribution are not required for evaluating the predictive density. Assuming that the future covariates $\mathbf{x}_{T+1:T+h}$ are observed at time T, the predictive distribution of $y_{T+1:T+h}$ given a model M_z turns out to be a multivariate student distribution. Supplementary Appendix C.2.2 documents the analytical expression of $f(y_{T+1:T+h}|\mathbf{y}, M_z)$. Therefore, we can efficiently take into account model uncertainty in the predictive density since Equation (3.16) simplifies into

$$f(y_{T+1:T+h}|\mathbf{y}, \boldsymbol{\tau}) \quad = \quad \sum_{z=1}^{Z} f(y_{T+1:T+h}|\mathbf{y}, \boldsymbol{\tau}, M_z) f(M_z|\mathbf{y}, \boldsymbol{\tau}). \tag{3.17}$$

3.4.2 How to choose the values of λ and $\boldsymbol{a}$

When the number of models to consider is too large to directly explore the model space using the criterion (3.11) (i.e., when $(m-1)K > 10$, see remark 3.1), we rely on the SELO penalty function to uncover the promising explanatory variables. While the asymptotic result of Proposition 3.1 is reassuring, it only applies if the parameters λ and $\boldsymbol{a}$ are adequately chosen. Similar to what is generally done in the penalized regression literature, we propose to explore many values of λ and $\boldsymbol{a}$ and consider each couple as a model that would be ultimately discriminated via criterion (3.11). For the parameter $\boldsymbol{a}$, we use a value of $a_i = \kappa \times \text{std}(\hat{\beta}_{j1})$ for each j of the K parameters per regime where $\text{std}(\hat{\beta}_{j1})$ stands for the standard deviation of the OLS estimate $\hat{\beta}_{j1}$ when we assume no break in the linear regression (i.e., $\mathbf{X} = \mathbf{X}_{\tau_0}$). We test several values for the parameter κ, namely $\kappa \in \{0.1, 1\}$. Regarding the penalty parameter λ, we test 50 different values uniformly spaced in the interval $(0, \bar{\lambda}]$ in which $\bar{\lambda} = 2\ln T$. The penalty imposes by the upper bound $\bar{\lambda}$ is conservative enough as it is stronger than standard information criteria such as the BIC (that corresponds to a penalty of $\frac{1}{2}\ln T$) and the modified BIC.

[2] Using the full marginal likelihood for weighting the models' predictions could raise concerns as only the last segment matters in CP processes. However, as marginal likelihood is frequently used for selecting the number of regimes in the literature and because it is also informative about the fit of the last regime, this average should give large weights to the models exhibiting a good fit at the end of the sample. Nevertheless, we could also weight the models' predictions using the predictive marginal likelihood $f(y_{t_1+1:T}|y_{1:t_1}, \boldsymbol{\tau})$ in which t_1 is a user-defined value.

3.5 Break date detection

In this Section, we present one approach to obtain a set of potential break dates. Before going into details, it is worth emphasizing that our method for detecting which parameters vary when a break occurs is independent of the segmentation detection procedure used in the first phase. To build the break date set, we could, for instance, adapt the dynamic programming method of Bai and Perron (2003) for the marginal likelihood given by Equation (3.11) and therefore propose our own CP detection method. We could also detect the locations of the segments using one of the standard segmentation approaches such as Bai and Perron (1998), Killick et al. (2012) or Korkas and Fryzlewiczv (2017). Even better, we could apply several CP detection algorithms and discriminate between the sets of breakpoints by comparing their marginal likelihoods *once* the SELO optimization has been carried out on each set. However, as the emphasis of the paper is not on the break detection, we prefer relying on one break detection procedure, the one documented in Yau and Zhao (2016), because i) it delivers a set of potential break dates with a computational complexity of $\mathcal{O}(T\,(\log(T))^2)$ (which is faster than $\mathcal{O}(T^2)$, i.e., the complexity of the dynamic programming method of Bai and Perron (2003)) and because ii) we slightly improve their CP detection procedure. In particular, their estimated breakpoints depend on one tuning parameter, the radius h. Instead of fixing it, we use multiple values of h and we also adapt their approach to end up with a potential breakpoint set.

It is worth noting that, as the paper combines model selection and CP detection methods, our approach only requires a set of potential break dates that includes the correct break dates. By penalizing the parameter variation between two consecutive regimes, the spurious break dates are consistently deleted (see remarks 3.2 and 3.3).

3.5.1 Segmentation procedure

Yau and Zhao (2016) propose a likelihood ratio scan method in three steps for estimating multiple break dates in piecewise stationnary processes. They also establish the consistency of the estimated number and location fractions of the CPs. We apply their three steps to detect the break dates but we modify them to reduce the computational burden and to keep at the end of the procedure a potential break date set (that could overestimate the true number of regimes). We now detail the three steps that we use to segment the data.

First step. Fix a window radius $h \in [K+1, T-K]$. For $t = h$ to $T-h$, compute the likelihood ratio scan statistic given by,

$$S_h(t) = \frac{1}{h} L_{t-h+1:t}(\hat{\boldsymbol{\beta}}, \hat{\sigma}) + \frac{1}{h} L_{t+1:t+h}(\hat{\boldsymbol{\beta}}, \hat{\sigma}) - \frac{1}{h} L_{t-h+1:t+h}(\hat{\boldsymbol{\beta}}, \hat{\sigma}), \tag{3.18}$$

where $\hat{L}_{t_1:t_2}(\hat{\boldsymbol{\beta}}, \hat{\sigma})$ denotes the maximum value of the log-likelihood of model (3.1) over the segment $t \in [t_1, t_2]$, assuming that $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. Then, the set $\Gamma(h)$ of potential break dates

is given by,

$$\Gamma(h) = \left\{ j \in \{h+h+1, \ldots, T-h\}; S_h(j) = \max_{t \in [j-h, j+h]} S_h(t) \right\}, \tag{3.19}$$

where $S_h(t) = 0$ for $t < h$ and $t > T - h$. As the window radius h is crucial, we differ from Yau and Zhao (2016) by using a grid of M values uniformly-spaced in the interval $[\frac{h_{\text{YZ}}}{2}, 2h_{\text{YZ}}]$ in which h_{YZ} denotes their advocated value that is $h_{\text{YZ}} = \max\left\{25, (\log(T))^2\right\}$ when $T < 800$ and $h_{\text{YZ}} = \max\left\{50, 2\,(\log(T))^2\right\}$ otherwise. So, at the end of the first step, we end up with M potential break date sets, i.e., $\Gamma(h_1), \ldots, \Gamma(h_M)$.

Second step. For every $z \in [1, M]$ and $i \in [1, m_{h_z} - 1]$ where $m_{h_z} = |\Gamma(h_z)| + 1$, we re-estimate each break date location $\tau_i^{(z)} \in \Gamma(h_z)$ as follows

$$\hat{\tau}_i^{(z)} \quad = \quad \text{argmax}_{t \in [\tau_i^{(z)} - h_z, \tau_i^{(z)} + h_z]} L_{\tau_i^{(z)} - \lfloor 1.5 h_z \rceil : t}(\hat{\boldsymbol{\beta}}, \hat{\sigma}) + L_{t+1 : \tau_i^{(z)} + \lfloor 1.5 h_z \rceil}(\hat{\boldsymbol{\beta}}, \hat{\sigma}),$$

in which $\lfloor x \rceil$ stands for the nearest integer to x. Gathering all the new locations in the set $\hat{\Gamma}(h_z) = \{\hat{\tau}_1^{(z)}, \ldots, \hat{\tau}_{m_{h_z}}^{(z)}\}$, it is clear from Theorems 1 to 3 in Yau and Zhao (2016) that for any $j \in \{1, \ldots, m-1\}$, there exist $\hat{\tau}_i^{(z)} \in \hat{\Gamma}(h_z)$ with $i \in [1, m_{h_z} - 1]$ such that $\hat{\tau}_i^{(z)} - \tau_j = \mathcal{O}_p(1)$.

Third step. We select the best breakpoints among the M potential break date sets by minimizing the Minimum Description Length (MDL) defined by, for $z \in [1, M]$,

$$\begin{aligned} \text{MDL}(h_z) &= \ln^+ (m_{h_z} - 1) + m_{h_z} \ln (T) \\ &\quad + \sum_{j=1}^{m_{h_z}} \left(\frac{K+1}{2} \log(\hat{\tau}_j^{(z)} - \hat{\tau}_{j-1}^{(z)}) - L_{\hat{\tau}_{j-1}^{(z)} + 1 : \hat{\tau}_j^{(z)}}(\hat{\boldsymbol{\beta}}, \hat{\sigma}) \right), \end{aligned} \tag{3.20}$$

where $\hat{\tau}_0 = 0$, $\hat{\tau}_{m_{h_z}} = T$ and $\{\hat{\tau}_j\}_{j=2,\ldots,m_{h_z}-1} = \hat{\Gamma}(h_z)$. In practice, we fix $M = 30$.

3.5.2 Break uncertainty

Given a set of break dates obtained either from the procedure described in Section 3.5.1 or from any other existing break detection method such as the one of Bai and Perron (1998), our method to uncover the partial structural changes can be undertaken. Let us denote by $M_* = (\boldsymbol{a}_*, \lambda_*)$ the SELO parameters maximizing the marginal likelihood criterion (3.11) and their corresponding break dates $J = \{\bar{\tau}_0 = 0, \bar{\tau}_1, \ldots, \bar{\tau}_{\hat{m}-1}, \bar{\tau}_{\hat{m}} = T\}$. To provide break uncertainty, we shall infer the posterior distribution of the structural breaks; i.e., $\boldsymbol{\tau} \equiv \tau_1, \ldots, \tau_{\hat{m}-1} | \mathbf{y}, M_*$. To do so, we first assume uninformative priors for the break dates using the set J. For $i = 1, ..., \hat{m} - 1$, the break parameter τ_i is driven by a Uniform distribution as follows

$$\tau_i \quad \sim \quad \mathcal{U}[\left\lfloor \frac{\bar{\tau}_{i-1} + \bar{\tau}_i}{2} \right\rfloor + \gamma, \left\lfloor \frac{\bar{\tau}_{i-1} + \bar{\tau}_i}{2} \right\rfloor - \gamma],$$

in which $\lfloor x \rfloor$ stands for the nearest integer less than or equal to x and $\gamma = (K+1)$ is a minimum duration parameter ensuring that the marginal likelihood criterion (3.11) can be

computed for any break parameters complying with the prior distributions given by Equation (3.13). The posterior density is proportional to

$$f(\boldsymbol{\tau}|\mathbf{y}, M_*) \propto f(\mathbf{y}|M_*, \boldsymbol{\tau}) f(\mathbf{y}|M_*, \boldsymbol{\tau}) \left(\prod_{i=1}^{\hat{m}-1} \mathbf{1}_{\{\boldsymbol{\tau}_i \in [\lfloor \frac{\bar{\tau}_{i-1}+\bar{\tau}_i}{2} \rfloor - \gamma, \lfloor \frac{\bar{\tau}_{i-1}+\bar{\tau}_i}{2} \rfloor + \gamma]\}} \right). \tag{3.21}$$

As shown in Appendix C.2, the marginal likelihood $f(\mathbf{y}|M_*, \boldsymbol{\tau}) = f(\mathbf{y}|\boldsymbol{a}_*, \lambda_*, \boldsymbol{\tau})$ exhibits a closed form expression. Several solutions exist to sample the break parameters (see, e.g., Stephens, 1994; Liao, 2008). In this paper, we use the D-DREAM algorithm developed in Bauwens et al. (2011). It builds a symmetric proposal distribution inspired by the Differential Evolution optimization literature and draws from this proposal distribution are accepted or rejected through a Metropolis step in a Markov-chain Monte Carlo (MCMC) algorithm. As shown in Bauwens et al. (2011), the D-DREAM algorithm complexity is $\mathcal{O}(T)$ and leads to a rapidly mixing MCMC algorithm since the break parameters are jointly sampled from the proposal distribution. To infer the break parameters, we apply the following steps:

- Sample $R = 2m$ initial structural break vectors $\{\boldsymbol{\tau}_i\}_{i=1}^{R}$ from the prior distribution.
- At each MCMC iteration, for each $j = 1, ..., R$, apply the D-DREAM Metropolis move:
  1. Propose a new draw of the break parameter as follows

$$\hat{\boldsymbol{\tau}}_j = \boldsymbol{\tau}_j + \left\lfloor \gamma(\delta, m)(\sum_{g=1}^{\delta} \boldsymbol{\tau}_{r_1(g)} - \sum_{h=1}^{\delta} \boldsymbol{\tau}_{r_2(h)}) + \xi \right\rceil, \tag{3.22}$$

  with $\xi \sim \mathcal{N}(0, (0.0001)I)$ and $\forall g, h = 1, 2, ..., \delta,\ j \neq r_1(g),\ r_2(h)$; $r_1(.)$ and $r_2(.)$ stand for random integers uniformly distributed on the support $[1, R]$. We set $\gamma(\delta, m) = \frac{2.38}{\sqrt{2\delta m}}$ and $\delta \sim \mathcal{U}[1, 3]$.[3]
  2. Accept the proposal $\hat{\boldsymbol{\tau}}_j$ according to the probability

$$\alpha(\boldsymbol{\tau}_j, \hat{\boldsymbol{\tau}}_j) = \min \left\{ \frac{f(\mathbf{y}|M_*, \hat{\boldsymbol{\tau}}_j)}{f(\mathbf{y}|M_*, \boldsymbol{\tau}_j)}, 1 \right\}.$$

In practice, we set the number of MCMC iterations to 4000 and start collecting the draws after round$[\frac{M}{2}]$ MCMC iterations. In addition, we assess the convergence of the MCMC algorithm using the multivariate Potential Scale Reduction Factor test proposed in Brooks and Gelman (1998). For the two in-sample applications below in which credible intervals of the breakpoints are computed, the convergence statistics amount to 1.014 and 1.052, respectively. These values meet the threshold of 1.1 commonly used to validate the convergence of MCMCs.

[3] When the posterior distribution is a multivariate normal one, Ter Braak (2006) proves that choosing $\gamma = \frac{2.38}{\sqrt{\delta m}}$ leads to the optimal acceptance rate of the Metropolis ratio. As shown in Ter Braak (2006), the proposal distribution works when the number of chains, i.e. δ, is equal to one. However, the mixing of the MCMC algorithm can be improved by increasing δ as illustrated with simulation exercises in Vrugt et al. (2009).

3.6 Monte Carlo study

In this Section, we document a Monte Carlo study to assess the accuracy of the SELO approach. We first rely on nine different data generating processes (DGPs) that are documented in Table 3.1. For each DGP, we simulate 1000 series with a sample size equal to $T = 1024$ and we investigate i) the performance of detecting the break dates using the approach in Section 3.5.1 and ii) the performance of the SELO method for detecting which parameter truly varies when a break occurs. The nine DGPs differ in their mean parameter specifications. For each of them, we study the SELO performance when the innovation is either homoskedastic or driven by a GARCH process.

Regarding the DGPs, the first six DGPs are piecewise stationary AR models directly taken from Yau and Zhao (2016) while the others cover situations with exogenous explanatory variables. DGP A and E do not exhibit any breakpoint. They aim at showing the performance of the SELO approach when only spurious break dates are detected. DGPs B and C are weakly persistent piecewise stationary AR models exhibiting three regimes. Simulated series from DGP D experience a break after 50 observations. This DGP should highlight the performance in a short regime context. DGPs E and F are highly persistent piecewise stationary AR models but DGP F differs by exhibiting breaks in the mean parameters. Eventually, DGPs G, H and I include exogenous variables. While DGPs G only exhibits exogeneous regressors, DGPs H and I stand for ARX processes by mixing the parameters of the DGPs B and G.

Table 3.2 documents the percentage of detecting a number of regimes per model parameter over the 1000 simulated series per DGP for the SELO method. Overall, the detection rates of identifying the true number of regimes per parameter are excellent and besides DGP F, they are at least equal to 86.4%. Interestingly, this detection rate does not deteriorate when the innovation is driven by a GARCH process. The worst detection rates arise for the DGP F. Even though this DGP is highly persistent with an autocorrelation structure that barely varies over time, the SELO method correctly identifies that the intercept does not experience abrupt switches 69.7% of the times. Note that the potential breakpoint sets for this DGP poorly identify the true breakpoints since only 25.5% of the sets exhibit at least one potential CP close to every true breakpoints. Therefore, the SELO detection rate could hardly exceed this bound. As exemplified by DGPs G, H and I, the detection rates of the SELO method remain excellent when exogenous variables kick in even in the presence of heteroscedasticity. The Table also documents the rate of detecting the true model (i.e. jointly the correct number of regimes) with a posterior probability of at least 10%.[4] For all the DGPs but DGP F, the correct detection amounts to at least 83.1% and 85.1% for the constant and the GARCH

[4]For this simulation study, the number of explanatory variables ranges from 2 to 5 and the maximum number of potential regimes observed for each DGP is as follows: DGP A (5), DGP B (5), DGP C (6), DGP D (6), DGP E (6), DGP F (6), DGP G (5), DGP H (9), DGP I (9). It can thus lead to a number of models amounting to 2^{40}. In such a case, the set of the models exhibiting a probability equal or greater than 10% has a prior probability of containing the true model that is approximately equal to $\frac{1000}{2^{40}}\%$.

Table 3.1 – Data Generating Processes of sample size amounting to $T = 1024$

	DGP A	**DGP B**	**DGP C**
Breaks	-	[512, 768]	[400, 612]
Intercept	[0]	[0, 0, 0]	[0, 0, 0]
AR_1	[- 0.7]	[0.9, 1.69, 1.32]	[0.4, - 0.6, 0.5]
AR_2	-	[0, - 0.81, - 0.81]	-
	DGP D	**DGP E**	**DGP F**
Breaks	[50]	-	[400, 750]
Intercept	[0, 0]	[0]	[0, 0, 0]
AR_1	[0.75, - 0.5]	[0.999]	[1.399, 0.999, 0.699]
AR_2	-	-	[- 0.4, 0, 0.3]
	DGP G	**DGP H**	**DGP I**
Breaks	[400, 750]	[400,750]	[512, 768]
Intercept	[1, 0, 0]	[0, 0, 0]	[0,0,0]
AR_1	-	[0.9, 1.69, 1.32]	[0.9, 1.69, 1.32]
AR_2	-	[0, -0.81, -0.81]	[0, -0.81, -0.81]
V	[1.5, 0.9, 2.2]	[1.5, 0.9, 2.2]	[1.5, 0.9, 2.2]
W	[- 0.6, - 0.6, - 1]	[- 0.6, - 0.6, - 1]	[- 0.6, - 0.6, - 1]
	Dynamic of the variance of $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2)$		
Constant	$\sigma_t^2 = 1,\ \forall t \in [1, T]$		
GARCH	$\sigma_t^2 = 0.05 + 0.05\epsilon_{t-1}^2 + 0.9\sigma_{t-1}^2,\ \forall t \in [1, T]$ and $\sigma_0^2 = \dfrac{0.05}{1 - 0.95} = 1$		

This Table summarizes the DGPs from which 1000 series are simulated for the Monte Carlo study. The variables V and W stand for exogenous variables such that, $V_t \sim \mathcal{N}(0, 3^2)$ and $W_t \sim \mathcal{N}(0, 4^2)$. For instance, DGP B is an AR(2) model that exhibits two breakpoints at $t = 512$ and $t = 768$. The true values of the first AR term for the first two regimes are equal to 0.9 and 1.69, respectively. The dynamic of the variance is either homoskedastic ('Constant') or heteroskedastic ('GARCH').

innovation dynamics, respectively. These excellent results highlight that model uncertainty should be taken into account since several models often exhibit high posterior probabilities.

DGPs from Table 3.1 are frequently used in the CP literature to assess the performance of a new segmentation method (see, e.g., Cho and Fryzlewicz (2015), Yau and Zhao (2016) and Korkas and Fryzlewiczv (2017)). Nevertheless, our empirical exercise implies more explanatory variables and a smaller sample size. To assess the SELO performance in such environment, we also consider fourteen variants of an 'empirical DGP' given by

$$y_t = \begin{cases} \mathbf{x}_t'\boldsymbol{\beta}_1 + \sigma_t\epsilon_t, & \text{if } 1 \leq t \leq 132, \\ \mathbf{x}_t'\boldsymbol{\beta}_2 + \sigma_t\epsilon_t, & \text{if } 133 \leq T, \end{cases} \tag{3.23}$$

where $T = 256$ as in the application, $\epsilon_t \sim \mathcal{N}(0, 1)$ and $\mathbf{x}_t = (1, x_{t,1}, \ldots, x_{t,12})'$. The explanatory variables are close to the risk factors used in our empirical exercise. In particular, they are generated from AR models whose coefficients and AR orders are estimated using the risk factors of the application. The parameter values of $\boldsymbol{\beta}_1$ are equal to the OLS estimates of

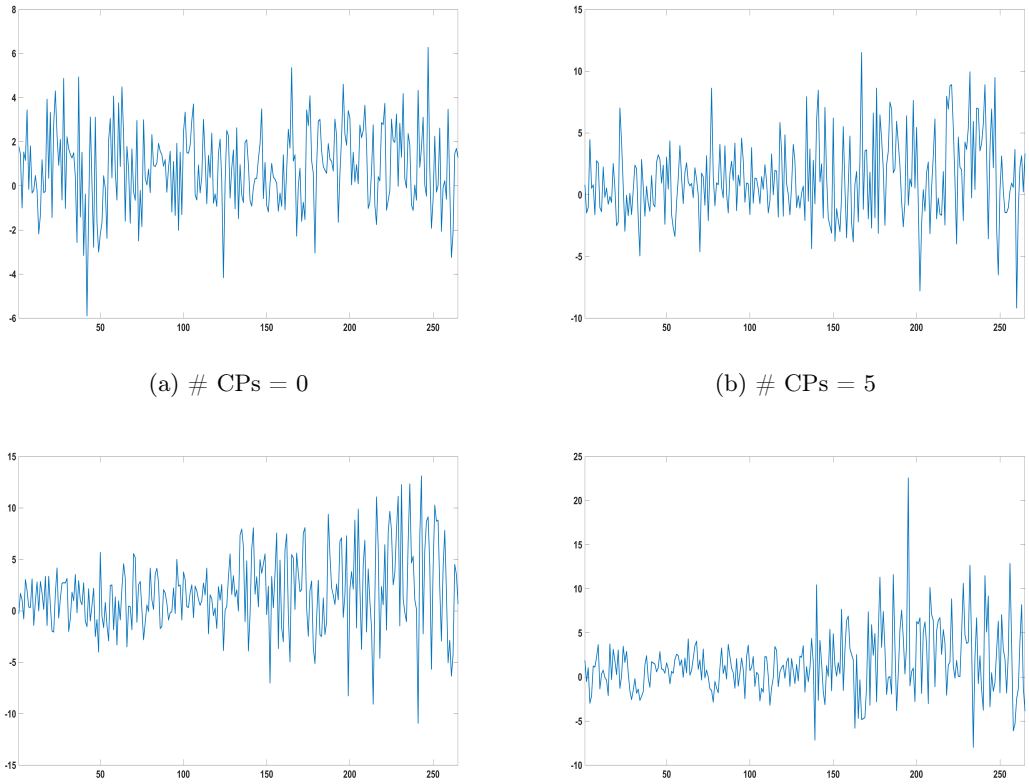
Table 3.2 – Break estimates : SELO approach.

		Constant Variance								GARCH Variance							
		Number of regimes						Break	Exact	Number of regimes						Break	Exact
DGP		1	2	3	4	5	6			1	2	3	4	5	6		
A	Intercept	**99.4**	0.6	0	0	0	0	—	99.9	**99.2**	0.8	0	0	0	0	—	99.2
	AR1	**99.5**	0.5	0	0	0	0			**99.4**	0.6	0	0	0	0		
B	Intercept	**98.6**	1.4	0	0	0	0			**97.3**	2.7	0	0	0	0		
	AR1	0	0	**100**	0	0	0	100	99.7	0	0.2	**99.4**	0.4	0	0	99.3	99.5
	AR2	0	**98.8**	1.2	0	0	0			0	**98.3**	1.7	0	0	0		
C	Intercept	**97.9**	2	0.1	0	0	0	99.8	99.7	**97.6**	2.4	0	0	0	0	99.8	99.1
	AR1	0	0	**100**	0	0	0			0	0	**99.7**	0.3	0	0		
D	Intercept	**97.4**	2.6	0	0	0	0	99.8	99.5	**97.6**	2.2	0.2	0	0	0	99.7	99.1
	AR1	0.1	**99.4**	0.5	0	0	0			0.2	**99.3**	0.4	0.1	0	0		
E	Intercept	**86.4**	12.4	1.2	0	0	0	—	94.6	**84.8**	12.5	2.5	0.2	0	0	—	91.5
	AR1	**93.6**	6.1	0.3	0	0	0			**91**	8.2	0.5	0.3	0	0		
F	Intercept	**69.7**	23.9	6.2	0.1	0.1	0			**65.3**	27.5	6.7	0.5	0	0		
	AR1	0	68.7	**31**	0.1	0.2	0	25.5	23.2	0	69.5	**29.6**	0.8	0.1	0	22.4	22.1
	AR2	0	71.5	**28.3**	0.2	0	0			0	73.2	**26.4**	0.4	0	0		
G	Intercept	0	**99.3**	0.7	0	0	0			0	**99.2**	0.8	0	0	0		
	V	0	0	**99.8**	0.2	0	0	100	99.8	0	0	**99.7**	0.3	0	0	100	99.8
	W	0	**99.2**	0.8	0	0	0			0	**99**	0.9	0.1	0	0		
H	Intercept	**88.9**	11	0.1	0	0	0			**92.9**	6.9	0.2	0	0	0		
	AR1	0	0	**92.7**	7.3	0	0			0	0	**94.7**	5.3	0	0		
	AR2	0	**92.6**	7.4	0	0	0	100	83.1	0	**94.1**	5.9	0	0	0	100	86.8
	V	0	0	**87.7**	12.3	0	0			0	0	**89.6**	10.4	0	0		
	W	0	**88**	12	0	0	0			0	**90.4**	9.6	0	0	0		
I	Intercept	**91.6**	8.4	0	0	0	0			**91**	8.9	0.1	0	0	0		
	AR1	0	0	**94.3**	5.7	0	0			0	0	**95**	4.9	0.1	0		
	AR2	0	**94.6**	5.4	0	0	0	100	85.7	0	**94.9**	5	0.1	0	0	100	85.1
	V	0	0	**89.8**	10.2	0	0			0	0	**89.4**	10.6	0	0		
	W	0	**88.7**	11.1	0.2	0	0			0	**90**	10	0	0	0		

Based on 1000 replications, this Table presents several metrics for assessing the performance of the SELO method on DGPs detailed in Table 3.1. **Number of regimes** is the rate of detecting a specific number of regimes per model parameter. Bold values correspond to the true number of regimes. **Break** documents the rate of having at least one breakpoint in the potential CP set located in the neighborhood of 50 observations of the true breakpoints. We use '—' when the DGP exhibits no breakpoint. **Exact** denotes the rate of detecting the true number of breakpoints for all the model parameters with a posterior probability of at least 10%.

the Hedge fund Index (HFI) regression without breakpoints (see Table 3.7 in the empirical application). We consider 14 variants of the DGP given by Equation (3.23) that differ by the number of parameters experiencing a breakpoint. Defining $\boldsymbol{\beta}_2 = (\beta_{2,1}, \ldots, \beta_{2,13})'$ and considering the ith DGP, with $i = 1, \ldots, 14$, we have $\beta_{2,j} = \beta_{1,j} + 3\omega_j \text{sign}(\beta_{1,j})$ for $j < i$ and $\beta_{2,j} = \beta_{1,j}$ for $j \geq i$. The size of the break given by ω_j is equal to the standard deviation of the jth OLS estimate of the Hedge fund Index (HFI) regression without breakpoints (see Table 3.7). To summarize, the first DGP does not exhibit a breakpoint while the 14th one exhibits a structural change in all its parameters. As before, we consider homoskedastic errors with $\sigma_t^2 \equiv \bar{\omega}^2 = 1.7 \ \forall t$ and heteroskedastic ones with $\sigma_t^2 = 0.05\bar{\omega}^2 + 0.05\epsilon_{t-1}^2 + 0.9\sigma_{t-1}^2$ for $t > 1$ in which $\bar{\omega}^2$ stands for the OLS variance estimate of the Hedge fund Index (HFI) regression without breakpoints (see Table 7 of the paper). Figure 3.2 displays one simulated series drawn from some of the fourteen variants with heteroskedastic errors.

To carry out the Monte Carlo study, we have drawn 100 series from the 14 variants of the empirical DGP. For each simulated series, we have also generated the explanatory variables using the AR models. Table 3.3 provides the percentage of the number of regimes detected



(a) # CPs = 0

(b) # CPs = 5

(c) # CPs = 10

(d) # CPs = 13

Figure 3.2 – One simulated series with a GARCH dynamic from the DGP based on the empirical data

CPs' stands for the number of parameters that are experiencing a breakpoint at time $t = 133$.

by the selective segmentation and by the Lasso methods. The Lasso approach consists in using a Lasso penalty function instead of the SELO penalty function and in choosing the best Lasso penalty value and the corresponding model using the marginal likelihood proposed in Section 3.4.[5] First, we observe that the selective segmentation approach delivers high detection rates whatever the dynamic of the variance. In addition, the detection rate does not deteriorate when the parameter experiences a break. Note also that all the average detections are above 89%. In contrast, the Lasso method does not provide good results when only a subset of the parameters exhibits a CP. In fact, the average detection rates follow a 'U-shape'

[5] We replace the SELO with the Lasso penalty function. In particular, conditional on the potential breakpoints $\boldsymbol{\tau}$, we minimize the following objective function:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=2}^{m} \sum_{k=1}^{K} |\Delta\beta_{jk}|. \tag{3.24}$$

Using the Matlab Lasso toolbox of Mcilhagga (2016). We also tested the matlab glmnet toolbox of Qian et al. (2013) which leads to similar results. In particular, we also observe that the Lasso estimates over-estimate the number of regimes.

function meaning that the Lasso method is good at detecting no breakpoint or when all the parameters are experiencing a break. When partial breaks occur, the Lasso approach typically over-estimates the number of regimes.

To further illustrate the issue with the Lasso method, Table 3.4 shows the detailed results based on 100 simulated series when the first five parameters exhibit a CP (equivalent to the variant called '# CPs = 5' in Table 3.3). While the selective segmentation method accurately detects the number of regimes for each parameter, the Lasso approach finds two regimes for most of the parameters that are constant over the sample.

We end this simulation section with a "big data" example motivated by the fact that when the number of explanatory variables is large, the current Bayesian alternatives do not work (see Giordani and Kohn, 2008; Eo, 2016; Huber et al., 2019; Dufays and Rombouts, 2020) (see Appendix C.5 for more details). To do so, we propose the DGP J that is specified by 100 explanatory variables and one CP as follows:

DGP J: piecewise linear model with big data

$$Y_t = \begin{cases} \mathbf{x}_t'\boldsymbol{\beta_1} + \varepsilon_t & \text{if } 1 \leq t \leq 499, \\ \mathbf{x}_t'\boldsymbol{\beta_2} + \varepsilon_t & \text{if } 500 \leq t \leq T, \end{cases}$$

where $T = 1024$, $\forall\ t \in [1, T]$ and for $i = 1, ..., 100$, $\mathbf{x}_{t,i} \sim \mathcal{N}(0, 1)$ and $\varepsilon_t \sim \mathcal{N}(0, 1)$. The parameter values of $\boldsymbol{\beta_1}$ are uniformly and randomly set to -1 or 1. In the second regime, the parameter values of $\boldsymbol{\beta_2}$ are equal to $\boldsymbol{\beta_1}$ except for 10 of them randomly chosen that are set to the opposite value (i.e. $-\boldsymbol{\beta}_1$). Thus, 10 parameters of DGP J does experience a break at observation 500.
We simulate 100 series from DGP J to assess the SELO performance in detecting which parameters experience a breakpoint. For every simulation, the selective segmentation approach identifies 10 parameters that experience one breakpoint in the sample while the others remain constant. In addition, the exact model specification was always among the specification exhibiting a posterior probability of at least 10%.

3.7 Empirical application

We illustrate the selective segmentation method with 14 monthly Credit Suisse HF indices spanning from March 1994 to March 2016. These indices are the weighted average of HF returns following specific trading strategies. Fung and Hsieh (2004) suggested a risk-based approach to model HF returns and identified seven factors on which HF strategies are generally exposed (see also Fung and Hsieh, 2001). Since this seminal work, many other risk factors have been uncovered. So, we include five other risk factors that are also popular in the literature.

Table 3.3 – Break detection rates - Selective segmentation and Lasso approaches

DGP	**Correct detection rate**													**Avg.**
	Inter	PMKT	SMB	TERM	DEF	SBD	SFX	SCOM	UMD	SIR	STK	CPI	NAREIT	Avg. detection
Selective segmentation - Constant variance														
# CPs = 0	100	99	99	100	98	100	100	100	100	100	99	99	100	99.5
# CPs = 1	47	100	98	99	100	99	98	98	98	98	97	99	95	94.3
# CPs = 2	98	100	98	100	96	100	100	96	99	98	97	97	96	98.1
# CPs = 3	98	100	88	100	96	98	100	97	98	99	94	98	98	97.2
# CPs = 4	99	100	97	98	98	98	97	99	100	98	96	97	98	98.1
# CPs = 5	98	100	93	94	98	97	100	98	98	96	98	94	95	96.8
# CPs = 6	99	100	95	90	100	91	97	100	98	97	93	99	95	96.5
# CPs = 7	97	100	89	94	99	91	97	97	100	97	97	97	99	96.5
# CPs = 8	95	100	91	96	100	88	97	66	98	98	98	95	96	93.7
# CPs = 9	99	98	95	93	99	95	97	68	100	98	99	96	100	95.2
# CPs = 10	97	99	94	96	100	91	97	71	100	76	96	98	93	92.9
# CPs = 11	96	100	90	97	100	98	98	81	100	79	95	99	98	94.7
# CPs = 12	93	99	90	98	98	92	96	68	100	78	95	96	98	92.4
# CPs = 13	94	99	93	93	98	93	96	70	99	85	96	98	92	92.8
Lasso - Constant variance														
# CPs = 0	100	100	100	100	100	100	100	100	100	100	99	100	100	99.9
# CPs = 1	10	99	99	100	100	93	91	91	95	91	89	100	95	88.7
# CPs = 2	28	99	96	99	100	77	70	77	90	59	61	100	83	79.9
# CPs = 3	51	100	71	100	100	52	51	55	76	47	37	100	75	70.4
# CPs = 4	71	100	84	52	99	37	33	35	56	30	27	97	47	59.1
# CPs = 5	85	92	84	78	87	20	17	20	28	20	19	89	27	51.2
# CPs = 6	88	91	81	75	85	86	19	18	25	15	14	92	30	55.3
# CPs = 7	89	83	84	80	93	78	78	14	19	14	16	90	19	58.2
# CPs = 8	87	86	82	81	87	81	83	77	28	18	21	92	23	65.1
# CPs = 9	87	89	84	84	90	84	84	83	86	19	15	92	25	70.9
# CPs = 10	93	88	90	82	92	86	86	82	92	81	25	92	22	77.8
# CPs = 11	95	87	85	84	93	83	81	85	90	85	84	97	27	82.8
# CPs = 12	95	86	88	93	99	86	83	84	87	83	83	94	17	82.9
# CPs = 13	93	85	81	90	97	74	79	77	81	77	77	94	79	83.4
Selective segmentation - GARCH variance														
# CPs = 0	100	99	99	98	98	98	98	100	99	99	98	100	98	98.8
# CPs = 1	54	95	96	98	98	96	96	97	94	95	96	96	98	93.0
# CPs = 2	99	100	95	100	97	99	98	99	99	99	97	98	99	98.4
# CPs = 3	99	100	93	99	99	98	99	96	97	98	92	96	100	97.4
# CPs = 4	98	99	93	98	99	99	96	95	98	98	97	95	95	96.9
# CPs = 5	96	100	92	97	100	97	96	97	97	97	94	98	98	96.8
# CPs = 6	95	100	87	93	99	96	96	99	99	96	99	96	96	96.2
# CPs = 7	97	100	92	98	100	91	92	97	99	97	98	98	100	96.8
# CPs = 8	98	99	90	87	99	91	92	76	98	98	98	97	93	93.5
# CPs = 9	98	100	89	94	100	91	95	67	100	99	98	99	99	94.5
# CPs = 10	96	99	96	97	98	90	97	77	100	84	96	95	99	94.2
# CPs = 11	94	99	93	97	100	89	93	69	99	78	94	97	97	92.2
# CPs = 12	91	96	89	93	99	92	97	72	99	84	94	97	100	92.5
# CPs = 13	85	96	91	92	100	90	96	77	99	82	86	95	77	89.7
Lasso - GARCH variance														
# CPs = 0	100	100	100	100	100	99	99	100	100	99	98	100	100	99.6
# CPs = 1	20	93	98	100	99	87	84	81	88	85	79	99	94	85.2
# CPs = 2	26	99	97	100	100	80	82	80	89	66	63	100	90	82.5
# CPs = 3	49	99	64	100	100	56	52	46	76	51	48	100	77	70.6
# CPs = 4	68	96	75	50	99	37	38	41	51	26	28	100	51	58.5
# CPs = 5	85	90	82	75	84	19	19	24	25	17	15	93	29	50.5
# CPs = 6	86	93	87	77	86	87	19	21	21	16	17	94	23	55.9
# CPs = 7	89	92	85	81	88	91	89	19	30	10	21	95	25	62.7
# CPs = 8	87	87	83	70	87	83	86	77	23	16	21	97	26	64.8
# CPs = 9	88	85	86	81	90	83	85	80	87	16	15	95	29	70.8
# CPs = 10	97	88	88	91	96	85	85	86	91	84	9	90	28	78.3
# CPs = 11	92	89	91	88	93	79	81	75	88	77	78	97	18	80.5
# CPs = 12	93	88	84	85	91	80	83	82	86	78	80	90	26	80.5
# CPs = 13	92	82	85	86	92	80	82	75	85	77	80	88	81	83.5

Based on 100 replications, this Table assesses the break detection performance of the selective segmentation and the Lasso methods on the 14 variants of the empirical DGP detailed in Equation (3.23). **Correct detection rate** is the rate of detecting the true number of regimes per model parameter. Underlined values correspond to the detection rates when the parameter experiences a breakpoint at $t = 133$. **Avg.** documents the average rate of detecting the true number of regimes for each variant.

Table 3.4 – Empirical DGP with 5 CPs - Break detection rates of the Selective segmentation and the Lasso approaches

	Constant variance												GARCH variance											
	Sel. segmentation						Lasso						Sel. segmentation						Lasso					
# of regimes	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Inter.	2	**98**	0	0	0	0	12	**85**	3	0	0	0	3	**96**	1	0	0	0	15	**85**	0	0	0	0
PMKT	0	**100**	0	0	0	0	0	**92**	8	0	0	0	0	**100**	0	0	0	0	0	**90**	10	0	0	0
SMB	5	**93**	2	0	0	0	10	**84**	6	0	0	0	6	**92**	2	0	0	0	9	**82**	9	0	0	0
TERM	6	**94**	0	0	0	0	22	**78**	0	0	0	0	2	**97**	1	0	0	0	25	**75**	0	0	0	0
DEF	0	**98**	1	1	0	0	12	**87**	1	0	0	0	0	**100**	0	0	0	0	16	**84**	0	0	0	0
SBD	**97**	3	0	0	0	0	**20**	70	10	0	0	0	**97**	2	1	0	0	0	**19**	76	4	1	0	0
SFX	**100**	0	0	0	0	0	**17**	73	10	0	0	0	**96**	4	0	0	0	0	**19**	70	10	1	0	0
SCOM	**98**	2	0	0	0	0	**20**	71	9	0	0	0	**97**	3	0	0	0	0	**24**	69	7	0	0	0
UMD	**98**	2	0	0	0	0	**28**	66	6	0	0	0	**97**	2	1	0	0	0	**25**	68	6	1	0	0
SIR	**96**	3	1	0	0	0	**20**	69	11	0	0	0	**97**	2	1	0	0	0	**17**	71	11	1	0	0
STK	**98**	2	0	0	0	0	**19**	71	10	0	0	0	**94**	6	0	0	0	0	**15**	77	7	1	0	0
CPI	**94**	6	0	0	0	0	**89**	11	0	0	0	0	**98**	2	0	0	0	0	**93**	7	0	0	0	0
NAREIT	**95**	4	1	0	0	0	**27**	67	6	0	0	0	**98**	1	1	0	0	0	**29**	68	3	0	0	0

Based on 100 replications, this Table assesses the break detection performance of the selective segmentation and the Lasso methods on the fifth variant of the empirical DGP detailed in Equation (3.23). **Number of regimes** is the rate of detecting a specific number of regimes per model parameter. Bold values correspond to the true number of regimes.

We add two Fung and Hsieh trend following risk factors, PTFSIR, returns on PTFS short term interest rate lookback straddle, and PTFSSTK, returns on PTFS stock index lookback straddle. Following Agarwal and Naik (2004), among many others, we also use the Up-minus-Down (UMD) factor (see Carhart, 1997). As suggested by Chen et al. (1986), we include the expected inflation, the log relative of US Consumer Price Index (CPI). Finally, we also take into account a factor for real estate risk relevant to explain HF and stocks returns (see, e.g., Ambrose and D'Lima, 2016; Carmichael and Coen, 2018). Table 3.5 documents the fourteen strategies on which we focus as well as the twelve factors.

Table 3.5 – Description of the HF returns and the risk factors

Credit Suisse Hedge fund indices		Risk factors		
Name	Description	Name	Description	Paper
HFI	Hedge Fund Index	PMKT	Market factor (S&P 500)	FH
CNV	Convertible Arbitrage	SMB	Small firm minus big firm	FH
DSB	Dedicated Short Bias	TERM	Change in 10-year treasury yields	FH
EME	Emerging Markets	DEF	Change in the yield spread of	FH
EMN	Equity Market Neutral		10-year treasury and Moody's Baa bonds	
EDR	Event Driven	PTFSBD	Lookback options on Bonds	FH
EDD	Event Driven Distressed	PTFSFX	Lookback options on currencies	FH
EDM	Event Driven Multi-Strategy	PTFSCOM	Lookback options on commodities	FH
EDRA	Event Driven Risk Arbitrage	UMD	Momentum (Up-minus-Down)	C
FIA	Fixed Income Arbitrage	PTFSIR	Lookback options on short term interest rate	FH
GMA	Global Macro	PTFSSTK	Lookback options on Stock Index	FH
LES	Long/Short Equity	CPI	Consumer price index	CRR
MFU	Managed Futures	NAREIT	Real estate investment trust index	AD
MUS	Multi-Strategy			

The column 'Paper' highlights a paper in which the factor has already been used. FH, C, CRR and AD refer to Fung and Hsieh (2004), Carhart (1997), Chen et al. (1986) and Ambrose and D'Lima (2016), respectively.

It is well acknowledged in the financial literature that HF strategies (or trading techniques)

are time-varying. Their changing risk exposures are directly related to market events and economic fluctuations (see, e.g., Agarwal and Naik (2004), Fung et al. (2008) or Patton et al. (2015) among others). Hedge fund time-varying risk dynamics has important implications for performance appraisal. As pointed out by Mitchell and Pulvino (2001), the changes can be in response to arbitrage opportunities. The cycles of mergers and acquisitions in the 1990s and the 2000s and the corresponding level of risk arbitrage led by HF are illustrations of these changing dynamics. In standard linear asset pricing models, the intercept and risk factor loadings are not constant but time-varying. Moreover, HF returns exhibit significant non-linearities. Therefore, there is a need of dynamic models able to capture non-linearities and changes in risk exposures.

Following Meligkotsidou and Vrontos (2008), we suggest the use of CP risk factor models. This class of models is suited for studying the changes in risk exposures and their time-varying parameters. However, instead of directly focusing on the twelve factors, we take a slightly different approach since we additionally take into account autocorrelations of the returns.[6] To do so, we first look at the best autoregressive model that fits the returns. In particular, for each HF returns, we estimate ARX(q) models with q ranging from 0 to 4 and in which the explanatory variables are the twelve factors (and an intercept) and we select the best AR order using the Bayesian information criterion (BIC). Table 3.6 documents the best order for each strategy.

Table 3.6 – Order of the optimal ARX-model for each HF strategy

Strat.	HFI	CNV	DSB	EME	EMN	EDR	EDD
Lag order	0	1	0	1	0	1	2
Strat.	EDM	EDRA	FIA	GMA	LES	MFU	MUS
Lag order	1	1	1	0	1	0	0

The optimal AR order is chosen by maximizing the Bayesian information criterion over the whole sample. When looking for the best autoregressive lag order, the explanatory variables include the seven factors and an intercept.

As reported by Fung and Hsieh (2004), composites obtained from the individual funds may be contaminated with severe survivorship, selection and instant history biases. Therefore, to avoid these problems, we use the Credit Suisse indices that provide full transparency about their constituents.

Section 3.7.1 discusses in-sample results of our selective segmentation method and we compare them to those of standard CP models and time-varying parameter models. We then illustrate the difference of our approach with the CP method of Meligkotsidou and Vrontos (2008) in Section 3.7.2. Section 3.7.3 documents a forecasting exercise in which we assess the predictive performance of the selective segmentation approach with respect to flexible alternatives. Im-

[6] As reported by Getmansky et al. (2004), the analysis of serial dependence of returns is a reasonable way of assessing the liquidity of hedge fund investments.

portantly, all the subsequent results include the optimal AR order documented in Table 3.6 as additional explanatory variables.

3.7.1 Hedge funds strategies evolve over time

Fung and Hsieh (2004) focus on linear models. However, as the period covers critical events such as the Long Term Capital Management (LTCM) collapse, the dot-com crisis and the global financial crisis (GFC), one could argue that CP models are more appropriate. In this Section, we focus on two specific indices, namely the Hedge Fund Index (HFI) and the HF returns that are applying a Fixed-Income Arbitrage (FIA) strategy. Results for all the other returns are available upon request.

Tables 3.7 and 3.8 show how the selective segmentation method can improve the interpretation of CP models. The Tables document how the results evolve from a standard linear risk model to a selective segmentation model passing by a standard CP process. As expected, for the two HF returns, ignoring breakpoints can be misleading as the CP results emphasize that they modify the risk exposition of the returns. Also, although one can study in details the results of the standard CP model, the selective segmentation model offers a straightforward picture of the relevant risk factors and how the risk exposition evolves. It also estimates more accurately the parameters that do not change when a break occurs. As the CP model detects three breakpoints for the HFI and six abrupt changes for the FIA strategy, the number of models to consider amounts to 2^{36} and 2^{84} respectively. Our selective segmentation strategy explores these large model spaces and find the most promising configurations in several minutes on a standard laptop. Let us now discuss in more details the results of the two returns.

Hedge Fund Index

As documented in Table 3.7, the CP model with breakpoints determined by the approach in Section 3.5.1 finds four regimes (hereafter CP-YZ). The relevant breakpoints occur in April 2000, in December 2001 and in August 2014. Interestingly, these dates coincide with the dot-com crash that spanned from March 2000 to October 2002 and when stocks suffer steepest drop in 2014 (introducing the first significant stock market scares after the GFC). It is well acknowledged in the financial literature that the end of the dot-com bubble had important consequences for financial markets in the early 2000s. While all the parameters change for the CP model, the selective segmentation mainly identifies that the factors related to the breaks are the market factor (PMKT), the default risk factor (DEF), the momentum risk factor (UMD) and the real estate risk factor (NAREIT). Moreover, it discards two spurious breaks occurring in December 2001 and in August 2014 making the model even more parsimonious. We can notice that the market factor decreases from 0.34 during the first period to 0.21 during the second period. HFI is indeed more conservative during the 2000s and less correlated with the financial markets. We observe the same trend for the default risk factor increasing from

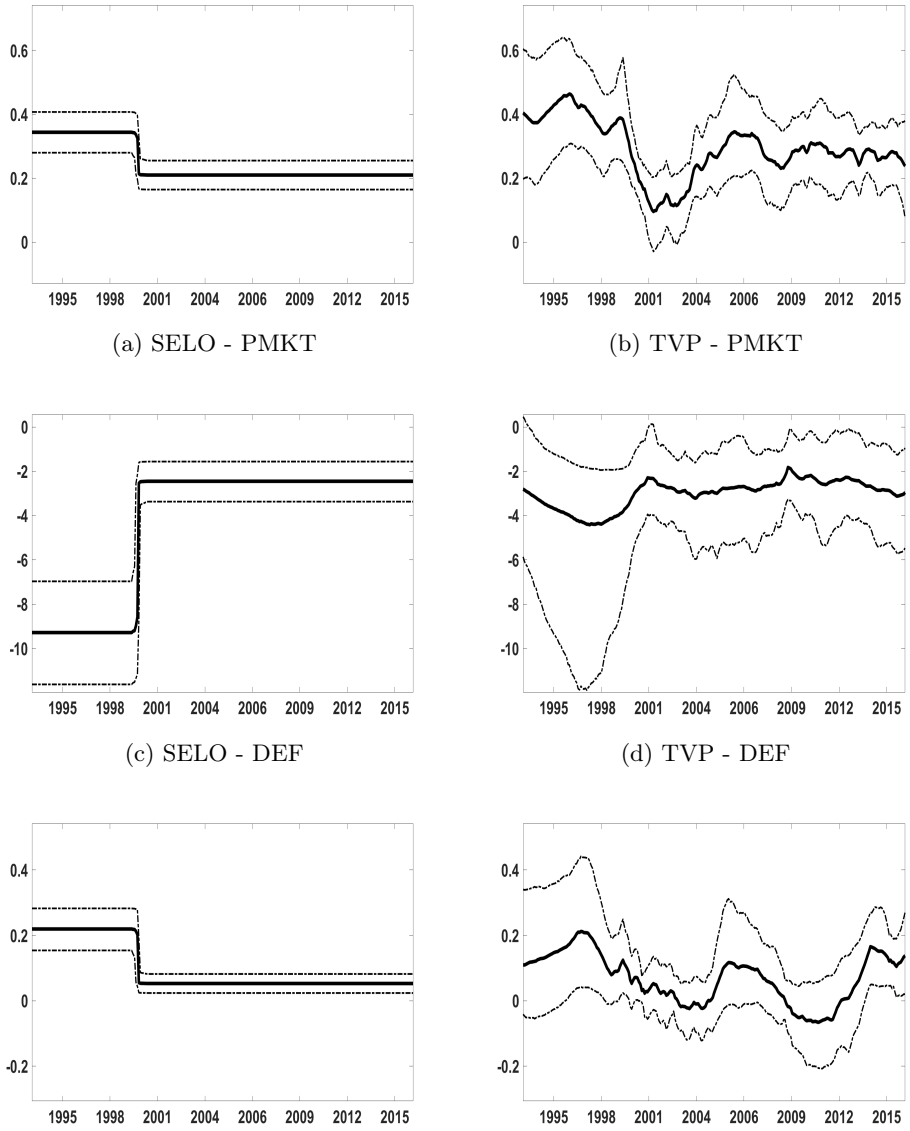
-9.15 to -2.42. The momentum factor, UMD, sharply declines after the nineties known as very volatile as reported by Shiller (2015) and documented by Campbell (2000) who both highlight the bullish market of this decade. The momentum is still significant since 2000 but its impact has significantly weakened. It decreases from 0.22 to 0.05. The real estate risk factor also exhibits a breakdown in the early 2000s. It is indeed well acknowledged in the real estate economics literature that the 1990s are considered as the new era of real estate investment trusts (REITs) and the period beginning in the early 2000s as the maturity REITs era (Pagliari et al. (2005), Ambrose et al. (2007) and Carmichael and Coen (2018) among others). From the new REITs era to the maturity REITs era, the real estate risk factor declines from 0.15 to -0.01. These results are consistent with the important variations of interest rates during these two sub-periods and the important increase of credit risk in the 2000s. As a final note, the credible interval of the breakpoint is narrow which indicates a sharp change in the risk exposition in March 2000 (see also Figure 3.3).

Table 3.7 – Hedge Fund Index: linear, CP and selective segmentation regression models

Period	Int.	PMKT	SMB	TERM	DEF	PTFSBD	PTFSFX	PTFSCOM	UMD	PTFSIR	PTFSSTK	CPI	NAREIT
	Standard linear risk model												
03-1994 to 03-2016	0.33	0.27	0.07	-0.86	-3.03	-0.01	0.01	0.00	0.11	-0.00	0.02	1.07	-0.01
	(0.10)	(0.02)	(0.03)	(0.45)	(0.59)	(0.01)	(0.00)	(0.01)	(0.02)	(0.00)	(0.01)	(0.32)	(0.02)
	CP-YZ risk model												
03-1994 to 04-2000	0.59	0.34	0.03	-1.46	-10.24	-0.02	0.02	0.02	0.22	-0.02	0.04	0.26	0.16
	(0.37)	(0.05)	(0.05)	(1.35)	(2.52)	(0.01)	(0.01)	(0.02)	(0.06)	(0.01)	(0.02)	(1.56)	(0.07)
05-2000 to 12-2001	-0.12	0.17	0.10	-1.19	-0.54	0.00	0.03	-0.04	0.11	0.03	-0.03	0.32	0.00
	(0.84)	(0.10)	(0.15)	(3.17)	(6.08)	(0.03)	(0.04)	(0.10)	(0.11)	(0.04)	(0.06)	(2.48)	(0.13)
01-2002 to 08-2014	0.22	0.22	-0.00	-0.71	-2.17	-0.01	0.01	0.00	0.05	-0.01	0.02	1.21	-0.01
	(0.17)	(0.05)	(0.06)	(0.76)	(0.83)	(0.01)	(0.01)	(0.01)	(0.03)	(0.01)	(0.01)	(0.45)	(0.03)
09-2014 to 03-2016	-0.40	0.39	-0.06	-1.95	-3.35	0.01	0.03	-0.02	0.19	-0.01	-0.00	-0.93	-0.16
	(0.68)	(0.21)	(0.19)	(5.53)	(5.64)	(0.04)	(0.03)	(0.03)	(0.18)	(0.03)	(0.03)	(1.87)	(0.20)
	Selective segmentation risk model (77%)												
03-1994 to 04-2000	0.24	0.34	0.04	-1.05	-9.15	-0.01	0.01	0.01	0.22	-0.01	0.02	1.20	0.15
[02-2000 05-2000]	(0.09)	(0.04)	(0.02)	(0.42)	(1.35)	(0.01)	(0.00)	(0.01)	(0.04)	(0.00)	(0.01)	(0.29)	(0.05)
05-2000 to 03-2016	—	0.21	—	—	-2.42	—	—	—	0.05	—	—	—	-0.01
	—	(0.03)	—	—	(0.54)	—	—	—	(0.02)	—	—	—	(0.02)

The Table details the parameter estimates of the linear model, of the CP model and of the selective segmentation process with HFI returns as the dependent variable. Parentheses and brackets indicate standard deviations and 90% credible intervals, respectively. A cell filled with '—' indicates that the parameter does not vary over the related period. The posterior probability of the selective segmentation model amounts to 77%.

Figure 3.3 shows the posterior medians over time and their corresponding credible intervals of the parameters related to the MKT, DEF and UMD factors given by our method (see Section 3.5.2 for the related Bayesian model and how the breakpoints are integrated out) and the time-varying parameter (TVP) model (see Appendix C.4 for the model specification). As with the CP model, one can easier interpret the time-varying dynamics of the parameters given by the selective segmentation method than those of the TVP model. For instance, while the exposition to the default factor seems fixed over the sample due to the smooth transition of the parameter, it is clear that the exposition is changing when we look at the selective segmentation results. Regarding the market factor, we also observe with the TVP model that the exposition seems different before and after the dot-com crash but the credible intervals are too wide to confirm the statement.



(a) SELO - PMKT

(b) TVP - PMKT

(c) SELO - DEF

(d) TVP - DEF

(e) SELO - UMD

(f) TVP - UMD

Figure 3.3 – HFI returns - Selective segmentation (SELO) model and Time-varying parameter (TVP) model

Posterior medians (black) and the 90% credible intervals (dotted black lines) of the model parameters over time. For the SELO method, we take the break uncertainty into account using the MCMC algorithm presented in Section 3.5.2.

Fixed Income Arbitrage (FIA) strategy

The FIA strategy is based on the exploitation of inefficiencies in the pricing of bonds and interest rate derivatives (including futures, options, swaps and also mortgage back securities). It was very appreciated among hedge fund managers until the collapse of the LTCM fund in September 1998. After this incident, a change of behavior among managers has been observed for this strategy on financial markets.

The results of the FIA returns from a standard linear regression to the selective segmentation model are documented in Table 3.8. Focusing on the latter model, breakpoints are detected in September 1997, in March 1999, August 2000 which are related to the Russia financial crisis and the dot-com crash. In addition, three other breakpoints capture the financial crisis: the beginning of the global financial crisis in late June 2007, the turmoil of September and October 2008 after the collapse of Lehman Brothers and the sovereign debt crisis in the euro zone. The selective segmentation specification highlights the role played by the market factor (i.e., PMKT, before and after the global financial crisis with estimates of 0.02 and 0.08 respectively), the variation of the size effect, the default risk factor, the momentum factor, the inflation factor and four trend following risk factors, especially during the sub-prime crisis from late June 2007 to October 2008. The size effect changes from 0.02 before the crisis to -0.05 afterwards. The bond trend following factor, PTFSBD, the currency trend following factor, PTFSFX, the commodity trend following factor, PTFSCOM and the stock index trend following factor, PTFSSTK are highly significant during the GFC. The credit spread factor, DEF, is highly significant and constant after the dot com crisis (estimate amounting to -1.18). Significant during the first and the second periods related to the bullish market of the 1990s, the momentum effect, UMD, dramatically changes during the GFC, with a negative estimate of -0.34 (as expected). Our results also report the impact of inflation risk, CPI, during the crisis (with an estimate of 2.69) until the sovereign debt crisis (fifth and sixth periods). The real estate risk factor, NAREIT, is significant and positive during the 1990s (first, second and third periods) and during the GFC, with a negative estimate (as expected) of -0.16. As shown by Figures C.2 to C.4 in Appendix C.4.1, the selective segmentation method allows easier detection of the relevant factors as compared to the TVP model.

3.7.2 Comparison with advanced CP models

We now compare our results with those of Meligkotsidou and Vrontos (2008). Meligkotsidou and Vrontos (2008) rely on CP models to capture the risk exposition of HF returns over time. In particular, they consider the 4096 distinct combinations of the twelve risk factors and for each of them, they estimate a CP model exhibiting several numbers of segments m (from one to ten). Eventually, they use the marginal likelihood to select the best model among the set of $m \times 2^K$ estimated processes (i.e., 40960 models since $m = 10$ and $K = 12$). Their approach consists therefore in first selecting the relevant factors and then, in investigating if the exposition to them is time-varying.

Table 3.8 – Fixed Income Arbitrage: linear, CP and selective segmentation risk models

Period	Int.	AR1	PMKT	SMB	TERM	DEF	PTFSBD	PTFSFX	PTFSCOM	UMD	PTFSIR	PTFSSTK	CPI	NAREIT
	Standard linear risk model													
03-1994 to 03-2016	0.12	0.24	0.02	-0.03	-0.99	-3.60	-0.00	-0.01	0.01	-0.00	-0.01	0.01	1.04	0.03
	(0.08)	(0.04)	(0.02)	(0.02)	(0.33)	(0.46)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.24)	(0.02)
	CP-YZ risk model													
03-1994 to 09-1997	0.57	0.43	0.03	-0.05	0.23	0.63	-0.01	-0.00	0.00	0.00	-0.00	0.00	-1.12	0.09
	(0.22)	(0.13)	(0.03)	(0.03)	(0.68)	(2.13)	(0.01)	(0.00)	(0.01)	(0.04)	(0.01)	(0.01)	(0.79)	(0.03)
10-1997 to 03-1999	-0.21	0.24	0.13	-0.18	-2.09	-14.47	0.04	-0.03	-0.04	0.18	-0.03	0.02	-0.24	0.08
	(0.30)	(0.07)	(0.06)	(0.10)	(2.79)	(4.18)	(0.01)	(0.01)	(0.02)	(0.06)	(0.01)	(0.01)	(1.13)	(0.07)
04-1999 to 08-2000	1.99	-1.68	0.15	0.10	1.79	-1.31	-0.00	0.03	-0.01	-0.08	0.00	-0.04	-3.13	0.04
	(0.90)	(0.90)	(0.06)	(0.05)	(1.66)	(1.84)	(0.02)	(0.02)	(0.02)	(0.06)	(0.02)	(0.03)	(1.95)	(0.07)
09-2000 to 05-2007	0.39	0.29	-0.01	0.02	-0.55	-2.41	-0.00	0.01	0.00	-0.02	0.00	0.01	0.25	-0.03
	(0.09)	(0.08)	(0.02)	(0.02)	(0.37)	(0.72)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.01)	(0.20)	(0.02)
06-2007 to 10-2008	-1.89	-0.79	0.49	0.86	-17.65	-9.84	-0.00	-0.10	0.16	-0.10	-0.01	0.00	3.14	-0.66
	(0.33)	(0.20)	(0.10)	(0.20)	(3.48)	(1.94)	(0.02)	(0.01)	(0.02)	(0.06)	(0.00)	(0.02)	(0.56)	(0.10)
11-2008 to 08-2010	0.99	0.28	0.13	0.16	-0.10	-0.14	0.02	-0.04	0.05	-0.01	-0.03	0.05	3.97	-0.10
	(0.21)	(0.06)	(0.05)	(0.07)	(0.89)	(0.70)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.73)	(0.04)
09-2010 to 03-2016	0.23	0.36	0.05	-0.05	-0.24	-1.52	-0.01	-0.00	0.00	0.01	0.01	-0.00	0.22	-0.01
	(0.11)	(0.11)	(0.03)	(0.03)	(0.48)	(0.71)	(0.01)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.31)	(0.02)
	Selective segmentation risk model (100%)													
03-1994 to 09-1997	0.23	0.45	0.02	0.02	-0.24	-1.35	-0.00	-0.01	0.00	0.12	-0.00	0.01	0.17	0.07
[10-1996 09-1997]	(0.03)	(0.07)	(0.01)	(0.01)	(0.25)	(1.62)	(0.00)	(0.00)	(0.00)	(0.03)	(0.01)	(0.00)	(0.22)	(0.02)
10-1997 to 03-1999	—	—	—	—	—	-10.56	—	—	—	—	-0.03	—	—	—
[01-1999 06-1999]	—	—	—	—	—	(0.97)	—	—	—	—	(0.01)	—	—	—
04-1999 to 08-2000	—	—	—	—	—	-1.18	—	0.01	—	-0.01	0.00	—	—	—
[07-2000 05-2003]	—	—	—	—	—	(0.34)	—	(0.00)	—	(0.01)	(0.00)	—	—	—
09-2000 to 05-2007	—	—	—	—	—	—	—	—	—	—	—	—	—	-0.03
[04-2007 06-2007]	—	—	—	—	—	—	—	—	—	—	—	—	—	(0.02)
06-2007 to 10-2008	—	-1.35	0.08	-0.05	—	—	-0.09	-0.06	0.10	-0.34	—	-0.08	2.69	-0.16
[10-2008 10-2008]	—	(0.24)	(0.02)	(0.03)	—	—	(0.02)	(0.01)	(0.01)	(0.04)	—	(0.01)	(0.40)	(0.03)
11-2008 to 08-2010	—	0.63	—	—	—	—	-0.00	-0.01	0.00	-0.00	—	0.00	—	-0.01
[06-2010 11-2010]	—	(0.14)	—	—	—	—	(0.01)	(0.00)	(0.00)	(0.01)	—	(0.00)	—	(0.02)
09-2010 to 03-2016	—	0.19	—	—	—	—	—	—	—	—	—	—	0.21	—
	—	(0.09)	—	—	—	—	—	—	—	—	—	—	(0.35)	—

The Table details the parameter estimates of the linear model, of the CP model and of the selective segmentation process with FIA returns as the dependent variable. Parentheses indicate standard deviations and brackets [-] document the 95% credible intervals of the breakpoints that are computed from the Bayesian model given in Section 3.5.2. A cell filled with '—' indicates that the parameter does not vary over the related period. The posterior probability of the selective segmentation model amounts to 99.6%.

There is a striking difference with our approach since, for each breakpoint, our procedure can detect what are the time-varying factors. In fact, our approach discriminates between $2^{m\times K}$ models; a number of models that exponentially increases with the amount of breaks. Note that we could also search for the best regressors to include by considering all the 4096 distinct combinations of the twelve factors. In such a case, the number of models to consider would reach $2^{(m+1)\times K}$.

We reproduce the results of Meligkotsidou and Vrontos (2008) on our data by additionally taking the autocorrelation structure into account. Fixing the AR order q to the value given in Table 3.6, for each possible combination of the factors, we estimate CP-ARX(q) models with different numbers of breaks (ranging from 1 to 10) by (globally) minimizing the MDL criterion. Then, we report the combination of factors exhibiting the best MDL value. Hereafter, we denote this model by CP-MV.[7]

[7]Our approach is slightly different as the one used in Meligkotsidou and Vrontos (2008) since we minimize the MDL criterion instead of maximizing the marginal likelihood of a Bayesian CP model for finding the best combination of the factors and the breakpoints. This is motivated by the fact that the MDL criterion consistently selects the true number of regimes while there is no equivalent proof for the marginal likelihood used in Meligkotsidou and Vrontos (2008). In addition, Ardia et al. (2019) show that the MDL criterion is equal to minus the marginal log-likelihood of a CP Bayesian model with particular g-prior distributions. So, our approach can be understood as the method of Meligkotsidou and Vrontos (2008) with different hyper-parameters. We globally minimize the MDL criterion using the dynamic programming of Bai and Perron

Table 3.9 documents the factors of the CP-MV model for the two HF strategies. It also reports the factors which exhibit significant parameter estimates at least at one period for the TVP model and the selective segmentation approach. For the HFI, the selected factors by the selective segmentation methods seem more complete than those of the TVP model in light of the current literature. While the TVP selects eight risk factors, the selective segmentation reports the same factors, except the SMB factor, and adds two trend following risk factors, PTFSBD and PTFSSTK, and the real estate risk factor, NAREIT. For the FIA, the results are much more contrasted. All risk factors are selected by the selective segmentation methods whereas five factors are omitted by the TVP (SMB, TERM, PTFSFX, PTFSCOM and UMD). The CP-MV approach does not select SMB, NAREIT and four trend following risk factors (PTFSBD, PTFSCOM, PTFSSIR and PTFSSTK) for the HFI. The analysis of the single strategy, Fixed Income Arbitrage (FIA), also highlights important and significant differences as far as only two factors are selected, DEF and PTFSFX, by the CP-MV process. It may be very surprising since the market premium is (almost) always used in linear asset pricing models as pointed out by Fung and Hsieh (2001). We may also note that the look-back straddles on commodities, PTFSCOM, on bond, PTFSBD and on stock index and PTFSSTK designed to capture non-linearities especially during changes in international economic policies, are not selected whereas the phenomenon is observed just after the GFC.

Table 3.9 – HFI and FIA strategies: Selected factors given several time-varying parameter models

	HFI			FIA		
	TVP	Sel. Seg.	CP-MV	TVP	Sel. Seg.	CP-MV
PMKT	√	√	√	√	√	
SMB	√				√	
TERM	√	√	√		√	
DEF	√	√	√	√	√	√
PTFSBD		√		√	√	
PTFSFX	√	√	√		√	√
PTFSCOM					√	
UMD	√	√	√		√	
PTFSIR	√	√		√	√	
PTFSSTK		√		√	√	
CPI	√	√	√	√	√	
NAREIT		√		√	√	

Selected factors by the TVP, the selective segmentation process and the CP-MV model of Meligkotsidou and Vrontos (2008). The factors of the latter process are chosen by minimizing the MDL criterion while for the TVP and the selective segmentation model, a factor is selected if its related parameter estimate is significant at least at one period over the sample.

Table 3.9 does not inform on the dynamic of the selected factors by the CP-MV process. Although the preferred specification of the CP-MV model does not include all the factors, the

(2003).

risk exposure of the HF strategies is still abruptly changing over time. Regarding the HFI, two breakpoints are detected and occur in April 2000 and in March 2005, respectively. Table 3.10 shows how the selective segmentation method improves the interpretation of the CP-MV results. It also illustrates the improvement in economic modelling as far as it highlights the relative role played by static and dynamic parameters. First, we observe that the alpha, the currency lookback straddle, PTFSFX, and the consumer prince index, CPI, are quite static during the full period. Interestingly, the selective segmentation also reports the CP in the bullish market in the early 2000s. As mentioned earlier, the financial markets were indeed very volatile during the 1990s. This break is clearly reported by the dynamic risk factors, PMKT, UMD and DEF. PMKT declines from 0.41 to 0.21 during the 2000s and afterwards. The trend is more striking for the momentum factor, UMD, with a decline from 0.18 to 0.05, and for the credit default risk factor, DEF, rising from -11.71 to -2.21. The term structure risk factor, TERM, is not statistically significant after the rise of the housing price index in 2005-2006, announcing the collapse of the financial markets in 2008 (as anticipated by Shiller (2015) among others) and the following quantitative easing policies with very low inters rates.

The FIA strategy exhibits seven regimes which makes the CP-MV model heavily parametrized (i.e., $K \times m = 35$ parameters). This large number of regimes is probably related to the fact that more breakpoints are needed to adequately fit the FIA returns since the CP-MV specification includes only two risk factors, DEF and PTFSFX and/or because in this specific economic modelling the breakpoints also capture the variance dynamic. Using the selected factors and the breakpoints of the best CP-MV specification, we estimate the selective segmentation model to uncover what are the static and the dynamic parameters. First, we must acknowledge that the best specification selected by the CP-MV model is doubtful with regards to the financial literature and the practice. Nevertheless, we compare the CP-MV approach with the selected segmentation to highlight the contribution of the latter. In particular, we observe that the 'alpha' is varying and statistically positive during the 1990s until LTCM collapse. As expected, the default risk factor is negative, time-varying and high during crises (-9.49 during the LTCM collapse and -6.17 during the GFC). After the GFC, the default factor is constant, negative and not statistically significant. This result is consistent with the trend observed on financial markets (especially on fixed incomes markets after the GFC). The currency trend following factor, PTFSFX, is very low, time-varying and statistically significant during the global financial crisis (as expected) and before the impact of the quantitative easing policies starting in the late 2010. After this date (11/2010) the factor is not statistically significant. This is an illustration of the impact of quantitative easing on fixed income arbitrage.

As illustrated by these empirical results, our method uncovers which parameters truly vary when a CP is detected. This technical improvement induces financial consequences and especially cost reductions.[8] Moreover, our method should imply more accurate and thus less

[8] We thank an anonymous referee for this relevant comment.

frequent portfolio rebalancing strategies. Investor could indeed change his timing by using our approach and decide to rebalance parsimoniously (and thus efficiently) his investment when a break is detected, with a special focus on the relevant benchmark.

Table 3.10 – Hedge Fund Index: Best CP-MV model and best selective segmentation model

	Preferred CP-MV model							**Selective segmentation (77%)**						
Period	Int.	PMKT	TERM	DEF	PTFSFX	UMD	CPI	Int.	PMKT	TERM	DEF	PTFSFX	UMD	CPI
03-1994 to 04-2000	0.42	0.41	-2.47	-12.13	0.02	0.18	0.52	0.22	0.41	-2.44	-11.71	0.01	0.18	1.17
	(0.38)	(0.05)	(1.28)	(2.62)	(0.01)	(0.06)	(1.63)	(0.09)	(0.03)	(0.52)	(1.44)	(0.00)	(0.04)	(0.29)
05-2000 to 03-2005	0.44	0.19	-2.20	-2.57	0.02	0.05	0.64	—	0.21	—	-2.21	—	0.05	—
	(0.34)	(0.06)	(1.21)	(2.21)	(0.01)	(0.04)	(1.11)	—	(0.02)	—	(0.54)	—	(0.02)	—
04-2005 to 03-2016	0.11	0.23	0.28	-1.82	0.01	0.05	1.25	—	—	0.29	—	—	—	—
	(0.18)	(0.04)	(0.85)	(0.91)	(0.01)	(0.03)	(0.51)	—	—	(0.55)	—	—	—	—

The Table details the parameter estimates of the preferred CP-MV model and of the selective segmentation process given the selected factors and the breakpoints found by the CP-MV model. Parentheses indicate standard deviations. A cell filled with '—' indicates that the parameter does not vary over the related period. The posterior probability of the selective segmentation model amounts to 77%.

Table 3.11 – Fixed Income Arbitrage: Best CP-MV model and best selective segmentation model

	Preferred CP-MV model				**Selective segmentation (52%)**			
Period	Int.	AR1	DEF	PTFSFX	Int.	AR1	DEF	PTFSFX
03-1994 to 05-1995	0.23	0.40	7.57	-0.03	0.55	0.36	-0.14	-0.01
	(0.23)	(0.19)	(3.14)	(0.01)	(0.13)	(0.06)	(2.04)	(0.01)
06-1995 to 08-1997	-0.13	1.28	-0.68	0.00	—	—	—	—
	(0.44)	(0.53)	(2.57)	(0.01)	—	—	—	—
09-1997 to 11-1998	0.10	-0.23	-7.52	-0.07	0.20	—	-9.49	-0.05
	(0.08)	(0.22)	(1.47)	(0.02)	(0.04)	—	(1.52)	(0.02)
12-1998 to 02-2008	0.29	0.38	-1.46	0.00	—	—	-1.52	0.00
	(0.09)	(0.08)	(0.48)	(0.00)	—	—	(0.59)	(0.00)
03-2008 to 05-2009	0.04	-0.41	-6.97	-0.03	—	—	-6.17	-0.04
	(0.05)	(0.20)	(0.51)	(0.01)	—	—	(0.54)	(0.01)
06-2009 to 10-2010	0.06	1.23	0.09	-0.05	—	—	-0.99	—
	(0.21)	(0.35)	(0.94)	(0.01)	—	—	(0.56)	—
11-2010 to 03-2016	0.28	0.24	-1.91	-0.01	—	—	—	-0.01
	(0.16)	(0.10)	(0.73)	(0.00)	—	—	—	(0.01)

The Table details the parameter estimates of the preferred CP-MV model and of the selective segmentation process given the selected factors and the breakpoints found by the CP-MV model. Parentheses indicate standard deviations. A cell filled with '—' indicates that the parameter does not vary over the related period. The posterior probability of the selective segmentation model amounts to 52%.

3.7.3 Out-of-sample

Sections 3.7.1 and 3.7.2 highlight the in-sample advantages of detecting which parameter truly varies when a break is detected. In addition to that, since the selective segmentation method can more accurately estimate parameters that do not change when a break occurs, we could also expect some prediction gains with respect to the standard CP model. In this Section, we investigate this aspect using the root mean squared forecast errors (RMSFE) and the

cumulative log predictive density (CLPD), two standard loss functions specified as,

$$\text{RMSFE} = \sqrt{\frac{1}{T-\underline{t}}\sum_{t=\underline{t}+1}^{T}(y_t-\hat{y}_t)^2}, \text{ and CLPD} = \sum_{t=\underline{t}+1}^{T}\log f(y_t|y_{1:t-1},\mathbf{x}_t),$$

in which $\hat{y}_t$ is the conditional mean of y_t given the information up to period t, $f(y_t|y_{1:t-1},\mathbf{x}_t)$ denotes the predictive density of the model and $\underline{t}+1$ denotes the beginning of the out-of-sample forecasting period. In our prediction exercise, the training set is fixed to 20% of the sample size and the 80% remaining observations are used to assess the forecast performance (i.e., $\underline{t}=0.2T$). Since our data comprise 265 monthly returns, the out-of-sample set of observations amounts to 212 months. Each time we move forward by one month, all the considered models are re-estimated and a forecast for the next period is produced.

As competitors to our model, we consider three other processes: i) a linear regression, ii) a standard CP model with breakpoints determined by the modified method of Yau and Zhao (2016) documented in Section 3.5.1 (hereafter CP-YZ), iii) a CP model with the number and the locations of the breakpoints selected by minimizing the MDL criterion (hereafter CP-MDL).[9] The minimization of the MDL criterion is carried out using the dynamic programming of Bai and Perron (2003).[10] In addition to the factors and an intercept, we also account for the autocorrelation of the HF returns by fixing the AR order to the value given in Table 3.6. Regarding the CLPD metric, we assume a normal distribution for the error term and we also use the prior distributions given in Equation (3.13) for the linear and the full CP models.

Table 3.12 documents the RMSFE and the CLPD criteria for all the Credit Suisse HF returns. For both metrics, we observe that the linear model dominates at least half of the times. Overall, the selective segmentation method improves the RMSFE and the CLPD for 6 and 5 out of 14 HF returns, respectively. Importantly, Table 3.12 highlights that the selective segmentation process provides the most robust predictions. In particular, our approach delivers at least the second best predictive performance for all the HF returns. This is evidence that model averaging stabilizes the forecast by reducing its variance as argued in Rapach et al. (2009). Interestingly, our method compares extremely well with respect to the two CP models since it outperforms them 13 out of 14 HF returns for both metrics. Since the CP models are based on the same breakpoints as the selective segmentation processes, it is remarkable that the latter models almost systematically dominate CP models where all the parameters are time-varying. From this small sample of series, we could argue that the selective segmentation approach should replace the CP process as it would likely improve the forecast performance.

[9] We do not compare with the CP model of Meligkotsidou and Vrontos (2008) since the model is computationally too involved due to the number of explanatory variables. When an AR(2) model is selected, the number of models to consider at each iteration of the prediction exercise amounts to $10 \times 2^{15} = 327680$.

[10] See Eckley et al. (2011) for a discussion on the implementation of the algorithm for the MDL criterion. Minimum regime duration is set to $\frac{3}{2}(K+1)$ to avoid capturing outliers. This choice is in favor of the standard CP model as the parameter estimates of the new regimes are based on at least $\frac{3}{2}(K+1)$ observations.

Table 3.12 – RMSFE and CLPD for the fourteen HF strategies ($\underline{t} = 0.2T$)

				RMSFE			
Series	HFI	CNV	DSB	EME	EMN	EDR	EDD
Linear	1.41	**1.63**	**2.80**	2.99	**2.95**	**1.40**	1.54*
CP-MDL	1.41	1.92	2.95	**2.88**	4.12	1.62	1.84
SELO-MDL	1.30*	1.84*	2.88*	2.91*	3.95*	1.52	**1.53**
CP-Yau	1.70	2.63	4.03	3.90	6.06	3.24	2.89
SELO-Yau	**1.28**	2.07	3.08	2.93	5.26	1.50*	1.55
Series	EDM	EDRA	FIA	GMA	LES	MFU	MUS
Linear	**1.56**	1.06*	**1.22**	2.57	1.52	3.31*	**1.21**
CP-MDL	1.72	1.12	1.36	2.57	1.64	3.50	1.40
SELO-MDL	1.61	1.07	1.33*	**2.39**	1.49*	**3.31**	1.30*
CP-Yau	2.29	2.39	5.66	3.23	2.69	4.28	1.83
SELO-Yau	1.60*	**1.06**	1.57	2.45*	**1.47**	3.31	1.40

				CLPD			
Series	HFI	CNV	DSB	EME	EMN	EDR	EDD
Linear	-365.63	**-401.02**	**-513.39**	-527.76	**-547.17**	**-356.96**	-360.72
CP-MDL	-359.08	-444.16	-526.48	**-515.71**	-713.12	-433.56	-389.30
SELO-MDL	**-347.80**	-434.44	-519.29*	-516.35*	-688.23*	-402.20	-353.99*
CP-Yau	-366.11	-483.20	-570.98	-565.77	-916.15	-482.06	-439.75
SELO-Yau	-347.83*	-434.19*	-529.15	-519.49	-932.26	-395.13*	**-346.94**
Series	EDM	EDRA	FIA	GMA	LES	MFU	MUS
Linear	**-375.00**	-301.93*	**-329.73**	-490.10	-382.79	**-553.71**	**-330.48**
CP-MDL	-445.12	-307.96	-377.46	-484.06	-387.78	-564.47	-343.62
SELO-MDL	-389.83	-303.77	-352.53*	**-467.38**	-369.13*	-555.02	-339.35
CP-Yau	-460.95	-426.12	-472.99	-492.14	-464.27	-606.26	-368.72
SELO-Yau	-384.86*	**-301.57**	-380.37	-472.48*	**-364.02**	-554.74*	-337.58*

The Table details the RMSFE and the CLPD for five processes. Bold values indicate the model that delivers the best prediction performance. A star points out the second best model.

3.8 Conclusion

Since the seminal work of Chernoff and Zacks (1964), many CP detection methods for linear models have been proposed. Most of these CP models have in common to assume, at least in practice, that all the model parameters have to change when a break is detected. In this paper, we propose to go beyond this standard framework by capturing which parameters vary when a structural break occurs. Even when conditioning to the break dates, detecting the parameters that vary from one segment to the next is not straightforward since the number of possibilities grows exponentially with the number of breaks and the number of explanatory variables. To solve this dimensional problem, we propose a penalized regression method to explore the model space and we select the best specification by maximizing a criterion that can be interpreted as a marginal likelihood in the Bayesian paradigm.

To carry out the model space exploration, we use an almost unbiased penalty function, a desirable property in CP frameworks that is not exhibited by standard penalty functions (e.g., LASSO and Ridge estimators). Also, we prove the consistency of our estimator and we show how to estimate it using the DAEM algorithm. To apply the DAEM algorithm in our context, we transform the penalty function into a mixture of Normal distributions. This simple transformation greatly speeds the estimation as the DAEM algorithm iterates over closed-form expressions.

Once the promising models have been uncovered by the penalized regression approach, selecting the parameters of the penalty function is carried out by maximizing a marginal likelihood. Thanks to the Bayesian interpretation of this consistent criterion, we can take model uncertainty into account and can do Bayesian model averaging, a feature that generally improves forecast performance. A simulation study highlights that our selective segmentation method works well in practice for a range of diversified data generating processes.

We illustrate our approach with HF returns. The selective segmentation model has two main advantages. First, as the standard CP models, it detects the breakpoints and the corresponding regimes. Second, it highlights the time-varying dynamics of the changing risk factors. When we compare our model with previous advanced CP models, we observe that it is particularly appealing to capture the time-varying dynamics of risk exposures. Then, we test the predictive performance of the selective segmentation approach with respect to the linear regression and standard CP processes. We note that our method produces the most robust forecasts and almost systematically dominates the CP processes based on the same breakpoints. These encouraging results suggest promising developments and applications in financial economics.

Conclusion

This dissertation contributes to the literature on social interactions and time series modelling. It develops new promising models to estimate peer effects and analyse long time series. It is structured in three (03) separate and independent chapters. The first chapter studies a method for estimating peer effects through social networks when researchers do not observe the network structure. The second chapter presents a structural model of peer effects in which the dependent variable is counting. The third chapter presents a time series modelling approach through the linear-in-means specification by relaxing the assumption that a break triggers a change in all the model parameters.

The first chapter shows that the estimation of the linear-in-means peer effects model is possible when the researchers have (a consistent estimator of) the *distribution* of the network and not the network data itself. It presents an estimation strategy that adapts the instrumental variables procedure used to estimate this model when the network structure is observed. Indeed, one can construct valid instruments using the distribution of the network. However, the instrumental variable estimator is only valid if the researcher also observes peer characteristics. For example, consider a binary variable (e.g. gender). One can obtain information on peers by asking questions such as "What fraction of your friends are female?". For continuous variables (e.g., age), one can obtain peer characteristics by asking about the average age of one's friends. This chapter also presents a Bayesian estimator which does not require observing peer characteristics. In this case, the assumed distribution of the network acts as a prior distribution, and the inferred network structure is updated through the MCMC algorithm.
There remains one important challenge about the approach developed in this chapter, in particular with respect to the study of compatible models to estimate the distribution of the network.

The second chapter studies a social network model for count data. Counting variables are present in most survey data (e.g., Number of cigarettes smoked, frequency of restaurant visits, frequency of participation in activities). The analysis of peer effects using count data are often carried out using standard approaches, such as the linear-in-means spatial autoregressive (SAR) model or the SAR Tobit (SART) model. I develop a structural network model in which the outcome is a counting variable. I show that estimating peer effects on this counting

variable using the SAR or SART model asymptotically underestimates the peer effects. The estimation bias decreases when the range of the dependent variable increases.
Moreover, It is shown that the model generalizes the rational expectations model of peer effects for binary data. The model estimation is straightforward and does not require computing the equilibrium of the game. I also show that the model is flexible in terms of dispersion fitting and allows equidispersion, overdispersion, and underdispersion as the generalized Poisson model. However, the model does not consider zero inflated specifications for data that contain structural and sampled zeros. One important example is the number of cigarettes smoked during a period time, where the zeros might denote non-smokers (structural zeros) or smokers who did not smoke during the period (sampled zeros).

The third chapter addresses an issue in time series modelling. Most change point (CP) models developed in the literature have in common to assume, at least in practice, that all the model parameters have to change when a break is detected. This chapter goes beyond this standard framework by capturing which parameters vary when a structural break occurs. Even when conditioning on the break dates, detecting the parameters that vary from one segment to the next is not straightforward since the number of possibilities grows exponentially with the number of breaks and the number of explanatory variables. To solve this dimensional problem, a penalized regression method is used to explore the model space and select the best specification by maximizing a criterion that can be interpreted as a marginal likelihood in the Bayesian paradigm.
A penalty function is used to carry out the model space exploration. This penalty function is almost unbiased, a desirable property in CP frameworks that is not exhibited by standard penalty functions (e.g., LASSO and Ridge estimators). Once the promising models have been uncovered by the penalized regression approach, selecting the parameters of the penalty function is carried out by maximizing a marginal likelihood. Thanks to the Bayesian interpretation of this consistent criterion, one can also take the model uncertainty into account and can do Bayesian model averaging, a feature that generally improves forecast performance.
This approach is illustrated with Hedge Funds returns. It produces the most robust forecasts and almost systematically dominates the CP processes based on the same breakpoints. However, although the approach works well for many time series models including heteroskedastic processes in practice, all the theoretical results are based on the assumption of homoskedasticity.

Appendix A

Chapter 1 of appendix

A.1 Proof of Proposition 1.1

The model can be written as:

$$\mathbf{y} = [\mathbf{I} - \alpha\mathbf{G}]^{-1}[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}].$$

Or, using the geometric expansion:

$$\mathbf{y} = \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^k \mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^k \boldsymbol{\varepsilon}$$

or

$$\mathbf{G}\mathbf{y} = \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^{k+1} \mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{\infty} \alpha^k \mathbf{G}^{k+1} \boldsymbol{\varepsilon}.$$

As such, any variable correlated with $\mathbf{GX}$, $\mathbf{G}^2\mathbf{X}$,... is also correlated with $\mathbf{Gy}$, conditional on $\mathbf{X}$. It remains to show that such variables are valid. For the first part of Proposition 1.1, we need:

$$\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{HX}, \mathbf{H}^2\mathbf{X}, ...] = 0,$$

which is true by assumption. For the second part of Proposition 1.1, we need:

$$\mathbb{E}[\boldsymbol{\varepsilon} + \boldsymbol{\eta}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...] = 0.$$

By Assumption A.5., this is equivalent to:

$$\mathbb{E}[\boldsymbol{\eta}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...] = 0,$$

which is equivalent to:

$$\mathbb{E}[\mathbf{Gy}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...] = \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X}, \hat{\mathbf{G}}\mathbf{X}, \hat{\mathbf{G}}^2\mathbf{X}, ...].$$

Using iterated expectations:

$$\mathbb{E}_{\mathbf{y}}\mathbb{E}[\mathbf{G}\mathbf{y}|\mathbf{X},\hat{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},...,\mathbf{y}]|\mathbf{X},\hat{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},...=\mathbb{E}[\mathbf{G}|\mathbf{X},\hat{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},...]\mathbb{E}[\mathbf{y}|\mathbf{X},\hat{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},...],$$

and so $\mathbb{E}[\mathbf{G}\mathbf{y}|\mathbf{X},\hat{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},...] = \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X},\hat{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},...]$ follows since $\mathbf{G}$ $\hat{\mathbf{G}}$ and $\tilde{\mathbf{G}}$ are iid, which implies that $\mathbb{E}[\mathbf{G}|\mathbf{X},\hat{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},...] = \mathbb{E}[\tilde{\mathbf{G}}|\mathbf{X},\hat{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},...]$.

A.2 Proof of Proposition 1.2

The model can be written as:

$$\mathbf{y} = [\mathbf{I} - \alpha\mathbf{G}]^{-1}[\mathbf{X}\boldsymbol{\beta} + \mathbf{G}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}].$$

Or, using the geometric expansion:

$$\mathbf{y} = \sum_{k=0}^{\infty}\alpha^k\mathbf{G}^k\mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{\infty}\alpha^k\mathbf{G}^k\mathbf{G}\mathbf{X}\boldsymbol{\gamma} + \sum_{k=0}^{\infty}\alpha^k\mathbf{G}^k\boldsymbol{\varepsilon}$$

or

$$\mathbf{G}\mathbf{y} = \sum_{k=0}^{\infty}\alpha^k\mathbf{G}^{k+1}\mathbf{X}\boldsymbol{\beta} + \sum_{k=0}^{\infty}\alpha^k\mathbf{G}^{k+2}\mathbf{X}\boldsymbol{\gamma} + \sum_{k=0}^{\infty}\alpha^k\mathbf{G}^{k+1}\boldsymbol{\varepsilon}.$$

As such, any variable correlated with $\mathbf{G}^2\mathbf{X}$, $\mathbf{G}^3\mathbf{X}$,... is also correlated with $\mathbf{G}\mathbf{y}$, conditional on $\mathbf{X}$ and $\mathbf{G}\mathbf{X}$. It remains to show that such variables are valid. For the first part of Proposition 1.2, we need:

$$\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{X},\mathbf{G}\mathbf{X},\mathbf{H}^2\mathbf{X},\mathbf{H}^3\mathbf{X},...] = 0,$$

which is true by assumption. For the second part of Proposition 1.2, we need:

$$\mathbb{E}[\boldsymbol{\varepsilon}+\boldsymbol{\eta}|\mathbf{X},\mathbf{G}\mathbf{X},\tilde{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},\hat{\mathbf{G}}^3\mathbf{X},...] = 0.$$

By Assumption A.5., this is equivalent to:

$$\mathbb{E}[\boldsymbol{\eta}|\mathbf{X},\mathbf{G}\mathbf{X},\tilde{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},\hat{\mathbf{G}}^3\mathbf{X},...] = 0,$$

which is equivalent to:

$$\mathbb{E}[\mathbf{G}\mathbf{y}|\mathbf{X},\mathbf{G}\mathbf{X},\tilde{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},\hat{\mathbf{G}}^3\mathbf{X},...] = \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X},\mathbf{G}\mathbf{X},\tilde{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},\hat{\mathbf{G}}^3\mathbf{X},...].$$

Using iterated expectations:

$$\begin{aligned}&\mathbb{E}_{\mathbf{y}}\mathbb{E}[\mathbf{G}\mathbf{y}|\mathbf{X},\mathbf{G}\mathbf{X},\tilde{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},\hat{\mathbf{G}}^3\mathbf{X},...,\mathbf{y}]|\mathbf{X},\mathbf{G}\mathbf{X},\tilde{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},\hat{\mathbf{G}}^3\mathbf{X},...\\&=\mathbb{E}[\mathbf{G}|\mathbf{X},\mathbf{G}\mathbf{X},\tilde{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},\hat{\mathbf{G}}^3\mathbf{X},...]\mathbb{E}[\mathbf{y}|\mathbf{X},\mathbf{G}\mathbf{X},\tilde{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},\hat{\mathbf{G}}^3\mathbf{X},...],\end{aligned}$$

and so $\mathbb{E}[\mathbf{G}\mathbf{y}|\mathbf{X},\mathbf{G}\mathbf{X},\tilde{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},\hat{\mathbf{G}}^3\mathbf{X},...] = \mathbb{E}[\tilde{\mathbf{G}}\mathbf{y}|\mathbf{X},\mathbf{G}\mathbf{X},\tilde{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},\hat{\mathbf{G}}^3\mathbf{X},...]$ follows since $\mathbf{G}$ $\hat{\mathbf{G}}$ and $\tilde{\mathbf{G}}$ are iid, which implies that
$\mathbb{E}[\mathbf{G}|\mathbf{X},\mathbf{G}\mathbf{X},\tilde{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},\hat{\mathbf{G}}^3\mathbf{X},...] = \mathbb{E}[\tilde{\mathbf{G}}|\mathbf{X},\mathbf{G}\mathbf{X},\tilde{\mathbf{G}}\mathbf{X},\hat{\mathbf{G}}^2\mathbf{X},\hat{\mathbf{G}}^3\mathbf{X},...]$.

A.3 Additional Monte-Carlo Results

Table A.1 – Simulation results without contextual effects (2)

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - **Gy is Observed**					
Estimation results					
$Intercept = 2$	2.011	0.263	1.844	2.007	2.182
$\alpha = 0.4$	0.399	0.014	0.390	0.400	0.409
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	1,302.788	232.211	1,145.790	1,289.808	1,445.453
Hausman	1.210	1.713	0.099	0.509	1.642
Sargan	0.970	1.382	0.095	0.420	1.300
$N = 50$, $M = 100$ - **Gy is not observed - same draw**					
Estimation results					
$Intercept = 2$	4.843	0.324	4.630	4.828	5.044
$\alpha = 0.4$	0.244	0.017	0.234	0.245	0.256
$\beta_1 = 1$	1.002	0.003	1.000	1.002	1.004
$\beta_2 = 1.5$	1.503	0.007	1.498	1.503	1.507
Tests					
F-test	26,588.232	21,13.817	25,076.054	26,534.073	27,959.250
Hausman	339.782	44.092	310.264	338.783	368.245
Sargan	2.261	3.229	0.237	1.067	3.085
Validity					
$cor(\eta_i, \hat{x}_{i,1})$	−0.403	0.019	−0.416	−0.403	−0.390
$cor(\eta_i, \hat{x}_{i,2})$	−0.296	0.017	−0.307	−0.296	−0.284
$N = 50$, $M = 100$ - **Gy is not observed - different draws**					
Estimation results					
$Intercept = 2$	2.018	0.302	1.828	2.020	2.213
$\alpha = 0.4$	0.399	0.016	0.389	0.399	0.409
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.007	1.495	1.500	1.505
Tests					
F-test	1,311.739	231.331	1,150.204	1,305.262	1,457.470
Hausman	71.466	18.214	58.611	70.960	82.160
Sargan	1.000	1.365	0.095	0.456	1.389
Validity					
$cor(\eta_i, \hat{x}_{i,1})$	−0.001	0.015	−0.011	0.000	0.009
$cor(\eta_i, \hat{x}_{i,2})$	−0.001	0.014	−0.010	−0.001	0.009

This table presents additional Monte Carlo results where the model does not include contextual effects (see Table 1.1). The number of simulations is 1,000 and $\lambda = +\infty$. The instruments used are **GX** if **Gy** is observed, $\hat{\mathbf{G}}\mathbf{X}$ if **Gy** is not observed and approximated by $\tilde{\mathbf{G}}\mathbf{y}$.

Table A.2 – Simulation results without contextual effects (3)

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - **Gy is Observed**					
Estimation results					
$Intercept = 2$	2.001	0.188	1.867	1.990	2.125
$\alpha = 0.4$	0.400	0.010	0.393	0.401	0.407
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	2,776.682	416.894	2,489.415	2,755.846	3,036.762
Hausman	1.673	2.285	0.157	0.756	2.307
Sargan	2.889	2.356	1.156	2.274	3.959
$N = 50$, $M = 100$ - **Gy is not observed - same draw**					
Estimation results					
$Intercept = 2$	3.719	0.274	3.520	3.702	3.905
$\alpha = 0.4$	0.306	0.015	0.296	0.307	0.316
$\beta_1 = 1$	1.001	0.003	0.999	1.001	1.003
$\beta_2 = 1.5$	1.502	0.006	1.498	1.502	1.506
Tests					
F-test	38,566.204	5,520.495	34,806.901	38,162.018	42,221.999
Hausman	21.208	11.433	13.058	19.173	28.217
Sargan	247.860	32.667	225.385	246.456	268.697
Validity					
$cor(\eta_i, \hat{\tilde{x}}_{i,1})$	0.000	0.014	−0.009	0.000	0.009
$cor(\eta_i, \hat{\tilde{x}}_{i,2})$	0.000	0.014	−0.010	0.000	0.010
$N = 50$, $M = 100$ - **Gy is not observed - different draws**					
Estimation results					
$Intercept = 2$	2.002	0.202	1.857	1.999	2.137
$\alpha = 0.4$	0.400	0.011	0.392	0.400	0.408
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	2,798.114	418.630	2,508.203	2,795.093	3,071.257
Hausman	218.584	34.913	192.694	217.089	240.687
Sargan	2.828	2.232	1.112	2.293	3.888
Validity					
$cor(\eta_i, \hat{\tilde{x}}_{i,1})$	0.000	0.014	−0.009	0.000	0.009
$cor(\eta_i, \hat{\tilde{x}}_{i,2})$	0.000	0.014	−0.010	0.000	0.010

This table presents additional Monte Carlo results where the model does not include contextual effects (see Table 1.1). The number of simulations is 1,000 and $\lambda = 1$. The instruments used are $\left\{(\tilde{\mathbf{G}})^k \mathbf{X}, k = 1, 2\right\}$.

Table A.3 – Simulation results without contextual effects (4)

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - **Gy is Observed**					
Estimation results					
$Intercept = 2$	2.003	0.190	1.877	1.997	2.126
$\alpha = 0.4$	0.400	0.010	0.393	0.400	0.407
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	2,582.805	396.149	2,310.803	2,566.537	2,834.415
Hausman	1.648	2.135	0.175	0.803	2.350
Sargan	2.962	2.547	1.153	2.313	4.007
$N = 50$, $M = 100$ - **Gy is not observed - same draw**					
Estimation results					
$Intercept = 2$	4.088	0.312	3.885	4.068	4.283
$\alpha = 0.4$	0.285	0.017	0.275	0.287	0.296
$\beta_1 = 1$	1.001	0.003	0.999	1.001	1.004
$\beta_2 = 1.5$	1.502	0.007	1.498	1.502	1.507
Tests					
F-test	37,084.837	5,430.758	33,178.701	36,580.801	40,451.244
Hausman	24.804	12.850	15.008	23.435	32.737
Sargan	351.747	38.676	327.470	349.879	376.039
Validity					
$cor(\eta_i, \hat{\hat{x}}_{i,1})$	0.000	0.015	−0.011	−0.001	0.010
$cor(\eta_i, \hat{\hat{x}}_{i,2})$	0.000	0.014	−0.010	0.000	0.010
$N = 50$, $M = 100$ - **Gy is not observed - different draws**					
Estimation results					
$Intercept = 2$	2.006	0.216	1.866	2.010	2.143
$\alpha = 0.4$	0.400	0.012	0.392	0.400	0.407
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.007	1.495	1.500	1.505
Tests					
F-test	2,605.493	395.108	2,339.618	2,608.021	2,860.739
Hausman	298.171	43.582	267.427	296.419	325.031
Sargan	3.001	2.464	1.170	2.417	4.076
Validity					
$cor(\eta_i, \hat{\hat{x}}_{i,1})$	0.000	0.015	−0.011	−0.001	0.010
$cor(\eta_i, \hat{\hat{x}}_{i,2})$	0.000	0.014	−0.010	0.000	0.010

This table presents additional Monte Carlo results where the model does not include contextual effects (see Table 1.1). The number of simulations is 1,000 and $\lambda = +\infty$. The instruments used are $\left\{(\tilde{\mathbf{G}})^k \mathbf{X}, k = 1, 2\right\}$.

Table A.4 – Simulation results with contextual effects (2)

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - Instrument: $(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is observed					
Estimation results					
$Intercept = 2$	2.004	0.181	1.880	2.007	2.126
$\alpha = 0.4$	0.400	0.003	0.398	0.400	0.402
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
$\gamma_1 = 5$	5.001	0.021	4.987	5.001	5.015
$\gamma_2 = -3$	−3.001	0.029	−3.020	−3.001	−2.981
Tests					
F-test	16,233.492	1,898.917	14,917.312	16,163.443	17,491.015
Hausman	1.239	1.768	0.102	0.504	1.627
Sargan	1.005	1.364	0.104	0.487	1.289
$N = 50$, $M = 100$ - Instrument: $(\hat{\mathbf{G}})^2\mathbf{X}$ - $\boldsymbol{Gy}$ is not observed					
Estimation results					
$Intercept = 2$	2.002	0.217	1.855	2.005	2.150
$\alpha = 0.4$	0.400	0.004	0.397	0.400	0.403
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
$\gamma_1 = 5$	5.357	0.021	5.343	5.357	5.371
$\gamma_2 = -3$	−2.380	0.037	−2.405	−2.381	−2.357
$\hat{\gamma}_1 = 0$	−0.356	0.024	−0.372	−0.356	−0.340
$\hat{\gamma}_2 = 0$	−0.620	0.035	−0.642	−0.621	−0.597
Tests					
F-test	10,741.676	1,124.978	9,978.928	10,687.760	11,475.206
Hausman	22.119	10.011	14.423	21.247	28.586
Sargan	0.956	1.304	0.107	0.464	1.263

This table presents additional Monte Carlo results where the model includes contextual effects (see Table 1.2). The number of simulations is 1,000 and $\lambda = +\infty$.

Table A.5 – Simulation results with subpopulation unobserved fixed effects

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - Instrument: $\mathbf{J}(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\hat{\mathbf{y}}$ is observed					
Estimation results					
$\alpha = 0.4$	0.379	1.703	−0.122	0.398	0.938
$\beta_1 = 1$	1.000	0.005	0.997	1.000	1.003
$\beta_2 = 1.5$	1.500	0.009	1.496	1.500	1.505
$\gamma_1 = 5$	5.018	1.318	4.584	5.001	5.415
$\gamma_2 = -3$	−2.967	2.710	−3.861	−2.998	−2.156
Tests					
F-test	2.860	2.270	1.132	2.364	4.045
Hausman	1.029	1.398	0.105	0.484	1.392
Sargan	0.820	1.184	0.066	0.335	1.078
$N = 50$, $M = 100$ - Instrument: $\mathbf{J}(\hat{\mathbf{G}})^2\mathbf{X}$ - $\hat{\mathbf{y}}$ is not observed					
Estimation results					
$\alpha = 0.4$	0.345	1.609	−0.246	0.265	0.869
$\beta_1 = 1$	1.000	0.005	0.997	1.000	1.002
$\beta_2 = 1.5$	1.500	0.008	1.495	1.500	1.505
$\gamma_1 = 5$	5.351	0.170	5.284	5.343	5.408
$\gamma_2 = -3$	−2.378	0.123	−2.416	−2.373	−2.331
$\hat{\gamma}_1 = 0$	−0.307	1.442	−0.769	−0.236	0.237
$\hat{\gamma}_2 = 0$	−0.534	2.490	−1.346	−0.407	0.398
Tests					
F-test	2.773	2.189	1.072	2.304	3.773
Hausman	1.114	1.563	0.105	0.525	1.483
Sargan	0.889	1.329	0.083	0.381	1.206

This table presents additional Monte Carlo results where the model includes fixed effects (see Section 1.3.1). The number of simulations is 1,000 and $\lambda = 1$. In each group, the fixed effects are generated as $0.3x_{1,1} + 0.3x_{3,2} - 1.8x_{50,2}$.

Table A.6 – Simulation results with subpopulation unobserved fixed effects (2)

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 50$, $M = 100$ - Instrument: $\mathbf{J}(\tilde{\mathbf{G}})^2\mathbf{X}$ - $\hat{\mathbf{y}}$ is observed					
Estimation results					
$\alpha = 0.4$	0.623	4.670	−0.475	0.341	1.345
$\beta_1 = 1$	1.001	0.015	0.997	1.000	1.003
$\beta_2 = 1.5$	1.499	0.019	1.494	1.500	1.504
$\gamma_1 = 5$	4.826	3.644	4.263	5.042	5.688
$\gamma_2 = -3$	−3.357	7.485	−4.491	−2.913	−1.572
Tests					
F-test	1.025	1.023	0.293	0.710	1.451
Hausman	0.979	1.527	0.090	0.396	1.264
Sargan	0.665	1.184	0.042	0.203	0.801
$N = 50$, $M = 100$ - Instrument: $\mathbf{J}(\hat{\mathbf{G}})^2\mathbf{X}$ - $\hat{\mathbf{y}}$ is not observed					
Estimation results					
$\alpha = 0.4$	−0.071	3.561	−0.949	−0.047	0.876
$\beta_1 = 1$	0.999	0.010	0.996	0.999	1.002
$\beta_2 = 1.5$	1.500	0.018	1.495	1.501	1.506
$\gamma_1 = 5$	5.305	0.377	5.208	5.306	5.406
$\gamma_2 = -3$	−2.364	0.192	−2.411	−2.355	−2.306
$\hat{\gamma}_1 = 0$	0.062	3.183	−0.777	0.051	0.846
$\hat{\gamma}_2 = 0$	0.117	5.512	−1.360	0.062	1.459
Tests					
F-test	1.063	1.046	0.329	0.741	1.486
Hausman	0.993	1.499	0.095	0.409	1.255
Sargan	0.670	1.117	0.041	0.207	0.817

This table presents additional Monte Carlo results where the model includes fixed effects (see Section 1.3.1). The number of simulations is 1,000 and $\lambda = \infty$. In each group, the fixed effects are generated as $0.3x_{1,1} + 0.3x_{3,2} - 1.8x_{50,2}$.

Table A.7 – Simulation results with ARD: without contextual effects (1,000 replications)

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 250$, $M = 20$ - Instrument: $\tilde{\mathbf{G}}\mathbf{X}$ - $\mathbf{G}\mathbf{y}$ is observed					
Estimation results					
$Intercept = 2$	2.031	0.941	1.471	2.048	2.608
$\alpha = 0.4$	0.398	0.051	0.368	0.397	0.429
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	152.722	63.626	104.971	145.232	189.435
Hausman	1.037	1.467	0.098	0.435	1.352
Sargan	1.003	1.407	0.096	0.422	1.335
$N = 250$, $M = 20$ - Instrument: $\left\{\left(\tilde{\mathbf{G}}\right)^k \mathbf{X}, k = 1, 2\right\}$ -$\mathbf{G}\mathbf{y}$ is observed					
Estimation results					
$Intercept = 2$	1.999	0.429	1.721	1.985	2.254
$\alpha = 0.4$	0.400	0.023	0.386	0.400	0.415
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
Tests					
F-test	427.225	146.605	319.567	417.901	517.949
Hausman	1.026	1.467	0.116	0.491	1.339
Sargan	3.000	2.442	1.218	2.342	4.089
$N = 250$, $M = 20$ - Instrument: $\hat{\mathbf{G}}\mathbf{X}$ - $\mathbf{G}\mathbf{y}$ is not observed					
Estimation results					
$Intercept = 2$	1.854	1.117	1.182	1.879	2.569
$\alpha = 0.4$	0.408	0.061	0.368	0.407	0.445
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.007	1.496	1.500	1.505
Tests					
F-test	144.813	60.270	100.244	138.238	178.694
Hausman	30.339	14.849	19.275	27.985	39.404
Sargan	1.105	1.576	0.109	0.506	1.450
$N = 250$, $M = 20$ - Instrument: $\left\{\left(\hat{\mathbf{G}}\right)^k \mathbf{X}, k = 1, 2\right\}$ - $\mathbf{G}\mathbf{y}$ is not observed					
Estimation results					
$Intercept = 2$	1.969	0.531	1.599	1.962	2.318
$\alpha = 0.4$	0.402	0.029	0.383	0.402	0.422
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.007	1.496	1.500	1.505
Tests					
F-test	419.464	142.058	314.376	407.311	505.246
Hausman	171.162	53.370	133.847	169.651	205.461
Sargan	3.274	2.575	1.346	2.613	4.443

Table A.8 – Simulation results with nuclear ARD: without contextual effects, and $\hat{\mathbf{y}}$ is observed (1,000 replications)

Weight[†]		Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 250$, $M = 20$ - Instrument: $\tilde{\mathbf{G}}\mathbf{X}$ - $\mathbf{G}\mathbf{y}$ is observed						
$\tau = 200$	$\alpha = 0.4$	0.403	0.088	0.350	0.406	0.456
$\tau = 600$	$\alpha = 0.4$	0.396	0.139	0.322	0.396	0.468
$\tau = 1374$	$\alpha = 0.4$	0.399	0.242	0.279	0.396	0.516
$N = 250$, $M = 20$ - Instrument: $\left\{\left(\hat{\mathbf{G}}\right)^k \mathbf{X}, k = 1, 2\right\}$ - $\mathbf{G}\mathbf{y}$ is observed						
$\tau = 200$	$\alpha = 0.4$	0.400	0.036	0.378	0.401	0.422
$\tau = 600$	$\alpha = 0.4$	0.396	0.050	0.368	0.398	0.427
$\tau = 1374$	$\alpha = 0.4$	0.403	0.111	0.346	0.402	0.461

The parameter τ corresponds to the parameter λ in Alidaee et al. (2020) and represents the weight associated to the nuclear norm. In our context, the recommended value of Alidaee et al. (2020) is given by $\tau = 1374$. The optimal value (in terms of RMSE) found through cross-validation is $\tau = 600$, while $\tau = 200$ gives a RMSE similar to the recommended value.

Table A.9 – Simulation results with nuclear ARD: $\hat{\mathbf{y}}$ is observed, $\tau = 600$ (1,000 replications)

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 250$, $M = 20$ - Instrument: $\tilde{\mathbf{G}}\mathbf{X}^2$ - $\mathbf{G}\mathbf{y}$ is observed					
Estimation results					
$Intercept = 2$	2.001	0.267	1.825	1.999	2.181
$\alpha = 0.4$	0.400	0.011	0.393	0.400	0.406
$\beta_1 = 1$	1.000	0.003	0.998	1.000	1.002
$\beta_2 = 1.5$	1.500	0.006	1.496	1.500	1.504
$\gamma_1 = 5$	5.001	0.027	4.982	5.001	5.020
$\gamma_2 = -3$	−3.000	0.033	−3.022	−2.999	−2.977
Tests					
F-test	810.356	338.648	570.099	774.973	1005.605
Hausman	0.956	1.281	0.113	0.467	1.310
Sargan	1.033	1.417	0.107	0.515	1.395

Table A.10 – Simulation results with nuclear ARD: $\hat{\mathbf{y}}$ is not observed, $\tau = 600$ (1,000 replications)

Statistic	Mean	Std. Dev.	Pctl(25)	Median	Pctl(75)
$N = 250$, $M = 20$ - Instrument: $\tilde{\mathbf{G}}\mathbf{X}$ - $\mathbf{G}\mathbf{y}$ is not observed					
Estimation results					
$Intercept = 2$	5.215	2.934	4.008	5.423	6.748
$\alpha = 0.4$	0.225	0.161	0.141	0.214	0.291
$\beta_1 = 1$	1.001	0.004	0.998	1.001	1.003
$\beta_2 = 1.5$	1.501	0.007	1.497	1.501	1.505
Tests					
F-test	21.033	13.187	11.490	18.861	28.535
Hausman	7.626	6.980	2.377	5.937	10.976
Sargan	2.488	3.432	0.258	1.191	3.313
$cor(\eta_i, \hat{x}_{i,1})$	−0.006	0.022	−0.020	−0.006	0.008
$cor(\eta_i, \hat{x}_{i,2})$	−0.036	0.029	−0.055	−0.036	−0.016
$N = 250$, $M = 20$ - Instrument: $\left\{\left(\hat{\mathbf{G}}\right)^k \mathbf{X}, k = 1, 2\right\}$ - $\mathbf{G}\mathbf{y}$ is not observed					
Estimation results					
$Intercept = 2$	4.481	1.503	3.469	4.493	5.474
$\alpha = 0.4$	0.266	0.082	0.211	0.264	0.321
$\beta_1 = 1$	1.001	0.004	0.998	1.001	1.003
$\beta_2 = 1.5$	1.501	0.007	1.496	1.501	1.506
Tests					
F-test	50.332	24.549	31.909	46.716	64.013
Hausman	57.141	40.279	27.257	48.421	79.545
Sargan	11.075	9.786	4.209	7.877	15.298
$cor(\eta_i, \hat{\hat{x}}_{i,1})$	−0.011	0.023	−0.025	−0.011	0.005
$cor(\eta_i, \hat{\hat{x}}_{i,2})$	−0.064	0.040	−0.092	−0.064	−0.036

A.4 ARD Simulations Setting

This section provides details about ARD simulation and model estimation using a MCMC method. We simulate the network for a population of 5000 individuals divided into $m = 20$ groups of $n = 250$ individuals. Within each group, the probability of a link is:

$$\mathbb{P}(a_{ij} = 1) \propto \exp\{\nu_i + \nu_j + \zeta \mathbf{z}_i' \mathbf{z}_j\}. \tag{A.1}$$

Since there is no connection between the groups, the networks are simulated and estimated independently. We first present how we simulate the data following the model (1.7).

A.5 ARD Simulation

The parameters are defined as follows: $\zeta = 1.5$, $\nu_i \sim \mathcal{N}(-1.25, 0.37)$, and $\mathbf{z}_i$ are distributed uniformly according to a von Mises–Fisher distribution. We use a hypersphere of dimension 3. We set the same values for the parameter for the 20 groups. We generate the probabilities of links in each network following Breza et al. (2020).

$$\mathbb{P}(a_{ij} = 1 | \nu_i, \nu_j, \zeta, \mathbf{z}_i, \mathbf{z}_j) = \frac{\exp\{\nu_i + \nu_j + \zeta \mathbf{z}_i' \mathbf{z}_j\} \sum_{i=1}^{N} d_i}{\sum_{ij} \exp\{\nu_i + \nu_j + \zeta \mathbf{z}_i' \mathbf{z}_j\}}, \tag{A.2}$$

where d_i is the degree defined by $d_i \approx \dfrac{C_p(0)}{C_p(\zeta)} \exp(\nu_i) \displaystyle\sum_{i=1}^{N} \exp(\nu_i)$, the function $C_p(.)$ is the normalization constant in the von Mises–Fisher distribution density function. After computing the probability of a link for any pair in the population, we sample the entries of the adjacency matrix using a Bernoulli distribution with probability (A.2).

To generate the ARD, we require the "traits" (e.g. cities) for each individual. We set $K = 12$ traits on the hypersphere. Their location $\mathbf{v}_k$ is distributed uniformly according to the von Mises–Fisher distribution. The individuals having the trait k are assumed to be generated by a von Mises–Fisher distribution with the location parameter $\mathbf{v}_k$ and the intensity parameter $\eta_k \sim |\mathcal{N}(4, 1)|$, $k = 1, \ldots, 12$.

We attribute traits to individuals given their spherical coordinates. We first define N_k, the number of individuals having the trait k:

$$N_k = \left\lfloor r_k \frac{\sum_{i=1}^{N} f_{\mathcal{M}}(\mathbf{z}_i | \mathbf{v}_k, \eta_k)}{\max_i f_{\mathcal{M}}(\mathbf{z}_i | \mathbf{v}_k, \eta_k)} \right\rfloor,$$

where $\lfloor x \rfloor$ stands for the greatest integer less than or equal to x, r_k is a random number uniformly distributed over $(0.8; 0.95)$, and $f_{\mathcal{M}}(\mathbf{z}_i | \mathbf{v}_k, \eta_k)$ is the von Mises–Fisher distribution density function evaluated at $\mathbf{z}_i$ with the location parameter $\mathbf{v}_k$ and the intensity parameter η_k.

The intuition behind this definition for N_k is that when many $\mathbf{z}_i$ are close to $\mathbf{v}_k$, many individuals should have the trait k.

We can finally attribute trait k to individual i by sampling a Bernoulli distribution with the probability f_{ik} given by:

$$f_{ik} = N_k \frac{f_{\mathcal{M}}(\mathbf{z}_i|\mathbf{v}_k, \eta_k)}{\sum_{i=1}^{N} f_{\mathcal{M}}(\mathbf{z}_i|\mathbf{v}_k, \eta_k)}.$$

The probability of having a trait depends on the proximity of the individuals to the trait's location on the hypersphere.

Once the traits for each individual and the network are generated, we can build the ARD.

A.6 Model Estimation

In practice, we only have the ARD and the traits for each individual. McCormick and Zheng (2015) propose a MCMC approach to infer the parameters in model (A.1).

However, the spherical coordinates and the degrees in this model are not identified. The authors solve this issue by fixing some $\mathbf{v}_k$ and use the fixed positions to rotate the latent surface back to a common orientation at each iteration of the MCMC using a Procrustes transformation. In addition, the total size of a subset b_k is constrained in the MCMC.

As discussed by McCormick and Zheng (2015), the numbers of $\mathbf{v}_k$ and b_k to be set as fixed depends on the dimension of hypersphere. In our simulations, $\mathbf{v}_1$, $\mathbf{v}_2$, ..., $\mathbf{v}_5$ are set as fixed to rotate back the latent space. When simulating the data, we let $\mathbf{v}_1 = (1, 0, 0)$, $\mathbf{v}_2 = (0, 1, 0)$, and $\mathbf{v}_3 = (0, 0, 1)$. This ensures that the fixed positions on the hypersphere are spaced, as suggested by the authors, to use as much of the space as possible, maximizing the distance between the estimated positions. We also constrain b_3 to its true value. The results do not change when we constrain a larger set of b_k

Following Breza et al. (2020), we estimate the link probabilities using the parameters' posterior distributions. The gregariousness parameters are computed from the degrees d_i and the parameter ζ using the following equation:

$$\nu_i = \log(d_i) - \log\Big(\sum_{i=1}^{N} d_i\Big) + \frac{1}{2}\log\Big(\frac{C_p(\zeta)}{C_p(0)}\Big).$$

A.7 Network Sampling

This section explains how we sample the network in Algorithm 1.1. using Gibbs sampling. As discussed above, a natural solution is to update only one entry of the adjacency matrix at every step t of the MCMC. The entry (i, j) is updated according to its conditional posterior

distribution:

$$a_{ij} \sim P(\cdot|\mathbf{A}_{-ij}, \mathbf{y}) = \frac{\mathcal{P}(\mathbf{y}|a_{ij}, \mathbf{A}_{-ij})P(a_{ij}|\mathbf{A}_{-ij})}{\mathcal{P}(\mathbf{y}|1, \mathbf{A}_{-ij})P(a_{ij} = 1|\mathbf{A}_{-ij}) + \mathcal{P}(\mathbf{y}|0, \mathbf{A}_{-ij})P(a_{ij} = 0|\mathbf{A}_{-ij})}.$$

However, for each entry, we need to compute $\mathcal{P}(\mathbf{y}|0, \mathbf{A}_{-ij})$ and $\mathcal{P}(\mathbf{y}|1, \mathbf{A}_{-ij})$, which are the respective likelihoods of replacing a_{ij} by 0 or by 1. The likelihood computation requires the determinant of $(\mathbf{I} - \alpha\mathbf{G})$, which has a complexity $O(N^3)$, where N is the dimension of $\mathbf{G}$. This implies that we must compute $2N(N-1)$ times $\det(\mathbf{I} - \alpha\mathbf{G})$ to update the adjacency matrix at each step of the MCMC. As $\mathbf{G}$ is row-normalized, alternating any off-diagonal entry (i, j) in $\mathbf{A}$ between 0 and 1 perturbs all off-diagonal entries of the row i in $(\mathbf{I} - \alpha\mathbf{G})$. We show that $\mathbf{A}_{ij}$ and $\det(\mathbf{I} - \alpha\mathbf{G})$ can be updated by computing a determinant of an auxiliary matrix that requires only updating two entries.

Assume that we want to update the entry (i, j). Let h be the function defined in $\mathbb{N}$ such that $\forall\ x \in \mathbb{N}^*$, $h(x) = x$, and $h(0) = 1$. Let $\mathbf{L}$ be an N $\times$ N diagonal matrix, where $\mathbf{L}_{ii} = h(n_i)$, and n_i stands for the degree of i, while $\mathbf{L}_{kk} = 1$ for all $k \neq i$, and $\mathbf{W}$ is the matrix $\mathbf{G}$ where the row i of $\mathbf{W}$ is replaced by the row i of $\mathbf{A}$. Then, since the determinant is linear in each row, we can obtain $\mathbf{I} - \alpha\mathbf{G}$ by dividing the row i of $\mathbf{L} - \alpha\mathbf{W}$ by $h(n_i)$. We get:

$$\det(\mathbf{I} - \alpha\mathbf{G}) = \frac{1}{h(n_i)}\det(\mathbf{L} - \alpha\mathbf{W}).$$

When a_{ij} changes (from 0 to 1, or 1 to 0), note that only the entries (i, i) and (i, j) change in $\mathbf{L} - \alpha\mathbf{W}$. Two cases can be distinguished.

- If $a_{ij} = 0$ before the update, then the new degree of i will be $n_i + 1$. Thus, the entry (i, i) in $\mathbf{L} - \alpha\mathbf{W}$ will change from $h(n_i)$ to $h(n_i + 1)$ (since the diagonal of $\mathbf{W}$ equals 0) and the entry (i, j) from 0 to $-\alpha$. The new determinant is therefore given by:

  $$\det(\mathbf{I} - \alpha\mathbf{G}^*) = \frac{1}{h(n_i + 1)}\det(\mathbf{L}^* - \alpha\mathbf{W}^*),$$

  where $\mathbf{G}^*$, $\mathbf{L}^*$, and $\alpha\mathbf{W}^*$ are the new matrices, once a_{ij} has been updated.
- If $a_{ij} = 1$ before the update, then the new degree of k will be $n_i - 1$. Thus the entry (i, i) in $\mathbf{L} - \alpha\mathbf{W}$ will change from $h(n_i)$ to $h(n_i - 1)$ and the entry (i, j) from $-\alpha$ to 0. The new determinant is therefore given by:

  $$\det(\mathbf{I} - \alpha\mathbf{G}^*) = \frac{1}{h(n_i - 1)}\det(\mathbf{L}^* - \alpha\mathbf{W}^*).$$

Then, to update $\det(\mathbf{L} - \alpha\mathbf{W})$ when only the entries (i, i) and (i, j) change, we adapt the Lemma 1 in Hsieh et al. (2019) as follows:

Proposition A.1. *Let $\mathbf{e}_i$ be the i'th unit basis vector in $\mathbb{R}^N$. Let $\mathbf{M}$ denote an $N \times N$ matrix and $\mathbf{B}_{ij}(\mathbf{Q}, \epsilon)$ an $N \times N$ matrix as function of an $N \times N$ matrix $\mathbf{Q}$ and a real value ϵ, such*

that:

$$\mathbf{B}_{ij}(\mathbf{Q}, \epsilon) = \frac{\mathbf{Q}\mathbf{e}_i\mathbf{e}_j'\mathbf{Q}}{1 + \epsilon\mathbf{e}_j'\mathbf{Q}\mathbf{e}_i}. \tag{A.3}$$

Adding a perturbation ϵ_1 in the (i, i)th position and a perturbation ϵ_2 in the (i, j)th position to the matrix $\mathbf{M}$ can be written as $\tilde{\mathbf{M}} = \mathbf{M} + \epsilon_1\mathbf{e}_i\mathbf{e}_i' + \epsilon_2\mathbf{e}_i\mathbf{e}_j'$.

(a) The inverse of the perturbed matrix can be written as:

$$\tilde{\mathbf{M}}^{-1} = \mathbf{M}^{-1} - \epsilon_1\mathbf{B}_{ii}(\mathbf{M}^{-1}, \epsilon_1) - \epsilon_2\mathbf{B}_{ij}\left(\mathbf{M}^{-1} - \epsilon_1\mathbf{B}_{ii}(\mathbf{M}^{-1}, \epsilon_1), \epsilon_2\right).$$

(b) The determinant of the perturbed matrix can be written as:

$$det\left(\tilde{\mathbf{M}}\right) = \left(1 + \epsilon_2\mathbf{e}_j'\left(\mathbf{M}^{-1} - \epsilon_1\mathbf{B}_{ii}(\mathbf{M}^{-1}, \epsilon_1)\mathbf{e}_i\right)\right)(1 + \epsilon_1\mathbf{e}_i'\mathbf{M}^{-1}\mathbf{e}_i)det\left(\mathbf{M}\right).$$

Proof. (a) By the Sherman–Morrison formula (Mele, 2017), we have:

$$\left(\mathbf{M} + \epsilon\mathbf{e}_i\mathbf{e}_j'\right)^{-1} = \mathbf{M}^{-1} - \epsilon\frac{\mathbf{M}^{-1}\mathbf{e}_i\mathbf{e}_j'\mathbf{M}^{-1}}{1 + \epsilon\mathbf{e}_j'\mathbf{M}^{-1}\mathbf{e}_i} = \mathbf{M}^{-1} - \epsilon\mathbf{B}_{ij}(\mathbf{M}, \epsilon).$$

Thus,

$$\begin{aligned}
\tilde{\mathbf{M}}^{-1} &= \left((\mathbf{M} + \epsilon_1\mathbf{e}_i\mathbf{e}_i') + \epsilon_2\mathbf{e}_i\mathbf{e}_j'\right)^{-1}, \\
\tilde{\mathbf{M}}^{-1} &= (\mathbf{M} + \epsilon_1\mathbf{e}_i\mathbf{e}_i')^{-1} - \epsilon_2\mathbf{B}_{ij}((\mathbf{M} + \epsilon_1\mathbf{e}_i\mathbf{e}_i')^{-1}, \epsilon_2), \\
\tilde{\mathbf{M}}^{-1} &= \mathbf{M}^{-1} - \epsilon_1\mathbf{B}_{ii}(\mathbf{M}^{-1}, \epsilon_1) - \epsilon_2\mathbf{B}_{ij}\left(\mathbf{M}^{-1} - \epsilon_1\mathbf{B}_{ii}(\mathbf{M}^{-1}, \epsilon_1), \epsilon_2\right).
\end{aligned}$$

(b) By the matrix determinant lemma (Johnson and Horn, 1985), we have:

$$\det\left(\mathbf{M} + \epsilon\mathbf{e}_i\mathbf{e}_j'\right) = (1 + \epsilon\mathbf{e}_j'\mathbf{M}^{-1}\mathbf{e}_i)\det\left(\mathbf{M}\right).$$

It follows that:

$$\begin{aligned}
\det\left(\tilde{\mathbf{M}}\right) &= \det\left((\mathbf{M} + \epsilon_1\mathbf{e}_i\mathbf{e}_i') + \epsilon_2\mathbf{e}_i\mathbf{e}_j'\right), \\
\det\left(\tilde{\mathbf{M}}\right) &= (1 + \epsilon_2\mathbf{e}_j'(\mathbf{M} + \epsilon_1\mathbf{e}_i\mathbf{e}_i')^{-1}\mathbf{e}_i)\det\left(\mathbf{M} + \epsilon_1\mathbf{e}_i\mathbf{e}_i'\right), \\
\det\left(\tilde{\mathbf{M}}\right) &= \left(1 + \epsilon_2\mathbf{e}_j'\left(\mathbf{M}^{-1} - \epsilon_1\mathbf{B}_{ii}(\mathbf{M}^{-1}, \epsilon_1)\mathbf{e}_i\right)\right)(1 + \epsilon_1\mathbf{e}_i'\mathbf{M}^{-1}\mathbf{e}_i)\det\left(\mathbf{M}\right).
\end{aligned}$$

□

The method proposed above becomes computationally intensive when many entries must be updated simultaneously. We also propose an alternative method that allows updating the block for entries in $\mathbf{A}$. Let $\mathbf{D} = (\mathbf{I} - \alpha\mathbf{G})$; we can write:

$$\det(\mathbf{D}) = \sum_{j=1}^{N}(-1)^{i+j}\mathbf{D}_{\mathrm{ij}}\delta_{\mathrm{ij}}, \tag{A.4}$$

where i denotes any row of $\mathbf{D}$ and δ_{ij} the minor[1] associated with the entry (i, j). The minors of row i do not depend on the values of entries in row i. To update any block in row i, we therefore compute the N minors associated to i and use this minor within the row. We can then update many entries simultaneously without increasing the number of times that we compute $\det(\mathbf{D})$.

One possibility is to update multiple links simultaneously by randomly choosing the number of entries to consider and their position in the row. As suggested by Chib and Ramamurthy (2010), this method would help the Gibbs to converge more quickly. We can summarize how we update the row i as follows:

(a) Compute the N minors $\delta_{i1}, \ldots, \delta_{in}$.

(b) Let $\Omega_{\mathbf{G}}$ be the entries to update in the row i, and $n_{\mathbf{G}} = |\Omega_{\mathbf{G}}|$ the number of entries in $\Omega_{\mathbf{G}}$.

(b.1) Choose r, the size of the block to update, as a random integer number such that $1 \leq r \leq n_{\mathbf{G}}$. In practice, we choose $r \leq \min(5, n_{\mathbf{G}})$ since the number of possibilities of links to consider grows exponentially with r.

(b.2) Choose the r random entries from $\Omega_{\mathbf{G}}$. These entries define the block to update.

(b.3) Compute the posterior probabilities of all possibilities of links inside the block and update the block (there are 2^r possibilities). Use the minors calculated at (a) and the formula (A.4) to quickly compute $det(\mathbf{D})$.

(b.4) Remove the r drawn positions from $\Omega_{\mathbf{G}}$ and let $n_{\mathbf{G}} = n_{\mathbf{G}} - r$. Replicate (b.1), (b.2), and (b.3) until $n_{\mathbf{G}} = 0$.

A.8 Posterior Distributions for Algorithm 1.1..

To compute the posterior distributions, we set prior distributions on $\tilde{\alpha}$, $\mathbf{\Lambda}$, and σ^2, where $\tilde{\alpha} = \log(\frac{\alpha}{1-\alpha})$ and $\mathbf{\Lambda} = [\boldsymbol{\beta}, \boldsymbol{\gamma}]$. In Algorithm 1.1., we therefore sample $\tilde{\alpha}$ and compute α, such that $\alpha = \frac{\exp(\tilde{\alpha})}{1+\exp(\tilde{\alpha})}$. Using this functional form for computing α ensures that $\alpha \in (0, 1)$. The prior distributions are set as follows:

$$\begin{aligned}
\tilde{\alpha} &\sim \mathcal{N}(\mu_{\tilde{\alpha}}, \sigma^2_{\tilde{\alpha}}), \\
\mathbf{\Lambda}|\sigma^2 &\sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{\Lambda}}, \sigma^2 \mathbf{\Sigma}_{\mathbf{\Lambda}}), \\
\sigma^2 &\sim IG(\frac{a}{2}, \frac{b}{2}).
\end{aligned}$$

For the simulations and estimations in this paper, we set: $\mu_{\tilde{\alpha}} = -1$, $\sigma_{\tilde{\alpha}}^{-2} = 2$, $\boldsymbol{\mu}_{\mathbf{\Lambda}} = \mathbf{0}$, ${\mathbf{\Sigma}_{\mathbf{\Lambda}}}^{-1} = \frac{1}{100}\boldsymbol{I}_K$, $a = 4$ and $b = 4$, where $\boldsymbol{I}_K$ is the identity matrix of dimension K and $K = \dim(\mathbf{\Lambda})$.

[1] The determinant of the submatrix of $\mathbf{M}$ by removing row i and column j.

Following Algorithm 1.1., α is updated at each iteration t of the MCMC by drawing $\tilde{\alpha}^*$ from the proposal $\mathcal{N}(\tilde{\alpha}_{t-1}, \xi_t)$, where the jumping scale ξ_t is also updated at each t following Atchadé and Rosenthal (2005) for an acceptance rate of a^* targeted at 0.44. As the proposal is symmetrical, $\alpha^* = \dfrac{\exp(\tilde{\alpha}^*)}{1+\exp(\tilde{\alpha}^*)}$ is accepted with the probability:

$$\min\left\{1, \frac{\mathcal{P}(\mathbf{y}|\mathbf{A}_t, \mathbf{\Lambda}_{t-1}, \alpha^*)P(\tilde{\alpha}^*)}{\mathcal{P}(\mathbf{y}|\mathbf{A}_t, \boldsymbol{\theta}_{t-1})P(\tilde{\alpha}_t)}\right\}.$$

The parameters $\mathbf{\Lambda}_t = [\boldsymbol{\beta}_t, \boldsymbol{\gamma}_t]$ and σ_t^2 are drawn from their posterior conditional distributions, given as follows:

$$\mathbf{\Lambda}_t|\mathbf{y}, \mathbf{A}_t, \alpha_t, \sigma_{t-1}^2 \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathbf{\Lambda}_t}, \sigma_{t-1}^2\hat{\mathbf{\Sigma}}_{\mathbf{\Lambda}_t}),$$
$$\sigma_t^2|\mathbf{y}, \mathbf{A}_t, \boldsymbol{\theta}_t \sim IG\left(\frac{\hat{a}_t}{2}, \frac{\hat{b}_t}{2}\right),$$

where,

$$\begin{aligned}
\hat{\mathbf{\Sigma}}_{\mathbf{\Lambda}_t}^{-1} &= \mathbf{V}_t'\mathbf{V}_t + \mathbf{\Sigma}_{\mathbf{\Lambda}}^{-1}, \\
\hat{\boldsymbol{\mu}}_{\mathbf{\Lambda}_t} &= \hat{\mathbf{\Sigma}}_{\mathbf{\Lambda}_t}\left(\mathbf{V}_t'(\mathbf{y} - \alpha_t\boldsymbol{G}_t\mathbf{y}) + \mathbf{\Sigma}_{\mathbf{\Lambda}}^{-1}\boldsymbol{\mu}_{\mathbf{\Lambda}}\right), \\
\hat{a}_t &= a + N, \\
\hat{b}_t &= b + (\mathbf{\Lambda}_t - \boldsymbol{\mu}_{\mathbf{\Lambda}})'\mathbf{\Sigma}_{\mathbf{\Lambda}}^{-1}(\mathbf{\Lambda}_t - \boldsymbol{\mu}_{\mathbf{\Lambda}}) + (\mathbf{y} - \alpha_t\mathbf{G}_t\mathbf{y} - \mathbf{V}_t\mathbf{\Lambda}_t)'(\mathbf{y} - \alpha_t\mathbf{G}_t\mathbf{y} - \mathbf{V}_t\mathbf{\Lambda}_t), \\
\mathbf{V}_t &= [\mathbf{1},\ \mathbf{X},\ \mathbf{G}_t\mathbf{X}].
\end{aligned}$$

A.9 Empirical Application

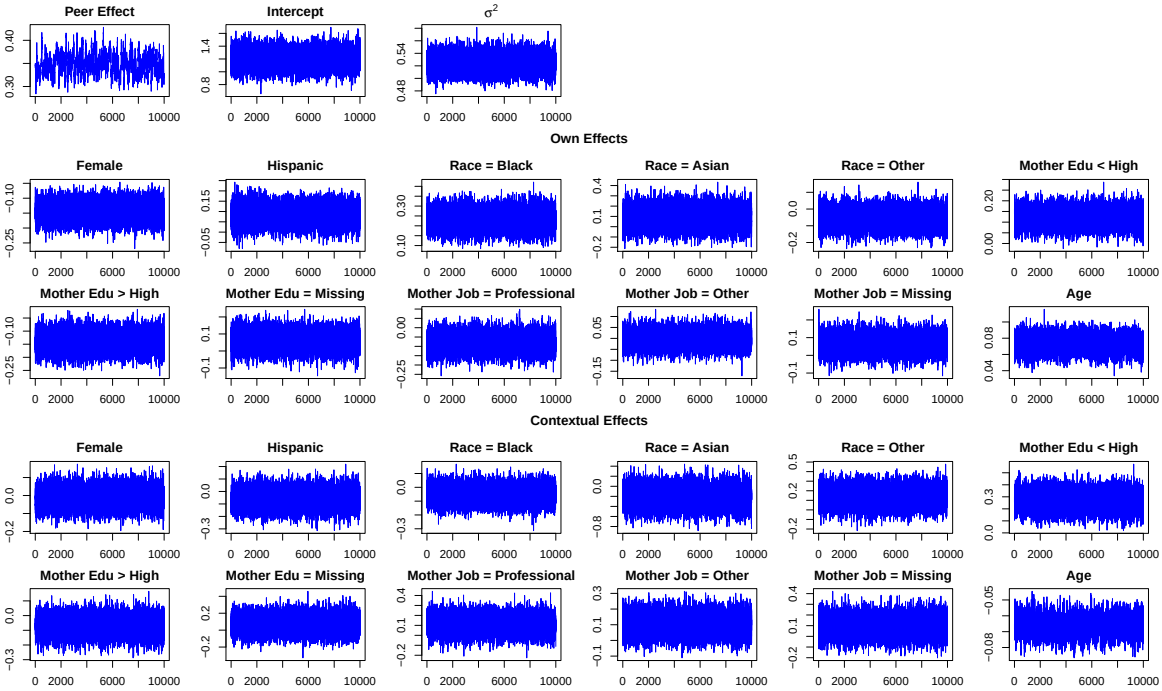


Figure A.1 – Simulations using the observed network

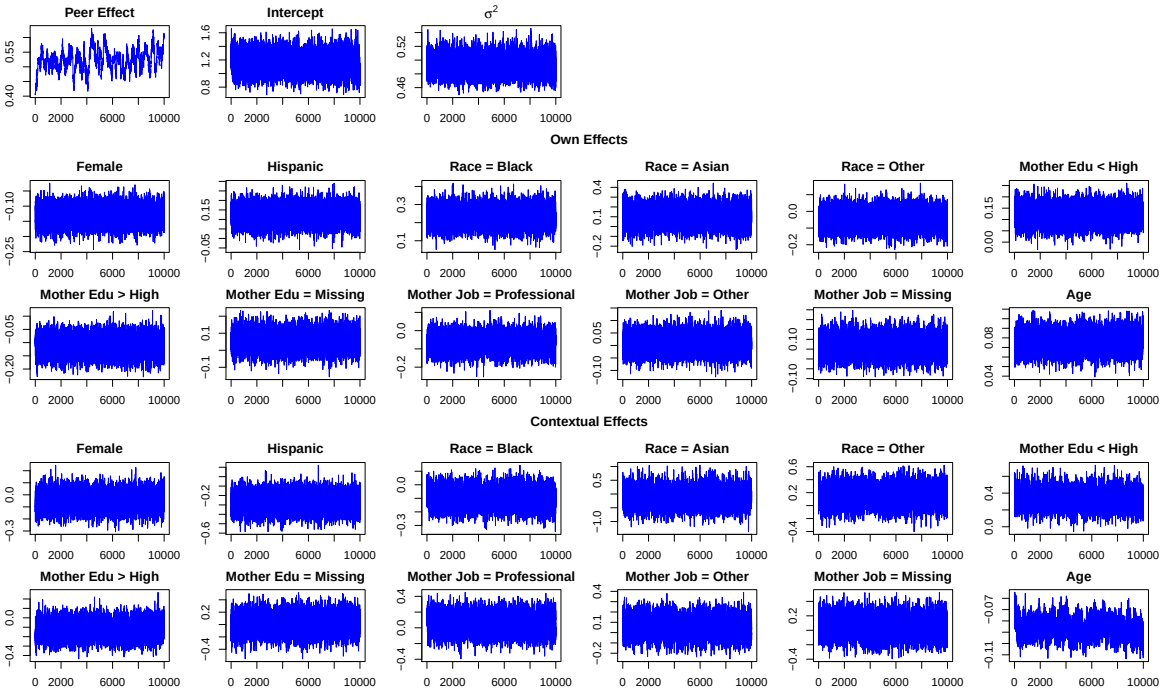


Figure A.2 – Simulations using the reconstructed network

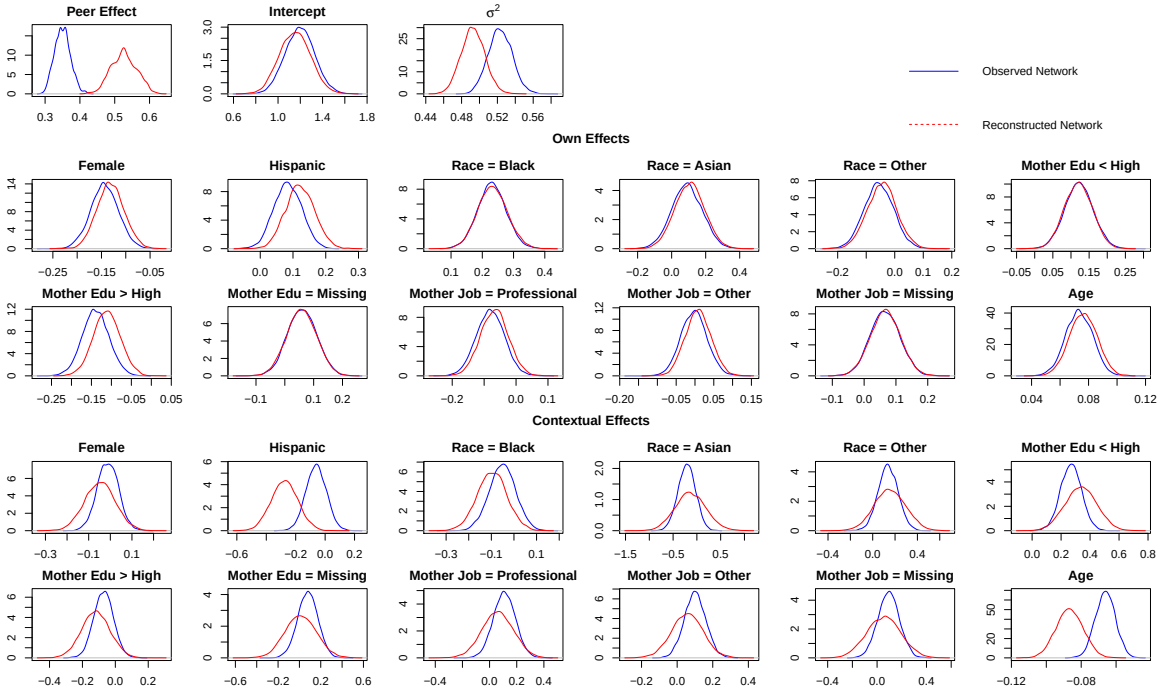


Figure A.3 – Posterior density

A.10 Expectation Maximization Algorithm

If $\mathbf{A}$ was observed, the objective would be to maximize the log-likelihood of the model with respect to $\boldsymbol{\theta}$:

$$\mathcal{P}(\mathbf{y}|\mathbf{A}, \mathbf{X}; \boldsymbol{\theta}).$$

Since $\mathbf{A}$ is unobserved, we propose to treat it as a latent variable. We therefore look for the value of $\boldsymbol{\theta}$ that maximizes:

$$\mathcal{P}(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta}) = \sum_{\mathbf{A}} \mathcal{P}(\mathbf{y}|\mathbf{A}, \mathbf{X}; \boldsymbol{\theta}) P(\mathbf{A}).$$

Since the number potential network structures is huge, evaluating this expectation is unfeasible.[2] We therefore propose to maximize $\mathcal{P}(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta})$ using an expectation maximization algorithm:

[2]For a population of only 5 individuals, the number of network structures is 2^{20}.

Algorithm A.1. Expectation maximization algorithm

Initialize $\boldsymbol{\theta}_0$, and for $t = 0, ..., T$, do the following;

Use a Metropolis–Hastings algorithm (see Algorithm 1.1.) to obtain the draws $(\mathbf{A}_1, ..., \mathbf{A}_R)$ from $P(\mathbf{A}|\mathbf{y}, \boldsymbol{\theta}_t)$;

Evaluate $Q_t(\boldsymbol{\theta}) \approx \sum_{r=1}^{R} \mathcal{P}(\mathbf{y}|\mathbf{A}_r, \mathbf{X}; \boldsymbol{\theta}_t)$;

Set $\boldsymbol{\theta}_{t+1} = \arg\max Q_t(\boldsymbol{\theta})$;

Appendix B

Chapter 2 of appendix

B.1 Proof of the Bayesian Nash Equilibrium (BNE)

B.1.1 Proof of Proposition 2.1

For any $\bar{\mathbf{y}} \in \mathbb{R}_+^N$, $\mathbf{L}(\bar{\mathbf{y}}) = (\ell_1(\bar{\mathbf{y}}) \ \dots \ell_n(\bar{\mathbf{y}}))'$, where $\ell_i(\bar{\mathbf{y}}) = \sum_{r=1}^{\infty} F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_r)$ for all $i \in \mathcal{V}$.

At the equilibrium, (p_{iq}) verifies (2.5),

$$\begin{aligned} p_{iq} &= F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_q) - F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_{q+1}), \\ \bar{y}_i &= \sum_{r=0}^{\infty} r p_{ir} = \underbrace{\sum_{r=0}^{\infty} r F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_r)}_{S_1} - \underbrace{\sum_{r=0}^{\infty} r F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_{r+1})}_{S_2}. \end{aligned} \tag{B.1}$$

Equation (B.1) holds because $S_1 < \infty$ and $S_2 < \infty$. To prove this, let $x < 0$ with $|x|$ being sufficiently large. By Assumption C, $f_\varepsilon = o(1/x^\alpha)$ at ∞ for some $\alpha > 3$. Then $F_\varepsilon = O(1/x^{\alpha-1})$ at $-\infty$, and $F_\varepsilon = o(1/x^{\alpha/2})$ at $-\infty$. Hence, $S_1 < \infty$. Analogously, $S_2 < \infty$.

$$\bar{y}_i = \sum_{r=0}^{\infty} r F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_r) - \sum_{r=0}^{\infty} (r+1) F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_{r+1}) + \sum_{r=0}^{\infty} F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_{r+1}),$$

$$\bar{y}_i = \sum_{r=1}^{\infty} r F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_r) - \sum_{r=1}^{\infty} r F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_r) + \sum_{r=0}^{\infty} F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_{r+1}),$$

$$\bar{y}_i = \sum_{r=1}^{\infty} F_\varepsilon(\lambda \mathbf{g}_i \bar{\mathbf{y}} + \psi_i - a_r) = \ell_i(\bar{\mathbf{y}}).$$

Hence, $\bar{\mathbf{y}} = \mathbf{L}(\bar{\mathbf{y}})$.

B.1.2 Proof of Theorem 2.1

From the BNE (2.5), the key determinant of the proof is to establish that the vector of equilibrium beliefs $\mathbf{p}$ exists (which implies the existence of an expected outcome $\bar{\mathbf{y}}$ at equilibrium) and that there is at most one expected equilibrium outcome $\bar{\mathbf{y}}$. This implies that there is a unique expected equilibrium outcome and thus, a unique vector of equilibrium beliefs.
Let $\mathbb{R}^{\infty}$ be the space of infinite-dimensional real vectors.[1] Let us denote by $\mathbf{p}_{\mathrm{q}} = (p_{1q}, \dots, p_{nq})'$, an n-dimensional vector for any $q \in \mathbb{N}$, $\mathbf{p} = \left(\mathbf{p}_0', \mathbf{p}_1', \mathbf{p}_2', \mathbf{p}_3', \dots\right)'$, $\mathbf{h}_1 = (a_0, a_1, a_2, a_3, \dots)'$, $\mathbf{h}_2 = (a_1, a_2, a_3, a_4, \dots)'$ infinite-dimensional vectors, and $\mathbf{1}_d$, the d-dimensional vector of ones for any $d \in \mathbb{N}^*$ or $d = \infty$. Let also $\mathbf{J} = (0, 1, 2, 3, \dots)$, an infinite-dimensional row-vector, and $\mathbf{B} = \mathbf{1}_{\infty} \otimes \mathbf{J} \otimes \mathbf{G}$. Equation (2.5) in matrix form is given by

$$\mathbf{p} = \mathbf{F}_{\varepsilon}\left(\lambda \mathbf{B}\mathbf{p} + \mathbf{1}_{\infty} \otimes \Psi - \mathbf{h}_1 \otimes \mathbf{1}_n\right) - \mathbf{F}_{\varepsilon}\left(\lambda \mathbf{B}\mathbf{p} + \mathbf{1}_{\infty} \otimes \Psi - \mathbf{h}_2 \otimes \mathbf{1}_n\right), \tag{B.2}$$

where $\mathbf{F}_{\varepsilon}$ is defined for any $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots) \in \mathbb{R}^{\infty}$ as $\mathbf{F}_{\varepsilon}(\boldsymbol{\omega}) = (F_{\varepsilon}(\omega_1), F_{\varepsilon}(\omega_2), \dots)$.
Assumption C implies that $F_{\varepsilon} = o(1/x)$ at $-\infty$. Therefore, $\exists\ M > 0$, such that $\forall\ i \in \mathcal{V}$, $q \in \mathbb{N}$, $p_{iq} \leq \dfrac{M}{q+1}$. Let $\mathbf{C_M}$ be a subset of $\mathbb{R}^{\infty}$ defined by

$$\mathbf{C_M} := \left\{ \mathbf{p} \in \mathbb{R}^{\infty} \ / \forall\ i \in \mathcal{V} \text{ and } q \in \mathbb{N},\ p_{iq} \geq 0 \text{ and } p_{iq} \leq \frac{M}{q+1} \right\}.$$

For any $M > 0$, $\mathbf{C_M}$ is a compact and convex nonempty subset of the infinite dimensional space $\mathbb{R}^{\infty}$.
Let also $\mathbf{H}$ be a mapping from $\mathbf{C_M}$ to itself, such that $\forall\ \mathbf{p} \in \mathbf{C_M}$,

$$\mathbf{H}(\mathbf{p}) = \mathbf{F}_{\varepsilon}\left(\lambda \mathbf{B}\mathbf{p} + \mathbf{1}_{\infty} \otimes \Psi - \mathbf{h}_1 \otimes \mathbf{1}_n\right) - \mathbf{F}_{\varepsilon}\left(\lambda \mathbf{B}\mathbf{p} + \mathbf{1}_{\infty} \otimes \Psi - \mathbf{h}_2 \otimes \mathbf{1}_n\right). \tag{B.3}$$

Any $\mathbf{p} \in \mathbf{C_M}$ is an equilibrium belief of the incomplete information network game with the utility (2.1) if $\mathbf{p} = \mathbf{H}(\mathbf{p})$. $\mathbf{H}$ is a continuous mapping from $\mathbf{C_M}$ to itself. By Schauder's fixed-point theorem (generalization of Brouwer's fixed-point theorem to an infinite dimensional space, see Smart, 1980, Chapter 2), there exists $\mathbf{p}^e \in \mathbf{C_M}$, such that $\mathbf{p}^e = \mathbf{H}(\mathbf{p}^e)$. By Proposition 2.1, the expected outcome $\bar{\mathbf{y}}^e = (\bar{y}_1^e \dots \bar{y}_n^e)$, where $\bar{y}_i^e = \sum_{r=0}^{\infty} r p_{ir}^e$, verifies $\bar{\mathbf{y}}^e = \mathbf{L}(\bar{\mathbf{y}}^e)$.
Let us show that $\mathbf{u} = \mathbf{L}(\mathbf{u})$ has at least one solution. By the contraction mapping theorem, it is sufficient to prove that $\forall\ \mathbf{u} = (u_1, \dots, u_n) \in \mathbb{R}^n$, $\left|\left|\dfrac{\partial \mathbf{L}(\mathbf{u})}{\partial \mathbf{u}'}\right|\right|_{\infty} < \bar{\kappa}$ for some $\bar{\kappa} < 1$ not depending on $\mathbf{u}$.
For all i and j,

$$\frac{\partial \ell_i(\mathbf{u})}{\partial u_j} = \lambda g_{ij} \underbrace{\sum_{r=1}^{\infty} f_{\varepsilon}\left(\lambda \mathbf{g}_i \mathbf{u} + \psi_i - a_r\right)}_{f_i^*} = \lambda g_{ij} f_i^*. \tag{B.4}$$

[1] A natural generalization of $\mathbb{R}^k$, $k \in \mathbb{N}^*$ (see Halmos, 2012).

From Equation (B.4), $\dfrac{\partial \mathbf{L}(\mathbf{u})}{\partial \mathbf{u}'}$ is defined by

$$\frac{\partial \mathbf{L}(\mathbf{u})}{\partial \mathbf{u}'} = \lambda \begin{pmatrix} g_{11}f_1^* & \cdots & g_{1n}f_1^* \\ \vdots & \vdots & \vdots \\ g_{n1}f_n^* & \cdots & g_{nn}f_n^* \end{pmatrix}.$$

It follows that

$$\begin{aligned} \left|\left|\frac{\partial \mathbf{L}(\mathbf{u})}{\partial \mathbf{u}'}\right|\right|_\infty &= |\lambda| \max_i \left\{ f_i^* \sum_{j=1}^n g_{ij} \right\}, \\ \left|\left|\frac{\partial \mathbf{L}(\mathbf{u})}{\partial \mathbf{u}'}\right|\right|_\infty &\leq |\lambda| \left(\max_i f_i^*\right) \max_i \left\{ \sum_{j=1}^n g_{ij} \right\} = |\lambda| \left(\max_i f_i^*\right) ||\mathbf{G}||_\infty. \end{aligned} \tag{B.5}$$

I will now focus on the term f_i^*.

$$f_i^* = \sum_{r=1}^{\infty} f_\varepsilon \left(\lambda \mathbf{g}_i \mathbf{u} + \psi_i - a_r\right) = \sum_{r=1}^{\infty} f_\varepsilon \left(m_i + a_1 - a_r\right),$$

where $m_i^* = \lambda \mathbf{g}_i \mathbf{u} + \psi_i - a_1$. As $a_q = a_1 + \gamma(q-1)$ for any $q \in \mathbb{N}^*$,

$$\begin{aligned} f_i^* &= \sum_{r=1}^{\infty} f_\varepsilon \left(m_i - \gamma(r-1)\right) < \sum_{k=-\infty}^{\infty} f_\varepsilon \left(m_i + \gamma k\right), \\ f_i^* &< \max_{u \in \mathbb{R}} \sum_{k=-\infty}^{\infty} f_\varepsilon \left(u + \gamma k\right) = \frac{1}{C_\gamma} \end{aligned} \tag{B.6}$$

From Equations (B.5) and (B.6),

$$\left|\left|\frac{\partial \mathbf{L}(\mathbf{u})}{\partial \mathbf{u}'}\right|\right|_\infty < \frac{|\lambda| ||\mathbf{G}||_\infty}{C_\gamma} < 1 \quad \text{by Assumption D.} \tag{B.7}$$

Hence, $\mathbf{u} = \mathbf{L}(\mathbf{u})$ has a unique solution $\bar{\mathbf{y}}^e$.

By Equation (2.5) and Proposition 2.1, it follows that $\mathbf{p} = \mathbf{H}(\mathbf{p})$ also has a unique solution $\mathbf{p}^e$, such that $p_{iq}^e = F_\varepsilon\left(\lambda \mathbf{g}_i \bar{\mathbf{y}}_j^e + \psi_i - a_q\right) - F_\varepsilon\left(\lambda \mathbf{g}_i \bar{\mathbf{y}}_j^e + \psi_i - a_{q+1}\right)$.

As a result, the incomplete information network game with the utility (2.1) has a unique pure strategy BNE with the equilibrium strategy profile $\mathbf{y}^{e*}$ given by $\mathbf{y}^{e*} = \lambda \mathbf{G} \bar{\mathbf{y}}^e + \boldsymbol{\psi} + \boldsymbol{\varepsilon}$, where the equilibrium belief system $\left(p_{iq}^e\right)_{\substack{i \in \mathcal{V} \\ q \in \mathbb{N}}}$ is such that $\bar{y}_i^e = \sum_{r=0}^{\infty} r p_{ir}^e$ is the unique solution of $\mathbf{u} = \mathbf{L}(\mathbf{u})$.

B.1.3 BNE when the increment of the sequence varies

Assume that the increment of the sequence in Assumption B varies; that is, there is a strictly increasing sequence $\left(a_q^i\right)_{q \in \mathbb{N}}$ such that if $y_i^* \in (a_q^i, a_{q+1}^i]$, then $y_i = q$ and $a_{q+1}^i - a_q^i = \gamma_q^i$ varies. Note that Proposition 2.1 is still true as long as Assumption C holds. To prove the uniqueness of the BNE, I consider the following assumption as an alternative to Assumption D.

Assumption D'. $|\lambda| < \frac{C_\gamma}{||\mathbf{G}||_\infty}$, *where* $C_\gamma = \left(\max_{u \in \mathbb{R}} \sum_{r=1}^{\infty} f_\varepsilon \left(u - a_r^i \right) \right)^{-1}$.

With the new definition of the sequence $\left(a_q^i\right)_{q \in \mathbb{N}}$, the mapping $\mathbf{L}$ is contracting under Assumption C and D'. Indeed, from Equation (B.5),

$$\left|\left| \frac{\partial \mathbf{L}(\mathbf{u})}{\partial \mathbf{u}'} \right|\right|_\infty \leq |\lambda| \left(\max_i f_i^* \right) ||\mathbf{G}||_\infty, \tag{B.8}$$

where $f_i^* \sum_{r=1}^{\infty} f_\varepsilon \left(\lambda \mathbf{g}_i \mathbf{u} + \psi_i - a_r^i \right)$.

It follows that $\max_i f_i^* \leq \max_{u \in \mathbb{R}} \sum_{r=1}^{\infty} f_\varepsilon \left(u - a_r^i \right) = C_\gamma^{-1}$. Hence, $\left|\left| \frac{\partial \mathbf{L}(\mathbf{u})}{\partial \mathbf{u}'} \right|\right|_\infty < \frac{|\lambda| ||\mathbf{G}||_\infty}{C_\gamma} < 1$ by Assumption D'.

Importantly, $\max_{u \in \mathbb{R}} \sum_{r=1}^{\infty} f_\varepsilon \left(u - a_r^i \right) < \infty$ as long as $\lim_{q \to \infty} a_{q+1}^i - a_q^i > 0$ because f_ε is continuous and $o(1/x^\alpha)$ at ∞ for some $\alpha > 3$. If $\lim_{q \to \infty} a_{q+1}^i - a_q^i > 0$ does not hold and $\max_{u \in \mathbb{R}} \sum_{r=1}^{\infty} f_\varepsilon \left(u - a_r^i \right) = \infty$, then Assumption D' would imply that $|\lambda| < 0$ and the BNE would not be unique for any λ. For instance, this is the case when $a_0 = -\infty$, $a_q = \sqrt{\log(q)}\ \forall\ q \in \mathbb{N}^*$, and $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$.

B.1.4 Upper bound of the peer effects under Assumptions B' and E

Assume that $\varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$, $a_1 = 0$, and $\gamma = 1$. Let us compute the upper bound of $|\lambda|$, $\frac{C_\gamma}{||\mathbf{G}||_\infty}$, where $C_\gamma = \frac{1}{\max_{u \in \mathbb{R}} \sum_{k=-\infty}^{\infty} f_\varepsilon (u + k)}$.
By the Poisson summation formula (see Bellman, 2013, Section 6)

$$\sum_{k=-\infty}^{\infty} f_\varepsilon (u + k) = \sum_{k=-\infty}^{\infty} \hat{f}_\varepsilon (u + k), \tag{B.9}$$

where $\hat{f}_\varepsilon$ is the Fourier transform of f_ε given by

$$\hat{f}_\varepsilon (u + k) = \int_{-\infty}^{\infty} f_\varepsilon(x + u) e^{-2\pi i k x} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} e^{-\frac{1}{2\sigma_\varepsilon^2}(x+u)^2 - 2\pi i k x} dx. \tag{B.10}$$

In Equation (B.10), i is the pure imaginary complex number ($i^2 = -1$).

$$\begin{aligned}
\hat{f}_\varepsilon (u + k) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} e^{-\frac{1}{2\sigma_\varepsilon^2}\left(x^2 + 2ux + u^2 + 4\pi i \sigma_\varepsilon^2 k x\right)} dx, \\
\hat{f}_\varepsilon (u + k) &= e^{\frac{1}{2\sigma_\varepsilon^2}\left(u + 2\pi i \sigma_\varepsilon^2 k\right)^2 - \frac{1}{2\sigma_\varepsilon^2} u^2} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} e^{-\frac{1}{2\sigma_\varepsilon^2}\left(x + u + 2\pi i \sigma_\varepsilon^2 k\right)^2} dx}_{=1}, \\
\hat{f}_\varepsilon (u + k) &= e^{-2\pi^2 k^2 \sigma_\varepsilon^2 + 2\pi i k u}.
\end{aligned} \tag{B.11}$$

By replacing the Fourier transform (B.11) in Equation (B.9),

$$\begin{aligned}
\sum_{k=-\infty}^{\infty} f_{\varepsilon}(u+k) &= \sum_{k=-\infty}^{\infty} e^{-2\pi^2 k^2 \sigma_{\varepsilon}^2 + 2\pi i k u}, \\
\sum_{k=-\infty}^{\infty} f_{\varepsilon}(u+k) &= 1 + \sum_{k=1}^{\infty} e^{-2\pi^2 (-k)^2 (\sigma_{\varepsilon})^2} e^{-2\pi i k u} + \sum_{k=1}^{\infty} e^{-2\pi^2 k^2 \sigma_{\varepsilon}^2} e^{2\pi i k u}, \\
\sum_{k=-\infty}^{\infty} f_{\varepsilon}(u+k) &= 1 + \sum_{k=1}^{\infty} e^{-2\pi^2 k^2 \sigma_{\varepsilon}^2} \left(e^{-2\pi i k u} + e^{2\pi i k u} \right). \qquad \text{(B.12)}
\end{aligned}$$

By Euler's formula,

$$\begin{aligned}
e^{-2\pi i k u} + e^{2\pi i k u} &= \cos(-2\pi k u) + i \sin(-2\pi k u) + \cos(2\pi k u) + i \sin(2\pi k u), \\
e^{-2\pi i k u} + e^{2\pi i k u} &= 2\cos(2\pi k u). \qquad \text{(B.13)}
\end{aligned}$$

By replacing (B.13) in (B.12),

$$\sum_{k=-\infty}^{\infty} f_{\varepsilon}(u+k) = 1 + 2\sum_{k=1}^{\infty} e^{-2\pi^2 k^2 \sigma_{\varepsilon}^2} \cos(2\pi k u).$$

Therefore,

$$\max_{u \in \mathbb{R}} \sum_{k=-\infty}^{\infty} f_{\varepsilon}(u+k) = 1 + 2\sum_{k=1}^{\infty} e^{-2\pi^2 k^2 \sigma_{\varepsilon}^2} = \sum_{k=-\infty}^{\infty} f_{\varepsilon}(k), \qquad \text{(B.14)}$$

$$\max_{u \in \mathbb{R}} \sum_{k=-\infty}^{\infty} f_{\varepsilon}(u+k) = \frac{\phi(0) + 2\sum_{k=1}^{\infty} \phi\left(\frac{k}{\sigma_{\varepsilon}}\right)}{\sigma_{\varepsilon}}. \qquad \text{(B.15)}$$

The quantity $\sum_{k=-\infty}^{\infty} f_{\varepsilon}(k)$ can also be computed using the third Theta function (see Bellman, 2013, Section 2). From (B.14), it follows that

$$\sum_{k=-\infty}^{\infty} f_{\varepsilon}(k) = \theta_3\left(0, e^{-2\pi^2 \sigma_{\varepsilon}^2}\right),$$

where for any complex z and $q \in \mathbb{R}_+$, $\theta_3(z, q)$ is the third Theta function evaluated at (z, q). As a result,

$$C_{\gamma} = C_{1,\sigma_{\varepsilon}} = \frac{\sigma_{\varepsilon}}{\phi(0) + 2\sum_{k=1}^{\infty} \phi\left(\frac{k}{\sigma_{\varepsilon}}\right)} = \frac{1}{\theta_3\left(0, e^{-2\pi^2 \sigma_{\varepsilon}^2}\right)}. \qquad \text{(B.16)}$$

B.2 Supplementary note on the econometric model

B.2.1 Proof of Proposition 2.2

The pseudo likelihood is given by

$$\mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}}) = \sum_{i=1}^{n} \sum_{r=0}^{\infty} d_{ir} \log\left\{ \Phi\left(\frac{\mathbf{z}_i' \boldsymbol{\Lambda} - a_r}{\sigma_{\varepsilon}} \right) - \Phi\left(\frac{\mathbf{z}_i' \boldsymbol{\Lambda} - a_{r+1}}{\sigma_{\varepsilon}} \right) \right\}, \qquad \text{(B.17)}$$

where $\mathbf{z}_i' = (\mathbf{g}_i\bar{\mathbf{y}},\ \mathbf{x}_i')$, $\mathbf{\Lambda} = (\lambda,\ \boldsymbol{\beta}')'$, and $\boldsymbol{\theta} = (\mathbf{\Lambda},\ \sigma_\varepsilon)'$. Let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$, and $\bar{\mathbf{y}}_0$ be the expected outcome associated with $\boldsymbol{\theta}$. The first-order conditions of the pseudo likelihood maximization give

$$\begin{cases} \dfrac{\partial \mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}})}{\partial \mathbf{\Lambda}} = \displaystyle\sum_{i=1}^{n}\sum_{r=0}^{\infty} d_{ir}\frac{f_{ir} - f_{i(r+1)}}{F_{ir} - F_{i(r+1)}}\mathbf{z}_i = 0, \\ \dfrac{\partial \mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}})}{\partial \sigma_\varepsilon} = -\displaystyle\sum_{i=1}^{n}\sum_{r=0}^{\infty} d_{ir}\frac{m_{ir}f_{ir} - m_{i(r+1)}f_{i(r+1)}}{\sigma_\varepsilon\left(F_{ir} - F_{i(r+1)}\right)} = 0, \end{cases} \tag{B.18}$$

where $\forall\ i \in \mathcal{V},\ q \in \mathbb{N},\ m_{iq} = \mathbf{z}_i'\mathbf{\Lambda} - a_q,\ f_{iq} = \dfrac{1}{\sigma_\varepsilon}\phi\left(\dfrac{m_{iq}}{\sigma_\varepsilon}\right)$, and $F_{iq} = \Phi\left(\dfrac{m_{iq}}{\sigma_\varepsilon}\right)$. As $\mathcal{L}$ is continuous, the consistency of the NPL estimator is ensured by the fact that plim $\left(\dfrac{1}{n}\mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}})\right)$ is maximized at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\bar{\mathbf{y}} = \bar{\mathbf{y}}_0$, where plim stands for the probability limit.
Let us focus on the limiting distribution. The Taylor expansion of $\dfrac{\partial \mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}})}{\partial \boldsymbol{\theta}}$ around $\boldsymbol{\theta}_0$ gives

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}})}{\partial \boldsymbol{\theta}} = \frac{\partial \mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}_0} + \left(\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}_0, \bar{\mathbf{y}})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_0} + \frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}_0, \bar{\mathbf{y}})}{\partial \boldsymbol{\theta}\partial \bar{\mathbf{y}}'}\bigg|_{\boldsymbol{\theta}_0}\frac{\partial \bar{\mathbf{y}}}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_0}\right)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + O_p(1).$$

To simplify the notations of the partial derivatives, I will use $\dfrac{\partial \mathcal{L}(\boldsymbol{\theta}_0, \bar{\mathbf{y}})}{\partial \boldsymbol{\theta}}$ to mean $\dfrac{\partial \mathcal{L}(\boldsymbol{\theta}, \bar{\mathbf{y}})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}_0}$ (this notation is also applied to the second partial derivatives) and $\dfrac{\partial \bar{\mathbf{y}}_0}{\partial \boldsymbol{\theta}'}$ to mean $\dfrac{\partial \bar{\mathbf{y}}}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}_0}$. It follows that

$$\sqrt{n}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = -\left(\frac{1}{n}\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}_0, \bar{\mathbf{y}})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} + \frac{1}{n}\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta}_0, \bar{\mathbf{y}})}{\partial \boldsymbol{\theta}\partial \bar{\mathbf{y}}'}\frac{\partial \bar{\mathbf{y}}_0}{\partial \boldsymbol{\theta}'}\right)^{-1}\left(\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}(\boldsymbol{\theta}_0, \bar{\mathbf{y}})}{\partial \boldsymbol{\theta}} + O_p\left(\frac{1}{\sqrt{n}}\right)\right). \tag{B.19}$$

Let us first apply the central Theorem limit to the term $\dfrac{1}{\sqrt{n}}\dfrac{\partial \mathcal{L}(\boldsymbol{\theta}_0, \bar{\mathbf{y}})}{\partial \boldsymbol{\theta}}$.

$$\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}(\boldsymbol{\theta}_0, \bar{\mathbf{y}})}{\partial \boldsymbol{\theta}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\underbrace{\begin{pmatrix} \displaystyle\sum_{r=0}^{\infty} d_{ir}\frac{f_{ir}^0 - f_{i(r+1)}^0}{F_{ir}^0 - F_{i(r+1)}^0}\mathbf{z}_i \\ -\displaystyle\sum_{r=0}^{\infty} d_{ir}\frac{m_{ir}^0 f_{ir}^0 - m_{i(r+1)}^0 f_{i(r+1)}^0}{\sigma_\varepsilon\left(F_{ir}^0 - F_{i(r+1)}^0\right)} \end{pmatrix}}_{\mathbf{v}_i^0} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{v}_i^0,$$

where $\forall\ i \in \mathcal{V},\ q \in \mathbb{N},\ m_{iq}^0$, f_{iq}^0, and F_{iq}^0 are defined as in (B.18) but with $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

$$\mathbf{E}\left(\mathbf{v}_i^0|\mathbf{X}, \mathbf{G}\right) = \begin{pmatrix} \displaystyle\sum_{r=0}^{\infty}\left(f_{ir}^0 - f_{i(r+1)}^0\right)\mathbf{z}_i \\ -\dfrac{1}{\sigma_\varepsilon}\displaystyle\sum_{r=0}^{\infty}\left(m_{ir}^0 f_{ir}^0 - m_{i(r+1)}^0 f_{i(r+1)}^0\right) \end{pmatrix} = 0, \text{ thus } \mathbf{E}\left(\mathbf{v}_i^0\right) = 0.$$

Let denote by $A_i = \displaystyle\sum_{r=0}^{\infty}\frac{\left(f_{ir}^0 - f_{i(r+1)}^0\right)^2}{F_{ir}^0 - F_{i(r+1)}^0}$, $B_i = \displaystyle\sum_{r=0}^{\infty}\frac{\left(m_{ir}^0 f_{ir}^0 - m_{i(r+1)}^0 f_{i(r+1)}^0\right)^2}{\sigma_\varepsilon^2\left(F_{ir}^0 - F_{i(r+1)}^0\right)}$, and

$$C_i = -\sum_{r=0}^{\infty} \frac{\left(f_{ir}^0 - f_{i(r+1)}^0\right)\left(m_{ir}^0 f_{ir}^0 - m_{i(r+1)}^0 f_{i(r+1)}^0\right)}{\sigma_\varepsilon\left(F_{ir}^0 - F_{i(r+1)}^0\right)}.$$

$$\mathbf{Var}\left(\mathbf{v}_i^0|\mathbf{X},\mathbf{G}\right) = \mathbf{E}\left(\mathbf{v}_i^0\mathbf{v}_i^{0\prime}|\mathbf{X},\mathbf{G}\right) = \underbrace{\begin{pmatrix} A_i\mathbf{z}_i\mathbf{z}_i' & C_i\mathbf{z}_i \\ C_i\mathbf{z}_i' & B_i \end{pmatrix}}_{\boldsymbol{\Sigma}_i} = \boldsymbol{\Sigma}_i. \tag{B.20}$$

By the law of large numbers (LLN) applied to independent and non-identical variables (see Chow and Teicher, 2003, p. 124), assume that $\operatorname{plim}\left(\frac{1}{n}\sum_i^n \boldsymbol{\Sigma}_i\right)$ exists and is equal to $\boldsymbol{\Sigma}_0$. It follows by the Lindeberg–Feller central Theorem limit (see Chow and Teicher, 2003, p. 314) that,

$$\frac{1}{\sqrt{n}}\frac{\partial\mathcal{L}(\boldsymbol{\theta}_0,\bar{\mathbf{y}})}{\partial\boldsymbol{\theta}} \xrightarrow{d} \mathcal{N}\left(0,\boldsymbol{\Sigma}_0\right). \tag{B.21}$$

Let us now focus on $\operatorname{plim}\left(\frac{1}{n}\frac{\partial^2\mathcal{L}(\boldsymbol{\theta}_0,\bar{\mathbf{y}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)$ and $\operatorname{plim}\left(\frac{1}{n}\frac{\partial^2\mathcal{L}(\boldsymbol{\theta}_0,\bar{\mathbf{y}})}{\partial\boldsymbol{\theta}\partial\bar{\mathbf{y}}'}\frac{\partial\bar{\mathbf{y}}_0}{\partial\boldsymbol{\theta}'}\right)$.
By the LLN, $\operatorname{plim}\left(\frac{1}{n}\frac{\partial^2\mathcal{L}(\boldsymbol{\theta}_0,\bar{\mathbf{y}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right) = \operatorname{plim}\left(\frac{1}{n}\mathbf{E}_d\left(\frac{\partial^2\mathcal{L}(\boldsymbol{\theta}_0,\bar{\mathbf{y}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)\right)$, where $\mathbf{E}_d$ is the expectation with respect to d_{ir}.

$$\mathbf{E}_d\left(\frac{\partial^2\mathcal{L}(\boldsymbol{\theta}_0,\bar{\mathbf{y}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right) = -\sum_{i=1}^n \boldsymbol{\Sigma}_i \implies \operatorname{plim}\left(\frac{1}{n}\frac{\partial^2\mathcal{L}(\boldsymbol{\theta}_0,\bar{\mathbf{y}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right) = -\operatorname{plim}\left(\frac{1}{n}\sum_i^n \boldsymbol{\Sigma}_i\right) = -\boldsymbol{\Sigma}_0. \tag{B.22}$$

Analogously, $\operatorname{plim}\left(\frac{1}{n}\frac{\partial^2\mathcal{L}(\boldsymbol{\theta}_0,\bar{\mathbf{y}})}{\partial\boldsymbol{\theta}\partial\bar{\mathbf{y}}'}\frac{\partial\bar{\mathbf{y}}_0}{\partial\boldsymbol{\theta}'}\right) = \operatorname{plim}\left(\frac{1}{n}\mathbf{E}_d\left(\frac{\partial^2\mathcal{L}(\boldsymbol{\theta}_0,\bar{\mathbf{y}})}{\partial\boldsymbol{\theta}\partial\bar{\mathbf{y}}'}\right)\frac{\partial\bar{\mathbf{y}}_0}{\partial\boldsymbol{\theta}'}\right)$.

$$\mathbf{E}_d\left(\frac{\partial^2\mathcal{L}(\boldsymbol{\theta}_0,\bar{\mathbf{y}})}{\partial\boldsymbol{\theta}\partial\bar{\mathbf{y}}'}\right) = -\lambda\sum_{i=1}^n \begin{pmatrix} A_i\mathbf{z}_i\mathbf{g}_i \\ B_i\mathbf{g}_i \end{pmatrix} \quad \text{and} \quad \frac{\partial\bar{\mathbf{y}}_0}{\partial\boldsymbol{\theta}'} = \mathbf{S}^{-1}\mathbf{M}, \tag{B.23}$$

where $\mathbf{S} = \mathbf{I}_n - \lambda\mathbf{D}\mathbf{G}$, $\mathbf{I}_n$ is the identity matrix of dimension n,
$\mathbf{D} = \operatorname{diag}\left(\sum_{r=1}^\infty f_{1r}^0,\ \ldots,\ \sum_{r=1}^\infty f_{nr}^0\right)$, $\mathbf{M} = (\mathbf{D}\mathbf{Z},\mathbf{b})$, $\mathbf{Z} = (\mathbf{G}\bar{\mathbf{y}},\mathbf{X})$, and
$\mathbf{b} = \left(-\sum_{r=1}^\infty \frac{f_{1r}^0 m_{1r}^0}{\sigma_\varepsilon},\ \ldots,\ -\sum_{r=1}^\infty \frac{f_{nr}^0 m_{nr}^0}{\sigma_\varepsilon}\right)'$.
The partial derivative $\frac{\partial\bar{\mathbf{y}}_0}{\partial\boldsymbol{\theta}'}$ is computed using the implicit definition of $\bar{\mathbf{y}}$; that is, $\bar{\mathbf{y}} = \mathbf{L}(\bar{\mathbf{y}},\boldsymbol{\theta})$.
Assuming that $\operatorname{plim}\left(\frac{\lambda}{n}\sum_{i=1}^n \begin{pmatrix} A_i\mathbf{z}_i\mathbf{g}_i\mathbf{S}^{-1}\mathbf{M} \\ B_i\mathbf{g}_i\mathbf{S}^{-1}\mathbf{M}\end{pmatrix}\right)$ exists and is equal to $\boldsymbol{\Omega}_0$,

$$\operatorname{plim}\left(\frac{1}{n}\frac{\partial^2\mathcal{L}(\boldsymbol{\theta}_0,\bar{\mathbf{y}})}{\partial\boldsymbol{\theta}\partial\bar{\mathbf{y}}'}\frac{\partial\bar{\mathbf{y}}_0}{\partial\boldsymbol{\theta}'}\right) = -\boldsymbol{\Omega}_0. \tag{B.24}$$

From Equations (B.19), (B.21), (B.22), and (B.24), it follows that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(0,\ \left(\boldsymbol{\Sigma}_0 + \boldsymbol{\Omega}_0\right)^{-1}\boldsymbol{\Sigma}_0\left(\boldsymbol{\Sigma}_0' + \boldsymbol{\Omega}_0'\right)^{-1}\right). \tag{B.25}$$

In a finite sample, an estimator of the asymptotic variance of $\hat{\boldsymbol{\theta}}$ can be computed by

$$\widehat{AsyVar}\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{n}\left(\hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\Omega}}\right)^{-1} \hat{\boldsymbol{\Sigma}} \left(\hat{\boldsymbol{\Sigma}}' + \hat{\boldsymbol{\Omega}}'\right)^{-1}, \tag{B.26}$$

where $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i}^{n} \hat{\boldsymbol{\Sigma}}_i$, $\hat{\boldsymbol{\Omega}} = \frac{\hat{\lambda}}{n}\sum_{i=1}^{n} \begin{pmatrix} \hat{A}_i \mathbf{z}_i \mathbf{g}_i \hat{\mathbf{S}}^{-1}\hat{\mathbf{M}} \\ \hat{B}_i \mathbf{g}_i \hat{\mathbf{S}}^{-1}\hat{\mathbf{M}} \end{pmatrix}$, and $\hat{\boldsymbol{\Sigma}}_i$, $\hat{A}_i$, $\hat{B}_i$, $\hat{\mathbf{S}}$, $\hat{\mathbf{M}}$ are the estimates of $\boldsymbol{\Sigma}_i$, A_i, B_i, $\mathbf{S}$, $\mathbf{M}$, respectively by replacing $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}$.

B.2.2 Proof of Proposition 2.3

The likelihood of the linear-in-means model is

$$Q(\lambda, \boldsymbol{\beta}, \sigma_\nu) = \frac{n}{2}\log\left(2\pi\sigma_\nu^2\right) + \log|\mathbf{I}_n - \lambda\mathbf{G}| - \frac{(\mathbf{y} - \lambda\mathbf{G}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \lambda\mathbf{G}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma_\nu^2}.$$

Let $Q_0(\lambda, \boldsymbol{\beta}, \sigma_\nu) = \text{plim}\,\frac{1}{n}Q(\lambda, \boldsymbol{\beta}, \sigma_\nu)$. Assume that all the conditions of the MLE consistency set in Lee (2004) hold.
Let $\mathbf{B}(\lambda) = \mathbf{I}_n - \lambda\mathbf{G}$. It follows that

$$Q_0(\lambda, \boldsymbol{\beta}, \sigma_\nu) = \frac{\log\left(2\pi\sigma_\nu^2\right)}{2} + \text{plim}\left(\frac{\log|\mathbf{B}(\lambda)|}{n} - \frac{(\mathbf{B}(\lambda)\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{B}(\lambda)\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2n\sigma_\nu^2}\right).$$

By the LLN,

$$Q_0(\lambda, \boldsymbol{\beta}, \sigma_\nu) = \frac{\log\left(2\pi\sigma_\nu^2\right)}{2} + \text{plim}\left(\frac{\log|\mathbf{B}(\lambda)|}{n} - \frac{\mathbf{E}\left\{(\mathbf{B}(\lambda)\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{B}(\lambda)\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\middle|\,\mathbf{X}, \mathbf{G}\right\}}{2n\sigma_\nu^2}\right).$$

The first-order conditions (f.o.cs) with respect to λ of the maximization of $Q_0(\lambda, \boldsymbol{\beta}, \sigma_\nu)$ implies that,

$$\begin{aligned} &\frac{\partial Q_0(\lambda, \boldsymbol{\beta}, \sigma_\nu)}{\partial\lambda} = 0, \\ &\text{plim}\,\frac{\mathbf{E}\left\{(\mathbf{A}(\lambda)\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\nu}|\mathbf{X}, \mathbf{G}\right\}}{n\sigma_\nu^2} - \text{plim}\,\frac{\text{Tr}\left(\mathbf{A}(\lambda)\right)}{n} + \text{plim}\,\frac{\mathbf{E}\left\{\boldsymbol{\nu}'\mathbf{A}(\lambda)\boldsymbol{\nu}|\,\mathbf{X}, \mathbf{G}\right\}}{n\sigma_\nu^2} = 0, \end{aligned} \tag{B.27}$$

where $\mathbf{A}(\lambda) = \mathbf{G}(\mathbf{B}(\lambda))^{-1}$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)'$.
In addition, $\mathbf{E}(\boldsymbol{\nu}'\mathbf{A}(\lambda)\boldsymbol{\nu}|\mathbf{X}, \mathbf{G}) = \mathbf{E}(\text{Tr}\left(\boldsymbol{\nu}'\mathbf{A}(\lambda)\boldsymbol{\nu}|\mathbf{X}, \mathbf{G}\right) = \text{Tr}\left(\mathbf{A}(\lambda)\,\mathbf{E}(\boldsymbol{\nu}\boldsymbol{\nu}'|\mathbf{X}, \mathbf{G})\right)$.
One can express ν_i as function of ε_i.
From (2.14),

$$\begin{aligned} y_i + \underbrace{(y_i^* - y_i)}_{\zeta_i} &= \lambda\mathbf{g}_i(\mathbf{y} + \underbrace{(\bar{\mathbf{y}} - \mathbf{y})}_{\boldsymbol{\eta}}) + \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \\ y_i &= \lambda\mathbf{g}_i\mathbf{y} + \mathbf{x}_i'\boldsymbol{\beta} + \underbrace{\varepsilon_i + \lambda\mathbf{g}_i\boldsymbol{\eta} - \zeta_i}_{\nu_i}. \end{aligned}$$

Hence,

$$\nu_i = \varepsilon_i + \lambda\mathbf{g}_i\boldsymbol{\eta} - \zeta_i,$$

where $\boldsymbol{\eta} = \bar{\mathbf{y}} - \mathbf{y}$ and $\zeta_i = y_i^* - y_i$.
Let us consider the case where $\sigma_\varepsilon > 1$ and y_i takes values as large as possible. In this case, ζ_i is *approximately* distributed according to a uniform distribution over $[0, 1]$. In Equation (2.15), it is necessary to have $\mathbf{E}(\nu_i|\mathbf{X}, \mathbf{G}) = 0$. However, this condition is not verified. Nevertheless, without loss of generality, I can still assume that $\mathbf{E}(\zeta_i|\mathbf{X}, \mathbf{G}) = 0$ because the model includes an intercept and $\mathbf{E}(\zeta_i|\mathbf{X}, \mathbf{G})$ is a constant. Moreover, $\mathbf{E}(\boldsymbol{\eta}|\mathbf{X}, \mathbf{G}) = 0$.
Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$ and $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)'$. Then,

$$\begin{aligned}
\boldsymbol{\nu}\boldsymbol{\nu}' &= (\boldsymbol{\varepsilon} + \lambda \mathbf{G}\boldsymbol{\eta} - \boldsymbol{\zeta})(\boldsymbol{\varepsilon} + \lambda \mathbf{G}\boldsymbol{\eta} - \boldsymbol{\zeta})', \\
\boldsymbol{\nu}\boldsymbol{\nu}' &= \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' + \lambda^2 \mathbf{G}\boldsymbol{\eta}\boldsymbol{\eta}'\mathbf{G}' + \boldsymbol{\zeta}\boldsymbol{\zeta}' + \boldsymbol{\varepsilon}(\lambda \mathbf{G}\boldsymbol{\eta} - \boldsymbol{\zeta})' + \lambda \mathbf{G}\boldsymbol{\eta}(\boldsymbol{\varepsilon} - \boldsymbol{\zeta})' - \boldsymbol{\zeta}(\boldsymbol{\varepsilon} + \lambda \mathbf{G}\boldsymbol{\eta})'.
\end{aligned}$$

Therefore,

$$\mathbf{E}(\boldsymbol{\nu}\boldsymbol{\nu}'|\mathbf{X}, \mathbf{G}) = \left(\sigma_\varepsilon^2 + \frac{1}{12}\right) \mathbf{I}_n + \lambda^2 \mathbf{G}\, \mathbf{E}\left(\boldsymbol{\eta}\boldsymbol{\eta}'|\mathbf{X}, \mathbf{G}\right) \mathbf{G}'.$$

Given that, $\mathbf{E}(\zeta_i|\mathbf{X}, \mathbf{G}) = 0$, $\mathbf{E}(\varepsilon_i|\mathbf{X}, \mathbf{G}) = 0$, $\mathbf{E}(\boldsymbol{\eta}|\mathbf{X}, \mathbf{G}) = 0$,
and $\operatorname{plim} \frac{1}{n\sigma_\nu^2}(\mathbf{A}(\lambda)\mathbf{X}\boldsymbol{\beta})' \mathbf{E}(\boldsymbol{\nu}|\mathbf{X}, \mathbf{G}) = 0$, Equation (B.27) implies that

$$\begin{aligned}
\frac{\partial Q_0(\lambda, \boldsymbol{\beta}, \sigma_\nu)}{\partial \lambda} &= -\operatorname{plim} \frac{\operatorname{Tr}(\mathbf{A}(\lambda))}{n} + \operatorname{plim} \frac{1}{n\sigma_\nu^2} \operatorname{Tr}\left(\mathbf{A}(\lambda)\, \mathbf{E}(\boldsymbol{\nu}\boldsymbol{\nu}'|\mathbf{X}, \mathbf{G})\right), \\
\frac{\partial Q_0(\lambda, \boldsymbol{\beta}, \sigma_\nu)}{\partial \lambda} &= \frac{12(\sigma_\varepsilon^2 - \sigma_\nu^2) + 1}{12\sigma_\nu^2} \operatorname{plim} \frac{\operatorname{Tr}(\mathbf{A}(\lambda))}{n} + \frac{\lambda^2}{\sigma_\nu^2} \operatorname{plim} \frac{\operatorname{Tr}\left(\mathbf{A}(\lambda)\mathbf{G}\, \mathbf{E}\left(\boldsymbol{\eta}\boldsymbol{\eta}'|\mathbf{X}, \mathbf{G}\right) \mathbf{G}'\right)}{n}.
\end{aligned} \tag{B.28}$$

Equation (B.28) shows that $\frac{\partial Q_0(\lambda, \boldsymbol{\beta}, \sigma_\nu)}{\partial \lambda} \neq 0$ in general if $\boldsymbol{\eta} \neq 0$. Therefore, the MLE of $\left(\tilde{\lambda}, \tilde{\boldsymbol{\beta}}', \sigma_\nu^2\right)'$ is generally biased. Moreover, since the estimator of $\tilde{\boldsymbol{\beta}}$ and σ_ν^2 are given by $\hat{\tilde{\boldsymbol{\beta}}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{B}(\lambda)\mathbf{y}$ and $\hat{\sigma}_\nu^2 = \frac{1}{n}\left(\mathbf{B}(\lambda)\mathbf{y} - \mathbf{X}\hat{\tilde{\boldsymbol{\beta}}}\right)' \left(\mathbf{B}(\lambda)\mathbf{y} - \mathbf{X}\hat{\tilde{\boldsymbol{\beta}}}\right)$, respectively, this means that the estimator of $\tilde{\lambda}$ is necessarily biased. Indeed, if $\hat{\tilde{\lambda}}$ were consistent, then $\hat{\tilde{\boldsymbol{\beta}}}$ and $\hat{\sigma}_\nu^2$ would also be consistent. This is in contradiction with $\frac{\partial Q_0(\lambda, \boldsymbol{\beta}, \sigma_\nu)}{\partial \lambda} \neq 0$.
Note that the MLE is consistent if $\boldsymbol{\eta} = 0$. Indeed, in this case, $\nu_i = \varepsilon_i - \zeta_i$ and $\sigma_\nu^2 = \sigma_\varepsilon^2 + \frac{1}{12}$.
Hence, $\frac{\partial Q_0(\lambda, \boldsymbol{\beta}, \sigma_\nu)}{\partial \lambda} = 0$.

B.2.3 Proof of Proposition 2.4

The 2SLS estimator of $\tilde{\lambda}$ is

$$\hat{\tilde{\lambda}}_{2SLS} = \frac{\frac{1}{n}\sum_{i=1}^n \tilde{y}_i(\mathbf{g}_i\tilde{\mathbf{y}}) - \hat{\tilde{y}}(\hat{\mathbf{g}}\tilde{\mathbf{y}})}{\frac{1}{n}\sum_{i=1}^n (\mathbf{g}_i\tilde{\mathbf{y}})^2 - (\hat{\mathbf{g}}\tilde{\mathbf{y}})^2},$$

where $\tilde{y_i} = \mathbf{P_{Z}}_i\mathbf{y}$, $\mathbf{g}\tilde{_i\mathbf{y}} = \mathbf{P_{Z}}_i\mathbf{G}\mathbf{y}$, $\mathbf{P_Z} = \mathbf{Z}\left(\mathbf{Z}\mathbf{Z}\right)^{-1}\mathbf{Z}'$, $\mathbf{P_{Z}}_i$ is the i-th row of $\mathbf{P_Z}$, $\hat{\tilde{y}} = \frac{1}{n}\sum_{i=1}^{n}\tilde{y_i}$, and $\hat{\mathbf{g}\tilde{\mathbf{y}}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}\tilde{_i\mathbf{y}}$. It follows that

$$\hat{\tilde{\lambda}}_{2SLS} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\lambda(\mathbf{g}\tilde{_i\mathbf{y}}) + \lambda(\mathbf{g}\tilde{_i\boldsymbol{\eta}}) + \tilde{\varepsilon}_i + \tilde{\zeta}_i\right)(\mathbf{g}\tilde{_i\mathbf{y}}) - \left(\lambda(\hat{\mathbf{g}\tilde{\mathbf{y}}}) + \lambda(\hat{\mathbf{g}\tilde{\boldsymbol{\eta}}}) + \hat{\tilde{\varepsilon}} + \hat{\tilde{\zeta}}\right)(\hat{\mathbf{g}\tilde{\mathbf{y}}})}{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{g}\tilde{_i\mathbf{y}})^2 - (\hat{\mathbf{g}\tilde{\mathbf{y}}})^2},$$

where $\mathbf{g}\tilde{_i\boldsymbol{\eta}} = \mathbf{P_{Z}}_i\mathbf{g}_i\boldsymbol{\eta}$, $\tilde{\varepsilon}_i = \mathbf{P_{Z}}_i\boldsymbol{\varepsilon}$, $\tilde{\zeta}_i = \mathbf{P_{Z}}_i\boldsymbol{\zeta}$, $\hat{\mathbf{g}\tilde{\boldsymbol{\eta}}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}\tilde{_i\boldsymbol{\eta}}$, $\hat{\tilde{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\varepsilon}_i$, and $\hat{\tilde{\zeta}} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\zeta}_i$. By the LLN $\operatorname{plim}\hat{\tilde{\varepsilon}} = 0$, $\operatorname{plim}\hat{\mathbf{g}\tilde{\boldsymbol{\eta}}} = 0$ and $\operatorname{plim}\hat{\tilde{\zeta}} = 0$ (by assumption if $\sigma_\varepsilon > 1$ and y_i takes values as large as possible). Moreover, as $\mathbf{Z}$ is a valid instrument, $\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n}\tilde{\varepsilon}_i(\mathbf{g}\tilde{_i\mathbf{y}}) = 0$ and $\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n}\tilde{\zeta}_i(\mathbf{g}\tilde{_i\mathbf{y}}) = 0$ (by assumption if $\sigma_\varepsilon > 1$ and y_i takes values as large as possible). Then,

$$\begin{aligned}
\operatorname{plim}\hat{\tilde{\lambda}}_{2SLS} &= \lambda\frac{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{g}\tilde{_i\mathbf{y}})^2 + \frac{1}{n}\sum_{i=1}^{n}(\mathbf{g}\tilde{_i\boldsymbol{\eta}})(\mathbf{g}\tilde{_i\mathbf{y}}) - (\hat{\mathbf{g}\tilde{\mathbf{y}}})^2}{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{g}\tilde{_i\mathbf{y}})^2 - (\hat{\mathbf{g}\tilde{\mathbf{y}}})^2},\\
\operatorname{plim}\hat{\tilde{\lambda}}_{2SLS} &= \lambda + \lambda\operatorname{plim}\frac{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{g}\tilde{_i\boldsymbol{\eta}})(\mathbf{g}\tilde{_i\mathbf{y}})}{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{g}\tilde{_i\mathbf{y}})^2 - (\hat{\mathbf{g}\tilde{\mathbf{y}}})^2},\\
\operatorname{plim}\hat{\tilde{\lambda}}_{2SLS} &= \lambda + \lambda\operatorname{plim}\frac{\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left((\mathbf{g}\tilde{_i\boldsymbol{\eta}})(\mathbf{g}\tilde{_i\mathbf{y}})|\mathbf{X},\mathbf{G},\mathbf{Z}\right)}{\frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}(\mathbf{g}\tilde{_i\mathbf{y}})},\\
\operatorname{plim}\hat{\tilde{\lambda}}_{2SLS} &= \lambda - \lambda\operatorname{plim}\frac{\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left\{\left((\mathbf{g}\tilde{_i\mathbf{y}} - \mathbf{E}(\mathbf{g}\tilde{_i\mathbf{y}}|\mathbf{X},\mathbf{G},\mathbf{Z}))\right)(\mathbf{g}\tilde{_i\mathbf{y}})|\mathbf{X},\mathbf{G},\mathbf{Z}\right\}}{\frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}(\mathbf{g}\tilde{_i\mathbf{y}})},\\
\operatorname{plim}\hat{\tilde{\lambda}}_{2SLS} &= \lambda - \lambda\operatorname{plim}\frac{\frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}(\mathbf{g}\tilde{_i\mathbf{y}}|\mathbf{X},\mathbf{G},\mathbf{Z})}{\frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}(\mathbf{g}\tilde{_i\mathbf{y}})}.
\end{aligned} \tag{B.29}$$

B.2.4 Marginal effects and corresponding standard errors

The parameters $\boldsymbol{\theta}$ cannot be interpreted directly. Policy makers may be interested in the marginal effect of the explanatory variables on the expected outcome.
Let us recall the following notations: $\mathbf{z}_i' = (\mathbf{g}_i\bar{\mathbf{y}},\ \mathbf{x}_i')$ and $\boldsymbol{\Lambda} = (\lambda,\ \boldsymbol{\beta}')'$. For any $k = 1,\ldots,K+1$, let λ_k and z_{ik} be the k-th component in $\boldsymbol{\Lambda}$ and $\mathbf{z}_i$, respectively. The marginal effect of the

explanatory variable z_{ik} on $\bar{y}_i$, the expected outcome of the individual i is given by

$$\delta_{ik}(\boldsymbol{\theta}) = \frac{\partial \bar{y}_i}{\partial z_{ik}} = \frac{\lambda_k}{\sigma_\varepsilon} \sum_{r=1}^{\infty} \phi\left(\frac{\mathbf{z}_i'\boldsymbol{\Lambda} - a_r}{\sigma_\varepsilon}\right). \tag{B.30}$$

The standard error of $\delta_{ik}(\boldsymbol{\theta})$ can be computed using the Delta method.
The Taylor expansion of Equation (B.30) around $\boldsymbol{\theta}_0$ is

$$\delta_{ik}(\hat{\boldsymbol{\theta}}) = \delta_{ik}(\boldsymbol{\theta}_0) + \frac{\partial \delta_{ik}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + O_p(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\dfrac{\partial \delta_{ik}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}$ stands for the derivative of $\delta_{ik}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ applied to $\boldsymbol{\theta}_0$.
When n is sufficiently large,

$$\begin{aligned}
\delta_{ik}(\hat{\boldsymbol{\theta}}) &\approx \delta_{ik}(\boldsymbol{\theta}_0) + \frac{\partial \delta_{ik}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \\
\delta_{ik}(\hat{\boldsymbol{\theta}}) &\approx \delta_{ik}(\boldsymbol{\theta}_0) + \left(\frac{\partial \delta_{ik}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\Lambda}'}, \quad \frac{\partial \delta_{ik}(\boldsymbol{\theta}_0)}{\partial \sigma_\varepsilon}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).
\end{aligned} \tag{B.31}$$

It follows that a consistent estimator of the standard error of $\delta_{ik}(\hat{\boldsymbol{\theta}})$ is

$$Se\left(\delta_{ik}(\hat{\boldsymbol{\theta}})\right) = \sqrt{\left(\frac{\partial \delta_{ik}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\Lambda}'}, \quad \frac{\partial \delta_{ik}(\hat{\boldsymbol{\theta}})}{\partial \sigma_\varepsilon}\right) \widehat{AsyVar}\left(\hat{\boldsymbol{\theta}}\right) \left(\frac{\partial \delta_{ik}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\Lambda}'}, \quad \frac{\partial \delta_{ik}(\hat{\boldsymbol{\theta}})}{\partial \sigma_\varepsilon}\right)'}, \tag{B.32}$$

where

$$\frac{\partial \delta_{ik}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\Lambda}'} = \frac{\mathbf{e}_k}{\sigma_\varepsilon} \sum_{r=1}^{\infty} \phi\left(\frac{\mathbf{z}_i'\hat{\boldsymbol{\Lambda}} - a_r}{\sigma_\varepsilon}\right) - \frac{\lambda_k}{\sigma_\varepsilon^3}\mathbf{z}_i' \sum_{r=1}^{\infty} \left(\mathbf{z}_i'\hat{\boldsymbol{\Lambda}} - a_r\right) \phi\left(\frac{\mathbf{z}_i'\hat{\boldsymbol{\Lambda}} - a_r}{\sigma_\varepsilon}\right), \tag{B.33}$$

$$\frac{\partial \delta_{ik}(\hat{\boldsymbol{\theta}})}{\partial \sigma_\varepsilon} = \frac{\lambda_k}{\sigma_\varepsilon^4} \sum_{r=1}^{\infty} \left(\mathbf{z}_i'\hat{\boldsymbol{\Lambda}} - a_q\right)^2 \phi\left(\frac{\mathbf{z}_i'\hat{\boldsymbol{\Lambda}} - a_r}{\sigma_\varepsilon}\right) - \frac{\lambda_k}{\sigma_\varepsilon^2} \sum_{r=1}^{\infty} \phi\left(\frac{\mathbf{z}_i'\hat{\boldsymbol{\Lambda}} - a_r}{\sigma_\varepsilon}\right), \tag{B.34}$$

where $\mathbf{e}_k$ is a row vector of dimension K + 1 with the k-th term equal to one and the other terms equal to 0.
As in any non-linear model, the marginal effect depends on $\mathbf{z}_i$. I then report their average, $\dfrac{1}{n}\sum_{i=1}^{n} \delta_{ik}(\hat{\boldsymbol{\theta}})$, where

$$Se\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{ik}(\hat{\boldsymbol{\theta}})\right) = \sqrt{Q_{\boldsymbol{\theta}} * \widehat{AsyVar} * Q_{\boldsymbol{\theta}}'}, \tag{B.35}$$

and

$$Q_{\boldsymbol{\theta}} = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \delta_{ik}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\Lambda}'}, \quad \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \delta_{ik}(\hat{\boldsymbol{\theta}})}{\partial \sigma_\varepsilon}\right). \tag{B.36}$$

Table B.1 – Data summary

Variable	Mean	Sd.	Min	1st Qu.	Median	3rd Qu.	Max
Age	15.010	1.709	10	14	15	16	19
Sex							
Female	0.503	0.500	0	0	1	1	1
Male	0.497	0.500	0	0	0	1	1
Hispanic	0.168	0.374	0	0	0	0	1
Race							
White	0.625	0.484	0	0	1	1	1
Black	0.185	0.388	0	0	0	0	1
Asian	0.071	0.256	0	0	0	0	1
Other	0.097	0.296	0	0	0	0	1
Years at school	2.490	1.413	1	1	2	3	6
With both parents	0.727	0.445	0	0	1	1	1
Mother Educ.							
High	0.175	0.380	0	0	0	0	1
<High	0.302	0.459	0	0	0	1	1
>High	0.406	0.491	0	0	0	1	1
Missing	0.117	0.322	0	0	0	0	1
Mother job							
Stay at home	0.204	0.403	0	0	0	0	1
Professional	0.199	0.400	0	0	0	0	1
Other	0.425	0.494	0	0	0	1	1
Missing	0.172	0.377	0	0	0	0	1
Number of activities	2.353	2.406	0	1	2	3	33

B.3 Data summary

This section summarizes the data (see Table B.1). The categorical explanatory variables are discretized into several binary subvariables. For identification, the subvariables in italics are the omitted categories in the econometric models.

The dependent variable is the number of extracurricular activities in which students are enrolled. It varies from 0 to 33. However, most students declare that they participate in fewer than 10 extracurricular activities (see Figure B.1).

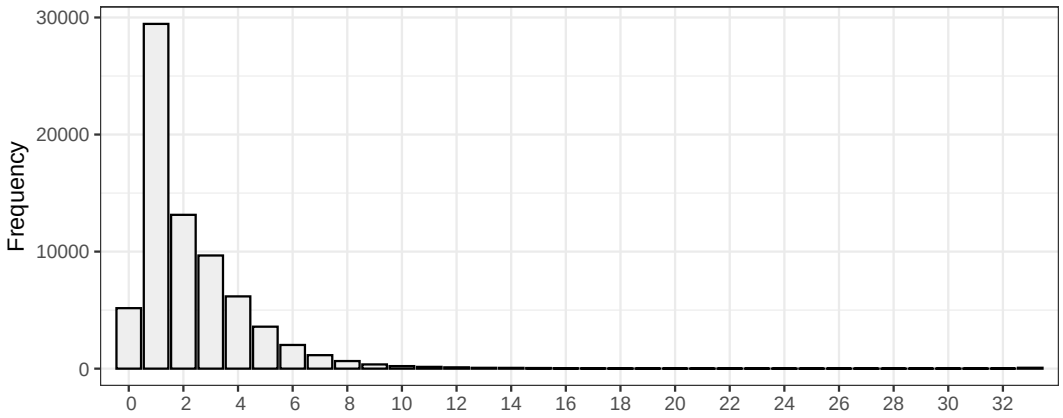


Figure B.1 – Distribution of the number of extracurricular activities

B.4 Supplementary note on network endogeneity

In this section, I present the posterior distribution of the dyadic linking model parameters and show how to simulate from this posterior distribution. I also present the new asymptotic variance of $\hat{\boldsymbol{\theta}}$, which includes the variability of $\tilde{\mu}_i$.

B.4.1 Posterior distribution of the dyadic linking model parameters

The likelihood of the model (2.18) is given by

$$\mathcal{L}(\mathbf{A}|\Delta\mathbf{X},\bar{\boldsymbol{\beta}},\boldsymbol{\mu}) = \prod_{s=1}^{S}\prod_{i\neq j}\frac{\exp\left(a_{ijs}(\Delta\mathbf{x}'_{ijs}\bar{\boldsymbol{\beta}}+\mu_{is}+\mu_{js})\right)}{1+\exp\left(\Delta\mathbf{x}'_{ijs}\bar{\boldsymbol{\beta}}+\mu_{is}+\mu_{js}\right)},$$

where $\mathbf{X}$ is the matrix of dyad-specific variables, $\boldsymbol{\mu}$ is the vector of unobserved individual-level attributes, and the subscript s is used to denote the school s. The number of schools is S. The joint distribution of $(\mathbf{A},\boldsymbol{\mu})$ conditionally on $\boldsymbol{\Theta} = \left(\Delta\mathbf{X},\bar{\boldsymbol{\beta}},u_{\mu 1},\sigma^2_{\mu 1},\ldots,u_{\mu S},\sigma^2_{\mu S}\right)$ can be defined by

$$\pi(\mathbf{A},\boldsymbol{\mu}|\boldsymbol{\Theta}) \propto \prod_{s=1}^{S}\left(\prod_{i\neq j}\frac{\exp\left(a_{ijs}(\Delta\mathbf{x}'_{ijs}\bar{\boldsymbol{\beta}}+\mu_{is}+\mu_{js})\right)}{1+\exp\left(\Delta\mathbf{x}'_{ijs}\bar{\boldsymbol{\beta}}+\mu_{is}+\mu_{js}\right)}\prod_{i=1}^{n_s}\frac{1}{\sigma_{\mu s}}\exp\left(-\frac{(\mu_{is}-u_{\mu s})^2}{\sigma^2_{\mu s}}\right)\right),$$

where n_s is the number of students in the school s.
I set a non-informative prior distribution on $\bar{\boldsymbol{\beta}}$ and conjugate prior on $\left(u_{\mu s},\sigma^2_{\mu s}\right)$; that is, $\pi\left(\bar{\boldsymbol{\beta}}\right)\propto 1$ and $\pi\left(u_{\mu s},\sigma^2_{\mu s}\right)\propto\frac{1}{\sigma_{\mu s}}$. Let Ξ be the vector containing $\bar{\boldsymbol{\beta}}$, $\boldsymbol{\mu}$, $u_{\mu 1},\sigma^2_{\mu 1},\ldots,u_{\mu S},\sigma^2_{\mu S}$. The posterior distribution of Ξ is

$$\pi(\Xi|\mathbf{A},\Delta\mathbf{X}) \propto \prod_{s=1}^{S}\left(\frac{1}{\sigma^{n_s+1}_{\mu s}}\prod_{i\neq j}\frac{\exp\left(a_{ijs}(\Delta\mathbf{x}'_{ijs}\bar{\boldsymbol{\beta}}+\mu_{is}+\mu_{js})\right)}{1+\exp\left(\Delta\mathbf{x}'_{ijs}\bar{\boldsymbol{\beta}}+\mu_{is}+\mu_{js}\right)}\prod_{i=1}^{n_s}\exp\left(-\frac{(\mu_{is}-u_{\mu s})^2}{\sigma^2_{\mu s}}\right)\right).$$

To simulate from this posterior distribution, I use a MCMC approach (see Algorithm B.1.) that combines a Metropolis–Hasting (Metropolis et al., 1953) and a Gibbs sampler

Algorithm B.1. MCMC to simulate the posterior distribution of the network formation model

Initialize $\bar{\boldsymbol{\beta}}$, $\boldsymbol{\mu}$, $u_{\mu 1}, \sigma^2_{\mu 1}, \ldots, u_{\mu S}, \sigma^2_{\mu S}$ to $\bar{\boldsymbol{\beta}}^{(0)}$, $\boldsymbol{\mu}^{(0)}$, $u^{(0)}_{\mu 1}, \sigma^{2(0)}_{\mu 1}, \ldots, u^{(0)}_{\mu S}, \sigma^{2(0)}_{\mu S}$, respectively;
for $t = 1, \ldots, T$, *where* T *is the number of simulations* **do**
  Draw the proposal $\bar{\boldsymbol{\beta}}^*$ from $\mathcal{N}\left(\bar{\boldsymbol{\beta}}^{(t-1)}, \text{jumping scale}\right)$. Update $\bar{\boldsymbol{\beta}}^{(t)}$ by accepting $\bar{\boldsymbol{\beta}}^*$ with the probability $\min\left\{1, \alpha_{\bar{\boldsymbol{\beta}}}\right\}$, where

$$\alpha_{\bar{\boldsymbol{\beta}}} = \prod_{s=1}^{S} \prod_{i \neq j} \frac{\exp\left(a_{ijs} \Delta \mathbf{x}'_{ijs} \bar{\boldsymbol{\beta}}^*\right) \left(1 + \exp\left(\Delta \mathbf{x}'_{ijs} \bar{\boldsymbol{\beta}}^{(t-1)} + \mu^{(t-1)}_{is} + \mu^{(t-1)}_{js}\right)\right)}{\exp\left(a_{ijs} \Delta \mathbf{x}'_{ijs} \bar{\boldsymbol{\beta}}^{(t-1)}\right) \left(1 + \exp\left(\Delta \mathbf{x}'_{ijs} \bar{\boldsymbol{\beta}}^* + \mu^{(t-1)}_{is} + \mu^{(t-1)}_{js}\right)\right)};$$

  for $s = 1, \ldots, S$ *and* $i = 1, \ldots, n_s$ **do**
    Draw the proposal μ^*_{is} from $\mathcal{N}\left(\mu^{(t-1)}_{is}, \text{jumping scale}\right)$. Update $\mu^{(t)}_{is}$ by accepting μ^*_{is} with the probability $\min\left\{1, \alpha_{\mu_{is}}\right\}$, where

$$\alpha_{\mu_{is}} = \exp\left(\frac{1}{\sigma^{2(t-1)}_{\mu s}} \left(\mu^{(t-1)}_{is} - u^{(t-1)}_{\mu s}\right)^2 - \frac{1}{\sigma^{2(t-1)}_{\mu s}} \left(\mu^*_{is} - u^{(t-1)}_{\mu s}\right)^2\right) \times$$

$$\prod_{j \neq i} \frac{\exp\left(a_{ijs} \mu^*_{is}\right) \left(1 + \exp\left(\Delta \mathbf{x}'_{ijs} \bar{\boldsymbol{\beta}}^{(t)} + \mu^{(t-1)}_{is} + \mu^*_{js}\right)\right)}{\exp\left(a_{ijs} \mu^{(t-1)}_{is}\right) \left(1 + \exp\left(\Delta \mathbf{x}'_{ijs} \bar{\boldsymbol{\beta}}^{(t)} + \mu^*_{is} + \mu^*_{js}\right)\right)}, \text{ and } \mu^*_{js} = \mu^{(t-1)}_{js}, \text{ if } i < j, \text{ and}$$

    $\mu^*_{js} = \mu^{(t)}_{js}$, if $i > j$;
  for $s = 1, \ldots, S$ **do**
    Use a Gibbs to update $u^{(t)}_{\mu s}$ from $\mathcal{N}\left(\frac{\sum_{i=1}^{n_s} \mu^{(t)}_{is}}{n_s}, \frac{\sigma^{2(t-1)}_{\mu s}}{n_s}\right)$;
  for $s = 1, \ldots, S$ **do**
    Use a Gibbs to update $\sigma^{2(t)}_{\mu s}$ from $Inv - \chi^2\left(n_s - 1, \sum_{i=1}^{n_s} \left(\mu^{(t)}_{is} - u^{(t)}_{\mu s}\right)^2\right)$;
  Update the jumping scales following Atchadé and Rosenthal (2005) to reach an acceptance rate equal to 0.27;

In practice the MCMC converges very quickly. I perform T = 20,000 simulations and keep the last 10,000. As the number of parameters in the model is large (72,291 parameters μ_i, 120 parameters $u_{\mu s}$, 120 parameters $\sigma^2_{\mu s}$ and, an eight-dimensional vector $\bar{\boldsymbol{\beta}}$), I randomly choose some parameters and present their posterior distribution in Figure B.2.

B.4.2 Correction of the asymptotic variance

As the estimation in done in two steps, the uncertainty related to $\tilde{\boldsymbol{\mu}}$ should be taken into account to correct the variance of the estimator at the second stage. The asymptotic variance, derived in Appendix B.2.1, is conditional on the explanatory variables, which include estimations of $\tilde{\boldsymbol{\mu}}$. In other words, the covariance of the estimator of $\hat{\boldsymbol{\theta}}$ resulting from the NPL approach is given by $\mathbf{Var}\left(\hat{\boldsymbol{\theta}} | \mathbf{G}, \mathbf{X}, \tilde{\boldsymbol{\mu}}\right)$ and not $\mathbf{Var}\left(\hat{\boldsymbol{\theta}} | \mathbf{G}, \mathbf{X}\right)$.
To simplify the notations, I omit conditioning on $\mathbf{G}$ and $\mathbf{X}$ in this section; that is, I write

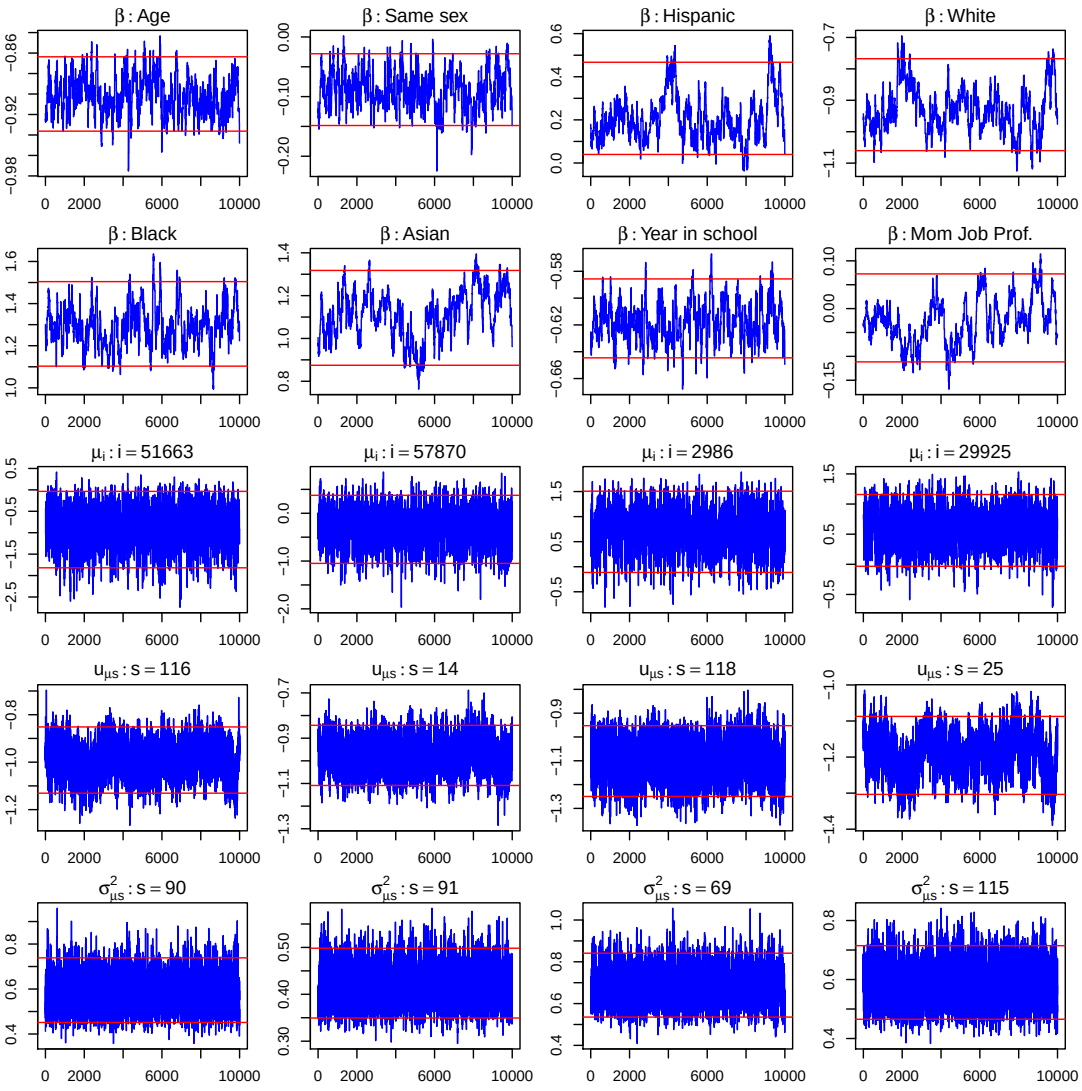


Figure B.2 – Posterior distribution of the network formation model parameters

This figure presents the posterior distribution of the coefficients of the observed dyad-specific variables as well as some other parameters chosen at random. Students of similar age, Hispanic, Black, and Asian students, as well as students who have spent a similar number of years at their current school are likely to form links. In contrast, students of the same sex and white students are not likely to form links.

$\mathbf{Var}\left(\hat{\theta}|\tilde{\boldsymbol{\mu}}\right)$ to mean $\mathbf{Var}\left(\hat{\theta}|\mathbf{G},\mathbf{X},\tilde{\boldsymbol{\mu}}\right)$ and $\mathbf{Var}\left(\hat{\theta}\right)$ to mean $\mathbf{Var}\left(\hat{\theta}|\mathbf{G},\mathbf{X}\right)$. Moreover, $\mathbf{E}_u$ (respectively $\mathbf{Var}_u$) means that the expectation (respectively variance) is taken with respect to $\tilde{\boldsymbol{\mu}}$. It follows that

$$
\begin{aligned}
\mathbf{Var}(\hat{\theta}) &= \mathbf{E}(\hat{\theta}\hat{\theta}') - \mathbf{E}(\hat{\theta})\,\mathbf{E}(\hat{\theta})', \\
\mathbf{Var}(\hat{\theta}) &= \mathbf{E}_u\left(\mathbf{E}(\hat{\theta}\hat{\theta}'|\tilde{\boldsymbol{\mu}})\right) - \mathbf{E}(\hat{\theta})\,\mathbf{E}(\hat{\theta})', \\
\mathbf{Var}(\hat{\theta}) &= \mathbf{E}_u\left(\mathbf{E}(\hat{\theta}\hat{\theta}'|\tilde{\boldsymbol{\mu}})\right) + \mathbf{E}_u\left(\mathbf{E}(\hat{\theta}|\tilde{\boldsymbol{\mu}})\,\mathbf{E}(\hat{\theta}|\tilde{\boldsymbol{\mu}})'\right) - \mathbf{E}_u\left(\mathbf{E}(\hat{\theta}|\tilde{\boldsymbol{\mu}})\,\mathbf{E}(\hat{\theta}|\tilde{\boldsymbol{\mu}})'\right) - \mathbf{E}(\hat{\theta})\,\mathbf{E}(\hat{\theta})', \\
\mathbf{Var}(\hat{\theta}) &= \mathbf{E}_u\Big(\underbrace{\mathbf{E}(\hat{\theta}\hat{\theta}'|\tilde{\boldsymbol{\mu}}) - \mathbf{E}(\hat{\theta}|\tilde{\boldsymbol{\mu}})\,\mathbf{E}(\hat{\theta}|\tilde{\boldsymbol{\mu}})'}_{\mathbf{Var}(\hat{\theta}|\tilde{\boldsymbol{\mu}})}\Big) + \underbrace{\mathbf{E}_u\left(\mathbf{E}(\hat{\theta}|\tilde{\boldsymbol{\mu}})\,\mathbf{E}(\hat{\theta}|\tilde{\boldsymbol{\mu}})'\right) - \mathbf{E}(\hat{\theta})\,\mathbf{E}(\hat{\theta})'}_{\mathbf{Var}_u\left(\mathbf{E}(\hat{\theta}|\tilde{\boldsymbol{\mu}})\right)}, \\
\mathbf{Var}(\hat{\theta}) &= \mathbf{E}_u\left(\mathbf{Var}(\hat{\theta}|\tilde{\boldsymbol{\mu}})\right) + \mathbf{Var}_u\left(\mathbf{E}(\hat{\theta}|\tilde{\boldsymbol{\mu}})\right). \qquad \text{(B.37)}
\end{aligned}
$$

In Equation (B.37), the first component of the variance, $\mathbf{E}_u\left(\mathbf{Var}(\hat{\theta}|\tilde{\boldsymbol{\mu}})\right)$ is the variance of $\hat{\theta}$ due to the NPL algorithm. This component does not include the uncertainty of $\tilde{\boldsymbol{\mu}}$. The second component of the variance $\mathbf{Var}_u\left(\mathbf{E}(\hat{\theta}|\tilde{\boldsymbol{\mu}})\right)$ is the variance due to the estimation of $\tilde{\boldsymbol{\mu}}$ at the first stage. To compute the second component of the variance, I make the following Assumption.

Assumption I. *Let $\tilde{\boldsymbol{\mu}}_s$ be a draw of $\tilde{\boldsymbol{\mu}}$ from its posterior distribution and $\hat{\theta}_s$ be the estimator of θ_0 associated with $\tilde{\boldsymbol{\mu}}_s$. $\hat{\theta}_s$ is a consistent estimator of $\mathbf{E}(\hat{\theta}_s|\tilde{\boldsymbol{\mu}}_s)$.*

Assumption I means that every estimator $\hat{\theta}_s$ associated with a draw $\tilde{\boldsymbol{\mu}}_s$ is a good estimator of $\mathbf{E}(\hat{\theta}_s|\tilde{\boldsymbol{\mu}}_s)$. This is useful because with many draws $\tilde{\boldsymbol{\mu}}_s$ the sample variance of $\hat{\theta}_s$ will be a good estimator of $\mathbf{Var}_u\left(\mathbf{E}(\hat{\theta}|\tilde{\boldsymbol{\mu}})\right)$. I also assume that the last 10,000 simulations from the posterior distribution at the first stage are sufficient to summarize well the posterior distribution of $\tilde{\boldsymbol{\mu}}_s$. Under these considerations, the variance of $\hat{\theta}_s$ is

$$
\widehat{AsyVar}\left(\hat{\theta}_s\right) = \frac{1}{S}\sum_{s=1}^{S}\mathbf{Var}(\hat{\theta}_s|\tilde{\boldsymbol{\mu}}_s) + \frac{1}{S-1}\sum_{s=1}^{T}\left(\hat{\theta}_s - \hat{\bar{\theta}}\right)\left(\hat{\theta}_s - \hat{\bar{\theta}}\right)', \qquad \text{(B.38)}
$$

where $\tilde{\boldsymbol{\mu}}_1, \ldots, \tilde{\boldsymbol{\mu}}_S$ are S draws of $\tilde{\boldsymbol{\mu}}$ with replacement from the population of the 10,000 simulations kept at the first stage, and $\hat{\bar{\theta}} = \frac{1}{S}\sum_{s=1}^{S}\hat{\theta}_s$. In practice, I set $S = 5,000$.

In Table 2.6, I present the average $\hat{\bar{\theta}}$ and the variance $\widehat{AsyVar}\left(\hat{\theta}_s\right)$ to summarize the distribution of $\hat{\theta}_s$. The same approach is used to compute the standard error of the marginal effects.

Appendix C

Chapter 3 of appendix

C.1 Proofs of the consistency of the Penalty function

In this appendix, we proof Proposition 3.1. To do so, we first state and prove two Lemmas.

Lemma C.1. *Under the conditions G.1-G.5 and let,*

$$f_T(\boldsymbol{\beta}) = \frac{1}{T}||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + \sum_{j=2}^{m}\sum_{k=1}^{K}\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda) \tag{C.1}$$

Then of every $\nu \in (0,1)$, there exists a constant $C_0 > 0$ such that

$$\liminf_{T\to\infty} \mathbf{P}\left[\underset{||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2 \leq C\sqrt{\frac{Km\sigma^2}{T}}}{\arg\min} f_T(\boldsymbol{\beta}) \subseteq \left\{\boldsymbol{\beta} \in \Re^{Km\times 1}; ||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2 < C\sqrt{\frac{Km\sigma^2}{T}}\right\}\right] > 1-\nu$$

for all $C \geq C_0$.

Proof. The proof is given in Appendix C.1.1. □

Lemma C.2. *Let $C > 0$ and f_T as defined by Equation (3.7). Under the conditions G.1-G.5,*

$$\liminf_{T\to\infty} \mathbf{P}\left[\underset{||\boldsymbol{\beta}-\boldsymbol{\beta}^*||_2 \leq C\sqrt{\frac{Km\sigma^2}{T}}}{\arg\min} f_T(\boldsymbol{\beta}) \subseteq \left\{\boldsymbol{\beta} \in \Re^{Km\times 1}; \boldsymbol{\beta}_{A^c} = 0\right\}\right] = 1$$

where $A^c = \{(j,k), j = 1,\ldots,m \text{ and } k = 0,\ldots,K-1\}\backslash A$ is the complement of A in $\{(j,k), j = 1,\ldots,m \text{ and } k = 0,\ldots,K-1\}$, $\boldsymbol{\beta}_{A^c} \in \Re^{|A^c|\times 1}$ is the $|A^c|$-dimensional sub-vector of $\boldsymbol{\beta}$ containing components subscripted by A^c.

Proof. See Appendix C.1.2 for the proof. □

C.1.1 Proof of Lemma 1

Proof. We consider the objective function

$$f_T(\boldsymbol{\beta}) = \frac{1}{T}||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 + \sum_{j=2}^{m}\sum_{k=1}^{K}\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k,\lambda)$$

Let $\alpha_T = \sqrt{\frac{Km\sigma^2}{T}}$ and $\nu \in (0,1)$. To prove the lemma C.1, It suffices to show that

$$\mathbf{P}\left(f_T(\boldsymbol{\beta}^*) < \inf_{||\mathbf{u}||_2=1} f_T(\boldsymbol{\beta}^* + C\alpha_T\mathbf{u})\right) = 1-\nu$$

for C>0 sufficiently large and for any T sufficiently large. In other words, we shall show that $H_T(\mathbf{u}) = f_T(\boldsymbol{\beta}^* + C\alpha_T\mathbf{u}) - f_T(\boldsymbol{\beta}^*)$ is positive for any T when C is large enough and for all $||\mathbf{u}||_2 = 1$, where $\mathbf{u} = (u_{11}, u_{12}\ \ldots, u_{1K},\ \ldots,\ u_{mK}) \in \mathbb{R}^{mK\times 1}$.

We can easily show that

$$\begin{aligned}
H_T(\mathbf{u}) &= \frac{1}{T}\left(C^2\alpha_T^2||\mathbf{X}_{\boldsymbol{\tau}}\mathbf{u}||_2^2 - 2C\alpha_T\boldsymbol{\varepsilon}'\mathbf{X}_{\boldsymbol{\tau}}\mathbf{u}\right) + \\
&\quad \sum_{j=2}^{m}\sum_{k=1}^{K}\left(\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^* + C\alpha_T u_{jk}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda)\right) \\
H_T(\mathbf{u}) &\geq \frac{1}{T}\left(C^2\alpha_T^2||\mathbf{X}_{\boldsymbol{\tau}}\mathbf{u}||_2^2 - 2C\alpha_T\boldsymbol{\varepsilon}'\mathbf{X}_{\boldsymbol{\tau}}\mathbf{u}\right) + \\
&\quad \sum_{(j,k)\in\mathbf{D}(\mathbf{u})}\left(\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda)\right)
\end{aligned}$$

where $\Delta\beta_{jk}^{*+} = \Delta\beta_{jk}^* + C\alpha_T u_{jk}$ and

$\mathbf{D}(\mathbf{u}) = \left\{(j,k); j \geq 2 \text{ and } \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda) < 0\right\}$.

For any $(j,k) \in \mathbf{D}(\mathbf{u})$, clearly $\Delta\beta_{jk}^* \neq 0$, otherwise $\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda) \geq 0$. Thus, if $C > 0$ is sufficiently large and fixed, as $\lim_{T\to\infty} C\alpha_T = 0$, we can consider that $\Delta\beta_{jk}^{*+}$ and $\Delta\beta_{jk}^*$ have the same sign for T sufficiently large; that is $0 \notin (c_T^-, c_T^+)$, where $c_T^- = \min(\Delta\beta_{jk}^{*+}, \Delta\beta_{jk}^*)$ and $c_T^+ = \max(\Delta\beta_{jk}^{*+}, \Delta\beta_{jk}^*)$. By the fact that $\mathcal{P}_{\mathbf{SELO}}(x|a_k,\lambda)$ is a concave function on $x \in (-\infty, 0]$ and on $x \in [0, +\infty)$, thus also on (c_T^-, c_T^+), we can establish the following conditions using the mean value theorem.

$$\begin{aligned}
\frac{\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda)}{C\alpha_T u_{jk}} &\leq \max\left(\mathcal{P}'_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda), \mathcal{P}'_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda)\right) \\
\frac{\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda)}{C\alpha_T u_{jk}} &\geq \min\left(\mathcal{P}'_{\mathbf{SELO}}(\Delta\beta_{jk}^{*+}|a_k,\lambda), \mathcal{P}'_{\mathbf{SELO}}(\Delta\beta_{jk}^*|a_k,\lambda)\right)
\end{aligned}$$

where $\mathcal{P}'_{\mathbf{SELO}}$ stands for the $\mathcal{P}_{\mathbf{SELO}}$ first derivative.

Let us note that $\forall\ (j,k) \in \mathbf{D}(\mathbf{u})$, $\Delta\beta^*_{jk} > 0 \implies u_{jk} < 0$ and $\Delta\beta^*_{jk} < 0 \implies u_{jk} > 0$ so that $\mathcal{P}_{\mathbf{SELO}}(\Delta\beta^{*+}_{jk}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta^*_{jk}|a_k,\lambda) < 0$ holds. Applying the mean value theorem in both cases, we end up with a common condition given by

$$\begin{aligned}\mathcal{P}_{\mathbf{SELO}}(\Delta\beta^{*+}_{jk}|a_k,\lambda) - \mathcal{P}_{\mathbf{SELO}}(\Delta\beta^*_{jk}|a_k,\lambda) &\geq -C\alpha_T|u_{jk}||\mathcal{P}'_{\mathbf{SELO}}(\Delta\beta^*_{jk} + C\alpha_T u_{jk}|a_k,\lambda)| \\ &\geq -\frac{C\lambda\alpha_T a_k\zeta}{\ln(2)(\rho^2 - 2\rho C\alpha_T)} \\ &\geq -\frac{C\lambda\alpha_T a_{max}\zeta}{\ln(2)(\rho^2 - 2\rho C\alpha_T)}\end{aligned}$$

where the last two inequalities come from the $|\mathcal{P}'_{\mathbf{SELO}}(\Delta\beta^*_{jk} + C\alpha_T u_{jk}|a_k,\lambda)|$ minimization with respect to $\Delta\beta^*_{jk}$. Then

$$H_T(\mathbf{u}) \geq \underbrace{\frac{C^2\alpha_T^2||\mathbf{X}_{\boldsymbol{\tau}}\mathbf{u}||_2^2}{T}}_{Q_1} - \underbrace{\frac{2C\alpha_T\boldsymbol{\varepsilon}'\mathbf{X}_{\boldsymbol{\tau}}\mathbf{u}}{T}}_{Q_2} - \underbrace{\frac{CKm\lambda\alpha_T a_{max}\zeta}{\ln(2)(\rho^2 - 2\rho C\alpha_T)}}_{Q_3}$$

Focusing on each term, we can show that

$$Q_1 \equiv \frac{C^2\alpha_T^2||\mathbf{X}_{\boldsymbol{\tau}}\mathbf{u}||_2^2}{T} = C^2\alpha_T^2\mathbf{u}'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}\mathbf{u} \geq C^2\alpha_T^2\lambda_{T,\min}$$

where $\lambda_{T,\min}$ is the smallest eigenvalue of $\dfrac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}$.

To show this condition, we can decompose $\dfrac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}$ into $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$ (by G.3). Moreover, any vector of Km dimension can be decomposed into a linear combination of the eigenvectors (i.e., $\boldsymbol{u} = \boldsymbol{U}\boldsymbol{\omega}$). Note that $\boldsymbol{u}'\boldsymbol{u} = \boldsymbol{\omega}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{\omega} = \boldsymbol{\omega}'\boldsymbol{\omega} = \sum_{i=1}^{Km}\omega_i^2 = 1$.

Thus $\mathbf{u}'\dfrac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}\mathbf{u} = \boldsymbol{\omega}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'\boldsymbol{U}\boldsymbol{\omega} = \boldsymbol{\omega}'\boldsymbol{\Lambda}\boldsymbol{\omega} = \sum_{i=1}^{Km}\omega_i^2\lambda_i \geq \lambda_{T,\min}$.

The second term is given by

$$\begin{aligned}Q_2 \equiv \frac{2C\alpha_T\boldsymbol{\varepsilon}'\mathbf{X}_{\boldsymbol{\tau}}\mathbf{u}}{T} &\leq \frac{2C\alpha_T|\boldsymbol{\varepsilon}'\mathbf{X}_{\boldsymbol{\tau}}\mathbf{u}|}{T} \\ &\leq \frac{2C\alpha_T||\boldsymbol{\varepsilon}'\mathbf{X}_{\boldsymbol{\tau}}||_2||\mathbf{u}||_2}{T} \text{ by Cauchy-Schwartz} \\ &\leq \frac{2C\alpha_T^2}{\sqrt{Km\sigma^2}}\sqrt{\frac{(\boldsymbol{\varepsilon}'(\mathbf{X}_{\boldsymbol{\tau}}\mathbf{X}'_{\boldsymbol{\tau}})\boldsymbol{\varepsilon})}{T}} \\ &\leq \mathcal{O}_p(C\alpha_T^2) \quad (\textit{By G.3 and G.4}).\end{aligned}$$

To show that $\sqrt{\dfrac{(\boldsymbol{\varepsilon}'(\mathbf{X}_{\boldsymbol{\tau}}\mathbf{X}'_{\boldsymbol{\tau}})\boldsymbol{\varepsilon})}{T}} = \mathcal{O}_p(1)$, we rely on the spectral theorem to decompose $\mathbf{X}_{\boldsymbol{\tau}}\mathbf{X}'_{\boldsymbol{\tau}}$ into two orthogonal matrices and a diagonal matrix of eigenvalues. With this decomposition,

we can show that $\epsilon' \frac{\mathbf{X_\tau X'_\tau}}{T} \epsilon \leq \max_i \lambda_i \frac{\epsilon'\epsilon}{T}$ which is $\mathcal{O}_p(1)$ under Assumption G.3 and the fact that the variance is bounded.

The last term Q_3 is defined by

$$\begin{aligned} Q_3 &\equiv \frac{CKm\lambda\alpha_T a_{max}\zeta}{\ln(2)(\rho^2 - 2\rho C\alpha_T)} \\ &= C\alpha_T^2 \frac{\lambda\zeta \frac{a_{\max}}{(Km)^{-1}\alpha_T^3}}{\ln(2)\left(\left(\frac{\rho}{\alpha_T}\right)^2 - 2C\left(\frac{\rho}{\alpha_T}\right)\right)} \end{aligned}$$

By G.2, $\left(\frac{\rho}{\alpha_T}\right)^2 - 2C\left(\frac{\rho}{\alpha_T}\right) \to \infty$ and by G.5, $\lim_{T\to\infty} \lambda\zeta \frac{a_{\max}}{(Km)^{-1}\alpha_T^3} < \infty$. Hence $Q_3 = o(C\alpha_T^2)$. Combining the conditions on Q_1, Q_2 and Q_3 we establish that

$$H_T(\mathbf{u}) \geq C^2\alpha_T^2\lambda_{T,\min} + \mathcal{O}_p(C\alpha_T^2) + o(C\alpha_T^2)$$

. It follows that there exists $C_0 > 0$ is large such that for all $C > C_0$, $\mathbf{P}\left(\inf_{||\mathbf{u}||_2=1} H_T(\mathbf{u}) > 0\right) = 1 - \nu$, for T sufficiently large. □

C.1.2 Lemma 2

Proof. Let $\boldsymbol{\beta} \in \Re^{Km\times 1}$ such that $||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_2 < C\sqrt{\frac{Km\sigma^2}{T}}$. We consider $\tilde{\boldsymbol{\beta}} \in \Re^{Km\times 1}$, where $\tilde{\boldsymbol{\beta}}_{A^c} = 0$ and $\tilde{\boldsymbol{\beta}}_A = \boldsymbol{\beta}_A$. We can notice that

$$\begin{aligned} &||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2 = ||\boldsymbol{\beta}_{A^c} - \tilde{\boldsymbol{\beta}}_{A^c}||_2 = ||\boldsymbol{\beta}_{A^c} - \boldsymbol{\beta}^*_{A^c}||_2 \\ &||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2 < C\alpha_T \end{aligned}$$

On the other hand

$$\begin{aligned} &||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}||_2 = ||\boldsymbol{\beta}^*_A - \tilde{\boldsymbol{\beta}}_A||_2 = ||\boldsymbol{\beta}^*_A - \boldsymbol{\beta}_A||_2 \\ &||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}||_2 < C\alpha_T \end{aligned}$$

Let us define $G_T(\boldsymbol{\beta}) = f_T(\boldsymbol{\beta}) - f_T(\tilde{\boldsymbol{\beta}})$. Similarly to the proof of the lemma C.1, it suffices to

show that $G_T(\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}) > 0$.

$$
\begin{aligned}
G_T(\boldsymbol{\beta}) &= \frac{1}{T}\left(||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\boldsymbol{\beta}||_2^2 - ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\tilde{\boldsymbol{\beta}}||_2^2\right) + \sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda) \\
&= \frac{1}{T}\left(||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\tilde{\boldsymbol{\beta}} - \mathbf{X}_{\boldsymbol{\tau}}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})||_2^2 - ||\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\tilde{\beta}||_2^2\right) + \sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda) \\
&= (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) - 2(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}(\mathbf{y} - \mathbf{X}_{\boldsymbol{\tau}}\tilde{\boldsymbol{\beta}})}{T} + \sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda) \\
&= (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) - 2(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\boldsymbol{\epsilon}}{T} - 2(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}) \\
&\quad + \sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda) \\
&= (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) - 2\alpha_T\frac{(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'}{\sqrt{Km\sigma^2}}\frac{\mathbf{X}'_{\boldsymbol{\tau}}\boldsymbol{\epsilon}}{\sqrt{T}} - 2(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}(\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}) \\
&\quad + \sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda)
\end{aligned}
$$

By G.1, G.3 and G.4, $\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T} = \mathcal{O}_p(1)$ and $\frac{\mathbf{X}'_{\boldsymbol{\tau}}\boldsymbol{\epsilon}}{\sqrt{T}} = \mathcal{O}_p(1)$ (as it is a martingale difference sequence following assumption G.4). Moreover $||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2 < C\alpha_T$ and $||\boldsymbol{\beta}^* - \tilde{\boldsymbol{\beta}}||_2 \leq C\alpha_T$. Then, for any T sufficiently large

$$
G_T(\boldsymbol{\beta}) = \mathcal{O}_p\left(||\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}||_2\alpha_T\right) + \sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda)
$$

As $\mathcal{P}_{\mathbf{SELO}}(x|a_k, \lambda)$ is a concave function on $x \in]-\infty, 0]$ and on $x \in [0, +\infty[$, for any $\nu_1 < \nu_2 \leq \nu_3 \leq 0$ (resp. $0 \leq \nu_1 \leq \nu_2 < \nu_3$), $\frac{\mathcal{P}_{\mathbf{SELO}}(\nu_1) - \mathcal{P}_{\mathbf{SELO}}(\nu_3)}{\nu_1 - \nu_3} \geq \frac{\mathcal{P}_{\mathbf{SELO}}(\nu_2) - \mathcal{P}_{\mathbf{SELO}}(\nu_3)}{\nu_2 - \nu_3}$ (resp. $\frac{\mathcal{P}_{\mathbf{SELO}}(\nu_3) - \mathcal{P}_{\mathbf{SELO}}(\nu_1)}{\nu_3 - \nu_1} \leq \frac{\mathcal{P}_{\mathbf{SELO}}(\nu_2) - \mathcal{P}_{\mathbf{SELO}}(\nu_1)}{\nu_2 - \nu_1}$).
$\forall\ (j,k) \in A^c$, $\Delta\beta^*_{jk} = 0$ and $\Delta\beta_{jk}$ is strictly positive or negative.

Thus, $-C\alpha_T \leq \beta_{jk} < 0$ or $0 < \Delta\beta_{jk} < C\alpha_T$, since $|\Delta\beta_{jk}| \leq ||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_2 < C\alpha_T$. In both cases, we end up with

$$
\begin{aligned}
\frac{\mathcal{P}_{\mathbf{SELO}}\left(C\alpha_T|a_k, \lambda\right)}{C\alpha_T} &\leq \frac{\mathcal{P}_{\mathbf{SELO}}\left(\Delta\beta_{jk}|a_k, \lambda\right)}{|\Delta\beta_{jk}|} \\
\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda) &\geq \frac{\lambda}{\ln(2)C\alpha_T}\ln\left(\frac{C\alpha_T}{C\alpha_T + a_k\zeta} + 1\right)|\Delta\beta_{jk}| \\
\mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k, \lambda) &\geq \frac{\lambda}{\ln(2)C}\ln\left(\frac{C\alpha_T}{C\alpha_T + a_{\max}\zeta} + 1\right)|\Delta\beta_{jk}|
\end{aligned}
$$

for any T sufficiently large. Thus

$$\sum_{\substack{(j,k)\in A^c \\ j\geq 2}} \mathcal{P}_{\mathbf{SELO}}(\Delta\beta_{jk}|a_k,\lambda) \geq \frac{\lambda}{\ln(2)C}\ln\left(\frac{C}{C+a_{\max}\zeta\sqrt{\frac{T}{Km\sigma^2}}}+1\right)||\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}||_2.$$

Furthermore, by G.5 $a_{\max} = \mathcal{O}_p\left(\sqrt{\frac{mK\sigma^2}{T}}\frac{\sigma_2}{T}\right)$. Thereby

$$a_{\max}\zeta\sqrt{\frac{T}{Km\sigma^2}} \xrightarrow{p} 0 \text{ and } \liminf_{T\to\infty}\left(\frac{C}{C+a_{\max}\zeta\sqrt{\frac{T}{Km\sigma^2}}}+1\right) > 0$$

It follows that, there exists $\tilde{C} > 0$ such that

$$\frac{G_T(\boldsymbol{\beta},\tilde{\boldsymbol{\beta}})}{||\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}||_2} \geq \tilde{C}\lambda + \mathcal{O}_p\left(\sqrt{\frac{Km\sigma^2}{T}}\right)$$

Thereby the result follows. □

C.1.3 Proof of the Proposition 3.1

Proof. The theorem is immediately given by the lemmas (C.1) and (C.2), in the sense that there exists a sequence of local minima $\hat{\boldsymbol{\beta}}$ of $f_T(\boldsymbol{\beta})$ such that $||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*|| = \mathcal{O}_p\left(T^{-\frac{1}{2}}\right)$ (since m, K and σ^2 are constants) and $\hat{\boldsymbol{\beta}}_{A^c} = \mathbf{0} \in \mathbb{R}^{|A^c|\times 1}$. Thus, as $T^{-\frac{1}{2}} \to 0$, it follows that $||\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}^*_A||_2 = o_P(1)$. □

C.1.4 Consequence of bounded eigenvalues

We show that bounded eigenvalues of the matrix $\frac{\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}}}{T}$ implies a fixed number of regimes. Note first that

$$\mathbf{X}'_{\boldsymbol{\tau}}\mathbf{X}_{\boldsymbol{\tau}} = \sum_{t=1}^{T}(\mathbf{1}_{\{t\}}\otimes\mathbf{x}_t)(\mathbf{1}_{\{t\}}\otimes\mathbf{x}_t)', \tag{C.2}$$

$$= \sum_{t=1}^{T}(\mathbf{x}_t\mathbf{x}'_t)\otimes(\mathbf{1}_{\{t\}}\mathbf{1}'_{\{t\}}), \tag{C.3}$$

where we define $\mathbf{1}_{\{t\}} = (\mathbf{1}_{\{t>\tau_0\}}, \mathbf{1}_{\{t>\tau_1\}}, \ldots, \mathbf{1}_{\{t>\tau_{m-1}\}})'$. Let us define $n_i = \sum_{t=1}^{T}\mathbf{1}_{\{t>\tau_{i-1}\}}$, i.e., the number of observations from the beginning of regime i to the end of the sample. Working

with $\mathbf{x}_t \equiv 1$, we have that

$$\frac{\mathbf{X}_{\boldsymbol{\tau}}'\mathbf{X}_{\boldsymbol{\tau}}}{T} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{1}_{\{t\}}\mathbf{1}_{\{t\}}'), \tag{C.4}$$

$$= \frac{1}{T}\begin{pmatrix} n_1 & n_2 & n_3 & \dots & n_m \\ n_2 & n_2 & n_3 & \dots & n_m \\ n_3 & n_3 & n_3 & \dots & n_m \\ & & & \dots & \\ n_m & n_m & n_m & \dots & n_m \end{pmatrix} \tag{C.5}$$

in which $n_1 = T$. It leads to the following determinant, when $m > 1$,

$$|\frac{\mathbf{X}_{\boldsymbol{\tau}}'\mathbf{X}_{\boldsymbol{\tau}}}{T}| = T^{-m}n_m\prod_{i=1}^{m-1}(n_i - n_{i+1}), \tag{C.6}$$

$$= \frac{n_m}{T}\prod_{i=1}^{m-1}\frac{(n_i - n_{i+1})}{T}. \tag{C.7}$$

Note that $n_i = T - \tau_{i-1} = \tau_m - \tau_{i-1}$, for $i = 1, \dots, m$. Thus,

$$|\frac{\mathbf{X}_{\boldsymbol{\tau}}'\mathbf{X}_{\boldsymbol{\tau}}}{T}| = \prod_{i=1}^{m}\frac{\tau_i - \tau_{i-1}}{T} \tag{C.8}$$

$$= \prod_{i=1}^{m}\delta_{\tau_i} > 0 \tag{C.9}$$

where $\delta_{\tau_i} = \frac{\tau_i - \tau_{i-1}}{T}$.
It shows that the number of segments cannot increase with T otherwise the determinant tends to zero. Let us assume that $m = \mathcal{O}(T^q)$ with $q > 0$. It is clear that when T tends to ∞, there exists $r \in \mathbb{N}$ such that $\delta_{\tau_r} = \mathcal{O}(T^x)$ where $x < 0$. If such a r does not exist, this would imply that $\forall i,\ \delta_{\tau_i}$ does not drift to 0 as $T \to \infty$ and then $m = \mathcal{O}(1)$ because $\inf_i\{\delta_{\tau_i}\}m \leq 1$. But, because $m = \mathcal{O}(T^q)$ with $q > 0$, there exists $r \in \mathbb{N}$ such that $\delta_{\tau_r} = \mathcal{O}(T^x)$ where $x < 0$. Therefore, we have that

$$|\frac{\mathbf{X}_{\boldsymbol{\tau}}'\mathbf{X}_{\boldsymbol{\tau}}}{T}| = \delta_{\tau_r}\prod_{\substack{i=1 \\ i\neq r}}^{m}\delta_{\tau_i} \to 0, \tag{C.10}$$

which contradicts Assumption G.3. Hence $m = \mathcal{O}(1)$. As a result G.3 implies that $m < \infty$. In addition, G.3 also implies that $\min_i\{\delta_{\tau_i}\} > 0$; that is δ_{τ_i} does not drift to 0 as $T \to \infty$.

C.1.5 Approximation of the penalty function with mixture of normal densities

To derive the DAEM algorithm, a mixture of two normal densities has been assumed for the mean parameter. We now provide a simple mixture approximation of the SELO penalty. Note

that in practice, one can use the output of the DAEM algorithm as a starting point to optimize the function of Equation (3.8). Due to the mixture approximation and the continuity of the SELO penalty function, the starting point would be in general very close to the value that globally minimizes the function (3.8). We now use a mixture of two normal densities that can be understood as a spike and slab prior in the Bayesian paradigm. In particular, we calibrate a spike and slab prior (see, e.g., George and McCulloch, 1993) to the SELO penalty function. Given a mean parameter β, the spike and slab prior is specified as,

$$\begin{aligned} \beta &\sim \mathcal{N}(0, r_z), \\ z &\sim \text{Bernoulli}(1-\omega), \end{aligned} \tag{C.11}$$

where $r_0 < r_1$ such that the spike distribution arises with probability $P[z = 0|\omega] = \omega$. By marginalizing out z, we get a mixture of two normal densities given by

$$f(\beta|\omega, r_0, r_1) = \omega(\frac{1}{r_0})^{\frac{1}{2}} f_Z(\frac{\beta}{\sqrt{r_0}}) + (1-\omega)(\frac{1}{r_1})^{\frac{1}{2}} f_Z(\frac{\beta}{\sqrt{r_1}}). \tag{C.12}$$

in which $f_Z(x)$ denotes the Normal density function evaluated at x and with expectation and variance equal to 0 and 1, respectively. The calibration is done as follows:

$$\begin{aligned} c \equiv \frac{r_1}{r_0} &= 10000, \\ \omega &= \frac{(\exp(\lambda) - 1)}{(\sqrt{c} + (\exp(\lambda) - 1))}, \\ r_0 &= \frac{a^2}{8} \frac{1 - c^{-1}}{|\ln(\exp(\lambda) - 1)|} \end{aligned} \tag{C.13}$$

We now detail how we come up with this simple calibration. Given β, note that the probability of being in the slab component is equal to

$$Pr(z = 1|\beta, \omega, r) = \frac{1}{\frac{\omega}{(1-\omega)} \frac{(\frac{1}{r_0})^{\frac{1}{2}} f_Z(\frac{\beta}{\sqrt{r_0}})}{(\frac{1}{r_1})^{\frac{1}{2}} f_Z(\frac{\beta}{\sqrt{r_1}})} + 1}. \tag{C.14}$$

To mimic the SELO penalty function, we impose the following constraints on the spike and slab hyper-parameters.

1. As standards in the spike and slab literature, we fix $c \equiv \frac{r_1}{r_0} = 10000$ (see, e.g. Malsiner-Walli and Wagner, 2016).
2. The SELO function imposes a penalty equal to $\mathcal{P}_{\mathbf{SELO}}(a) - \mathcal{P}_{\mathbf{SELO}}(0) = \lambda y$ with $y =$

0.99. To fix the same penalty value (neglecting y because $y \approx 1$), we set

$$\begin{aligned}
(-\lambda) &= \ln f(\beta = a|\omega, r_z) - \ln f(\beta = 0|\omega, r_z), \\
&= \ln \frac{f(\beta = a|\omega, r_1, z = 1)}{f(\beta = 0|\omega, r_1, z = 1)} + \ln \frac{Pr(z = 1|\beta = 0, \omega, r)}{Pr(z = 1|\beta = a, \omega, r)}, \\
&= -\frac{a^2}{2r_1} + \ln \frac{Pr(z = 1|\beta = 0, \omega, r)}{Pr(z = 1|\beta = a, \omega, r)}, \\
&\approx \ln \frac{Pr(z = 1|\beta = 0, \omega, r)}{Pr(z = 1|\beta = a, \omega, r)}, \text{ because } r_1 >> a^2, \\
&\approx \ln Pr(z = 1|\beta = 0, \omega, r), \text{ because } Pr(z = 1|\beta = a, \omega, r) \approx 1, \\
&= -\ln(\frac{\omega}{(1-\omega)}\sqrt{c} + 1).
\end{aligned} \tag{C.15}$$

3. Finally, we impose that $Pr(z = 1|\beta, \omega, r) = Pr(z = 0|\beta, \omega, r)$ when $\beta = \frac{a}{2}$ (this is called the intersection point in Ročková and George (2018)). This means that the slab component starts to dominate when $|\beta| > \frac{a}{2}$. It leads to the constraints:

$$\begin{aligned}
0 &= (2\pi)^{-\frac{1}{2}}[\frac{\omega}{\sqrt{r_0}} \exp(-\frac{a^2}{8r_0}) - \frac{1-\omega}{\sqrt{r_1}} \exp(-\frac{a^2}{8r_1})], \\
r_0 &= \frac{a^2}{8} \frac{1 - c^{-1}}{|\ln(\exp(\lambda) - 1)|}.
\end{aligned} \tag{C.16}$$

Figure C.1 shows the penalty imposed by the SELO function and by the calibrated spike and slab prior for several values of λ and a. We observe that the spike and slab prior provides a good approximation of the penalty function.

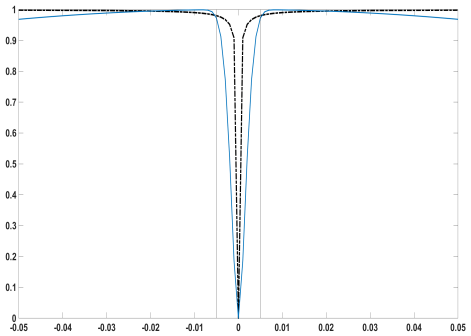
C.2 Marginal likelihood for the linear model

Let us derive the criterion (3.11). We first define $X_1 = \tilde{\mathbf{X}}_{\tau_0}$, $X_2 = X^{\hat{A}}_{\boldsymbol{\tau}}$ and $M_{X_1} = \mathbf{M}_{\tilde{\mathbf{X}}_{\tau_0}}$. Given the prior distributions in Equation (3.13), the marginal likelihood is given by,

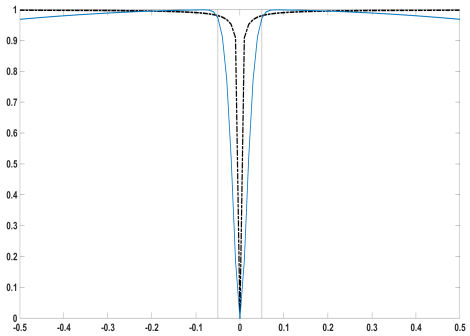
$$\begin{aligned}
f(\mathbf{y}|\mathbf{a}, \lambda, \boldsymbol{\tau}) = \int\int (2\pi)^{\frac{-(T+k_{\hat{A}})}{2}} (\sigma^2)^{-\frac{(T+2+k_{\hat{A}})}{2}} |g_{\hat{A}}(X_2)' M_{X_1} X_2)|^{1/2} \times \\
\exp \frac{-1}{2\sigma^2} \underbrace{\left\{ \begin{array}{c} (y - X_1\boldsymbol{\beta}_1 - X_2\Delta\boldsymbol{\beta})'(y - X_1\boldsymbol{\beta}_1 - X_2\Delta\boldsymbol{\beta}) \\ +\Delta\boldsymbol{\beta}' g_{\hat{A}}(X_2)' M_{X_1} X_2)\Delta\boldsymbol{\beta} \end{array} \right\}}_{B} d(\boldsymbol{\beta}_1, \Delta\boldsymbol{\beta}) d\sigma^2.
\end{aligned}$$

Focusing on the expression in the exponential, we have

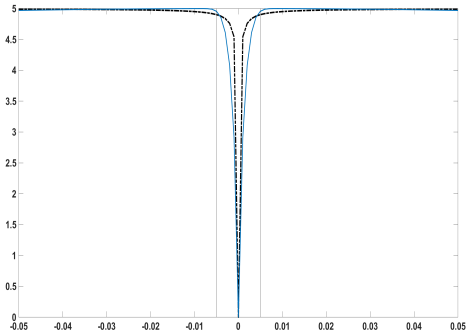
$$\begin{aligned}
B &= (y - X_1\boldsymbol{\beta}_1 - X_2\Delta\boldsymbol{\beta})'(y - X_1\boldsymbol{\beta}_1 - X_2\Delta\boldsymbol{\beta}) + \Delta\boldsymbol{\beta}' g_{\hat{A}} \underbrace{(X_2' M_{X_1} X_2)}_{\Sigma_X} \Delta\boldsymbol{\beta}, \\
&= (y - X_2\Delta\boldsymbol{\beta})'(y - X_2\Delta\boldsymbol{\beta}) + \Delta\boldsymbol{\beta}' g_{\hat{A}} \underbrace{(X_2' M_{X_1} X_2)}_{\Sigma_X} \Delta\boldsymbol{\beta} + \boldsymbol{\beta}_1' X_1' X_1 \boldsymbol{\beta}_1 \\
&\quad -2\boldsymbol{\beta}_1' X_1'(y - X_2\Delta\boldsymbol{\beta}), \\
&= (y - X_2\Delta\boldsymbol{\beta})'(y - X_2\Delta\boldsymbol{\beta}) + \Delta\boldsymbol{\beta}' g_{\hat{A}} \Sigma_X \Delta\boldsymbol{\beta} + (\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1)' \Omega^{-1} (\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1) - \bar{\boldsymbol{\beta}}_1{}' \Omega^{-1} \bar{\boldsymbol{\beta}}_1,
\end{aligned}$$



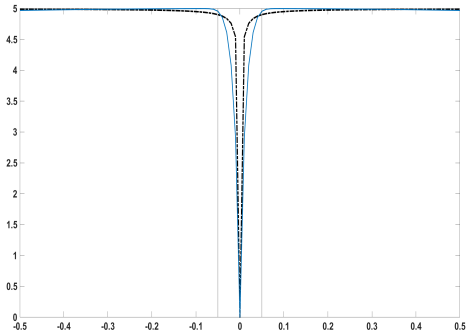
(a) $\lambda = 1$, $a = 0.01$



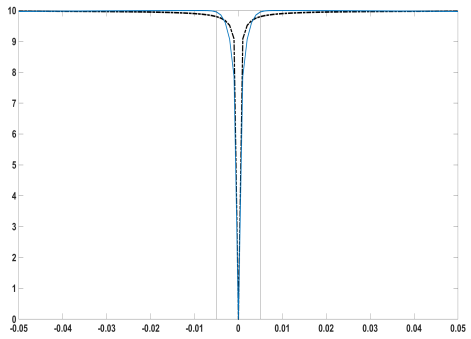
(b) $\lambda = 1$, $a = 0.1$



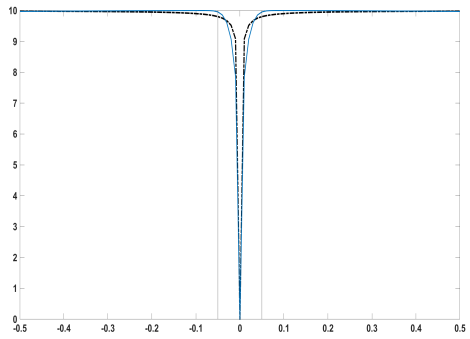
(c) $\lambda = 5$, $a = 0.01$



(d) $\lambda = 5$, $a = 0.1$



(e) $\lambda = 10$, $a = 0.01$



(f) $\lambda = 10$, $a = 0.1$

Figure C.1 – Penalty imposed by the SELO function and slab prior

Penalty imposed by the SELO function in dotted black line and slab prior in blue for several values of λ and a. Vertical lines are the intersection points of the spike and slab densities (at $|x| = \frac{a}{2}$).

where $\Omega^{-1} = X_1'X_1$, $\bar{\boldsymbol{\beta}}_1 = (X_1'X_1)^{-1}X_1'(y - X_2\Delta\boldsymbol{\beta})$
and $\bar{\boldsymbol{\beta}}_1{}'\Omega^{-1}\bar{\boldsymbol{\beta}}_1 = (y - X_2\Delta\boldsymbol{\beta})'X_1(X_1'X_1)^{-1}X_1'(y - X_2\Delta\boldsymbol{\beta}) = (y - X_2\Delta\boldsymbol{\beta})'P_{X_1}(y - X_2\Delta\boldsymbol{\beta})$.

The marginal likelihood can be simplifies as

$$f(\mathbf{y}|\mathbf{a},\lambda,\boldsymbol{\tau}) = |X_1'X_1|^{-\frac{1}{2}} \int\int (2\pi)^{\frac{-(T+k_{\hat{A}}-K)}{2}} (\sigma^2)^{-\frac{(T+2+k_{\hat{A}}-K)}{2}} |g_{\hat{A}}\Sigma_X|^{1/2} \times \exp\frac{-1}{2\sigma^2} \underbrace{\begin{Bmatrix} (y-X_2\Delta\boldsymbol{\beta})'(y-X_2\Delta\boldsymbol{\beta}) + \Delta\boldsymbol{\beta}' g_{\hat{A}}\Sigma_X\Delta\boldsymbol{\beta} \\ -(y-X_2\Delta\boldsymbol{\beta})'P_{X_1}(y-X_2\Delta\boldsymbol{\beta}) \end{Bmatrix}}_{C} d(\Delta\boldsymbol{\beta})d\sigma^2.$$

Again, focusing on the expression of the exponential, we obtain

$$\begin{aligned} C &= (y-X_2\Delta\boldsymbol{\beta})'(y-X_2\Delta\boldsymbol{\beta}) + \Delta\boldsymbol{\beta}' g_{\hat{A}}\Sigma_X\Delta\boldsymbol{\beta} - (y-X_2\Delta\boldsymbol{\beta})'P_{X_1}(y-X_2\Delta\boldsymbol{\beta}), \\ &= y'[I_T - P_{X_1}]y + \Delta\boldsymbol{\beta}'[g_{\hat{A}}X_2'M_{X_1}X_2 + X_2'X_2 - X_2'P_{X_1}X_2]\Delta\boldsymbol{\beta} - 2\Delta\boldsymbol{\beta}'X_2'[I_T - P_{X_1}]y, \\ &= y'M_{X_1}y + \Delta\boldsymbol{\beta}'[(1+g_{\hat{A}})X_2'M_{X_1}X_2]\Delta\boldsymbol{\beta} - 2\Delta\boldsymbol{\beta}'X_2'M_{X_1}y, \\ &= y'M_{X_1}y + (\Delta\boldsymbol{\beta}-\bar{\boldsymbol{\mu}})'\bar{\Sigma}^{-1}(\Delta\boldsymbol{\beta}-\bar{\boldsymbol{\mu}}) - \bar{\boldsymbol{\mu}}'\bar{\Sigma}^{-1}\bar{\boldsymbol{\mu}}, \end{aligned}$$

where $\bar{\Sigma}^{-1} = (1+g_{\hat{A}})X_2'M_{X_1}X_2 = (1+g_{\hat{A}})\Sigma_X$ and $\bar{\boldsymbol{\mu}} = \bar{\Sigma}X_2'M_{X_1}y$, $\bar{\boldsymbol{\mu}}'\bar{\Sigma}^{-1}\bar{\boldsymbol{\mu}} = (1+g_{\hat{A}})^{-1}y'M_{X_1}X_2[X_2'M_{X_1}X_2]^{-1}X_2'M_{X_1}y$.

Eventually, we find the following marginal likelihood

$$\begin{aligned} f(\mathbf{y}|\mathbf{a},\lambda,\boldsymbol{\tau}) &= (2\pi)^{\frac{-(T-K)}{2}} |X_1'X_1|^{-\frac{1}{2}} |g_{\hat{A}}\Sigma_X|^{1/2} |(1+g_{\hat{A}})\Sigma_X|^{\frac{-1}{2}} \int (\sigma^2)^{-\frac{(T+2-K)}{2}} \\ &\quad \exp\frac{-1}{2\sigma^2}\{y'M_{X_1}y - (1+g_{\hat{A}})^{-1}y'M_{X_1}X_2[X_2'M_{X_1}X_2]^{-1}X_2'M_{X_1}y\}d\sigma^2, \\ &= (\pi)^{\frac{-(T-K)}{2}}\Gamma(\frac{T-K}{2})|X_1'X_1|^{-\frac{1}{2}} \\ &\quad (\frac{g_{\hat{A}}}{1+g_{\hat{A}}})^{k_{\hat{A}}/2}[y'M_{X_1}y - (1+g_{\hat{A}})^{-1}y'M_{X_1}X_2[X_2'M_{X_1}X_2]^{-1}X_2'M_{X_1}y]^{-\frac{T-K}{2}}, \\ &= (\pi)^{\frac{-(T-K)}{2}}\Gamma(\frac{T-K}{2})|X_1'X_1|^{-\frac{1}{2}}(\frac{g_{\hat{A}}}{1+g_{\hat{A}}})^{k_{\hat{A}}/2} \\ &\quad [\frac{g_{\hat{A}}}{1+g_{\hat{A}}}y'M_{X_1}y + \frac{1}{(1+g_{\hat{A}})}[\tilde{y}'\tilde{y} - \tilde{y}'X_2[X_2'M_{X_1}X_2]^{-1}X_2\tilde{y}]]^{-\frac{T-K}{2}}, \\ &= (\pi)^{\frac{-(T-K)}{2}}\Gamma(\frac{T-K}{2})|X_1'X_1|^{-\frac{1}{2}}(\frac{g_{\hat{A}}}{1+g_{\hat{A}}})^{k_{\hat{A}}/2} \\ &\quad [\frac{g_{\hat{A}}}{1+g_{\hat{A}}}s_{X_1} + \frac{1}{(1+g_{\hat{A}})}s_{X_1,X_2}]^{-\frac{T-K}{2}}, \end{aligned}$$

where the penultimate equality comes from the Frisch-Waugh theorem.

C.2.1 Posterior distribution

$$f(\boldsymbol{\beta}_1,\Delta\boldsymbol{\beta},\sigma^2|\mathbf{y},\boldsymbol{\tau}) \propto (2\pi)^{\frac{-(T+k_{\hat{A}})}{2}}(\sigma^2)^{-\frac{(T+2+k_{\hat{A}})}{2}}|g_{\hat{A}}(X_2)'M_{X_1}X_2)|^{1/2} \exp\frac{-1}{2\sigma^2}\underbrace{\begin{pmatrix}(y-X_1\boldsymbol{\beta}_1-X_2\Delta\boldsymbol{\beta})'(y-X_1\boldsymbol{\beta}_1-X_2\Delta\boldsymbol{\beta}) \\ +\Delta\boldsymbol{\beta}'g_{\hat{A}}(X_2)'M_{X_1}X_2\Delta\boldsymbol{\beta}\end{pmatrix}}_{\text{Exp}}$$

Focusing on the expression of the exponential, we have

$$\begin{aligned}
\text{Exp} &= (y - X_2\Delta\boldsymbol{\beta})'(y - X_2\Delta\boldsymbol{\beta}) + \Delta\boldsymbol{\beta}' g_{\hat{A}} \Sigma_X \Delta\boldsymbol{\beta} + (\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1)'\Omega^{-1}(\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1) - \bar{\boldsymbol{\beta}}_1{}'\Omega^{-1}\bar{\boldsymbol{\beta}}_1, \\
&= y' M_{X_1} y - \bar{\boldsymbol{\mu}}'\bar{\Sigma}^{-1}\bar{\boldsymbol{\mu}} + (\Delta\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})'\bar{\Sigma}^{-1}(\Delta\boldsymbol{\beta} - \bar{\boldsymbol{\mu}}) + (\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1)'\Omega^{-1}(\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1), \\
&= \frac{g_{\hat{A}}}{1 + g_{\hat{A}}} s_{X_1} + \frac{1}{(1 + g_{\hat{A}})} s_{X_1, X_2} + (\Delta\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})'\bar{\Sigma}^{-1}(\Delta\boldsymbol{\beta} - \bar{\boldsymbol{\mu}}) \\
&\quad + (\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1)'\Omega^{-1}(\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1),
\end{aligned}$$

where $\bar{\Sigma}^{-1} = (1 + g_{\hat{A}}) X_2' M_{X_1} X_2 = (1 + g_{\hat{A}})\Sigma_X$ and $\bar{\boldsymbol{\mu}} = \bar{\Sigma} X_2' M_{X_1} y$, $\bar{\boldsymbol{\mu}}'\bar{\Sigma}^{-1}\bar{\boldsymbol{\mu}} = (1 + g_{\hat{A}})^{-1} y' M_{X_1} X_2 [X_2' M_{X_1} X_2]^{-1} X_2' M_{X_1} y$ and $\Omega^{-1} = X_1' X_1$, $\bar{\boldsymbol{\beta}}_1 = (X_1' X_1)^{-1} X_1'(y - X_2\Delta\boldsymbol{\beta})$. The posterior distribution can be decomposed as

$$\begin{aligned}
f(\boldsymbol{\beta}_1, \Delta\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \boldsymbol{\tau}) &= f(\sigma^2 | \mathbf{y}, \boldsymbol{\tau}) f(\Delta\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\tau}, \sigma^2) f(\boldsymbol{\beta}_1 | \mathbf{y}, \boldsymbol{\tau}, \sigma^2, \Delta\boldsymbol{\beta}) \\
&\propto (\sigma^2)^{-\frac{(T+2-K)}{2}} \exp \frac{-1}{\sigma^2} \{ \frac{\frac{g_{\hat{A}}}{1+g_{\hat{A}}} s_{X_1} + \frac{1}{(1+g_{\hat{A}})} s_{X_1, X_2}}{2} \} \\
&\quad (\sigma^2)^{-\frac{(k_{\hat{A}})}{2}} |g_{\hat{A}} (X_2)' M_{X_1} X_2)|^{1/2} \exp \frac{-1}{2\sigma^2} \{ (\Delta\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})'\bar{\Sigma}^{-1}(\Delta\boldsymbol{\beta} - \bar{\boldsymbol{\mu}}) \} \\
&\quad (\sigma^2)^{-\frac{(K)}{2}} \exp \frac{-1}{2\sigma^2} \{ (\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1)'\Omega^{-1}(\boldsymbol{\beta}_1 - \bar{\boldsymbol{\beta}}_1) \}.
\end{aligned}$$

It gives the following posterior distribution

$$\begin{aligned}
\sigma^2 | \mathbf{y}, \boldsymbol{\tau} &\sim \mathcal{IG}(\frac{T - K}{2}, \frac{\frac{g_{\hat{A}}}{1+g_{\hat{A}}} s_{X_1} + \frac{1}{(1+g_{\hat{A}})} s_{X_1, X_2}}{2}), \\
\Delta\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\tau}, \sigma^2 &\sim \mathcal{N}((1 + g_{\hat{A}})^{-1} [X_2' M_{X_1} X_2]^{-1} X_2' M_{X_1} y, \sigma^2 (1 + g_{\hat{A}})^{-1} [X_2' M_{X_1} X_2]^{-1}), \\
\boldsymbol{\beta}_1 | \mathbf{y}, \boldsymbol{\tau}, \sigma^2, \Delta\boldsymbol{\beta} &\sim \mathcal{N}((X_1' X_1)^{-1} X_1'(y - X_2\Delta\boldsymbol{\beta}), \sigma^2 (X_1' X_1)^{-1}).
\end{aligned}$$

C.2.2 Predictive density

In Appendix C.2.1, we derive the following posterior distributions:

$$\begin{aligned}
\sigma^2 | \mathbf{y}, \boldsymbol{\tau} &\sim \mathcal{IG}(\underbrace{\frac{T - K}{2}}_{a_{\sigma^2}}, \underbrace{\frac{\frac{g_{\hat{A}}}{1+g_{\hat{A}}} s_{X_1} + \frac{1}{(1+g_{\hat{A}})} s_{X_1, X_2}}{2}}_{b_{\sigma^2}}), \\
\Delta\boldsymbol{\beta} | \mathbf{y}, \boldsymbol{\tau}, \sigma^2 &\sim \mathcal{N}(\underbrace{(1 + g_{\hat{A}})^{-1} [X_2' M_{X_1} X_2]^{-1} X_2' M_{X_1} y}_{\boldsymbol{\mu}_{\Delta\boldsymbol{\beta}}}, \underbrace{\sigma^2 (1 + g_{\hat{A}})^{-1} [X_2' M_{X_1} X_2]^{-1}}_{\boldsymbol{\Sigma}_{\Delta\boldsymbol{\beta}}}), \\
\boldsymbol{\beta}_1 | \mathbf{y}, \boldsymbol{\tau}, \sigma^2, \Delta\boldsymbol{\beta} &\sim \mathcal{N}(\underbrace{(X_1' X_1)^{-1} X_1'(y - X_2\Delta\boldsymbol{\beta})}_{\boldsymbol{\mu}_{\boldsymbol{\beta}}}, \underbrace{\sigma^2 (X_1' X_1)^{-1}}_{\boldsymbol{\Sigma}_{\boldsymbol{\beta}}}).
\end{aligned}$$

Given these results, we can derive the joint posterior distribution of the variable $\boldsymbol{\psi} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \Delta\boldsymbol{\beta} \end{pmatrix}$. In particular, a standard algebraic calculus leads to

$$\begin{aligned} \boldsymbol{\psi}|\mathbf{y}, \boldsymbol{\tau}, \sigma^2 &\sim \mathcal{N}\left(\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \mathbf{B}\mu_{\Delta\boldsymbol{\beta}} \\ \boldsymbol{\mu}_{\Delta\boldsymbol{\beta}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{\Sigma}_{\beta}^{-1}\mathbf{B} \\ \mathbf{B}'\boldsymbol{\Sigma}_{\beta}^{-1} & [\boldsymbol{\Sigma}_{\Delta\boldsymbol{\beta}}^{-1} + \mathbf{B}'\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\mathbf{B}] \end{pmatrix}^{-1} \right), \\ &\sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\psi}}, \boldsymbol{\Sigma}_{\boldsymbol{\psi}}\right) \end{aligned} \tag{C.17}$$

with $\hat{\boldsymbol{\beta}}_1 = (X_1'X_1)^{-1}X_1'y$ and $\mathbf{B} = (X_1'X_1)^{-1}X_2$. Consequently, the predictive density is given by

$$y_{T+1}|\mathbf{y}, \boldsymbol{\tau}, \sigma^2 \sim \mathcal{N}\left(\mathbf{x}_{T+1}'\boldsymbol{\mu}_{\boldsymbol{\psi}}, \sigma^2(\mathbf{x}_{T+1}'\boldsymbol{\Sigma}_{\psi}\mathbf{x}_{T+1} + 1)\right). \tag{C.18}$$

Since $\sigma^2|\mathbf{y}, \boldsymbol{\tau}$ follows an inverse gamma distribution, the predictive distribution of $y_{T+1}|\mathbf{y}$ is a student distribution. Its density is given by

$$\begin{aligned} f(y_{T+1}|\mathbf{y}, \boldsymbol{\tau}) &= \frac{b_{\sigma^2}^{a_{\sigma^2}}}{\Gamma(a_{\sigma^2})}(2\pi(\mathbf{x}_{T+1}'\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{x}_{T+1} + 1))^{-\frac{1}{2}} \\ &\quad \int (\sigma^2)^{-(a_{\sigma^2}+1+0.5)} \exp(-\frac{(y_{T+1} - \mathbf{x}_{T+1}'\boldsymbol{\mu}_{\boldsymbol{\psi}})^2(\mathbf{x}_{T+1}'\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{x}_{T+1} + 1)^{-1} + 2b_{\sigma^2}}{2\sigma^2})d\sigma^2, \\ &= \frac{b_{\sigma^2}^{a_{\sigma^2}}}{\Gamma(a_{\sigma^2})}(2\pi(\mathbf{x}_{T+1}'\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{x}_{T+1} + 1))^{-\frac{1}{2}}\Gamma(a_{\sigma^2} + 0.5) \\ &\quad \left(\frac{(y_{T+1} - \mathbf{x}_{T+1}'\boldsymbol{\mu}_{\boldsymbol{\psi}})^2(\mathbf{x}_{T+1}'\boldsymbol{\Sigma}_{\boldsymbol{\psi}}\mathbf{x}_{T+1} + 1)^{-1} + 2b_{\sigma^2}}{2}\right)^{-(a_{\sigma^2}+0.5)}, \end{aligned} \tag{C.19}$$

The final expression in Equation (C.19) is equivalent to a student density with expectation $\mathbf{x}_{T+1}'\boldsymbol{\mu}_{\boldsymbol{\psi}}$, scale parameter $\frac{b_{\sigma^2}}{a_{\sigma^2}}(\mathbf{x}_{T+1}'\boldsymbol{\Sigma}_{\psi}\mathbf{x}_{T+1} + 1)$ and degree of freedom equal to $2a_{\sigma^2}$.

C.3 Consistency of the criterion

To prove the theorem, we focus on the ratio of the criterion for two different models $s = (a_s, \lambda_s)$ and $j = (a_j, \lambda_j)$ where s is considered as the true model. To simplify the notation, we denote by $\mathbf{X}_z$ the explanatory variable included by model z (i.e., $\mathbf{X}_z = \mathbf{X}_{\boldsymbol{\tau}}^{\hat{A}_z}$) for $z = s, j$ and $g_{\hat{A}} = g = \frac{1}{w(T)}$ and write the marginal likelihood as $f(\mathbf{y}|a_z, \lambda_z)$ instead of $f(\mathbf{y}|a_z, \lambda_z, \boldsymbol{\tau})$. We need to show that $\frac{f(\mathbf{y}|a_j, \lambda_j)}{f(\mathbf{y}|a_s, \lambda_s)} \to_p 0$ for any $j \neq s$. In particular, we have

$$\frac{f(\mathbf{y}|a_j, \lambda_j)}{f(\mathbf{y}|a_s, \lambda_s)} = \underbrace{\frac{(\frac{g}{1+g})^{k_{\hat{A}_j}/2}}{(\frac{g}{1+g})^{k_{\hat{A}_s}/2}}}_{C_{js}} \underbrace{[\frac{\frac{g}{1+g}s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g)}s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}}{\frac{g}{1+g}s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g)}s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_s}}]^{-\frac{T-K}{2}}}_{D_{js}}. \tag{C.20}$$

Focusing on the first term, it is easy to show that

$$\begin{aligned} C_{js} &= \frac{(1+g)^{k_{\hat{A}_s}/2}}{(1+g)^{k_{\hat{A}_j}/2}} g^{\frac{k_{\hat{A}_j}-k_{\hat{A}_s}}{2}} \\ &= \frac{(1+w(T)^{-1})^{k_{\hat{A}_s}/2}}{(1+w(T)^{-1})^{k_{\hat{A}_j}/2}} w(T)^{\frac{k_{\hat{A}_s}-k_{\hat{A}_j}}{2}} \\ &= \mathcal{O}(w(T)^{\frac{k_{\hat{A}_s}-k_{\hat{A}_j}}{2}}). \end{aligned}$$

When $T \to \infty$, we have

$$\begin{aligned} C_{js} &= 0 \text{ when } k_{\hat{A}_s} < k_{\hat{A}_j}, \\ &= 1 \text{ if } k_{\hat{A}_s} = k_{\hat{A}_j}, \\ &\to +\infty \text{ when } k_{\hat{A}_s} > k_{\hat{A}_j}. \end{aligned}$$

We now discuss three possible cases.

1. $k_{\hat{A}_s} < k_{\hat{A}_j}$ and the model j does not nest the model s. In such case, the term $C_{js} \to 0$. The second term also tends to zero since we have

$$\begin{aligned} D_{js} &= [\frac{\frac{g}{1+g} s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g)} s_{\tilde{\mathbf{X}}_{\tau_0}, \mathbf{X}_s}}{\frac{g}{1+g} s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g)} s_{\tilde{\mathbf{X}}_{\tau_0}, \mathbf{X}_j}}]^{\frac{T-K}{2}} \\ &= [\frac{g s_{\tilde{\mathbf{X}}_{\tau_0}} + s_{\tilde{\mathbf{X}}_{\tau_0}, \mathbf{X}_s}}{g s_{\tilde{\mathbf{X}}_{\tau_0}} + s_{\tilde{\mathbf{X}}_{\tau_0}, \mathbf{X}_j}}]^{\frac{T-K}{2}} \end{aligned}$$

   Using he fact that M_j does not nest M_s and the Frisch-Waugh theorem (see also Lemma A.1 in Fernandez et al. (2001)), we have that $\lim_{T\to\infty} \frac{s_{\tilde{\mathbf{X}}_{\tau_0}, \mathbf{X}_j}}{T} = \sigma^2 + b_j$ with $b_j > 0$. Combining with the fact that $g \to 0$, we end up with a limit of D_{js} given by

$$\lim_{T\to\infty} D_{js} = [\frac{\sigma^2}{\sigma^2 + b_j}]^{\frac{T-K}{2}} \to 0.$$

2. The model j does not nest the true model but $K_j < K_s$. In such case, the term $C_{js} \to +\infty$. However, we can show that $\lim_{T\to\infty} C_{js} w(T)^{-\frac{(K_s-K_j)}{2} + \frac{K_s-K_j}{T-K}} \to 1$. Indeed, we have that

$$\lim_{T\to\infty} C_{js} w(T)^{-\frac{(K_s-K_j)}{2} + \frac{K_s-K_j}{T-K}} = \frac{(1+w(T)^{-1})^{k_{\hat{A}_s}/2}}{(1+w(T)^{-1})^{k_{\hat{A}_j}/2}} w(T)^{\frac{K_s-K_j}{T-K}}.$$

   Let us define $q_T = w(T)^{\frac{K_s-K_j}{T-K}}$. We can compute the limit as follows $\lim_{T\to\infty} w(T)^{\frac{K_s-K_j}{T-K}} = \lim_{T\to\infty} \exp \ln q_T$. The limit of $\ln q_T$ is given by

$$\begin{aligned} \lim_{T\to\infty} q_T &= \lim_{T\to\infty} \frac{K_s - K_j}{T-K} \ln w(T), \\ &= \lim_{T\to\infty} \frac{w'(T)}{w(T)} \quad (= 0 \quad \text{by assumption}). \end{aligned}$$

We conclude that $\lim_{T\to\infty} w(T)^{\frac{K_s-K_j}{T-K}} = 1$.

Now, we need to show that $D_{js}w(T)^{\frac{(K_s-K_j)}{2}-\frac{K_s-K_j}{T-K}} \to 0$. In fact, we have

$$\begin{aligned} D_{js}w(T)^{\frac{\frac{(K_s-K_j)}{2}(T-K-2)}{T-K}} &= \lim_{T\to\infty} \underbrace{(\frac{\sigma^2}{\sigma^2+b_j})}_{a<1}^{\frac{T-K}{2}} w(T)^{\frac{(K_s-K_j)}{2}} \\ &= \lim_{T\to\infty} \frac{w(T)^{\frac{(K_s-K_j)}{2}}}{a^{-\frac{T-K}{2}}}, \end{aligned}$$

By applying $\left\lceil \frac{(K_s-K_j)}{2} \right\rceil$ times the Hospital's rule, we find that $a^{\frac{T-K}{2}}$ dominates and so $D_{js}w(T)^{\frac{\frac{(K_s-K_j)}{2}(T-K-2)}{T-K}} \to 0$.

3. We now consider the last case in which the model j nests the true model s. Consequently, we have $K_s < K_j$ and the term $C_{js} \to 0$. Regarding the other term, we can express it as

$$\begin{aligned} D_{js} &= [\frac{gs_{\tilde{\mathbf{X}}_{\tau_0}} + s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_s}}{gs_{\tilde{\mathbf{X}}_{\tau_0}} + s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}}]^{\frac{T-K}{2}}, \\ &= \underbrace{[\frac{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_s}}{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}}]^{\frac{T-K}{2}}}_{Q_1} \underbrace{[\frac{A_s + w(T)}{A_j + w(T)}]^{\frac{T-K}{2}}}_{Q2}, \end{aligned}$$

where $A_i = \frac{s_{\tilde{\mathbf{X}}_{\tau_0}}}{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_i}}$ for $i = j, s$. It is clear that the first term Q_1 has a limiting distribution related to the likelihood ratio test. In fact, we have that

$$\begin{aligned} \frac{T-K}{2} \ln \frac{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_s}}{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}} &= \underbrace{\frac{T-K}{2T}}_{\to\frac{1}{2}} \underbrace{T \ln \frac{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_s}}{s_{\tilde{\mathbf{X}}_{\tau_0},\mathbf{X}_j}}}_{\to_d \chi^2(\Delta_{js})}, \\ &\to_d \quad \text{Gamma}(\frac{\Delta_{js}}{2}, 1), \end{aligned}$$

in which $\Delta_{js} = |K_s - K_j|$. Since $Y \sim \text{Gamma}\left(\frac{\Delta_{js}}{2}, 1\right)$ is $\mathcal{O}_p(1)$, we have that $C_{js} \exp Y \to_p 0$.

Focusing on the second term Q_2, using assumption (iii), we have that

$$\begin{aligned} \frac{T-K}{2} \ln[\frac{A_s + w(T)}{A_j + w(T)}] &= \frac{T-K}{2} \ln[1 + \frac{A_s - A_j}{A_j + w(T)}], \\ &= \mathcal{O}_p(\frac{T}{w(T)}). \\ &\to_p [0, \infty). \end{aligned}$$

Since $C_{js} \to 0$, we conclude that $C_{js}Q_1Q_2 \to_p 0$.

C.3.1 Convergence to the BIC

The BIC of a linear regression model with K parameters is given by

$$\begin{aligned} BIC(K) &= -\frac{T}{2}\ln(\frac{s_{\mathbf{X}}}{T}) - \frac{K}{2}\ln T, \\ &= \underbrace{-\frac{T}{2}\ln(s_{\mathbf{X}}) - \frac{K}{2}\ln T}_{BIC^*(K)} + \frac{T}{2}\ln T, \end{aligned} \tag{C.21}$$

in which $s_{\mathbf{X}}$ denotes the sum of squared residuals given the $T \times K$ dimensional exogenous variables $\mathbf{X}$ evaluated at the OLS estimates. In this appendix, we show that the logarithm of the marginal likelihood and the BIC$^*(\alpha k_{\hat{A}})$ converges in probability to 0 when $g_{\hat{A}} = \frac{1}{T^\alpha}$ with $\alpha > 1$. In particular, the marginal likelihood is given by

$$f(\mathbf{y}|\mathbf{a},\lambda,\boldsymbol{\tau}) = (\frac{g_{\hat{A}}}{1+g_{\hat{A}}})^{k_{\hat{A}}/2}[\frac{g_{\hat{A}}}{1+g_{\hat{A}}}s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{1}{(1+g_{\hat{A}})}s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}]^{-\frac{T-K}{2}}. \tag{C.22}$$

We have the following results:

$$\begin{aligned} f(\mathbf{y}|\mathbf{a},\lambda,\boldsymbol{\tau}) &= T^{-\frac{\alpha k_{\hat{A}}}{2}}[\frac{1}{T^\alpha}s_{\tilde{\mathbf{X}}_{\tau_0}} + \frac{T^\alpha-1}{T^\alpha}s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}]^{-\frac{T-K}{2}}, \\ &= T^{-\frac{\alpha k_{\hat{A}}}{2}}[s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}]^{-\frac{T-K}{2}}[\frac{T^\alpha-1}{T^\alpha}]^{-\frac{T-K}{2}}[\frac{1}{T^\alpha-1}\frac{s_{\tilde{\mathbf{X}}_{\tau_0}}}{s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}}+1]^{-\frac{T-K}{2}}, \\ &= T^{-\frac{\alpha k_{\hat{A}}}{2}}[s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}]^{-\frac{T-K}{2}}\underbrace{[1-\frac{1}{T^\alpha}]^{-\frac{T-K}{2}}}_{C_1}\underbrace{[\frac{1}{T^\alpha-1}\frac{s_{\tilde{\mathbf{X}}_{\tau_0}}}{s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}}+1]^{-\frac{T-K}{2}}}_{C_2}, \end{aligned} \tag{C.23}$$

$$\ln f(\mathbf{y}|\mathbf{a},\lambda,\boldsymbol{\tau}) = -\frac{T-K}{T}\frac{T}{2}\ln s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}} - \frac{\alpha k_{\hat{A}}}{2}\ln T + \ln C_1 + \ln C_2$$

We now show that the two quantities, i.e. C_1 and C_2, tends to 1 when $T \to +\infty$:

$$\begin{aligned} C_1 &= \exp\left(-\frac{T-K}{2}\ln[1-\frac{1}{T^\alpha}]\right), \\ &\to_p 1 \text{ since } \alpha > 1, \\ C_2 &= \exp\left(-\frac{T-K}{2}\ln[\frac{1}{T^\alpha-1}\frac{s_{\tilde{\mathbf{X}}_{\tau_0}}}{s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}}+1]\right), \\ &\to_p 1 \text{ since } \alpha > 1 \text{ and } \frac{s_{\tilde{\mathbf{X}}_{\tau_0}}}{s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}}} = \mathcal{O}_p(1), \end{aligned} \tag{C.24}$$

It follows that

$$\begin{aligned} \ln f(\mathbf{y}|\mathbf{a},\lambda,\boldsymbol{\tau}) - \left(-\frac{T-K}{T}\frac{T}{2}\ln s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}} - \frac{\alpha k_{\hat{A}}}{2}\ln T\right) &\xrightarrow{p} 0 \\ \ln f(\mathbf{y}|\mathbf{a},\lambda,\boldsymbol{\tau}) - \left(-\frac{T}{2}\ln s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}} - \frac{\alpha k_{\hat{A}}}{2}\ln T\right) &\xrightarrow{p} 0 \\ \frac{f(\mathbf{y}|\mathbf{a},\lambda,\boldsymbol{\tau})}{\exp\left(-\frac{T}{2}\ln s_{\tilde{\mathbf{X}}_{\tau_0},\tilde{\mathbf{X}}^{\hat{A}}_{\boldsymbol{\tau}}} - \frac{\alpha k_{\hat{A}}}{2}\ln T\right)} &\xrightarrow{p} 1 \end{aligned} \tag{C.25}$$

As a result, the models selected using the marginal likelihood are equivalent (asymptotically) to those of the BIC with $\alpha k_{\hat{A}}$ parameters. In addition, the posterior probabilities computed using the marginal likelihood converge to the posterior probabilities that would have been computed using the BIC with $\alpha k_{\hat{A}}$ parameters (since the term $\frac{T}{2}\ln T$ cancels out).

C.4 Time-varying parameter model

We also consider a standard time-varying parameter process (see Primiceri (2005)). The model specification is given by

$$
\begin{aligned}
y_t &= \mathbf{x}_t'\boldsymbol{\beta}_0 + \mathbf{x}_t'\text{diag}(\boldsymbol{\omega})\boldsymbol{\beta}_t + \sigma_t\epsilon_t, && \text{(C.26)}\\
\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1} &\sim \mathcal{N}(\boldsymbol{\beta}_{t-1}, I_K), && \text{(C.27)}\\
\ln\sigma_t^2 &= \ln\sigma_{t-1}^2 + \eta_t, \text{for } t > 0, && \text{(C.28)}\\
\ln\sigma_0^2 &\sim N(0,1), && \text{(C.29)}
\end{aligned}
$$

where $\boldsymbol{\omega} = (\omega_1, ..., \omega_K)'$, $\eta_t \sim \mathcal{N}(0, q)$ with $q \sim \mathcal{IG}(\underline{q}_1 = 3, \underline{q}_2 = 20)$, $(\boldsymbol{\beta}_0', \boldsymbol{\omega}')' \sim \mathcal{N}(0, I_{2K})$ and K stands for the number of explanatory variables. We also define $\ln\sigma_{1:T} = (\ln\sigma_1^2, \ldots, \ln\sigma_T^2)'$ and $y_{1:T}$, $\boldsymbol{\beta}_{0:T}$ analogously. In order to take into account the autocorrelation structure, we use the same lag orders as exposed in Table 3.6. The other explanatory variables consist in an intercept and the seven factors.

The model parameters can be estimated with a standard Markov-chain Monte Carlo (e.g., Bitto and Frühwirth-Schnatter, 2019). In particular, the model parameters are estimated using an MCMC algorithm which consists of the following steps:

- Sampling $\ln\sigma_{1:T}$ using the offset mixture approach of Kim et al. (1998). In particular, given $\boldsymbol{\beta}_{0:T}$ and $\boldsymbol{\omega}$, we compute $y_t^* = \ln(\nu_t^2 + c)$, for all $t = 1, \ldots, T$ in which $\nu_t = y_t - \mathbf{x}_t(\boldsymbol{\beta}_0 + \text{diag}(\omega_1, ..., \omega_K)\boldsymbol{\beta}_t)$ and $c = 0.0001$. The model boils down to a standard TVP model with time-varying intercept since we have

$$
\begin{aligned}
y_t^* &= \ln\sigma_t^2 + \ln\epsilon_t^2, && \text{(C.30)}\\
\ln\sigma_t^2 &= \ln\sigma_{t-1}^2 + \eta_t. && \text{(C.31)}
\end{aligned}
$$

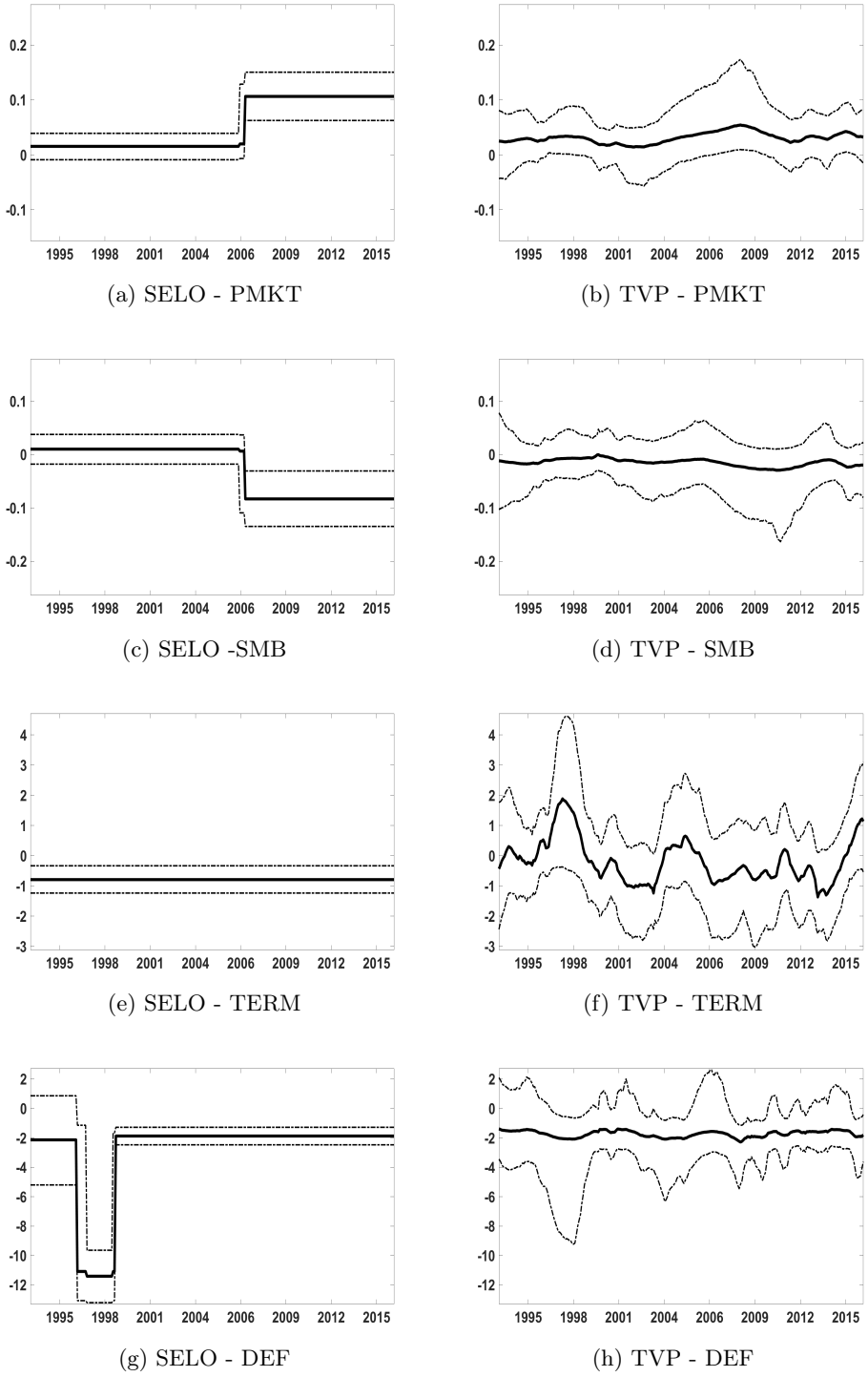
  Approximating the distribution $\ln\epsilon_t^2$ with a 8-component mixture of normal distributions, we sample the time-varying variance from a high-dimensional multivariate normal distribution using the sampler 'all without a loop' (AWOL) as suggested in McCausland et al. (2011) (see also Kastner and Frühwirth-Schnatter, 2014, for a simple exposition of the algorithm).
- Sampling $\ln\sigma_0^2|\ln\sigma_1^2, q \sim N((q^{-1}+1)^{-1}\frac{\ln\sigma_1^2}{q}, (q^{-1}+1)^{-1})$.
- Sampling $q|\ln\sigma_{0:T} \sim IG(\underline{q}_1 + \frac{T}{2}, \underline{q}_2 + \sum_{t=1}^{T}\frac{(\ln\sigma_t^2 - \ln\sigma_{t-1}^2)^2}{2})$.

- Sampling $\boldsymbol{\beta}_{1:T}|y_{1:T}, \boldsymbol{\beta}_0, \boldsymbol{\omega}, \ln \sigma_{1:T}$ using a Kalman filter. Note that this step could also be carried out with the AWOL approach.
- Sampling $(\boldsymbol{\beta}_0', \boldsymbol{\omega}')'|y_{1:T}, \boldsymbol{\beta}_{1:T}, \ln \sigma_{1:T}$ from a multivariate normal distribution. In fact, conditioning to $\boldsymbol{\beta}_{1:T}$ and $\ln \sigma_{1:T}$, the model boils down to a standard regression model since we have

$$\begin{aligned} y_t &= \mathbf{x}_t'\boldsymbol{\beta}_0 + (\mathbf{x}_t'\text{diag}(\boldsymbol{\beta}_t))\boldsymbol{\omega} + \sigma_t\epsilon_t, & \text{(C.32)} \\ &= \mathbf{x}_t'\boldsymbol{\beta}_0 + \mathbf{z}_t'\boldsymbol{\omega} + \sigma_t\epsilon_t, & \text{(C.33)} \end{aligned}$$

  in which $\mathbf{z}_t = (\mathbf{x}_t'\text{diag}(\boldsymbol{\beta}_t))'$.

C.4.1 Time-varying parameters computed with the FIA returns



(a) SELO - PMKT

(b) TVP - PMKT

(c) SELO -SMB

(d) TVP - SMB

(e) SELO - TERM

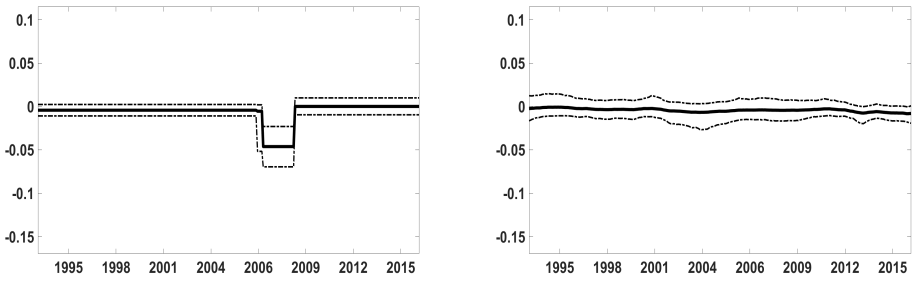
(f) TVP - TERM

(g) SELO - DEF

(h) TVP - DEF

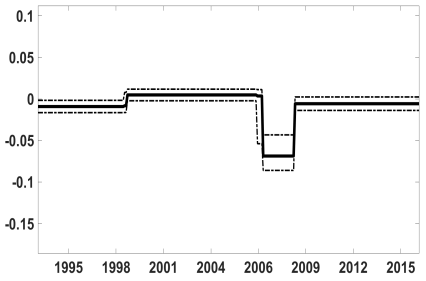
Figure C.2 – FIA returns - Selective segmentation (SELO) model and Time-varying parameter (TVP) model

Posterior medians (black) and the 90% credible intervals (dotted black lines) of the model parameters over time. For the SELO method, we take the break uncertainty into account using the MCMC algorithm presented in Section 3.5.2.

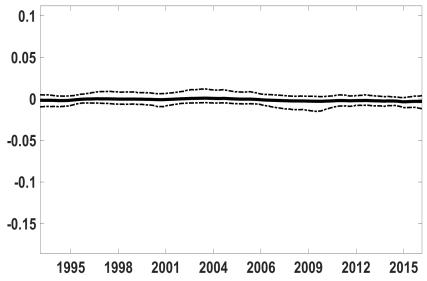


(a) SELO - PTFSBD

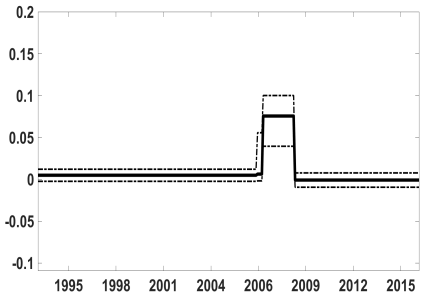
(b) TVP - PTFSBD



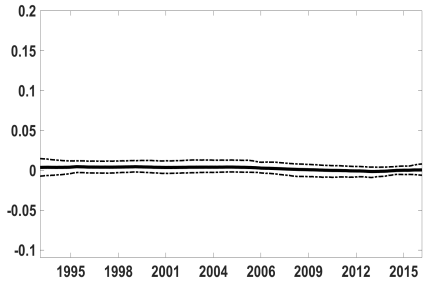
(c) SELO - PTFSFX



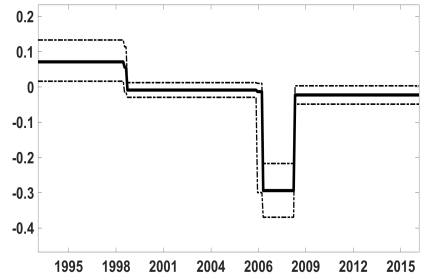
(d) TVP - PTFSFX



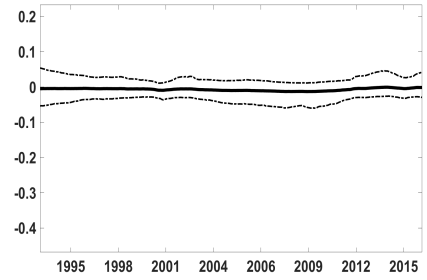
(e) SELO - PTFSCOM



(f) TVP - PTFSCOM



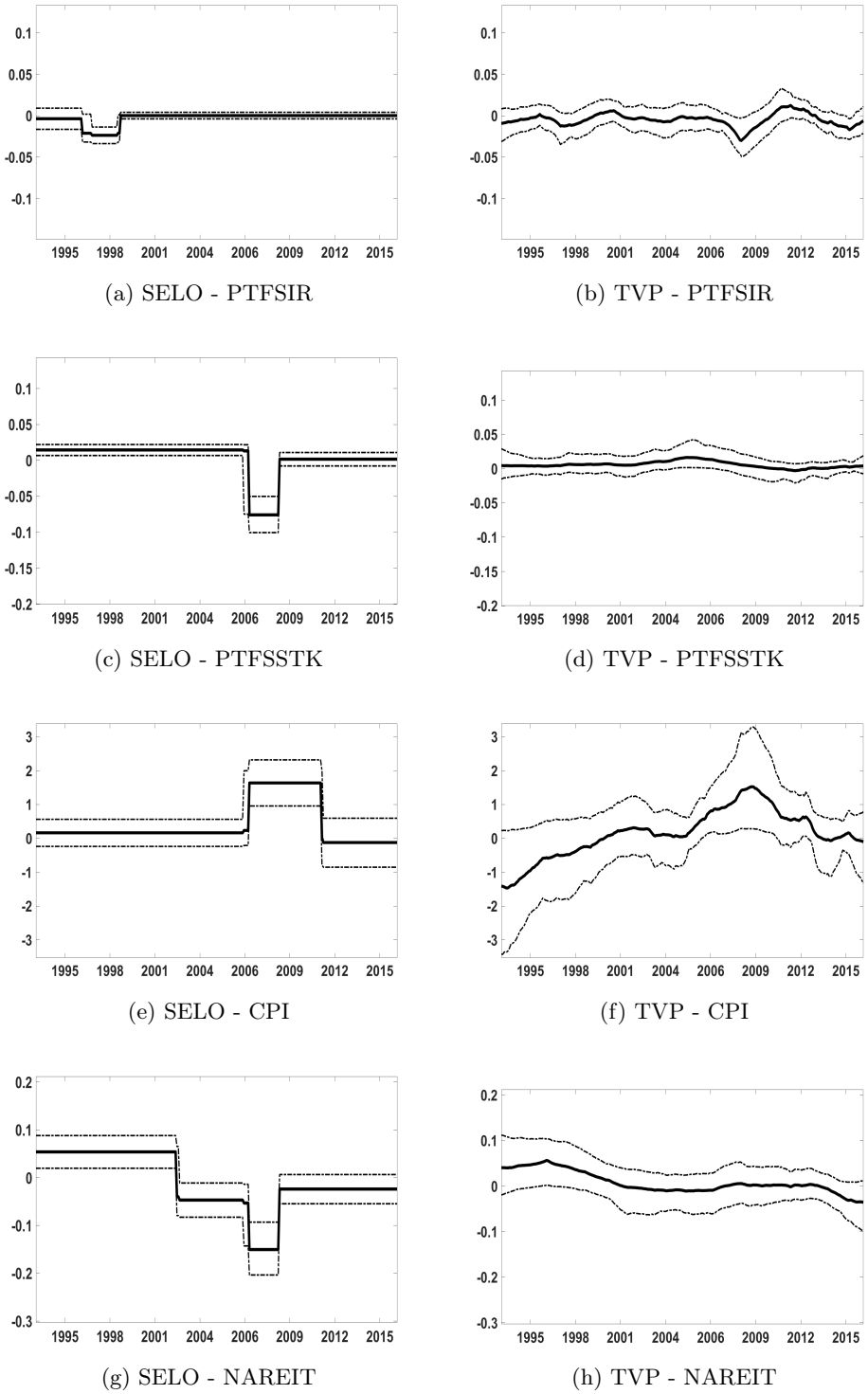
(g) SELO - UMD



(h) TVP - UMD

Figure C.3 – FIA returns - Selective segmentation (SELO) model and Time-varying parameter (TVP) model (2)

Posterior medians (black) and the 90% credible intervals (dotted black lines) of the model parameters over time. For the SELO method, we take the break uncertainty into account using the MCMC algorithm presented in Section 3.5.2



(a) SELO - PTFSIR

(b) TVP - PTFSIR

(c) SELO - PTFSSTK

(d) TVP - PTFSSTK

(e) SELO - CPI

(f) TVP - CPI

(g) SELO - NAREIT

(h) TVP - NAREIT

Figure C.4 – FIA returns - Selective segmentation (SELO) model and Time-varying parameter (TVP) model (3)

Posterior medians (black) and the 90% credible intervals (dotted black lines) of the model parameters over time. For the SELO method, we take the break uncertainty into account using the MCMC algorithm presented in Section 3.5.2.

C.5 Bayesian alternatives to the selective segmentation method

The DGP J exhibits 100 explanatory variables and one CP in ten parameters. To capture which model parameters are experiencing a break, we are aware of four Bayesian alternatives that are Giordani and Kohn (2008), Eo (2016), Huber et al. (2019) and Dufays and Rombouts (2020). We now discuss why these alternatives do not work when 100 explanatory variables are involved.

Giordani and Kohn (2008)

The model of Giordani and Kohn (2008) stands for a particular case of the general mixture state space model developed in Gerlach et al. (2000). They explain how to do inference for a Gaussian state space model with a latent variable K_t that determines the state of the model parameters. Modelling breaks in the mean parameters, their approach allows estimating the following state space model:

$$y_t = \beta_{t,1} + \beta_{t,2}x_{t,2} + \ldots + \beta_{t,N}x_{t,N} + \sigma\eta_t, \tag{C.34}$$

$$\beta_{t,i} = \beta_{t-1,i} + \gamma_{i,K_{t,i}}\nu_{t,i}, \text{ for } i = 1, \ldots, N, \text{ and } t > 1, \tag{C.35}$$

in which N is the number of explanatory variables, $\eta_t \sim N(0,1)$, $K_t = \{K_{t,1}, \ldots, K_{t,N}\}$ and $\nu_t = (\nu_{t,1}, \ldots, \nu_{t,N})' \sim N(0, I_N)$ (with I_N, the identity matrix of dimension N). To capture breakpoints, Giordani and Kohn (2008) suggest to set the states of the latent variable $K_{t,i}$ to $\{0,1\}$ such that we have $\gamma_{i,0} = 0$ (i.e., no break when $K_{t,i} = 0$) and $\gamma_{i,1} \in \Re^+$ (i.e., break in the ith parameter when $K_{t,i} = 1$). The model parameters are given by $\boldsymbol{\theta} = \{\gamma_{1,1}, \ldots, \gamma_{N,1}, \beta_{1,1}, \ldots, \beta_{1,N}, \sigma\}$.

To efficiently estimate the model, Giordani and Kohn (2008) relies on the algorithm of Gerlach et al. (2000). The main contribution of Gerlach et al. (2000) is to marginalize out the mean parameters $\beta_{1:T,1:N}$ and to provide an analytical formula for the latent variable distribution $f(K_t|y_{1:T}, K_{\neq t}, \boldsymbol{\theta})$ from which K_t is sampled in the MCMC algorithm. To normalize the posterior distribution $f(K_t|y_{1:T}, K_{\neq t}, \boldsymbol{\theta})$, it requires to sum over all the possible values of K_t. Because $K_t = \{K_{t,1}, \ldots, K_{t,N}\}$ and $K_{t,i} = \{0,1\}$ $\forall i \in [1, N]$, the number of possible values for K_t amounts to 2^N. Consequently, it increases geometrically with the number of explanatory variables. This is why Chan et al. (2012) on page 9 argue that the number of explanatory variables should be small (i.e., at least below fourteen) otherwise some structure on the break dynamic should be accounted for. With 100 regressors involved in DGP J, it is infeasible to compute the latent variable distribution $f(K_t|y_{1:T}, K_{\neq t}, \boldsymbol{\theta})$ because its normalization requires to sum over 2^{100} values.

Eo (2016)

Eo (2016) relies on the approach of Chib (1998) for finding which parameters are experiencing a break. The method consists of estimating all the possible models given several number of breakpoints. Then, the best specification is selected by maximizing the marginal likelihood that is computed, for each model, using the method of Chib (1995). Considering DGP J and its 100 exogenous variables, the number of models to estimate reaches $\sum_{i=0}^{\bar{m}} 2^{100i}$ in which $\bar{m}$ is the maximum number of breaks that can experience a parameter. In our context, the approach is computationally infeasible even when the upper bound of the number of break is equal to 1.

Huber et al. (2019)

Huber et al. (2019) propose a threshold approach to approximate the MCMC inference of mixture state space models. It generalizes the method of Giordani and Kohn (2008) because it is not limited by the number of explanatory variables. As illustrated in Appendix D of Dufays et al. (2020), the approximation makes the MCMC inference depending on the starting value and the estimated breakpoints are unstable from one estimation to another. Consequently, in-sample results and forecasting exercises will also depend on starting values. Because the question on how to choose the starting values is not addressed in the paper, the method does not provide reproducible results. However, it could be useful for exploring the space in order to find a promising starting value to be used in the MCMC algorithm of Giordani and Kohn (2008).

Dufays and Rombouts (2020)

Dufays and Rombouts (2020) rely on the standard CP model (see, e.g., Chib (1998), Pesaran et al. (2006) or Maheu and Song (2014)) to capture which parameters are time-varying when a break is detected. They specify the model parameters in first-difference with respect to the previous regime. By doing so, shrinkage priors can be used to infer which parameters are time-varying. The two main contributions of the paper are i) the introduction of a shrinkage prior that is a 2-component mixture of Uniform distributions (hereafter, 2MU) and ii) a method that operates for models exhibiting the path dependence issue such as ARMA and GARCH processes.

The new shrinkage prior mimics the standard information criteria such as the AIC and the BIC because one hyper-parameter of the 2MU distribution acts like a penalty on the log-likelihood. Consequently, the 2MU prior can be seen as a Bayesian alternative to the popular L_0 penalty functions used in classical statistics. However, the 2MU prior is not suited for high-dimensional regressions because it is not continuous. To mitigate this problem, Dufays and Rombouts (2020) propose a sequential Monte Carlo algorithm, which is known to explore multi-modal distributions more efficiently than MCMC algorithms based on a single chain

(see, e.g., Herbst and Schorfheide, 2014). Unfortunately, this algorithm is computationally intensive. While it takes around 10 minutes on a 6-CORE i5-8400 (2.8 Ghz) for estimating a series from DGPs A to F of our paper, it runs for 2.5 hours on the same computer for estimating one series similar to DGP J but with only 20 explanatory variables. Consequently, it is computationally infeasible to estimate the model of Dufays and Rombouts (2020) on 100 series from DGP J as we do with the selective segmentation approach. As a comparison, the selective segmentation method requires 20 minutes on the same computer for detecting which parameters are time-varying in one simulated series from DGP J.

Bibliography

Agarwal, V. and Naik, N. (2004). Risks and portfolio decisions involving hedge funds. *Review of Financial Studies*, 17:63–98.

Aguirregabiria, V. and Mira, P. (2007). Sequential estimation of dynamic discrete games. *Econometrica*, 75(1):1–53.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Alidaee, H., Auerbach, E., and Leung, M. P. (2020). Recovering network structure from aggregated relational data using penalized regression. *arXiv preprint arXiv:2001.06052*.

Ambrose, B. . W. and D'Lima, W. (2016). Real estate risk and hedge fund returns. *Journal of Real Estate Finance and Economics*, 52(3):197–225.

Ambrose, B. W., Lee, D. W., and Peek, J. (2007). Comovement after joining an index: Spillovers of nonfundamental effects. *Real Estate Economics*, 35(1):57–90.

Amemiya, T. (1981). Qualitative response models: A survey. *Journal of economic literature*, 19(4):1483–1536.

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica: Journal of the Econometric Society*, pages 821–856.

Ardia, D., Dufays, A., and Ordas, C. (2019). Change-point segmentation: the bayesian bridge. *Mimeo*.

Atchadé, Y. F. and Rosenthal, J. S. (2005). On adaptive markov chain monte carlo algorithms. *Bernoulli*, 11(5):815–828.

Baetschmann, G., Staub, K. E., and Winkelmann, R. (2015). Consistent estimation of the fixed effects ordered logit model. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 685–703.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural breaks. *Econometrica: Journal of the Econometric Society*, 66(1):47–48.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of applied econometrics*, 18(1):1–22.

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2006). Who's who in networks. wanted: The key player. *Econometrica*, 74(5):1403–1417.

Bauwens, L., Dufays, A., and De Backer, B. (2011). Estimating and forecasting structural breaks in financial time series. *Journal of Empirical Finance*, Forthcoming:DOI: 10.1016/j.jempfin.2014.06.008.

Bauwens, L., Koop, G., Korobilis, D., and Rombouts, J. (2015). The contribution of structural break models to forecasting macroecomic series. *Journal of Applied Econometrics*, 30(4):596–620.

Bellman, R. (2013). *A brief introduction to theta functions*. Courier Corporation.

Bhamidi, S., Bresler, G., and Sly, A. (2008). Mixing time of exponential random graphs. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 803–812. IEEE.

Bitto, A. and Frühwirth-Schnatter, S. (2019). Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics*, 210(1):75–97.

Blitzstein, J. and Diaconis, P. (2011). A sequential importance sampling algorithm for generating random graphs with prescribed degrees. *Internet mathematics*, 6(4):489–522.

Bollen, N. and Whaley, R. E. (2009). Hedge fund risk dynamics: Implications for performance appraisal. *Journal of Finance*, 64:985–1035.

Boucher, V. (2016). Conformism and self-selection in social networks. *Journal of Public Economics*, 136:30–44.

Boucher, V., Dedewanou, F. A., and Dufays, A. (2018). Peer-induced beliefs regarding college participation. *Working Paper*.

Boucher, V. and Fortin, B. (2016). Some challenges in the empirics of the effects of networks. *The Oxford Handbook on the Economics of Networks*, pages 277–302.

Boucher, V. and Houndetoungan, A. (2020). *Estimating peer effects using partial network data*. Centre de recherche sur les risques les enjeux économiques et les politiques.

Boucher, V. and Mourifié, I. (2017). My friend far, far away: a random field approach to exponential random graph models. *The econometrics journal*, 20(3):S14–S46.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2009). Identification of peer effects through social networks. *Journal of econometrics*, 150(1):41–55.

Bramoullé, Y., Djebbari, H., and Fortin, B. (2020). Peer effects in networks: A survey. *Annual Review of Economics*, 12:603–629.

Breza, E. (2016). Field experiments, social networks, and development. *The Oxford Handbook on the Economics of Networks*, pages 412–439.

Breza, E., Chandrasekhar, A. G., McCormick, T. H., and Pan, M. (2020). Using aggregated relational data to feasibly identify network structure without network data. *American Economic Review*, 110(8):2454–84.

Brock, W. A. and Durlauf, S. N. (2001). Discrete choice with social interactions. *The Review of Economic Studies*, 68(2):235–260.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455.

Calvó-Armengol, A., Patacchini, E., and Zenou, Y. (2009). Peer effects and social networks in education. *The Review of Economic Studies*, 76(4):1239–1267.

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

Campbell, J. Y. (2000). Asset pricing at the millennium. *Journal of Finance*, 50(4):1515–1567.

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance*, 52(1):57–82.

Carmichael, B. and Coen, A. (2018). Real estate and consumption growth as common risk factors in asset pricing models. *Real Estate Economics*, 46(4):936–970.

Chan, J. C., Koop, G., Leon-Gonzalez, R., and Strachan, R. W. (2012). Time varying dimension models. *Journal of Business & Economic Statistics*, 30(3):358–367.

Chan, N. H., Yau, C. Y., and Zhang, R.-M. (2014). Group lasso for structural break time series. *Journal of the American Statistical Association*, 109(506):590–599.

Chandrasekhar, A. and Lewis, R. (2011). Econometrics of sampled networks. *Unpublished manuscript, MIT.[422]*.

Chatterjee, S., Diaconis, P., et al. (2013). Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461.

Chen, N.-F., Roll, R., and Ross, S. A. (1986). Economic forces and the stock market. *Journal of business*, pages 383–403.

Chen, X., Chen, Y., and Xiao, P. (2013). The impact of sampling and network topology on the estimation of social intercorrelations. *Journal of Marketing Research*, 50(1):95–110.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, 35(3):999–1018.

Chernozhukov, V. and Hong, H. (2003). An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90:1313–1321.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86:221–241.

Chib, S. and Ramamurthy, S. (2010). Tailored randomized block mcmc methods with application to dsge models. *Journal of Econometrics*, 155(1):19–38.

Cho, H. and Fryzlewicz, P. (2015). Multiple change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507.

Chow, Y. S. and Teicher, H. (2003). *Probability theory: independence, interchangeability, martingales*. Springer Science & Business Media.

Conley, T. G. and Udry, C. R. (2010). Learning about a new technology: Pineapple in ghana. *American economic review*, 100(1):35–69.

Consul, P. C. and Jain, G. C. (1973). A generalization of the poisson distribution. *Technometrics*, 15(4):791–799.

De Paula, A. (2017). Econometrics of network models. In *Advances in Economics and Econometrics: Theory and Applications: Eleventh World Congress*, volume 1, pages 268–323. Cambridge University Press Cambridge.

De Paula, Á., Rasul, I., and Souza, P. (2018a). Recovering social networks from panel data: identification, simulations and an application. *LACEA Working Paper Series*.

De Paula, Á., Richards-Shubik, S., and Tamer, E. (2018b). Identifying preferences in networks with bounded degree. *Econometrica*, 86(1):263–288.

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.

Dicker, L., Huang, B., and Lin, X. (2013). Variable selection and estimation with the seamless-l 0 penalty. *Statistica Sinica*, 23:929–962.

Dufays, A. and Rombouts, J. V. (2020). Relevant parameter changes in structural break models. *Journal of Econometrics, https://doi.org/10.1016/j.jeconom.2019.10.008*.

Dufays, A., Zhuo, L., Rombouts, J., and Song, Y. (2020). Sparse change-point var models. *Available at SSRN: https://ssrn.com/abstract=3461692*.

Eckley, I. A., Fearnhead, P., and Killick, R. (2011). *Analysis of changepoint models*, chapter 10, pages 205–224. Cambridge University Press, Cambridge.

Eo, Y. (2016). Structural changes in inflation dynamics: multiple breaks at different dates for different parameters. *Studies in Nonlinear Dynamics & Econometrics*, 20(3):211–231.

Erbe, W. (1962). Gregariousness, group membership, and the flow of information. *American Journal of Sociology*, 67(5):502–516.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Fan, J., Peng, H., et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of Royal Statistical Society, Series B*, 69 (4):589–605.

Fernandez, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381–427.

Fortin, B. and Boucher, V. (2015). Some challenges in the empirics of the effects of networks. In *The Oxford Handbook of the Economics of Networks*. Oxford University Press.

Fortin, B. and Yazbeck, M. (2015). Peer effects, fast food consumption and adolescent weight gain. *Journal of health economics*, 42:125–138.

Fryzlewicz, P. et al. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281.

Fujimoto, K. and Valente, T. W. (2013). Alcohol peer influence of participating in organized school activities: a network approach. *Health Psychology*, 32(10):1084.

Fung, W. and Hsieh, D. (2001). The risk in hedge fund strategies: theory and evidence from trend followers. *The review of financial studies*, 14(2):313–341.

Fung, W., Hsieh, D., Naik, N., and Ramodarai, T. (2008). Hedge funds: Performance, risk and capital formation. *Journal of Finance*, 63:1777–1803.

Fung, W. and Hsieh, D. A. (2004). Hedge fund benchmarks: A risk-based approach. *Financial Analysts Journal*, 60(5):65–80.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gerlach, R., Carter, C., and Kohn, R. (2000). Efficient bayesian inference for dynamic mixture models. *Journal of the American Statistical Association*, 95(451):819–828.

Getmansky, M., Lo, A. W., and Makarov, I. (2004). An econometric model of serial correlation and illiquidity in hedge fund returns. *Journal of Financial Economics*, 74(3):529–609.

Geweke, J. and Porter-Hudak, S. (1983). The estimation and application of long memory time series models. *Journal of Time Series Analysis*, 4:221–238.

Giordani, P. and Kohn, R. (2008). Efficient bayesian inference for multiple change-point and mixture innovation models. *Journal of Business and Economic Statistics*, 26:66–77.

Glaser, S. (2017). A review of spatial econometric models for count data. Technical report, Hohenheim Discussion Papers in Business, Economics and Social Sciences.

Goldsmith-Pinkham, P. and Imbens, G. W. (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics*, 31(3):253–264.

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica*, 85(4):1033–1063.

Griffith, A. (2018). Random assignment with non-random peers: A structural approach to counterfactual treatment assessment. *Working Paper*.

Griffith, A. (2019). Name your friends, but only five? the importance of censoring in peer effects estimates using social network data. *Working Paper*.

Guerra, J.-A. and Mohnen, M. (2020). Multinomial choice with social interactions: occupations in victorian london. *Review of Economics and Statistics*, pages 1–44.

Hakim, S., Shefer, D., Hakkert, A.-S., and Hocherman, I. (1991). A critical review of macro models for road accidents. *Accident Analysis & Prevention*, 23(5):379–400.

Halmos, P. R. (2012). *A Hilbert space problem book*, volume 19. Springer Science & Business Media.

Hamilton, J. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384.

Hardy, M., Heath, R. M., Lee, W., and McCormick, T. H. (2019). Estimating spillovers using imprecisely measured networks. *arXiv preprint arXiv:1904.00136*.

Harsanyi, J. C. (1967). Games with incomplete information played by "bayesian" players, i–iii part i. the basic model. *Management science*, 14(3):159–182.

Herbst, E. and Schorfheide, F. (2014). Sequential monte carlo sampling for dsge models. *Journal of Applied Econometrics*, 29(7):1073–1098.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Hsieh, C.-S., Ko, S. I., Kovářík, J., and Logan, T. (2018). Non-randomly sampled networks: Biases and corrections. Technical report, National Bureau of Economic Research.

Hsieh, C.-S., König, M. D., and Liu, X. (2019). A structural model for the coevolution of networks and behavior. *Review of Economics and Statistics*, pages 1–41.

Hsieh, C.-S. and Lee, L. F. (2016). A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics*, 31(2):301–319.

Hsieh, C.-S., Lee, L.-F., and Boucher, V. (2020). Specification and estimation of network formation and network interaction models with the exponential probability distribution. *Quantitative Economics*, 11(4):1349–1390.

Hsieh, C.-S. and Van Kippersluis, H. (2018). Smoking initiation: Peers and personality. *Quantitative Economics*, 9(2):825–863.

Huber, F., Kastner, G., and Feldkircher, M. (2019). Should i stay or should i go? a latent threshold approach to large-scale mixture innovation models. *Journal of Applied Econometrics*.

Inouye, D. I., Yang, E., Allen, G. I., and Ravikumar, P. (2017). A review of multivariate distributions for count data derived from the poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3):e1398.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Johnson, C. R. and Horn, R. A. (1985). *Matrix analysis*. Cambridge University Press Cambridge.

Johnsson, I. and Moon, H. R. (2015). Estimation of peer effects in endogenous social networks: control function approach. *Review of Economics and Statistics*, pages 1–51.

Jones, A. M. (1989). A double-hurdle model of cigarette consumption. *Journal of applied econometrics*, 4(1):23–39.

Karlis, D. (2003). An em algorithm for multivariate poisson distribution and related models. *Journal of Applied Statistics*, 30(1):63–77.

Kasahara, H. and Shimotsu, K. (2012). Sequential estimation of structural models with a fixed point constraint. *Econometrica*, 80(5):2303–2319.

Kastner, G. and Frühwirth-Schnatter, S. (2014). Ancillarity-sufficiency interweaving strategy (asis) for boosting mcmc estimation of stochastic volatility models. *Computational Statistics & Data Analysis*, 76:408–423.

Kelejian, H. H. and Piras, G. (2014). Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes. *Regional Science and Urban Economics*, 46:140–149.

Kelejian, H. H. and Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, 17(1):99–121.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Kim, J. and Kim, H.-J. (2016). Consistent model selection in segmented line regression. *Journal of statistical planning and inference*, 170:106–116.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with arch models. *The review of economic studies*, 65(3):361–393.

Koop, G. and Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886.

Korkas, K. K. and Fryzlewiczv, P. (2017). Multiple change-point detection for non-stationary time series using wild binary segmentation. *Statistica Sinica*, 27:287–311.

Krinsky, I. and Robb, A. L. (1986). On approximating the statistical properties of elasticities. *The Review of Economics and Statistics*, pages 715–719.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of econometrics*, 95(2):391–413.

Lee, C. G., Kwon, J., Sung, H., Oh, I., Kim, O., Kang, J., and Park, J.-W. (2020a). The effect of physically or non-physically forced sexual assault on trajectories of sport participation from adolescence through young adulthood. *Archives of Public Health*, 78(1):1–10.

Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72(6):1899–1925.

Lee, L.-f., Li, J., and Lin, X. (2014). Binary choice models with social network under heterogeneous rational expectations. *Review of Economics and Statistics*, 96(3):402–417.

Lee, L.-f., Liu, X., and Lin, X. (2010). Specification and estimation of social interaction models with network structures. *The Econometrics Journal*, 13(2):145–176.

Lee, L.-f., Liu, X., Patacchini, E., and Zenou, Y. (2020b). Who is the key player? a network analysis of juvenile delinquency. *ournal of Business and Economic Statistics, forthcoming.*

Lewbel, A., Qu, X., and Tang, X. (2019). Social networks with misclassified or unobserved links. *Working Paper.*

Liao, W. (2008). Bayesian inference of structural breaks in time varying volatility models. *Working Paper, New-York University.*

Liesenfeld, R., Richard, J.-F., and Vogler, J. (2016). Likelihood evaluation of high-dimensional spatial latent gaussian models with non-gaussian response variables', spatial econometrics: Qualitative and limited dependent variables (advances in econometrics, volume 37).

Lin, X. and Lee, L.-f. (2010). Gmm estimation of spatial autoregressive models with unknown heteroskedasticity. *Journal of Econometrics*, 157(1):34–52.

Lin, Z. and Xu, H. (2017). Estimation of social-influence-dependent peer pressure in a large network game. *The Econometrics Journal*, 20(3):S86–S102.

Liu, J., Wu, S., and Zidek, J. V. (1997). On segmented multivariate regressions. *Statistica Sinica*, 7:497–525.

Liu, X. (2013). Estimation of a local-aggregate network model with sampled networks. *Economics Letters*, 118(1):243–246.

Liu, X. (2019). Simultaneous equations with binary outcomes and social interactions. *Econometric Reviews*, 38(8):921–937.

Liu, X., Patacchini, E., and Rainone, E. (2017). Peer effects in bedtime decisions among adolescents: a social network model with sampled data. *The econometrics journal*, 20(3):S103–S125.

Liu, X., Patacchini, E., and Zenou, Y. (2014). Endogenous peer effects: local aggregate or local average? *Journal of Economic Behavior & Organization*, 103:39–59.

Liu, X., Patacchini, E., Zenou, Y., and Lee, L.-F. (2012). Criminal networks: Who is the key player? *Unpublished manuscript, NOTA DI. LAVORO. [39.2012].*

Maddala, G. S. (1986). *Limited-dependent and qualitative variables in econometrics.* Cambridge university press.

Maheu, J. and Song, Y. (2013). A new structural break model, with an application to canadian inflation forecasting. *International journal of forecasting*, 30:144–160.

Maheu, J. M. and Song, Y. (2014). A new structural break model, with an application to canadian inflation forecasting. *International Journal of Forecasting*, 30(1):144–160.

Malsiner-Walli, G. and Wagner, H. (2016). Comparing spike and slab priors for bayesian variable selection. *Austrian Journal of Statistics*, 40(4):241–264.

Manresa, E. (2016). Estimating the structure of social interactions using panel data. *Working paper.*

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, 60(3):531–542.

McCausland, W. J., Miller, S., and Pelletier, D. (2011). Simulation smoothing for state–space models: A computational efficiency analysis. *Computational Statistics & Data Analysis*, 55(1):199–212.

McCormick, T. H. and Zheng, T. (2015). Latent surface models for networks using aggregated relational data. *Journal of the American Statistical Association*, 110(512):1684–1695.

Mcilhagga, W. H. (2016). penalized: A matlab toolbox for fitting generalized linear models with penalties. *Journal of Statistical Software*, 72 (6).

Mele, A. (2017). A structural model of dense network formation. *Econometrica*, 85:825–850.

Meligkotsidou, L. and Vrontos, I. D. (2008). Detecting structural breaks and identifying risk factors in hedge fund returns: A bayesian approach. *Journal of Banking & Finance*, 32(11):2471–2481.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Mitchell, M. and Pulvino, T. (2001). Characteristics of risk and return in risk arbitrage. *Journal of Finance*, 56:2135–2175.

Newton, N. J., Pladevall-Guyer, J., Gonzalez, R., and Smith, J. (2018). Activity engagement and activity-related experiences: The role of personality. *The Journals of Gerontology: Series B*, 73(8):1480–1490.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32.

Osborne, M. J. and Rubinstein, A. (1994). *A course in game theory*. MIT press.

Pagliari, J. L., Scherer, K. A., and Monopoli, R. T. (2005). Public versus private real estate equities: A more refined, long-term comparaison. *Real EState Economics*, 33(1):147–187.

Parise, F. and Ozdaglar, A. E. (2019). Graphon games: A statistical framework for network games and interventions. *Available at SSRN 3437293*.

Patacchini, E. and Zenou, Y. (2012). Juvenile delinquency and conformism. *The Journal of Law, Economics, & Organization*, 28(1):1–31.

Patton, A., Ramodarai, T., and Streatfield, M. (2015). Change you can believe in? hedge fund data revisions. *Journal of Finance*, 70:963–999.

Perron, P. et al. (2006). Dealing with structural breaks. *Palgrave handbook of econometrics*, 1(2):278–352.

Pesaran, M. H., Pettenuzzo, D., and Timmermann, A. (2006). Forecasting time series subject to multiple structural breaks. *Review of Economic Studies*, 73:1057–1084.

Pfeiffer, F. and Schulz, N. J. (2012). Gregariousness, interactive jobs and wages. *Journal for Labour Market Research*, 45(2):147–159.

Primiceri, G.-E. (2005). Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies*, 72:821–852.

Qian, J., Hastie, T., Friedman, J., Tibshirani, R., and Simon, N. (2013). Glmnet for matlab. $http://www.stanford.edu/\sim hastie/glmnet\_matlab/$.

Raftery, A. E., Kárnỳ, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52(1):52–66.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2009). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies*, 23(2):821–862.

Rigaill, G., Lebarbier, É., and Robin, S. (2012). Exact posterior distributions and model selection criteria for multiple change-point detection problems. *Statistics and computing*, 22(4):917–929.

Ročková, V. and George, E. I. (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Shiller, R. (2015). *Irrational Exuberance, Third edition*. Princeton University Press.

Smart, D. R. (1980). *Fixed point theorems*, volume 66. CUP Archive.

Snijders, T. A. (2002). Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40.

Soetevent, A. R. and Kooreman, P. (2007). A discrete-choice model with social interactions: with an application to high school teen behavior. *Journal of Applied Econometrics*, 22(3):599–624.

Souza, P. (2014). Estimating network effects without network data. *PUC-Rio Working Paper*.

Stephens, D. A. (1994). Bayesian retrospective multiple-changepoint identification. *Applied Statistics*, 1:159–178.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.

Ter Braak, C. J. (2006). A markov chain monte carlo version of the genetic algorithm differential evolution: easy bayesian computing for real parameter spaces. *Statistics and Computing*, 16(3):239–249.

Thirkettle, M. (2019). Identification and estimation of network statistics with missing link data. *Working Paper*.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

Tüzen, M. F. and Erbaş, S. (2018). A comparison of count data models with an application to daily cigarette consumption of young persons. *Communications in Statistics-Theory and Methods*, 47(23):5825–5844.

Ueda, N. and Nakano, R. (1998). Deterministic annealing em algorithm. *Neural networks*, 11(2):271–282.

Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M., and Higdon, D. (2009). Accelerating markov chain monte carlo simulation by differential evolution with self-adaptative randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulations*, 10:271–288.

Wang, W. and Lee, L.-F. (2013). Estimation of spatial autoregressive models with randomly missing data in the dependent variable. *The Econometrics Journal*, 16(1):73–102.

Winkelmann, R. and Zimmermann, K. F. (1995). Recent developments in count data modelling: theory and application. *Journal of economic surveys*, 9(1):1–24.

Xu, X. and Lee, L.-f. (2015a). Estimation of a binary choice game model with network links. *Submitted to Quantitative Economics*.

Xu, X. and Lee, L.-f. (2015b). Maximum likelihood estimation of a spatial autoregressive tobit model. *Journal of Econometrics*, 188(1):264–280.

Yau, C. Y. and Zhao, Z. (2016). Inference for multiple change points in time series via likelihood ratio scan statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):895–916.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, N. R. and Siegmund, D. O. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.

Zhang, Y., Li, R., and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323.

Zhao, Q., Hautamäki, V., Kärkkäinen, I., and Fränti, P. (2012). Random swap em algorithm for gaussian mixture models. *Pattern Recognition Letters*, 33(16):2120–2126.